Industrial Engineering & Management

Master Thesis

A Data-Driven Model for Financially Viable Fiber-Optic Network Expansion at KPN

A thesis submitted in fulfillment of the requirements for the degree of Master of Science in the field of:

Industrial Engineering & Management

with a specialisation in:

Production, Logistics & Management

Author:

Lars Harm-Willem Lubbers

Company Supervisors:

Danny de Graaf
Britt Gommans

University of Twente Supervisors:

Dr. Stephan Meisel
Dr. Wouter van Heeswijk

July 7, 2025

Management Summary

Management Summary

Over the past few years, Koninklijke PTT Nederland N.V., commonly known as KPN, has been accelerating the roll-out of its fiber-optic network across the Netherlands. While the coverage of fiber in the streets has increased significantly, a large number of buildings remain unconnected. This is particularly the case for high-rise and duplex buildings that were not connected during previous roll-out projects and are now included in so-called retrospective connection projects. In these cases, the fiber infrastructure is already present in the surrounding area, but individual buildings or addresses were skipped due to technical constraints, lack of consent, or other project limitations. As KPN shifts its strategic focus from roll-out speed to maximizing customer connections and capital efficiency, the need for a structured and financially informed selection process becomes more pressing.

Currently, building selection relies on the internally developed selection model, which is a machine learning model that predicts area-level returns without building-level granularity. This method was highly effective in the initial roll-out phase, but it does not account for address-specific costs or customer probabilities and offers limited transparency for planners. As a result, many decisions rely on manual interpretation, leading to inconsistencies and suboptimal capital allocation. This thesis addresses these challenges by developing a new, data-driven model that improves decision-making through more precise, transparent, and cost-effective prioritization.

The study follows the Managerial Problem-Solving Method (MPSM) and is conducted in collaboration with KPN's Fiber Connect Planning & Capacity department. It introduces an integrated model that combines predictive machine learning for customer uptake, routing-based heuristics for realistic trenching costs, and feasibility checks to capture operational constraints like demolition risk or competitor overbuild. These components are synthesized into a composite score used to rank buildings by financial and operational attractiveness. Clustering techniques then group prioritized addresses into coherent project areas, directly supporting practical formulation of rollout projects. Sensitivity analyses on weights, cost clipping, and clustering parameters confirm the robustness of the method across different strategic scenarios.

The models were trained and validated on KPN's historical datasets. The predictive model achieves a balanced macro F1 score of 0.65 and an AUC of 0.67, which, despite reflecting the inherent complexity of customer take-up behavior, clearly outperforms random baselines and improves on the current approach by providing address-level probabilities. Cost estimation experiments show that trench routing paths are on average 27% longer than straight-line distances, and that multi-unit buildings significantly lower per-address costs. These estimates were validated against actual historical project costs, confirming the credibility of the approach. The feasibility assessment filters out non-viable cases, ensuring realistic planning outcomes. All results are made directly actionable through a Power BI dashboard that supports planners in targeting investments and integrating the model into operational workflows.

Beyond its practical relevance, this thesis also contributes methodologically by showing how predictive modeling, cost heuristics, and feasibility assessments can be systematically combined into a unified framework for infrastructure prioritization. It is among the first studies to specifically address brownfield fiber-optic expansion. While tailored to KPN, the approach is generalizable to other utilities where network decisions rely on heterogeneous, location-based data. Future research could build on this by incorporating dynamic pricing, churn prediction, or competitive response scenarios to further optimize strategic roll-out

Preface

This thesis marks the final step of my master's program in Industrial Engineering and Management at the University of Twente, with a specialization in Production, Logistics & Management. It reflects not only the outcome of six months of dedicated research but also the culmination of years of academic and personal development.

The research presented here was conducted in collaboration with KPN, where I had the opportunity to contribute to the optimization of fiber-optic network expansion using data-driven decision-making. Working within KPN's Fiber Connect Planning & Capacity team, I explored how machine learning and cost heuristics can support strategic choices in a large-scale infrastructure context. I am grateful for the autonomy and trust I was given throughout this project and for the openness with which ideas and insights were shared.

I would like to thank my supervisors at KPN, Danny de Graaf and Britt Gommans, for their support, critical feedback, and willingness to engage in detailed discussions that improved both the relevance and the rigor of my work. Their pragmatic perspective helped me bridge theory and practice effectively. I also want to extend my appreciation to the entire Fiber Connect Planning & Capacity team and other colleagues at KPN who offered their time, expertise, and insights, which greatly enriched this research.

I am equally grateful to my academic supervisors, Dr. Stephan Meisel and Dr. Wouter van Heeswijk, for their constructive guidance and critical eye. Their input pushed me to clarify my research design, strengthen my modeling approach, and improve the academic contribution of this work.

Lastly, I want to thank my friends, family, and partner for their encouragement during this period. Their support helped me maintain perspective and motivation throughout the process.

I hope this thesis contributes to the ongoing development of data-informed infrastructure planning and that it proves valuable both within KPN and in the broader field of applied decision science.

Lars Harm-Willem Lubbers

Enschede, July 2025

Reading Guide

This thesis presents the development of a data-driven decision-support model for selecting financially viable fiber-optic connections within the Dutch telecommunications company KPN. The structure of the thesis reflects the logical flow of the research process. It follows the Managerial Problem-Solving Method (MPSM) [13]. Readers are guided through the problem context, theoretical foundation, model development, and practical implications.

Chapter 1 – Introduction

This chapter introduces the research context, the strategic relevance for KPN, and formulates the central problem and research questions. It explains the methodological approach based on MPSM and defines the scope and deliverables of the project.

Chapter 2 – Current Situation

This chapter describes the current operational environment at KPN. This includes performance indicators, stakeholder roles, connection processes, and the limitations of the existing selection method. This chapter motivates the need for a more data-driven approach.

Chapter 3 – Literature Review

This chapter provides an overview of relevant academic and technical literature on fiber-optic network planning, decision-making under uncertainty, and the application of machine learning for infrastructure prioritization. It positions the research within existing theoretical and applied work.

Chapter 4 – Solution Design

This chapter explains the step-by-step development of the selection model. This includes data preprocessing, feature engineering, model formulation, and validation strategy. It presents four model components: Connectivity Potential, Cost Estimation, Feasibility, and Composite Scoring.

Chapter 5 – Results

This chapter evaluates the performance of each model component using both quantitative metrics and sensitivity analyses. It discusses the implications of the model outputs and demonstrates how the results inform decision-making within a project selection dashboard.

Chapter 6 – Conclusion, Discussion, and Recommendations

This chapters summarizes the research findings and reflects on the model's effectiveness, limitations, and generalizability. It gives recommendations for KPN and gives recommendations for further research. The chapter concludes with an overview of the theoretical and practical contributions.

Contents

List of Figures

List of Tables

List of Abbreviations

The next list describes several abbreviations that will be later used within the body of the document.

AUC	Area Under the ROC Curve
BM	Business Market
CM	Consumer Market
DP	Distribution Point
FC	Fiber Connect
Ftth	Fiber to the Home
FTU	Fiber Termination Unit
HA	Homes Activated
HC	Homes Connected
HP	Homes Passed
KD-tree	K-dimensional Tree (used in nearest-neighbor searches)
KPN	Koninklijke PTT Nederland N.V.
LightGBM	Light Gradient Boosting Machine
MILP	Mixed Integer Linear Programming
MPSM	Managerial Problem Solving Method
PON	Passive Optical Network
PoP	Point of Presence
RAM	Random Access Memory
RNA	Reason Not Connected
ROCE	Return on Capital Employment
SHAP	Shapley Additive Explanations
XGBoost	Extreme Gradient Boosting

1 Introduction

The telecommunication sector is a large sector in the Netherlands with several competitors competing for a limited number of customers. Many of them play a large role in current society due to the digital world we live in. KPN is the biggest telecommunications company in the Netherlands. This thesis will explore and create a concept to improve the selection process of homes connected of the Fiber Connect Planning & Capacity team.

1.1 The company

Established in 1989, Koninklijke PTT Nederland N.V., commonly known as KPN, is a leading Dutch telecommunications and IT provider. With a strong presence in the Netherlands, the company offers a wide range of services which includes fixed and mobile telephony, data, and television. KPN serves both private customers and business users, from small enterprises to large corporations. In 2024, KPN generated revenues of €5.634 billion and solidified its position as the market leader in the Dutch telecommunications sector. Furthermore, KPN continued to lead the Dutch fiber market, expanding its footprint to cover 63% of the Netherlands, including through its joint venture Glaspoort[20]. This thesis focuses on the roll-out of fiberglass by KPN, specifically the activities performed by the Fiber Connect Planning & Capacity team.

1.2 Research Motivation

In 2019, KPN accelerated the expansion of its fiber-optic network. The reason for this was to replace the very old copper infrastructure with a higher-speed and more reliable fiber-optic network. The strategic goal behind this was to maintain and potentially improve its position as a premium telecommunications provider. Additionally, competitive pressure played a role, because alternative providers such as Delta and ODF were also deploying fiber-optic networks in the Netherlands. In this competitive environment, being the first to deploy fiber in an area was important, as customers typically do not feel the need to have multiple fiber connections in their homes. Therefore, it was important for KPN to be the first to connect areas to keep existing market share and potentially gain more market share.

KPN's fiber-optic network generates revenue through two primary channels. Firstly, by directly acquiring customers who subscribe to KPN services, which yields the highest financial return. Second, through wholesale agreements in which third-party providers, such as Odido, deliver services over KPN's network infrastructure. Although wholesale agreements generate revenue, this is significantly lower than revenue from direct customers. Consequently, maximizing the number of properties connected to the network, particularly those with high potential for direct customer acquisition, is essential to achieving both commercial and strategic objectives. KPN has set a target of having the Homes Passed (HP) status for 80% of private properties by 2026. HP means that the fiber-optic network is for example already in the street, but that the home itself is not connected to the fiber-optic network.

During the initial roll-out phase, the focus on rapid deployment led to incomplete connections within many areas. A significant number of properties were not connected during these projects, despite fiber infrastructure being present on their streets. This was primarily due to operational constraints such as homeowners being unavailable during connection attempts. These unconnected properties are particularly concentrated in high-rise buildings, which means KPN still has a lot of room to grow their network.

Consequently, KPN has shifted its focus more on actually connecting addresses and efficient use of their capital expenditure (CAPEX). For this reason, KPN is now revisiting past projects with the intention of connecting previously unconnected properties. However, the current process for selecting which building to connect is based on criteria developed for evaluating entire areas rather than individual buildings. This results in low accuracy and inefficiencies in resource allocation. Additionally, the process relies heavily on manual selection and subjective decision-making, which increases labor intensity and inconsistency. As a result, potentially viable buildings may be overlooked, while resources may be allocated to projects with limited financial return.

To address these limitations, there is a need for a structured, data-driven selection methodology that evaluates properties on an individual basis and incorporates financial viability into the decision-making process. This research aims to develop such a method, enabling KPN to prioritize properties for connection based on cost and revenue considerations, thereby improving efficiency and supporting the organization's strategic goals.

1.3 Problem Description

This section gives a clear explanation of the problem. Section 1.3.1 identifies the current problems at KPN. Section 1.3.2 shows the formed problem cluster and explains the chosen core problem of this thesis.

1.3.1 Problem Identification

At KPN, the current selection process for determining which buildings should be connected to its fiber-optic network has several problems. These problems have effect on both the operational effectiveness and financial performance. The management has identified that the main problems come from having an outdated selection criteria tool, manual decision making and a lack of inclusion of cost-based prioritization of buildings. As a result, these problems lead to inefficient resource allocation.

One of the main problems has it origin in the way buildings are selected to be connected. Currently, the selection is based on the machine learning scoring method, which is originally developed as tool to evaluate whether an area is profitable to connect. This is a developed in-house model that uses machine learning based on several features. More information about the exact techniques used to calculate the machine learning score can be found in Chapter 2.4.1. Nowadays, the method is also used as a support tool for evaluating individual buildings. However, as this method was never developed for this purpose, it does not give accurate assumptions of whether connecting a certain building is viable. Therefore, the method can lead to missed opportunities for KPN as potentially good buildings do not come up in the algorithm. At the same time, the selection process requires more time as potentially bad buildings might get a good score in the current method, which makes it seem like a good choice. As a result, the team currently relies heavily on manual selection of projects and decisions take longer and might be based more on opinion than being data-driven.

Secondly, another problem is that financial considerations are not explicitly incorporated into the selection process. As KPN aims to connect 80% of the Netherlands, achieving this in a cost-effective way is a priority. Every year there is a certain budget available for connecting houses for the department. As a result of not taking the costs of certain projects into account specifically, some projects may be selected despite possibly having high costs leading to low returns on investment, while others with potential higher returns might not be prioritized. As a result, the absence of a structured cost-benefit evaluation for all buildings makes it difficult to optimize KPN's investment in fiber-optic connections. Furthermore, the machine learning method does not show how it came to the score and of what elements the scores consist. This makes it hard for people working with the model to figure out why a certain property has been chosen and on what variables this is based.

Besides the inefficiencies in the selection process, KPN also has problems related to the accuracy of their contractor planning. The company relies on external contractors to execute the roll-out. However, management has noted several problems in their processes. One of these problems is the inconsistency in the Reason Not Connected (RNA) status. The RNA status is used to the record justifications for why certain buildings were not connected during the initial roll-out. A list of reasons for why buildings were not connected can be found in the Appendix A. As contractors get rewarded separately for different steps in the roll-out process, they may opt to only roll-out the fiberglass in the street and not connect the buildings. The reason for this is that connecting the homes is in many cases the most time-expensive and technical part of the project. So, contractors may choose to fulfill the first steps in the contract and get rewarded for those steps and then not complete the lasts steps. Due to this contract structure, many of the RNA statuses are incorrect or outdated, leading to misinformed decision-making when revisiting unconnected buildings. Some buildings classified as unfeasible to connect may in reality be easily connectable, while others marked as pending approval may have hidden obstacles that were not initially accounted for. These inaccuracies create inefficiencies in project planning and inefficient resource allocation.

Additionally, all contractors operate under different agreements. This makes it difficult to standardize cost expectations and accurately forecast expenses. Since KPN works with multiple contractors in different regions, each with their own pricing structures and service conditions, variations in cost estimates make cost predictions hard to achieve for every potential building. All in all, These problems with the contractors lead to delays, cost overruns, and inefficiencies, affecting KPN's goal of achieving 80% homes connected by 2026.

In summary, KPN management experiences multiple problems in efficiently selecting, planning, and prioritizing fiber-optic connections. The current approach lacks financial transparency, relies too heavily on manual intervention, and suffers from planning inaccuracies that create delays and inefficiencies.

1.3.2 Problem Cluster

Based on the problem identification, this section gives a logical analysis to identify the core problem. After further analysing the perceived problems, possible cause and effect relationships are identified ensuring that a problem cluster can be visualized. A problem cluster is a model used to identify problems and the interconnected relationships between them [13]. From this problem cluster, the core problem of the research can be be identified.

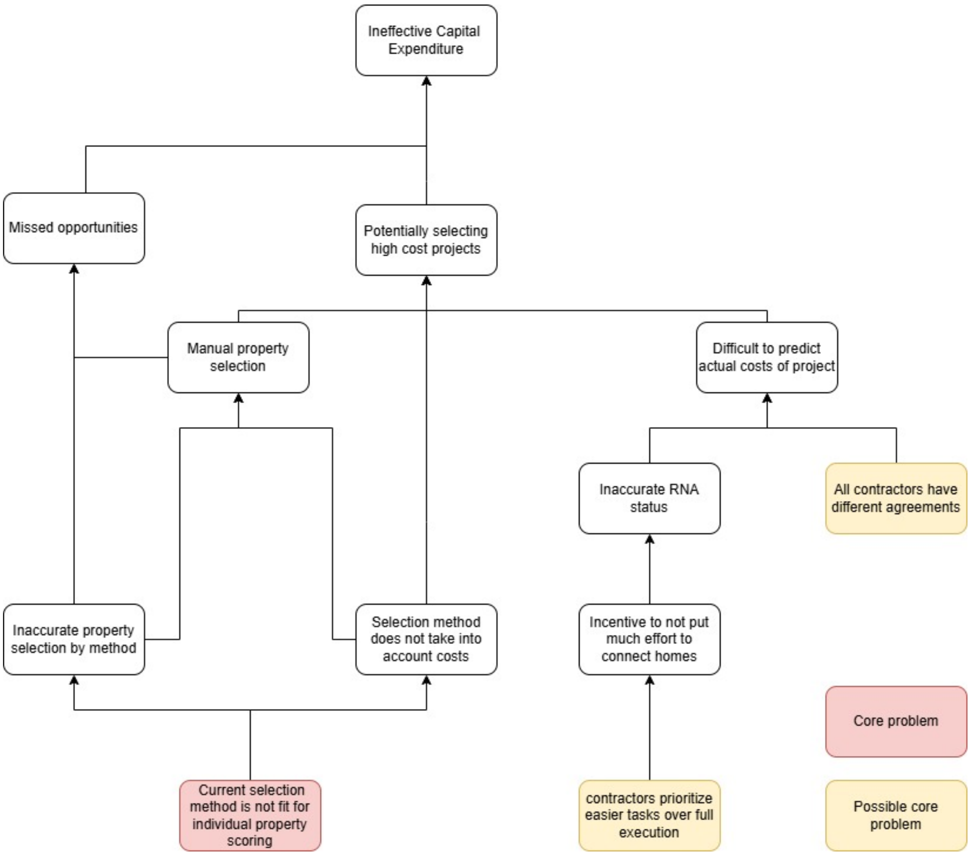

Ineffective Capital Expenditure
Missed opportunities
Potentially selecting high cost projects
Manual property selection
Difficult to predict actual costs of project
Inaccurate RNA status
All contractors have different agreements
Inaccurate property selection by method
Selection method does not take into account costs
Incentive to not put much effort to connect homes
Core problem
Current selection method is not fit for individual property scoring
contractors prioritize easier tasks over full execution
Possible core problem

Figure 1.1: *Problem cluster identifying the core problem in KPN's current fiber-optic selection process. It highlights how gaps in financial granularity, inconsistent manual decisions, and the limitations of the existing selection model interrelate.*

Core problems are those whose solutions will make a real difference at the origin [13]. The potential core problems are always at the beginning of a causation chain in a problem cluster. This leaves 2 possible options for the core problem. As the other tasks are both external problems that cannot be changed, the following core problem is chosen: The current selection method for determining which buildings should be connected to KPN's fiber-optic network is not fit for individual property scoring. This problem is highlighted in red in Figure 1.1.

This problem is selected because it has the greatest impact on KPN's strategic objectives. Improving the selection method will enable more accurate prioritization of buildings, ensuring that both operational efficiency and financial viability are considered. By addressing this core problem, KPN can reduce reliance on manual selection, minimize high-cost, low-return projects, and optimize the roll-out process toward achieving its 80% connectivity goal by 2026. The selected core problem is defined as:

In the current situation, KPN's selection process for determining which buildings to connect to its fiber-optic network is inaccurate, relying on a scoring model that is not suited for individual properties and failing to determine the financial viability of new connections, leading to inefficient capital expenditure.

1.4 Problem solving Approach and Research Design

This section explains the problem-solving approach and the research methodology. Section 1.4.1 gives the scope of the research. Section 1.4.2 explains the methodology and the research questions. Lastly, Section 1.4.3 shows the deliverables of the research

1.4.1 Research Scope

The research scope of this project focuses on step 1 of the Retrospective connections project process that can be seen in Figure 1.2. Retrospective connections project is defined as when a building has already been in a previous fiber-optic network expansion project, but not been connected in that project. Now, these buildings that have previously not been connected are evaluated again if they are financially viable to connect. If this is the case, they are retrospectively connected and follow the process shown in 1.2. The process itself is explained extensively in Chapter 2.3.2. Specifically, the goal of the thesis is to improve KPN's selection process for determining which buildings from the retrospective connections category should be connected to its fiber-optic network. The study aims to develop a data-driven selection model that replaces the current area based approach, which is not suited for evaluating individual buildings.

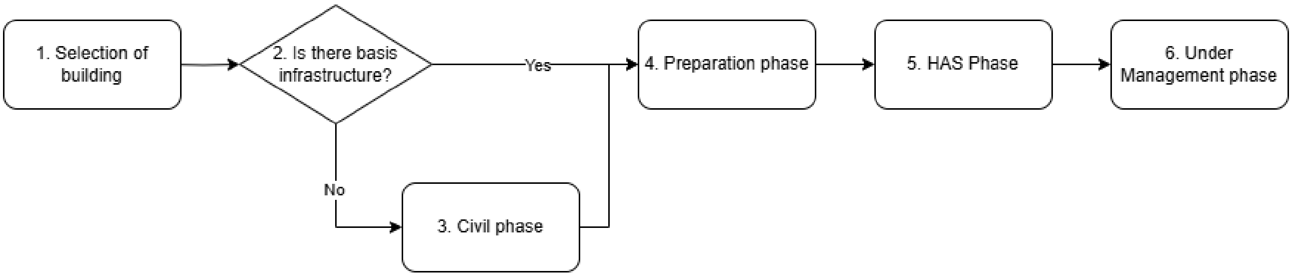


Figure 1.2: *Overview of the retrospective connections process, which targets buildings skipped in earlier rollouts. This schematic outlines the sequential steps from building identification to final integration into KPN's managed network.*

The research is limited to residential property connections from previous fiber-optic roll-out projects] that were not connected at the time. The primary focus is on high-rise and duplex buildings, though low-rise buildings are also included although they are less prevalent.

While the study evaluates the current selection process and proposes a more effective method, it does not focus on new network expansion strategies, contractor procurement policies, or technical installation processes. Instead, it aims to optimize the decision-making criteria for prioritizing buildings within existing previous project areas.

The proposed model will be validated through model experiments and expert evaluation, where industry professionals will assess whether the model correctly identifies buildings with strong connection potential. Their insights will refine the model to ensure its practical applicability. Furthermore, this research does not cover contractor incentives, installation challenges, or national fiber-optic expansion strategies.

1.4.2 Methodology and Research Questions

To achieve the research objective of the previous subsection, the main research question is formulated in the following way:

How can KPN develop a data-driven selection method to identify and prioritize financially viable buildings to connect to its fiber-optic network?

The goal of this research is to answer this main research question, thereby solving the core problem. This is done by dividing the research in different steps using the Managerial Problem-Solving Method (MPSM) of Heerkens & Van Winden[13]. This is method is a systematic problem-solving approach. It is applicable to various problems and takes a problem into account in the context of an organisation [13]. The MPSM is an approach visualized in Figure 1.3.

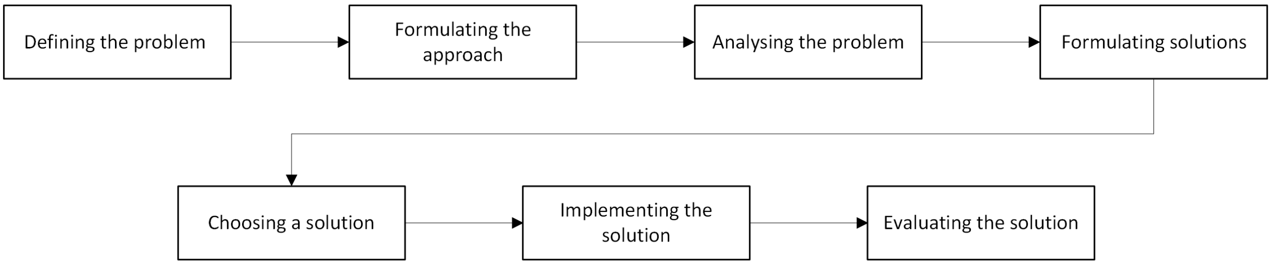


Figure 1.3: *Visualization of the Managerial Problem-Solving Method (MPSM) applied in this research. The diagram shows how the structured approach guides problem identification, solution design, implementation, and evaluation.*

With the identification of the core problem, the first step of the MPSM is completed. The next step of the MPSM is formulating the research approach. The following section describes the corresponding research questions and data collection methods of each step. Answering these research questions results in the answer to the main research question. Since the first chapter discusses the analysis of the problem and the methodology, only the chapters after the first get their own research questions. These can be found below.

Problem Analysis

1. What is the current situation regarding KPN's selection and connection of buildings to the fiber-optic network?

The answer of this research question is found in Chapter 2 with the analysis of the current situation, the current processes and the context of the problem. Furthermore, in Chapter 2 several other research questions are used to come to the answer of the research question. These are the following:

- *What are the key performance statistics of KPN's current selection process for connecting buildings from previous fiber-optic projects?*
- *Who are the stakeholders involved in the decision-making process for selecting buildings to be connected?*
- *What are the necessary process steps involved in selecting and connecting buildings to the fiber-optic network?*

- *What are the criteria currently used in the selection process, and how do they influence decision-making?*
- *How are the agreements with contractors currently structured? What are the data characteristics and decision requirements that make the selection problem suitable for machine learning?*

To answer the research question of Chapter 2 and it sub-research questions. The research conducts interviews, analyses organizational documents and gathers information by working with the involved stakeholders in the organization.

Formulation of Solutions

2. What is the theoretical background on data-driven decision-making and selection methodologies for prioritizing fiber-optic network connections?

The answer to this research question is given in Chapter 3 by conducting a literature study. The literature study will cover the following sub research questions to formulate an answer to the main research question of the chapter. The sub research questions are the following:

- *What does existing literature suggest about methods for prioritizing brownfield network expansions, and how do these differ from greenfield approaches?*
- *How are machine learning models leveraged to predict customer uptake or acquisition in telecom or similar network industries?*
- *What routing heuristics and cost estimation techniques are most commonly applied in infrastructure network planning, and how do they account for real-world trenching deviations?*
- *Which evaluation metrics are recommended in the literature for assessing predictive models in customer acquisition and infrastructure decision-making contexts?*

To answer this research question and the subsequent sub-research questions, the research uses a systematic literature review.

Solution Choice

3. How can data be evaluated and processed to develop a data-driven selection methodology for prioritizing fiber-optic network connections?

The answer to this research question is given in Chapter 4 by providing the data collection method, data-analysis, used calculations and the creation the prediction models for the inputs and the subsequent solution model.

- *What data sources and variables are relevant for evaluating fiber-optic network connection decisions?*
- *How should the collected data be structured and processed for use in the selection methodology?*

- *What selection models or decision-making frameworks are most effective for prioritizing fiber-optic connections?*
- *How can cost-based prioritization be incorporated into the selection methodology?*
- *How can the developed selection methodology be validated for accuracy and effectiveness?*

To answer this research question, and its applicable sub research questions, the research collects, processes and selects relevant data from different outputs. Furthermore, it describes the solution models that uses the data.

Solution Implementation & Evaluation

What are the results of the developed data-driven selection methodology, and how do they impact the performance of the fiber-optic connection planning process?

The answer of this research question in chapter 5 discusses the results of the solution design and how the model has an effect on the performance on the entire selection process. This is used as a validation of the model and as a consequence a validation of the entire research.

- *What are the results of the data-driven selection method in terms of prioritizing fiber-optic connections?*
- *What are the results of the predictive models used in the selection methodology?*
- *What are the results of the decision-making framework in optimizing fiber-optic selection?*
- *How does the developed selection methodology improve the current selection process?*
- *How sensitive is the selection methodology to changes in input parameters, such as cost factors, infrastructure availability, or competitor presence?*
- *Does the output correspond to what experts say? How does the output of the selection methodology compare with expert evaluations of fiber-optic connection prioritization?*

To answer this last research question, and its applicable sub research questions, the research develops the necessary models and solves them in an adequate software program (Python & PowerBI). Furthermore, the research evaluates different performance measures, compares the new with the current situation, and validates the results.

1.4.3 Deliverables

This thesis provides several key deliverables that together address the core problem identified and offer both immediate operational benefits and contributions to broader academic knowledge. The primary outcome is a data-driven selection model designed to optimize KPN's decision-making regarding fiber-optic connections. This integrated approach replaces the existing machine learning based method by incorporating predictive customer acquisition modeling, routing-based cost estimations, and feasibility assessments. It improves selection accuracy, transparency, and financial efficiency at the individual building level, directly supporting KPN's strategic objectives.

The thesis also includes a comprehensive evaluation of the developed model under different scenarios. Through extensive sensitivity analyses, the model's adaptability and robustness were tested against varying financial constraints, connection densities, and rollout strategies. This ensures that the method remains reliable and relevant across diverse project conditions.

Additionally, the research provides a set of practical recommendations to improve KPN's selection process, including automation opportunities, refinements to decision-making criteria, and ways to integrate expert validation. Implementation is supported by a Power BI dashboard that makes the results directly actionable and easy to incorporate into operational planning.

On a practical level, this work enhances KPN's ability to prioritize fiber-optic connections by offering precise, data-driven insights at the address level. It enables more accurate and financially informed decision-making, optimizes capital allocation, and improves day-to-day planning efficiency. From a theoretical perspective, the thesis contributes by showing how predictive modeling, routing-based cost heuristics, and feasibility assessments can be combined into a single framework. By specifically addressing brownfield fiber-optic expansion, it also adds to the limited literature on this topic and offers a methodological approach that can be generalized to similar infrastructure planning challenges.

2 Current Situation

This chapter covers the first research question:

What is the current situation regarding KPN's selection and connection of buildings to the fiber-optic network?

It frames the current situation in the context of the problem. This research question breaks down into the following sub-questions:

- *What are the key performance statistics of KPN's current selection process for connecting buildings from previous fiber-optic projects?*
- *Who are the stakeholders involved in the decision-making process for selecting buildings to be connected?*
- *What are the necessary process steps involved in selecting and connecting buildings to the fiber-optic network?*
- *What are the criteria currently used in the selection process, and how do they influence decision-making?*
- *How are the agreements with contractors currently structured?*
- *What are the data characteristics and decision requirements that make the selection problem suitable for machine learning?*

Section 2.1 evaluates the key performance indicators of KPN's current selection process, including fiber-optic network coverage, building types, and the reasons why certain buildings were not connected. Section 2.2 introduces the stakeholders involved in the selection process and classifies them based on their strategic, operational, or commercial role. Section 2.3 outlines the processes used by KPN for network expansion, distinguishing between the Overlay process and the Retrospective Connections process. Section 2.4 discusses the selection criteria applied in both processes, with a particular focus on the selection model and contractor capacity. Section 2.5 explains how the selection model functions, detailing both its cost and profit components. Section 2.6 describes the agreements with contractors and how they influence project planning and costs. Finally, Section 2.7 concludes the chapter.

2.1 Key performance Indicators

Currently, KPN has several Key Performance Indicators (KPIs) that show how they are currently performing. Below, an overview of these KPIs with explanation is shown of their current performance.

2.1.1 Fiber-optic network coverage

Since 2007, KPN has invested in replacing their copper network with the fiber-optic network. This has been done either by the acquisition of other parties that own fiber-optic networks or rolling it out via hired contractors. KPN has used their homes passed (HP) as a metric. This means in general that the fiber-optic network basis infrastructure is rolled out in the neighborhood, so all houses that have this state have access fiber-optic network in the street. The number of addresses achieving this status over the years can be seen in Figure 2.1. As can be seen in Figure 2.1, KPN decided in 2019 that the roll-out of the fiber-optic network was a high-priority and decided to speed up with the goal to reach 80% of dutch addresses having the HP status by the end of 2026.

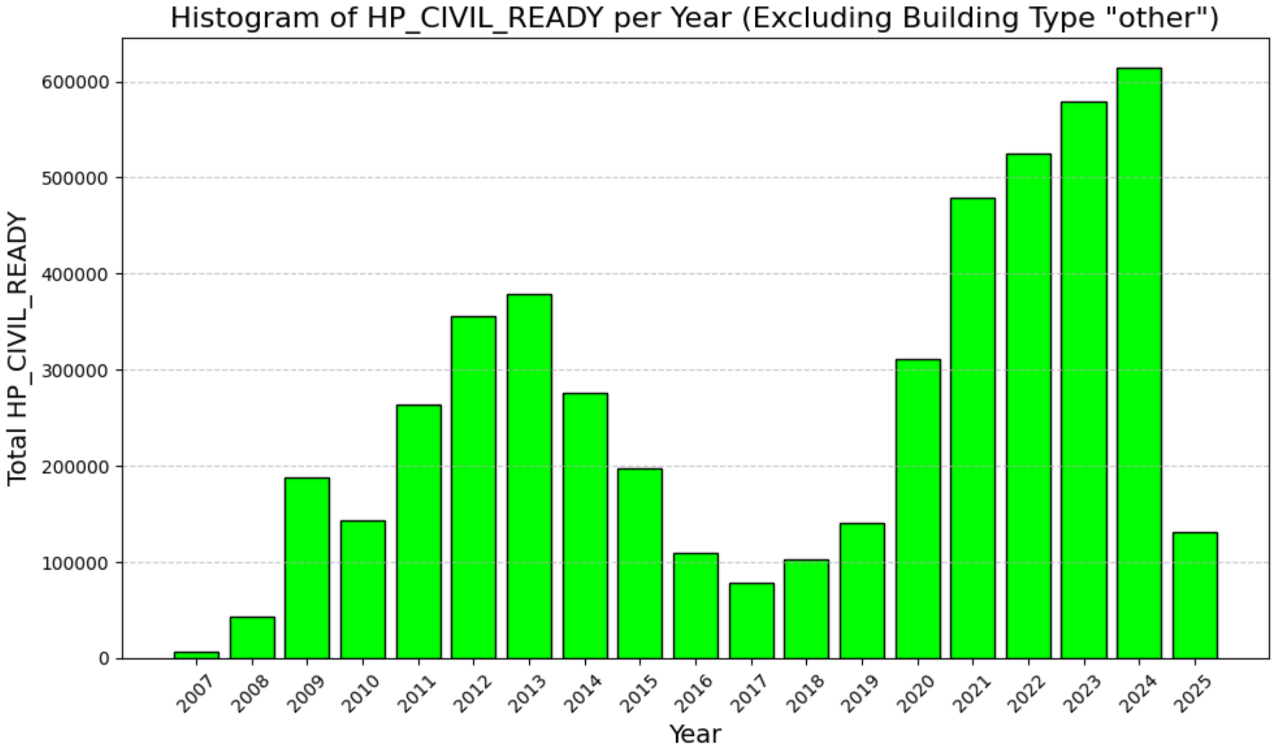


Figure 2.1: *PUBLIC EXAMPLE Annual production of homes passed (HP) by KPN's fiber-optic rollout. The figure highlights a sharp increase in deployment speed from 2019, illustrating KPN's strategic acceleration to achieve its target of 80% HP by 2026. This underscores the shift in focus toward broad network coverage.*

Within this group of Homes Passed, there are a lot of properties that have the Home Connected (HC) status. This means that the property itself is connected to the fiber-optic network and a service from KPN or another provider can potentially be delivered to the customer. The number of HC's is lower than the number of Homes Passed due to the extra efforts that have to take place to actually connect an address to the network. For example, it could be technically difficult to connect the address or the home-owner does not give permission to connect the address to the fiber optic network.

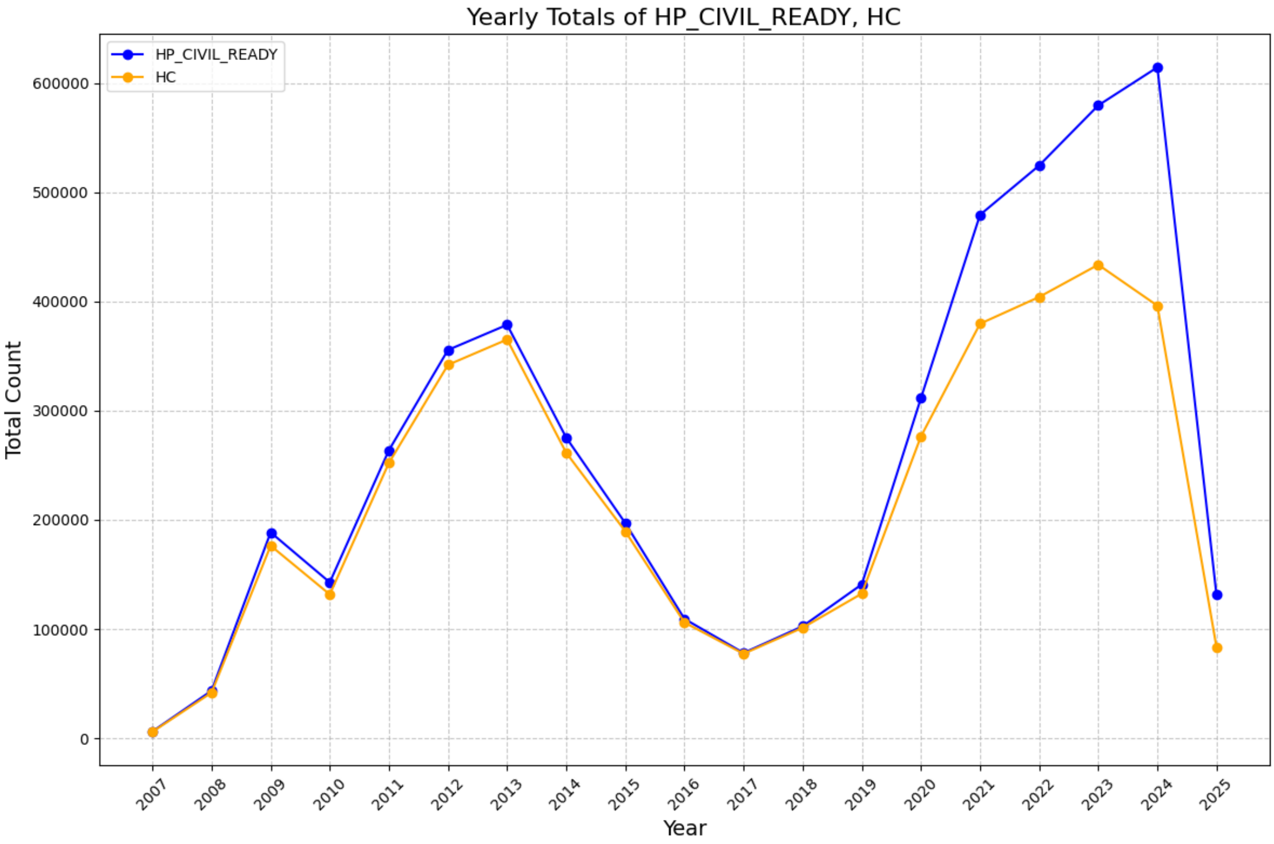


Figure 2.2: *PUBLIC EXAMPLE Yearly totals of homes passed (HP) versus homes connected (HC). The growing gap after 2019 shows that while street-level infrastructure expanded rapidly, actual home connections lagged behind, emphasizing the challenge of converting HP to HC.*

In Figure 2.2, it can be seen that since the acceleration of the roll-out, the ratio between HP and HC has become worse compared to earlier years. The reason for this is that due to the achieved speed of the roll-out and the used contract structure with the contractors, the time to effort reward has shifted more towards the HP status and less towards the HC status. At the time of the start of the acceleration, this was seen as acceptable, because rolling out faster also meant that other competitors most likely would not roll-out fiber-optic network in the same area. This could then result in a better position in competitive environment, as KPN can deliver their own services to the customers and can also earn money by offering competitors to sell services over their networks.

However, with KPN covering more and more of the Netherlands, their goals have shifted. Now they also desire to have as many HC's as possible. This is due to the reason that the number of HC's naturally has direct impact on the number of customers that they can deliver services to. In Figure 2.2, the current customers are noted as Homes Activated (HA). As can be seen, in recent years around 50% of the network is actively used at the moment. This means that either KPN or another provider delivers a service over the network of KPN. As, it requires another investment besides their initial investments. This directly leads into the core of this thesis, as KPN currently does not have an up to date selection method to determine the value of connecting a certain address, which in return is determined by a lot of factors. The different types of buildings will be covered in the next subsection.

2.1.2 Building types

At the moment, there are approximately 10.6 million addresses in the database of KPN. These are divided in 4 categories. Low-rise buildings, high-rise buildings, Other which are often offices of companies and Duplex, which means that one home is divided in two.

Table 2.1: *Distribution of Building Types, HP, and HC Percentages*

Building Type	Count	% of Total Buildings	% of HP	% of HC
Low-rise	5,378,513	55.58%	62.31%	53.69%
High-rise	3,117,651	32.22%	59.21%	40.82%
Other	876,163	9.05%	21.70%	10.99%
Duplex	305,152	3.15%	55.45%	36.10%

When filtered only to residential buildings approximately 8.4 million addresses remain. As can be seen no buildings with datatype other are present when filtering on residential buildings. The exact division can be seen in the table below. Concretely, this means that KPN has 1.28 million possible addresses that have been in a project, but are not connected to its fiber-optic network.

Table 2.2: *Distribution of Building Types, HP, and HC Percentages (residential buildings only)*

Building Type	Count	% of Total Buildings	% of HP	% of HC
Low-rise	5,200,869	62.15%	63.03%	54.53%
High-rise	2,895,050	34.60%	60.51%	42.13%
Duplex	271,968	3.25%	56.54%	37.11%

As shown in Table 2.2, high-rise and duplex buildings have a lower conversion rate from HP (Homes Passed) to HC (Homes Connected) compared to low-rise buildings. This difference is primarily caused by the increased complexity of connecting individual addresses within multi-unit buildings. One of the encountered problems is the throughput dependency present in buildings with multiple residences, where the fiber-optic cable must pass through lower-level apartments to connect the apartments on higher floors. If the resident or property owner of a lower-level unit denies access, the apartments above it cannot be connected. Additionally, in many cases, the consent of the homeowners' association or housing corporation is required to approve the connection for the entire building. If the permission is denied, none of the units in the building can be connected. These complications do not typically arise in low-rise buildings, which are often owned by an individual owner.

If a building is not connected they always get a Reason not Connected (RNA) status. Between the different building types there are several differences in the division of these statuses. These will be covered in the next subsection.

2.1.3 Reasons Not Connected

For high-rise buildings, the RNA distribution is dominated by R0, which represents planned connections. This indicates that a large portion of addresses in high-rise buildings are already scheduled to be connected and thus do not reflect an actual barrier to connection. Additionally, RNA 30 is notable here, as it signifies that no work is currently planned for the building—an issue specific to multi-unit structures such as high-rise complexes.

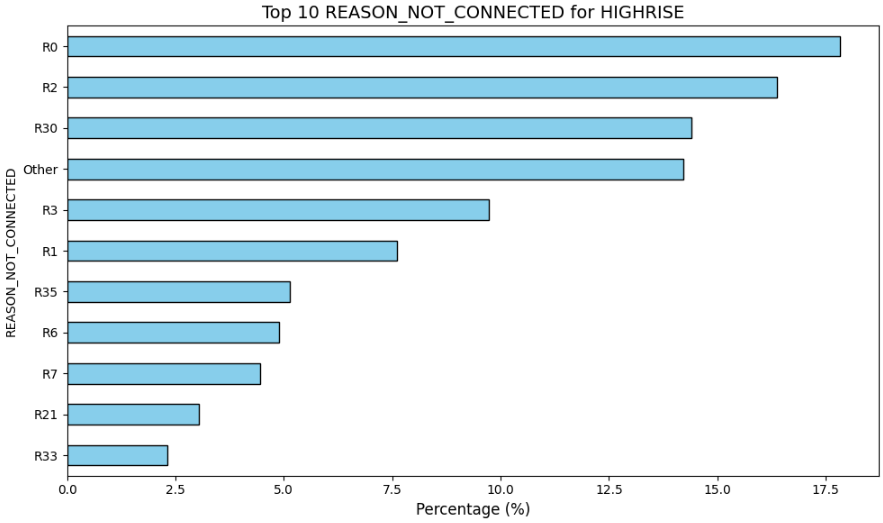


Figure 2.3: *PUBLIC EXAMPLE Distribution of the top 10 Reason Not Connected (RNA) codes for high-rise buildings. Most addresses are classified under R0 (planned connection) or R30 (no work yet done), indicating future opportunities but also current planning gaps.*

For duplex buildings, a similar pattern is observed. The RNA distribution also shows a clear prevalence of R0 and RNA 30. This reflects the fact that duplexes, much like high-rise buildings, often involve shared infrastructures and multiple residential units, leading to comparable connection challenges.

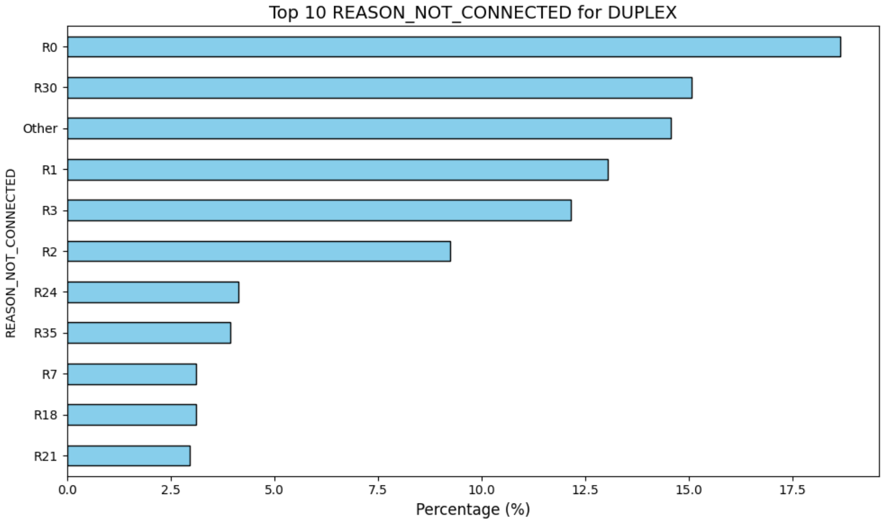


Figure 2.4: *PUBLIC EXAMPLE Top RNA codes for duplex buildings, which mirror the patterns observed in high-rises. The prevalence of R0 and R30 reflects similar challenges in multi-unit access and coordination.*

In contrast, low-rise buildings exhibit a different RNA profile. While R0 is still the most frequently listed reason, R1 and R3 appear more prominently than RNA 30. Here, R1 indicates situations where residents did not provide consent, and R3 points to cases where residents were not at home after three visits. These reasons highlight more household-specific obstacles to connection. The overall diversity of RNA codes is lower compared to high-rise and duplex buildings, which can be linked to the generally higher HP/HC conversion rates in low-rise structures.

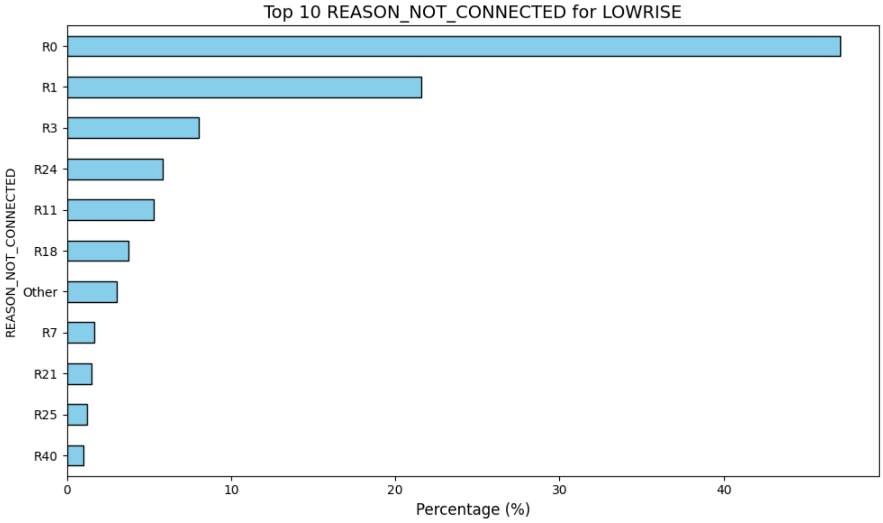


Figure 2.5: *PUBLIC EXAMPLE RNA code distribution for low-rise buildings. Compared to multi-unit structures, low-rise buildings show higher instances of consent-related issues (R1 and R3), highlighting different operational barriers.*

2.2 Stakeholders

Several stakeholders are involved in the decision-making process for selecting which buildings should be connected to KPN's network. These stakeholders have varying interests and roles, and can be classified as strategic, operational, or commercial which can be seen in Figure 2.6).

At the strategic level, the Fiber Connect department is responsible for overseeing the overall roll-out performance and ensuring that KPN's connectivity targets, such as the 80% homes passed goal by 2026, are met. Within this department, the Planning & Capacity team plays a central operational role. This team is directly responsible for evaluating, selecting, and prioritizing potential buildings for connection, and is the focus of this research project.

On the operational side, Roll-out Contractors and the Roll-out Department are external and internal stakeholders, respectively, who are responsible for executing the physical connection of buildings to the fiber-optic network. These stakeholders provide essential input regarding feasibility, technical constraints, and project costs, and they influence the selection process through the accuracy of their reporting, mostly in relation to RNA (Reason Not Connected) statuses.

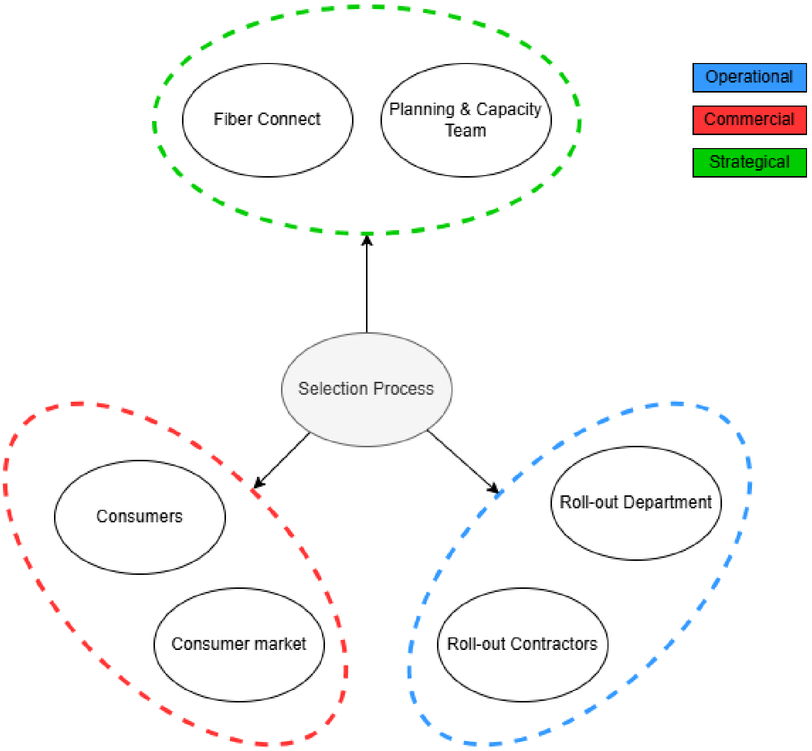


Figure 2.6: *Stakeholder involvement across strategic, operational, and commercial levels in the fiber connection decision process.*

From a commercial perspective, the Consumer Markets department focuses on revenue generation by maximizing customer acquisition for KPN services. This department influences selection by identifying areas or buildings with high potential customer value. Finally, consumers themselves are an external stakeholder group whose willingness and ability to be connected determine the viability and financial return of individual connection projects. Their presence and market potential are considered during the selection process, especially in competitive regions with multiple fiber providers.

2.3 Roll-Out and Retrospective Connection Processes

KPN categorizes its fiber-optic network expansion activities into two distinct project types, each with specific characteristics that influence the process of selecting properties for connection. The overlay process always precedes the retrospective connections process. For that reason, the outcomes of the overlay process have a big impact on the current state of the retrospective connections process. For that reason, the Overlay process, selection criteria and contract agreements are all discussed as well.

The first category, Overlay, refers to projects aimed at replacing the existing copper network infrastructure or installing entirely new fiber-optic infrastructure in areas without prior coverage. The second category is the retrospective connection of under management properties, which involves revisiting completed projects—now in the under management phase—to connect previously unconnected properties. This situation often involves high-rise buildings that were not connected during the initial roll-out. Additionally, new housing estates are automatically connected during their construction phase as part of KPN's standard procedure. These projects do not require a selection process.

Importantly, the selection process that forms the core of this research is directly applicable to the retrospective connection category and some of the insights and outcomes are also applicable to the Overlay process. The reason for this is, because some of the selection criteria are applicable for both. In this context, selecting which properties to connect requires structured evaluation and data-driven decision-making. The following sections detail the processes associated with these project types.

2.3.1 Overlay Process

The Overlay Process consists of 6 phases which can be seen below in Figure 2.7. These phases are shortly explained below.

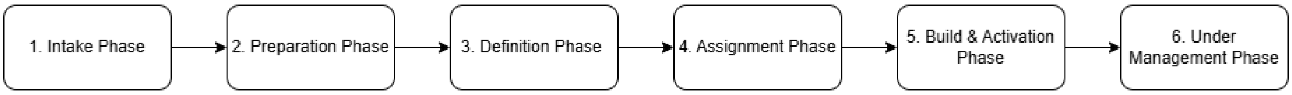


Figure 2.7: *Six phases of the Overlay process, from initial intake through to network management. Each stage involves distinct evaluations to ensure efficient rollout of new fiber areas.*

- Intake phase: In the intake phase, the goal is to determine which project from the next 12 quartiles planning should be started.
- Preparation phase: In the preparation phase, the feasibility of the project is determined.
- Definition phase: In the definition phase, the project is properly defined and the first actions towards installing the network infrastructure are taken.
- Assignment phase: In the assignment phase, the last details of the assignment for the contractor are administratively prepared, so the roll-out itself goes as smoothly as possible
- Build and Activation phase: In the build and activation phase, the fiber-optic network is rolled out and HP and HCs are made. Furthermore, these connections are also activated as part of the network, so that services can be delivered over the network.
- Under management phase: In the under management phase means that the new part of the network can be used and that the roll-out project is finished

2.3.2 Retrospective Connections Process

For the retrospective connections process, these always start with the selection of unconnected buildings that are already in the under management phase of the previous roll-out process. In Figure 2.8, the several process steps can be seen. The different steps are explained below the figure.

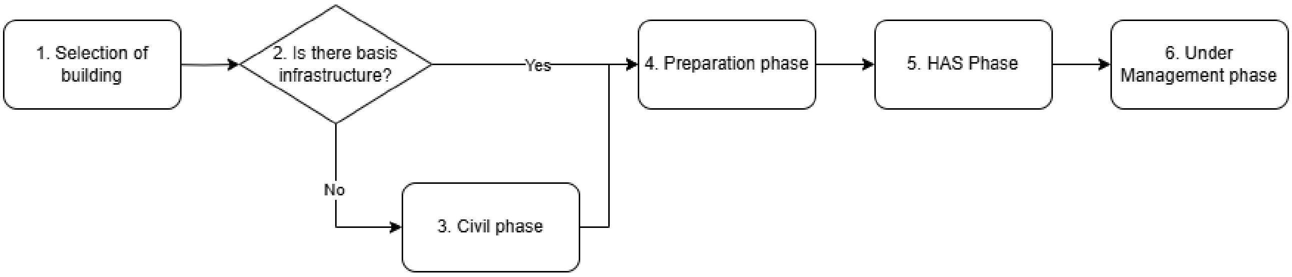


Figure 2.8: *Main steps in the retrospective connection process, beginning with building selection and concluding with the under management phase. This process specifically addresses previously skipped buildings.*

- Selection of building phase: Project selection of buildings based on connectivity and selection score
- Preparation phase: Getting permission to connect from building and municipality and prepare project.
- HAS phase: Building the new connections, sort of a small version of the Overlay process explained above
- Under management phase: In the under management phase means that the new part of the network can be used and that the project is finished

2.4 Selection Criteria

In this section, the selection methods for both the Overlay and retrospective connections categories are described in further detail.

2.4.1 General criteria

There are a couple of general criteria that are always taken into account when determining whether an area is viable for being connected to the fiber-optic network.

The key financial metric used in both selection processes is the selection score. Buildings with a high selection score are prioritized, meaning they are evaluated earlier than lower-scoring options. The selection model is a custom internally developed model by KPN that consists of several elements to predict the viability of an area. It was made at the start of the acceleration of the fiber-optic network to determine which areas were the best to roll-out first. This was based on expected costs and expected profits made of the network. The exact structure of the model can be found in Figure 2.9.

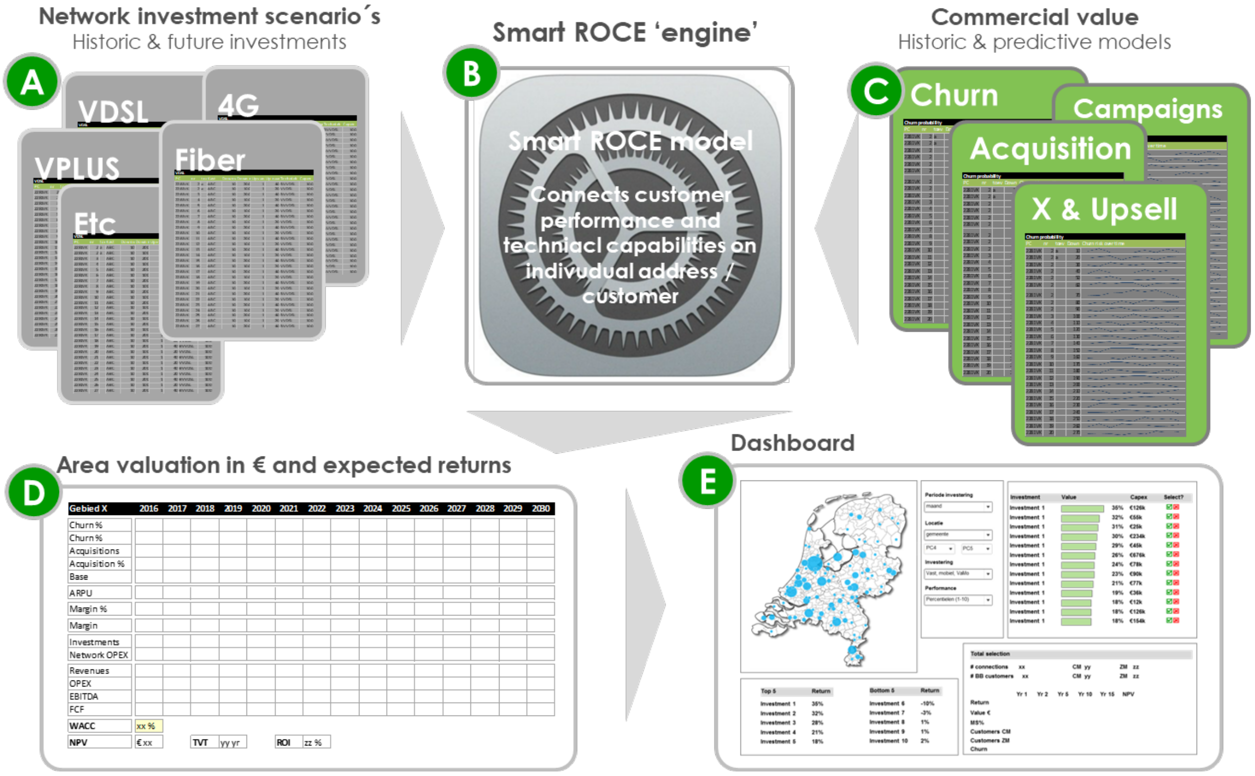


Figure 2.9: *Overview of the selection model, which integrates cost estimates and profit forecasts to assess the financial viability of area-level rollouts. It has historically guided KPN's strategic deployment.*

The cost side of the model is based on the investment made by rolling out the network. This is based on the following factors: Dutch road network, all addresses of an area that should be designed and business input for creating a Ftth (Fiber to the house) network. This generates the output of where to dig trenches, where to place which type of cables, where to place the POP (Point Of Presence) splitters and DP (Distribution Points) and detailed costs. The model itself first places the POP's in optimal places based on the minimum euclidean distance afterwards. Then it sums up all these steps and gives a total number for the costs made. A step by step overview of how the cost calculation algorithm works can be found in Figure 2.10.

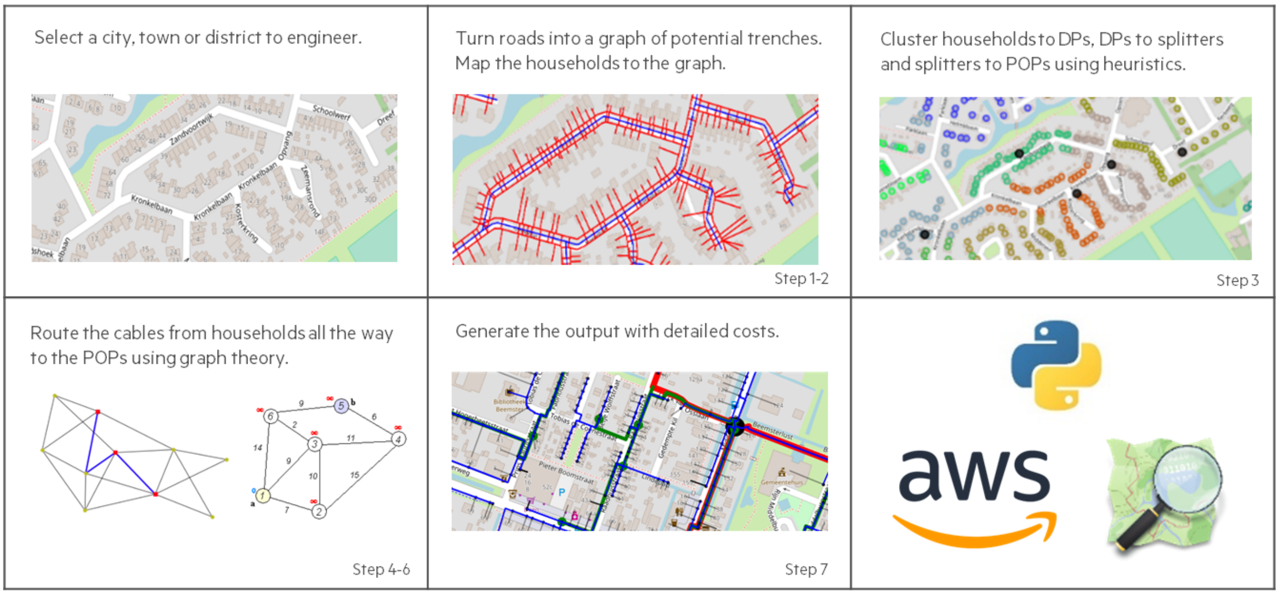

Select a city, town or district to engineer.
Turn roads into a graph of potential trenches. Map the households to the graph.
Step 1-2
Cluster households to DPs, DPs to splitters and splitters to POPs using heuristics.
Step 3
Route the cables from households all the way to the POPs using graph theory.
Step 4-6
Generate the output with detailed costs.
Step 7
aws

Figure 2.10: *Sequential calculation of costs within selection model, combining trench routing, cabling, and placement of network elements. This emphasizes how physical infrastructure planning underpins financial assessments.*

The profit side of the model is based on a logistic regression model with several input factors. The goal is to predict the number of households that will be connected extra compared to the current situation when KPN offers a fiber-optic connection instead of a copper connection. This is then simulated over the next 10 years. The outcome is calculated based on several factors. Examples of this are if the household is a current KPN customer, their internet speed, the possibility of a fiber connection, the probability that the customer leaves and much more. An overview can be found in Figure 2.11.

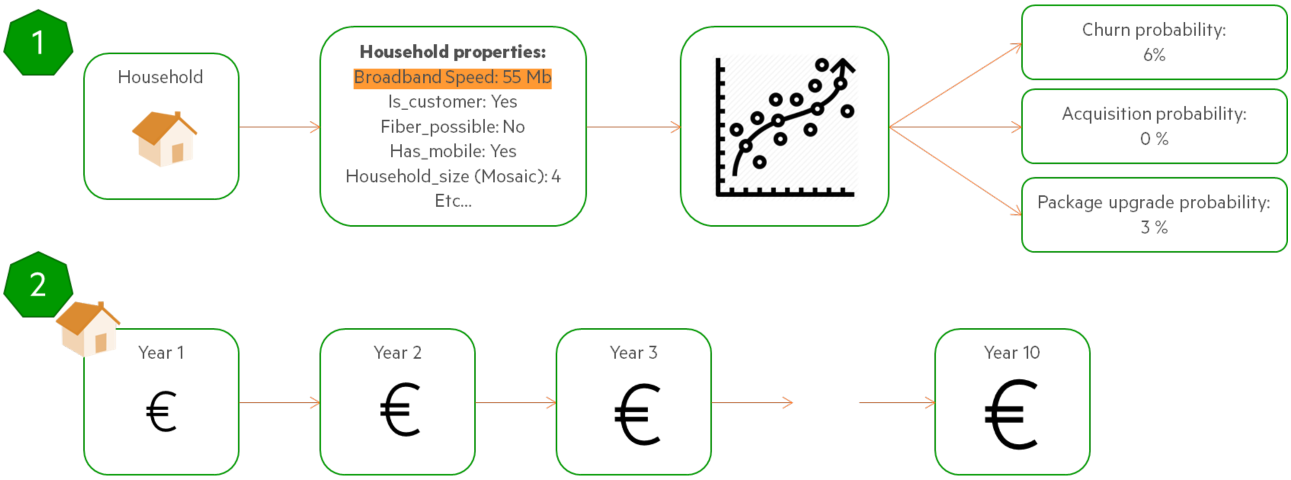


Figure 2.11: *Profit simulation in selection model using a logistic regression approach over a 10-year horizon. This captures expected incremental connections and revenue improvements from fiber upgrades.*

Furthermore, in the general criteria it is always important to take into account the current available capacity of the contractor. This to prevent delay in the roll-out process.

2.4.2 Overlay criteria

The selection of fiber-optic rollout areas under the *Overlay* category is based on a combination of geographic, technical, and economic criteria. These factors enable KPN to assess whether a specific area is suitable for efficient and profitable expansion of its fiber-optic network infrastructure.

KPN evaluates candidate areas using a predefined unit known as an *UrbanID*, which aggregates addresses at the neighborhood, district, or village level. The average selection model score across all addresses in the UrbanID serves as a key financial metric for viability assessment. Higher average scores indicate greater potential return on capital and thus prioritize the UrbanID for roll-out. In this decision, geographic proximity to the existing fiber-optic infrastructure also plays a critical role. Isolated areas, such as remote villages, typically incur higher connection costs and are therefore considered less attractive. In contrast, neighborhoods situated close to already connected areas or network nodes are favored due to lower marginal costs and ease of integration.

Another important consideration is the physical roll-out sequence. KPN typically initiates deployment at locations near a Point of Presence (PoP), which serves as a central distribution node, and then extends coverage outward. This method minimizes initial investment and leverages proximity for efficient resource allocation.

Contractor planning logistics are also integrated into the selection process. When possible, roll-out schedules are aligned such that contractors operate in adjacent or nearby regions. This spatial clustering of projects enhances operational efficiency and may reduce contracting costs, as dispersed work sites often lead to increased travel and mobilization expenses.

Finally, the availability and future viability of a suitable PoP location is assessed. In some cases, KPN may consider reusing an existing telephone exchange facility. However, this is contingent on whether the location is expected to remain operational in the long term. If the building is scheduled for decommissioning or presents high maintenance costs, a new site must be identified, which may affect rollout feasibility.

Table 2.3: *Summary of Overlay Selection Criteria*

Criterion	**Explanation**
UrbanID Score	Aggregated financial viability across grouped addresses using selection model.
Distance to Existing Network	Preference for areas adjacent to already connected regions to reduce infrastructure costs.
Logical Rollout Sequence	Rollout begins near the Point of Presence (PoP) and expands outward to optimize cost and planning.
Contractor Efficiency	Projects are clustered geographically to reduce operational inefficiencies and contractor expenses.
PoP Location Availability	Evaluation of whether an existing telephone exchange point can serve as a viable long-term PoP.
PoP Future Viability	Assessment of whether the PoP will remain usable, or if decommissioning or relocation is needed.

2.4.3 Retrospective connections criteria

For the retrospective connection category, a set of predefined selection criteria is used to determine which buildings are eligible for fiber-optic connection. These criteria help prioritize buildings with the highest potential for successful and cost-effective connection.

In addition to a high selection score and contractor capacity, a building must be in the under management phase and must not have been connected during the initial roll-out. Additionally, the building must fall under the responsibility of KPN or Glaspoort, which is a joint venture between KPN and the pension fund ABP, as only addresses managed by these entities are considered valid for retrospective connection.

Furthermore, the building must be classified as a high-rise building, which means it has a minimum of 10 residential units, to fall under the retrospective connection under management category. Finally, the RNA status (Reason Not Connected) or delivery status of the project can be an exclusion criterion. Buildings labeled with RNA 7 or delivery status 14 are typically excluded. RNA 7 means that there is technical obstruction, while delivery status 14 means that there is currently no fiber optic cable from the distribution point to the available building. A full overview of RNA statuses and delivery statuses is provided in Appendix A & B.

In Table 2.4 an overview of all selection criteria discussed in this section can be found.

Table 2.4: *Summary of Selection Criteria for Retrospective Connections*

Criterion	**Description**
Project Phase	The building must be in the under management phase.
Current Connection	The address must not have been connected during the initial roll-out.
Responsibility	The address must fall under the responsibility of either KPN or Glaspoort.
Building Type	The building must be a high-rise with at least 10 residential units.
Financial Viability	Prioritization is based on a high selection score.
RNA Status	RNA 7 (Technical obstruction) is excluded from selection.
Delivery Status	Delivery status 14 (No fiber available at building) is excluded from selection.

2.5 Contractor agreements

For the Overlay roll-out and the retrospective connection under management categories there are several contractors and contractual differences. The Overlay contractors and their agreements are highlighted in subsection 2.4.1. The retrospective connections under management which are handled in projects in subsection 2.4.2. Lastly the client-driven retrospective connections under management are covered in subsection 2.4.3.

2.5.1 Overlay agreements

KPN builds a fiber-optic network in areas in the entirety of the Netherlands. To facilitate all these areas they have contracts with several contractors that all are responsible for the roll-out of fiber-optic network in a specific area. These contractors all have slightly different contracts in the details, but in general they are the same. When making a new selection model the structure and the differences between the contracts should be taken into account. The Overlay contracts all have the following categories that are taken into account when determining the costs.

These are based on a standard project price that is always paid in combination with extra money that are project specific. There are several ledgers that are explained in the table 2.5.

Table 2.5: *Summary of Contract Criteria for Roll-out Projects*

Criterion	**Description**
Base Price	Base price for connecting an address.
Basis Infrastructure Length	The total length of fiber-optic cable required in public streets.
HAS Length	The length of fiber-optic cable from the public network to the property entrance.
Ground Type	Soil conditions such as sand, clay, or asphalt, which influence excavation difficulty and cost.
Street Covering	Types of paving or tiling that must be restored after digging.
Project Specifics	Unique on-site obstacles, such as trees, that require extra handling.
RNA Reductions	Compensation mechanisms if an address cannot be connected due to justifiable reasons.

2.5.2 Retrospective connections agreements

The first type of the retrospective connections agreements is the project-based type. These contracts are structured differently as they have previously already been in the Overlay phase and moved to the under management phase. There are two types of contracts available. These are the project-based type and the client-driven type. These are both very simple contracts. The project-based type uses the current delivery status of the address. There are only standard prices available for the delivery statuses 31, 33, 35, which are all for high-rise buildings. This is because most of the executed projects are for high-rise buildings All other project-based agreements are individual based and negotiated.

For the client-driven agreements there are standard prices for the most common delivery statuses. This is because the client-driven order often have majorly different starting points in terms of current delivery status. In general, the client-driven prices are higher than the project-based prices. This is largely due to the connection of the project-based type that is already closer to the building and the difference in contractors.

Table 2.6: *Key differences between Overlay and Retrospective connections agreements*

Aspect	**Overlay Agreements**	**Retrospective Connections Agreements**
Scope	New area-wide roll-out projects	Connecting previously skipped buildings (under management phase)
Pricing Structure	Standard contracts with base price plus variable costs (trench length, soil, obstacles)	Fixed prices only for certain delivery statuses (31, 33, 35); other cases negotiated individually
Main Building Focus	All building types in the area (low-rise, high-rise, duplex)	Primarily high-rise buildings ($\geq$ 10 residential units)
Contractor Role	Regional contractors with standardized area contracts	Contractors vary; costs depend on delivery status and project type (project-based or client-driven)
Typical Cost Drivers	Calculated from trench length, ground type, street covering, and site obstacles	Mostly determined by delivery status; routing costs estimated only if needed
RNA and Delivery Exclusions	RNA reductions apply as compensation for non-connections	Explicitly excludes RNA 7 (technical obstruction) and delivery status 14 (no fiber from DP)

2.6 Justification for a Predictive and Cost-Based Approach

The analysis in Chapters 1 and 2 shows that KPN's current approach for selecting buildings to connect to its fiber-optic network primarily relies on the selection method. This scoring tool, originally developed for area-level roll-out decisions, lacks the building-level granularity needed for today's brownfield environment. It neither explicitly predicts individual customer take-up probabilities nor provides detailed connection cost estimates, leading to inefficient capital allocation and potentially missed high-value opportunities.

The core challenge addressed in this thesis is to predict which buildings are most likely to convert to KPN fiber customers once connected. This problem spans a high-dimensional feature space involving socio-economic data, historical copper and fiber adoption, building typologies, and competitive infrastructure presence. Classical statistical methods, such as logistic regression, while transparent and interpretable, generally fail to capture the non-linear effects and complex feature interactions inherent in such data [15]. Similarly, rule-based scoring systems oversimplify decision boundaries, risking misclassification in heterogeneous urban contexts.

Simulation approaches, including discrete-event and agent-based models, are commonly employed in infrastructure planning and utility studies [12, 14]. However, these techniques require explicit behavioral assumptions and extensive calibration data, which are not available at the building level across millions of Dutch addresses. Moreover, executing such simulations at national scale would be computationally infeasible. Machine learning techniques, particularly ensemble models like LightGBM, are well-suited to this challenge. They efficiently process large, high-dimensional datasets, automatically uncover non-linear dependencies, and provide probabilistic outputs that support risk-informed decision-making. These advantages have been demonstrated in various telecom contexts, including customer churn and broadband adoption prediction [1, 23].

Identifying high-potential addresses is only part of the selection task. Effective fiber expansion also depends on the cost of physically connecting each location. While it is theoretically possible to formulate this as a Mixed Integer Linear Programming (MILP) problem to find globally optimal trenching and connection layouts [22], such formulations are computationally infeasible at national scale. MILPs are inherently NP-hard, and introducing each building-to-DP assignment or potential trench segment significantly enlarges the combinatorial space, quickly exceeding the capabilities of modern solvers [25]. Empirical studies show that even problems involving only a few thousand nodes can require hours to days to solve [11], making this approach impractical for continuous national-level planning. Stochastic or Monte Carlo simulations for assessing cost uncertainties [24] would similarly demand extensive local parameter data and computational effort.

Instead, this thesis employs heuristic routing approaches that approximate minimum-cost paths using distance metrics and domain-informed adjustments. Such methods provide sufficiently accurate cost estimates that align well with historical project data at KPN, while maintaining computational scalability [3]. This balance of predictive modeling for customer adoption with heuristic-based cost estimation forms a practical, interpretable, and scalable framework, well-matched to the financial and operational realities of brownfield fiber deployment.

Chapter 3 further explores the theoretical and methodological underpinnings of this data-driven strategy, situating it within the broader academic and technical literature on telecom infrastructure planning, machine learning applications, and heuristic network design.

2.7 Chapter Conclusion

This chapter has answered the research question by describing the current situation regarding KPN's selection and connection of buildings to its fiber-optic network. In the chapter, it is stated that the current selection process for both Overlay and Retrospective Connections is primarily based on the selection model, which combines cost and profit predictions to assess financial viability. However, this method is limited as it is not fit for predicting the financial viability of a specific building or street, which often happens in the retrospective connections category. Furthermore, the selection process varies between Overlay and Retrospective Connections, with each having its own process, criteria and contractor agreements.

Additionally, several factors such as contractor capacity, RNA statuses, and the complexity of connecting different building types have to be taken into account in the decision process for the new selection model. This is especially the case for high-rise and duplex properties as they play a significant role in the current gap between HP and HC. Furthermore, contractor agreements naturally have influence on the costs, with pricing partly driven by spatial factors like trench length and ground type.

Overall, the new selection process should be heavily data-driven, involving technical, financial, and organizational considerations. This further emphasizes the need for advanced analytical techniques to support decision-making. The next chapter builds on these insights by reviewing relevant literature on data-driven decision-making, machine learning, and operations research to search for solutions for the development of a new selection model for retrospective connections.

3 Literature Review

As established in Chapter 2, KPN's current approach to selecting buildings for fiber connection largely relies on the selection score. This score was designed for aggregated area-level decisions and does not explicitly account for individual customer acquisition probabilities, detailed connection costs, or feasibility constraints. This highlights a need for more granular, data-driven decision support tools that can guide roll-out investments in a brownfield environment where network infrastructure is only partially present.
This chapter addresses the second research question:

What is the theoretical background on data-driven decision-making and selection methodologies for prioritizing fiber-optic network connections?

It provides a theoretical framework based on a literature review to answer this research question. To arrive at an answer, the following sub-questions are formulated:

- *What does existing literature suggest about methods for prioritizing brownfield network expansions, and how do these differ from greenfield approaches?*
- *How are machine learning models leveraged to predict customer uptake or acquisition in telecom or similar network industries?*
- *What routing heuristics and cost estimation techniques are most commonly applied in infrastructure network planning, and how do they account for real-world trenching deviations?*
- *Which evaluation metrics are recommended in the literature for assessing predictive models in customer acquisition and infrastructure decision-making contexts?*

Section 3.1 discusses different selection methodologies with a particular emphasis on approaches tailored to brownfield contexts, highlighting the distinct challenges compared to new area (greenfield) roll-outs. Section 3.2 explores how machine learning models support customer uptake prediction and how routing heuristics contribute to more realistic cost estimations. Section 3.3 then reviews the evaluation metrics used to validate these approaches. Together, these insights shape the integrated solution approach presented in the next chapter.

3.1 Routing heuristics in fiber-optic network expansion

Building on the operational challenges identified in Chapter 2, this section reviews classical optimization approaches such as Minimum Spanning Tree (MST) and Steiner Tree models, as well as heuristic and rule-based strategies used in network planning. These insights clarify how different techniques address cost structures, routing complexities, and delivery feasibility, providing a foundation for the solution design presented later in this thesis.

There are several methods that have been created for the roll-out of a fiber-optic network. These are all mostly based on the green field approach [4]. The Greenfield approach means that you start the new product or model from scratch without having any previous product. In the case of the fiber-optic network papers this means that all papers and their models assume that there is currently no network. The first model of KPN, the selection model score, was also build on this principle. However, KPN has already covered 63% of the Netherlands with their network. This means that there already is an existing network. This corresponds to the Brownfield approach as described in [2], where new connections are established within areas that already have existing infrastructure [4]. This is the case for the current fiber-optic network situation in the Netherlands. Therefore, the methods found in the papers should be studied in a manner that takes into account the fact that the current network requires a Brownfield approach and not a Greenfield approach [5].

There are several known techniques for the expansion of a fiber-optic network in greenfield settings. These are discussed in the following subsections.

3.1.1 Minimum Spanning Tree

The Minimum Spanning Tree (MST) is a classical optimization approach in graph theory and has been widely applied in the design of telecommunication networks, including fiber-optic roll-outs. It provides a cost-effective means of interconnecting a set of nodes. For example, these nodes could represent connected addresses. The MST is performed by selecting a subset of edges that connects all nodes without forming any cycles and with the minimum possible total edge weight[26].

The MST problem is defined over an undirected, connected graph $G = (V, E)$, where V denotes the set of vertices and E the set of edges, each with an associated non-negative weight $w : E \to \mathbb{R}^+$. The objective is to find a tree $T \subseteq E$ such that all vertices in V are connected and the total cost is minimized:

$$\min_{T \subseteq E} \sum_{e \in T} w(e)$$

This is all subject to the constraint that T contains no cycles and spans all nodes in V.

This problem is often solved using greedy algorithms such as Prim's or Kruskal's algorithm. Either of these algorithms have polynomial-time solving times and are suitable for real-world deployment planning[2].

Overall, the MST approach has several advantages when applied to fiber-optic network design. It ensures cost efficiency by minimizing the total length (or cost) of fiber required to connect all points. Moreover, the simplicity and speed of MST algorithms make them attractive for use in automated planning tools, especially during the early phases of network design. However, MST also comes with some disadvantages. Since it results in a tree structure, it inherently lacks redundancy. This means that the failure of a single link can lead to disconnection of parts of the network. Additionally, MST does not account for capacity limitations, varying demand levels, or reliability requirements. All of these are important considerations in modern telecommunication infrastructure. As a result, while MST offers a strong baseline for cost-efficient planning, practical applications often extend or modify the model to incorporate these additional operational requirements.

Beyond MST, the Steiner Tree problem represents a more general formulation that allows for the inclusion of additional intermediate nodes (Steiner points) to potentially reduce the total network length. This approach can yield more cost-efficient layouts by optimizing shared infrastructure routes. However, the Steiner Tree problem is NP-hard and typically requires complex approximation algorithms or heuristics, which limits its practical applicability in large-scale or dynamically constrained environments such as brownfield fiber deployment[2].

3.1.2 Shortest Path Heuristics

The shortest path problem is a fundamental concept in graph theory and has been extensively studied in operations research and network optimization. It involves finding the path between two nodes in a graph such that the sum of the weights of its constituent edges is minimized. In the context of fiber-optic network planning, shortest path heuristics are commonly used to estimate the routing distance or cost from a building to the nearest distribution point (DP), particularly in brownfield environments where infrastructure is partially existing.

Formally, the shortest path problem is defined over a weighted, directed or undirected graph $G = (V, E)$, where V is the set of vertices and E is the set of edges, each associated with a non-negative weight function $w : E \rightarrow \mathbb{R}^+$. Given a source vertex $s \in V$ and a target vertex $t \in V$, the objective is to determine a path $P = (s, \ldots, t)$ such that the total cost is minimized:

$$\min_P \sum_{e \in P} w(e)$$

subject to P being a valid path connecting s and t.

This problem is typically solved using algorithms such as Dijkstra's algorithm for graphs with non-negative edge weights, or the A* algorithm when heuristic guidance is available to speed up computations[2]. These algorithms have polynomial-time complexity and are widely implemented in routing software and network planning tools.

Shortest path heuristics offer several advantages for fiber-optic network design, especially in retrospective connection scenarios. They are computationally efficient and can be applied independently for each building, making them well suited for large-scale datasets involving millions of addresses. Furthermore, they align with practical engineering approaches, where each building is often connected to the nearest feasible DP to minimize local construction costs.

However, shortest path approaches also have limitations. Unlike global optimization models such as the MST or Steiner Tree, they do not consider the aggregate cost across multiple buildings or potential cost savings from joint trenching. This means that while they provide quick and realistic estimates at the individual building level, they may overlook network-wide efficiencies. As a result, shortest path heuristics are most effective when used as part of a broader planning framework that also considers clustering or joint deployment opportunities.

3.1.3 Passive Optic Network Design

Passive Optical Network (PON) design is a part of fiber-optic infrastructure planning. It is mostly applicable in last-mile access networks. A PON is a point-to-multipoint network architecture in which a single optical fiber is split passively to serve multiple end-users. This eliminates the need for active components between the central office and users. The advantage of this is that it reduces maintenance and energy costs while increasing reliability. In network optimization literature, PON design is often formulated as a facility location or network design problem, where splitters (intermediate points) and routes must be strategically placed to minimize costs while satisfying demand and coverage constraints [21].

Mathematically, the PON design problem is typically modeled using integer or mixed-integer linear programming (MILP). Let N be the set of customer locations, F the set of potential splitter locations, and C_{ij} the cost of connecting customer $i \in N$ to splitter $j \in F$. Define binary decision variables x_{ij} indicating whether customer i is connected to splitter j, and y_j indicating whether splitter j is activated. The objective is to minimize total connection and deployment costs:

$$\min \sum_{j \in F} f_j y_j + \sum_{i \in N} \sum_{j \in F} C_{ij} x_{ij}$$

subject to:

$$\sum_{j \in F} x_{ij} = 1 \quad \forall i \in N$$

$$x_{ij} \leq y_j \quad \forall i \in N, j \in F$$

$$\sum_{i \in N} x_{ij} \leq K_j \quad \forall j \in F$$

where f_j is the fixed cost of deploying splitter j, and K_j is its capacity[18].

This modeling framework uses core planning objectives such as minimizing capital expenditures, enforcing customer coverage, and respecting splitter capacities. More advanced models can incorporate additional constraints, such as maximum distances between customers and splitters and hierarchical splitter levels[19].

The advantages of PON design lie in its scalability and cost-efficiency. Passive components lower operational costs, and the shared infrastructure model reduces the need for extensive cabling. Moreover, the ability to model the problem using MILP provides a clear optimization path with exact or near-exact solutions for moderate-sized instances.

However, PON design also presents challenges. The problem becomes computationally complex for large-scale networks, especially when dynamic or real-time data is introduced. In addition, while PONs reduce costs, they also reduce flexibility; failures in a shared segment affect all downstream users which makes fault tolerance a critical issue. As a result, Hybrid models and heuristics are increasingly explored to address these concerns in practical implementations [17].

3.1.4 Recommended approach for brownfield cost estimation

The reviewed techniques each have particular advantages for network planning and cost estimation. Minimum Spanning Trees (MST) and Steiner Trees are suitable for minimizing total network length in greenfield scenarios but assume that all nodes require connection, which does not correspond to brownfield environments with existing infrastructure and selective connections. Passive Optical Network (PON) heuristics can incorporate additional design considerations such as splitter placement and capacity constraints but introduce complexity that does not align with building-level selection. Shortest path heuristics provide a practical alternative by estimating the cost of connecting individual buildings to the nearest distribution point. These heuristics are computationally efficient and can be applied across large datasets, which makes them suitable for settings where cost differentiation at the address level is important. This aligns with the operational context outlined in Chapter 2 and supports the data-driven selection methodology developed in Chapter 4.

3.2 Usage of machine learning for customer acquisition predictions

To complement cost considerations, this section explores how machine learning models can improve selection methodologies by predicting customer acquisition probabilities. It focuses on applications relevant to fiber-optic network decision-making, highlighting why approaches such as decision trees, random forests, and gradient boosting are well-suited for capturing complex interactions in customer data. This directly supports the development of a predictive component that integrates with cost and feasibility assessments.

3.2.1 Regression models

Regression models are among the most widely used tools in data-driven analysis. There main advantages are in their mathematical simplicity and interpretability[8]. Linear regression models the relationship between a continuous dependent variable and one or more independent variables by fitting a linear equation to the data. The model can be expressed as:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n,$$

where $\hat{y}$ is the predicted output, $x_1, \ldots, x_n$ are the predictor variables, and $\beta_0, \ldots, \beta_n$ are the coefficients estimated via ordinary least squares minimization [8]. The interpretability of linear regression makes it a preferred baseline in many domains, allowing for straightforward evaluation of feature contributions and confidence interval estimation. It is computationally efficient and performs well when the underlying relationships are approximately linear[8].

Logistic regression extends the linear modeling framework to binary or categorical outcomes by modeling the log-odds of class membership as a linear function of the input variables. The predicted probability is derived using the logistic (sigmoid) function,

$$\sigma(z) = \frac{1}{1 + e^{-z}},$$

where $z = \beta_0 + \sum_{i=1}^{n} \beta_i x_i$. The model is typically estimated through maximum likelihood optimization [8]. Logistic regression is widely applied in classification tasks due to its probabilistic output, interpretability and the low computational cost.

Despite their advantages, both linear and logistic regression have limitations. They assume a linear relationship between predictors and the response, which may not capture complex or non-linear interactions in real-world data. Furthermore, multi-collinearity among predictors can distort coefficient estimates, and performance may degrade in high-dimensional or highly unstructured datasets. However, these models still provide a transparent and efficient first solution before applying more advanced techniques.

3.2.2 Random Forest

Random Forest is a widely adopted ensemble learning algorithm that enhances the performance of individual decision trees by aggregating the predictions of multiple trees to reduce variance and improve generalization. The method constructs a large number of decision trees during training. Each each tree is trained on a bootstrap sample of the data [9]. At each node within a tree a random subset of features is selected for splitting. This introduces both data and feature randomness. This stochasticity supports the robustness of the ensemble and reduces overfitting which is normally common with single decision trees. As seen above, In regression tasks the model outputs the average prediction of all trees. However, in classification, it returns the majority vote. The predictive function of a Random Forest regressor can be formalized as

$$\hat{f}(x) = \frac{1}{T} \sum_{t=1}^{T} f_t(x),$$

where f_t denotes the prediction of the t-th tree and T represents the total number of trees in the ensemble [9].

A key advantage of Random Forest lies in its ability to model complex, non-linear relationships without requiring extensive data pre-processing or manual feature engineering. The algorithm also offers built-in mechanisms for estimating feature importance. This can help with interpretability and variable selection. Furthermore, it is relatively robust to outliers and capable of handling large, high-dimensional datasets. However, Random Forests also have certain disadvantages. They can become computationally intensive with a large number of trees and high-dimensional data, particularly in prediction time. In addition, the ensemble structure may hide model interpretability compared to simpler models. Lastly, the presence of highly correlated trees may reduce the marginal gain from additional ensemble members.

Overall, Random Forest remains a versatile and powerful technique for predictive modeling across a range of domains. This particularly the case when the underlying relationships in the data are complex and not easily captured by parametric methods.

3.2.3 Extreme Gradient Boosting

Extreme Gradient Boosting (XGBoost) is a powerful and scalable machine learning algorithm based on the gradient boosting framework. It is designed to deliver high predictive performance with efficiency and robustness. Introduced by Chen and Guestrin (2016), XGBoost builds an ensemble of decision trees sequentially, where each tree corrects the residual errors of the aggregated predictions from previous iterations [10]. The algorithm optimizes a regularized objective function, which combines a differentiable convex loss function L (such as mean squared error) with a regularization term Ω that penalizes model complexity. The objective function at iteration t is typically defined as

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t),$$

where f_t represents the newly added tree and $\hat{y}_i^{(t-1)}$ the prediction from the ensemble before iteration t. XGBoost uses several improvements over traditional gradient boosting methods, including second-order optimization using Hessian information, column sub-sampling, shrinkage (learning rate), and efficient handling of missing values and sparse data[10]. These improvements help in its its ability to model complex, non-linear relationships while controlling overfitting.

Despite its evident advantages, XGBoost also has some disadvantages. The algorithm can be sensitive to hyperparameter tuning. This requires careful adjustment of learning rate, tree depth, and regularization parameters to achieve optimal results. Furthermore, due to it being a sequential algorithm training times may increase significantly on very large datasets or in cases where the number of boosting rounds is high. The model may also be less interpretable than simpler algorithms, such as linear regression or single decision trees. However, tools exist to mitigate this drawback through feature importance analysis.

Overall, XGBoost is well-suited for structured data prediction tasks where accuracy, scalability, and the ability to model feature interactions are critical, particularly in domains involving heterogeneous input features, cost sensitivity, and non-linear relationships.

3.2.4 Light Gradient Boosting Machine

Light Gradient Boosting Machine (LightGBM) is another efficient gradient boosting framework developed by Microsoft. It is designed to improve scalability and speed for large-scale and high-dimensional datasets [16]. Like other gradient boosting methods, LightGBM builds an ensemble of decision trees in a sequential order. Each tree aims to correct the residuals of the combined predictions of its predecessors. It is different from traditional approaches by using a histogram-based method for finding split points and a leaf-wise tree growth strategy. This often leads to better accuracy and faster training compared to level-wise growth [16].

The algorithm minimizes a regularized objective function composed of a convex loss function L and a regularization term Ω, defined similarly to XGBoost as

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t),$$

where f_t denotes the tree added at iteration t, and $\hat{y}_i^{(t-1)}$ is the current prediction of sample i. Optimization is performed using both first-order and second-order derivatives of the loss function. This allows for faster convergence and better handling of complex patterns in the data [16, 6].

One of the main advantages of LightGBM is its native support for categorical features, which are internally mapped to discrete bins. This often means that it is not necessary to use one-hot encoding and this feature also improves memory efficiency. This feature makes it suitable for structured datasets where categorical variables are present. [16]. Furthermore, LightGBM's histogram-based algorithm reduces computational complexity by grouping continuous values into bins, which accelerates the split-finding process.

However, LightGBM has limitations. Its aggressive leaf-wise tree growth can lead to overfitting, particularly in small or noisy datasets. To counteract this, hyperparameters such as max_depth must be tuned carefully. Furthermore, while the histogram-based splitting improves speed, it may lose some granularity compared to exact algorithms [6].

In summary, LightGBM is a robust and scalable algorithm for structured data prediction tasks. It works best in scenarios where training speed, memory efficiency, and predictive performance are important. Its balance of computational efficiency and modeling power makes it a strong choice for many industrial applications.

3.2.5 Recommended approach for customer acquisition prediction

The machine learning techniques reviewed each offer distinct strengths for customer acquisition prediction. Regression models, particularly logistic regression, provide clear interpretability and transparency, but their linear assumptions hinder accurate modeling of the complex and nonlinear relationships typically observed in customer data. Random Forest methods overcome linear limitations through ensemble learning, effectively capturing complex feature interactions, yet their black-box nature and computational demands may limit practical usability at scale. Similarly, XGBoost excels in predictive performance, especially for structured tabular data, but necessitates extensive hyperparameter tuning and computational resources, making it less practical for rapid iterative model development on large-scale datasets. Conversely, LightGBM addresses these concerns by efficiently handling high-dimensional and categorical data, demonstrating superior computational efficiency and scalability. Given these advantages, LightGBM aligns closely with the operational requirements of address-level customer prediction outlined in Chapter 2 and effectively supports the data-driven selection methodology detailed in Chapter 4.

3.2.6 Evaluation Metrics of Machine Learning Models

To evaluate the predictive performance of the machine learning models used in this thesis, multiple classification metrics are applied. As the number of addresses that do convert to KPN customers exceeds those that do not metrics such as overall accuracy are not suitable. Instead, the evaluation focuses on three key indicators: the AUC score, F1-score, and precision-at-k.

The Area Under the Receiver Operating Characteristic Curve (AUC) shows the model's ability to correctly rank positive examples above negative ones across all possible threshold values. An AUC score of 1.0 indicates perfect discrimination, while a score of 0.5 represents random guessing. This metric is threshold-independent and is widely used in machine learning applications for customer prediction tasks due to its robustness in imbalanced settings [27, 28].

The F1-score is a threshold-dependent metric that balances precision and recall. Precision refers to the proportion of predicted positives that are actually correct, while recall measures the proportion of true positives that are correctly identified. The F1-score is the harmonic mean of these two, providing a single measure that accounts for both false positives and false negatives. In this thesis, the macro F1-score is reported, which averages F1-scores across classes without weighting by class frequency. This is important when both the KPN and non-KPN classes are considered equally relevant from a business perspective [27].

Lastly, precision-at-k is used to assess how well the model ranks the top $k\%$ of addresses by predicted probability. This metric answers the practical question: “Of the top-k ranked addresses, how many are actual fiber customers?” It is particularly relevant in fiber optic network expansion situations like this one where limited budgets or operational constraints require selecting a fixed number of high-priority addresses. Precision-at-k is therefore a decision-oriented metric that bridges model output with real-world implementation [28].

Together, these metrics provide a comprehensive view of model performance. AUC captures overall discriminative ability, F1-score balances sensitivity and specificity at a chosen threshold, and precision-at-k reflects how well the model supports prioritization under constraints.

3.3 Chapter Conclusion

This chapter reviewed existing literature on fiber-optic network planning, focusing on two key decision-making dimensions: routing heuristics for cost estimation and machine learning models for predicting customer acquisition. Routing heuristics were critically compared, highlighting the suitability of shortest path heuristics for brownfield cost estimation due to their computational efficiency and adaptability to selective, address-level decision-making contexts. In contrast, Minimum Spanning Trees (MST) and Steiner Trees, while effective in greenfield scenarios due to their ability to minimize total network length, are limited by their assumption of mandatory connections for all nodes. Similarly, Passive Optical Network (PON) heuristics provide a comprehensive framework by incorporating additional design considerations such as splitter placement but introduce unnecessary complexity at the granular building level.

Regarding customer acquisition prediction, various machine learning methods were evaluated. Regression models such as logistic regression offer clarity and interpretability but are limited by linear assumptions. Ensemble methods like Random Forest and Extreme Gradient Boosting (XGBoost) address this limitation through robust handling of non-linear relationships and feature interactions, yet their complexity and computational intensity may hinder practical applicability at scale. Light Gradient Boosting Machine (LightGBM) effectively resolves these trade-offs, providing excellent predictive performance while maintaining computational efficiency and scalability for large datasets and high-dimensional feature spaces. However, its sensitivity to hyperparameter tuning and potential for overfitting necessitate careful validation and calibration.

Considering these factors, this research recommends shortest path heuristics for routing-based cost estimation and LightGBM, Random Forest or XGBoost will be tested for customer acquisition prediction. The combination of these methodologies offers a balanced approach, aligning closely with the operational requirements outlined in Chapter 2 and supporting the integrated, data-driven selection methodology developed in Chapter 4.

4 Solution Design

This chapter addresses the third research question:

How can a data-driven scoring methodology be developed to support strategic decision-making for fiber-optic network connections at the building level?

This chapter presents the design of a data-driven solution to improve the selection of buildings for fiber-optic connection. Based on the insights gained from the current situation and the theoretical background, this chapter outlines how data are collected, processed, and translated into a predictive and financially-informed selection model. To answer the main research question, the following sub-questions are addressed:

- *How can relevant data sources be collected and transformed into useful input variables for the selection model?*
- *How can the potential for customer acquisition be predicted at the address level?*
- *How can connection costs be estimated using existing infrastructure and project data?*
- *How can contextual feasibility constraints be quantified and integrated into the scoring methodology?*
- *How can the three model components be combined into a combined prioritization score?*
- *How can the selection model be implemented in a decision-support tool to support fiber-optic network expansion planning?*
- *How can the performance and robustness of the developed selection methodology be validated?*

Section 4.1 describes the data collection process and the identification of relevant input variables. Section 4.2 outlines the data preparation steps, including cleaning, feature selection, and transformation. Section 4.3 introduces the prediction models used to estimate financial viability and connection likelihood. Section 4.4 presents the design of the selection methodology that integrates these predictions into a decision-support tool. Section 4.5 explains how the methodology is validated and evaluated. Finally, Section 4.6 summarizes the outcomes of the solution design and its implications for the following results chapter.

4.1 Approach

To solve the core problem found in Chapter 1, this section explains the method used to develop a data-driven selection model for retrospective fiber-optic network connections. The goal is to capture all necessary elements that are also taken into consideration at various stages in the current scoring method as mentioned in section 2.6. Therefore, the model has to prioritize buildings on several aspects. To do this, a modular scoring system is developed consisting of three components. These are predicted costs based on the end-state of the previous project, the customer potential in a building and the contextual feasibility of connecting a building. Each component addresses a different dimension of the selection problem and contributes to a prioritization score that is used to rank buildings for project consideration. The Customer Uptake Prediction Models and Fiber Routing Cost Models are combined to indicate financial viability while the feasibility model is used to assess the operational feasibility of a residential address. The entire approach is also visualized in Figure 4.1

- **Connectivity potential** is estimated using a machine learning model that predicts the probability that an address becomes a customer after being connected. This model helps to take into account customer behavior and market dynamics into the selection process. The output is a probability score per address. Values closer to 1 show that there is a better chance of customer acquisition at an address.
- **Cost estimation** is done by using a heuristic method based on spatial routing, contract pricing, and previous project status. Each address has a predicted monetary connection cost. The cost of the connection depends on the state that the connection was left in at the previous project. This means that it is either a standard contractual agreement or that certain features still should be estimated based on the infrastructure of the network that still needs to be done.
- **Contextual feasibility** reflects the operational and technical conditions that may hinder successful connection. This includes demolition status, overbuild (presence of competitor fiber), and delivery or RNA-related constraints. A scoring system is developed to assign a colored feasibility score where scores from green to red are given out which indicate the roll-out risk or infeasibility.

The output of the connectivity potential and cost components are scaled to a common range, allowing the two parts to be combined into a combined prioritization score S_i, as described in Section 4.3.4. In Chapter 4.5, it is explained how this score and the operational feasibility indicator are integrated in to a Power BI dashboard. The goal of the dashboard is to give a clear overview of the value of buildings and also be able to automatically form viable projects by combining several good buildings in close proximity together into a project for efficiency purposes.

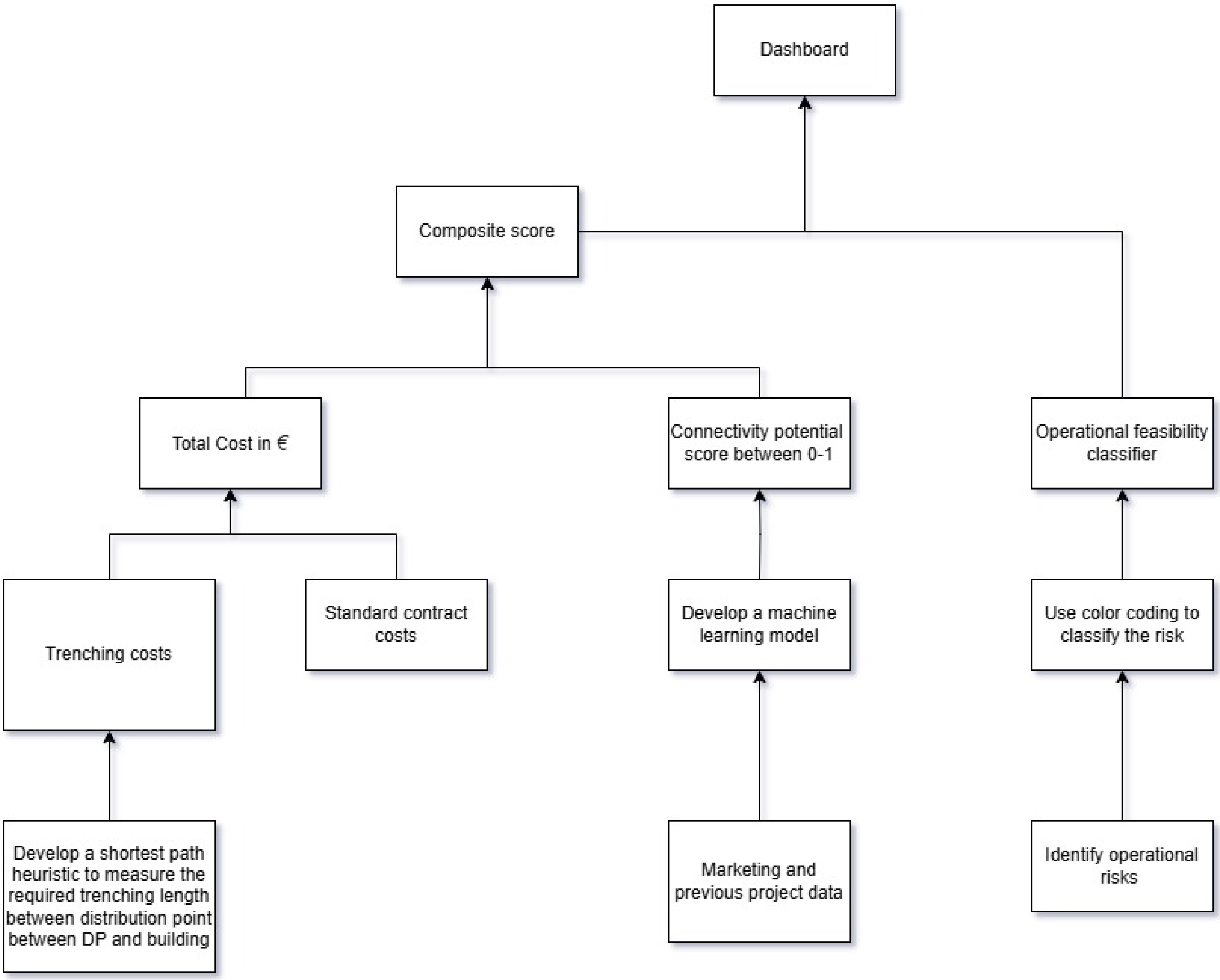

Dashboard
Composite score
Total Cost in €
Connectivity potential score between 0-1
Operational feasibility classifier
Trenching costs
Standard contract costs
Develop a machine learning model
Use color coding to classify the risk
Develop a shortest path heuristic to measure the required trenching length between distribution point between DP and building
Marketing and previous project data
Identify operational risks

Figure 4.1: *Integrated model approach combining a machine learning-based uptake prediction, a heuristic cost estimation, and a feasibility assessment. These components jointly inform a composite prioritization score for building-level roll-out.*

4.2 Data Processing & Feature Engineering

This section explains how the raw data is transformed into usable datasets as input for the selection model. Since the model consists of three components, the data preparation was performed such that each component has the required inputs from the different datasets. The exact variables used are listed in Appendix D. Below a detailed explanation of the processing and feature engineering strategy per component can be found.

4.2.1 Data Sources and Initial Processing

The raw datasets used in this thesis are retrieved from various internal KPN data warehouses. These include fiber roll-out data from previous projects, network delivery and infrastructure status per address, customer and subscription classifications, geographic coordinates, RNA codes, contractor zoning and a set of external socio-demographic indicators. Each dataset has a specific granularity. Most of the data is stored on address level, while a small part of the data is aggregated to building or region level.

The datasets were merged by either using the address or the building id as the merging key. To improve the data quality, several processing steps are performed over all sets: Addresses missing essential identifiers or coordinates were removed, duplicates were resolved, and addresses that are only consumer addresses, so no business-related properties. This ensures that the model only considers addresses that are relevant for retrospective connection projects.

After the processing of the datasets, the main dataset contains approximately 9 million address-level records. These are aggregated for some parts of the model such as the cost component. This aggregation is necessary, as retrospective fiber connections projects evaluate entire high-rise buildings instead of the individual addresses within the building. Afterwards for all components feature engineering was carried out with three goals in mind: To increase the predictive power, reflect the operational and commercial constraints and to maintain compatibility with the modular structure of the model.

4.2.2 Feature Engineering for Connectivity Potential

The first model component predicts whether a building will result in more customers using the fiber-optic network. This can be either KPN customers or wholesale customers. To enable this supervised learning model, features were engineered that capture historic customer behavior, address characteristics, and previous roll-out data.

- **Categorical variables** such as customer status, building type, and address typology were encoded either as ordinal categories or one-hot encoded variables, depending on their current data structure.
- **Demographic and financial indicators**, including income brackets, property value estimates (WOZ), and purchasing power segments, were linked at the address level. These variables were used to try to capture customer value and marketing attractiveness.
- **Interaction terms** were added to reflect nonlinear behavioral patterns. For example, overbuild presence combined with current customer type allows the model to detect areas where switching behavior is unlikely due to competitors already being present in that area. Similarly, income and property value were combined to represent socioeconomic potential more robustly than either metric in isolation.
- **Imbalance control and noise filtering** were applied through categorical grouping, thresholding of rare classes, and stratified sampling to prevent performance degradation during model training.

Feature inclusion was based on both theoretical reasoning and data analysis of the results of earlier stages of the model development. Features with low variance or poor correlation with the target variable were excluded prior to training. Additionally, early versions of the model were used to validate variable importance using SHAP values and F-statistics. This way non-informative features could be removed to speed up model performance.

4.2.3 Feature Engineering for Cost Estimation

The cost estimation component is based on the current available prices for the retrospective connections projects. Currently, there are only contractual prices for a few delivery statuses.

These are delivery status 30 until 35 which have fixed prices. For most of the other statuses the cost estimation must be determined by estimating the price. This is done with geo-spatial and contractual input features. Before any routing or cost calculation logic, the following processing and feature engineering steps are taken to get the right data formatting.

- **Building centroid coordinates:** For each building, geographic coordinates were derived by averaging the latitude and longitude of all addresses within a building. Some of the coordinates from the databases are formatted according to the Rijkdriehoekscoordinaten standard. These had to be formatted to longitude latitude formatting for further use. The coordinates are used as a spatial reference point for later distance computations.

- **Building size:** The total number of addresses per building was calculated and stored as a numeric feature. This is done to keep track of the addresses within a building and to be able to connect all model components in the end. Also, this makes it easier to compare the cost between high-rise buildings and low-rise buildings as they more often have only one address per building.

- **Delivery status:** Each address from a previous project has a delivery status code. This code indicates the present infrastructure present at an address or building. These codes should be the same for most buildings as the state should be same for all addresses within a building under normal conditions. However, in some cases this was not done. When this is the case the delivery statuses were encoded numerically or grouped into categories and aggregated to the building level. This is then used as the general indicator of the status of a building. Furthermore, this status also determines whether it is necessary to calculate the distance between the distribution and the building.

- **Contractor zone:** For each building a contractor zone is determined based on the delivery status of the building, as certain contractors often handle certain delivery status projects. With this is becomes easier to take into account the capacity of the contractor if the current project is added.

Together these features help with the Fiber Routing Cost Model.

4.2.4 Feature Engineering for Contextual Feasibility

The final component evaluates the technical and strategic feasibility of connecting a building. This part of the model incorporates business rules, known constraints, and risk signals that may affect a building's eligibility or desirability for connection. Examples of this are the following:

- **RNA codes** were evaluated within the department and grouped by severity. Certain codes, such as RNA 7 (technical obstruction) and RNA 16 (demolition), were flagged as hard blockers and used to assign feasibility scores of zero.

- **Delivery status values** were also reviewed to detect addresses that are technically unreachable due to missing infrastructure such as a vacant building status. This should always be checked as this status could also be given in case of a renovation which means the status is outdated

- **The overbuild feature** is included to indicate where competitors have already deployed fiber infrastructure. These features were used as soft penalties, reducing the feasibility score without outright exclusion.

This component ensures that buildings included in the final prioritization are not only financially and commercially viable but also technically feasible for retrospective connections project under realistic operational constraints. The entire overview of all the combinations can be found in Appendix D.

4.3 Model Formulation

In this section, the model formulation itself is discussed. The dataset resulting from the feature engineering steps in Section 4.2 is used as the input for the modeling components of the selection model.

4.3.1 Customer Uptake Prediction Model

The goal of the first component of the selection model is to predict whether a building will result in new fiber-optic customers once connected. These customers may be part of KPN's consumer or small-business market, or part of the wholesale network. The model assigns a probability score $P_i \in [0, 1]$ to each building, indicating the likelihood of customer acquisition from a building.

A binary classification model is trained to estimate this probability. The positive class ($y = 1$) includes all addresses that became either a KPN CM, KPN BM, or wholesale fiber customer after being connected to the fiber-optic network. The negative class ($y = 0$) includes addresses that were technically connected ($\texttt{HC} = 1$) but did not lead to customer acquisition. Restricting the training set to addresses with prior connection availability ensures the model reflects actual customer behavior rather than technical eligibility.

The prediction target y_i for each address i is therefore defined as:

$$y_i = \begin{cases} 1 & \text{if address } i \text{ became a fiber customer (KPN or wholesale)} \\ 0 & \text{otherwise} \end{cases}$$

The model will be tested using several machine learning techniques for comparison purposes. The tested models are Random Forest, XGBoost and LGBM. The starting version of the model is made with Random Forest due to its inherent simplicity and to function as a benchmark for the other models. Afterwards implementations for both LGBM and XGBoost are made. These are probably more suited due to their stronger performance working with large datasets. All processing and feature engineering steps described in Section 4.2 are applied within a single training pipeline, which includes encoding of categorical variables, imputation of missing numeric values, and processing of ordinal features.After training the base model, we apply post-hoc probability calibration using CalibratedClassifierCV with sigmoid calibration. This improves the reliability of predicted probabilities and is critical for evaluating roll-out risk based on predicted customer conversion likelihood. Rather than defaulting to 0.5, we tune the classification threshold using macro-averaged F1 score. This ensures the model balances performance across the minority (non-converting) and majority (converting) classes. The final threshold was selected as the one that maximized macro F1 on the validation set. Although early experiments used a full sklearn pipeline, we switched to native LightGBM training without preprocessing pipelines. This yielded better class performance (especially for class 0) and allowed native handling of categorical variables. We conclude that LightGBM's internal optimizations perform better when raw categorical types are used directly.

Let $X \in \mathbb{R}^{m\times n}$ be the feature matrix and $y \in \{0,1\}^m$ the target vector. Let $X \in \mathbb{R}^{m\times n}$ be the feature matrix, where m denotes the number of samples (addresses) and n the number of features (predictors), and let $y \in \{0,1\}^m$ be the target vector indicating customer acquisition. The pipeline π transforms raw features into model-ready inputs before training a classifier f_θ.

To improve model performance, hyperparameters are tuned using the Optuna framework. The optimization process maximizes the Area Under the Receiver Operating Characteristic Curve (AUC) on a stratified validation set. Parameters such as learning rate, number of estimators, number of leaves, and sampling ratios are optimized. The final configuration is selected based on validation AUC and used to train the model on the full training split.
Formally, the optimization objective is:

$$\theta^* = \arg\max_{\theta} \mathrm{AUC}(f_\theta(X_{\mathrm{val}}), y_{\mathrm{val}})$$

where θ denotes the set of hyperparameters selected through Optuna.

Model evaluation is based on AUC, precision, recall, k-precision and F1-score. A confusion matrix and ROC curve are also generated to visualize the model's performance at various thresholds. The final output consists of a probability score P_i per address, which is later aggregated to the building level and used as one of the inputs to the composite prioritization score described in Section 4.3.4.

4.3.2 Fiber Routing Cost Model

The goal of the cost estimation component is to approximate the monetary cost of connecting a given building based on its current delivery status. Some of these statuses can be approximated by simply referring to the contractual agreement for that delivery status. For other more difficult statuses it means that routing to the nearest available fiber distribution point (DP) is necessary for a good approximation. Since this is highly variable it is hard for KPN to estimate the associated costs for such buildings. For this reason, proxy variables are constructed based on spatial routing distance, infrastructure availability, and expected contractor pricing rules.

The primary approximation for the connection cost in case is the shortest path distance from each building to the nearest DP. This is computed over the national street network graph. For this purpose, a detailed OpenStreetMap (OSM) based walkable street graph of the Netherlands was loaded using `osmnx` package. This package has nodes and edges in most streets and from those nodes routes can be made between points.

The routing itself was conducted with the shortest path heuristic, for which several different algorithms are available within the `osmnx` package. The different routing methods are tested in Chapter 5 as part of the experiments. The methods tested are shortest path heuristic, Dijkstra's shortest path heuristic and minimum spanning tree. For computational efficiency, buildings and DPs were first mapped to the closest available point, and a KD-tree was used to identify the nearest five DPs to each building. A KD-tree is a data structure designed to organize spatial data for efficient nearest-neighbor searches [7]. The true shortest route was then computed using networkx, considering only those within a maximum 5 km cutoff.

Formally, for each building i with coordinates (x_i, y_i), a KD-tree is queried to return the five closest candidate DPs:

$$\mathcal{N}_i = \arg\min_{j\in\mathcal{J}}{}^{(5)} \|(x_i, y_i) - (x_j, y_j)\|_2$$

Dijkstra's algorithm is then applied from the building node to each DP node in $\mathcal{N}_i$, yielding path distances $d_{i,j}$. The minimum of these distances is selected:

$$d_i^{\min} = \min_{j\in\mathcal{N}_i} d_{i,j}$$

The minimal routing distance identified for each building–DP pair was multiplied by a unit trenching cost factor per meter. This rate is selected based on an internal contractor cost assumption and reflects both material and labor expenses under typical roll-out conditions. The estimated cost is calculated as:

$$C_i = d_i^{\min} \cdot \lambda$$

where $\lambda = x$ euros per meter.

The resulting value serves as the estimated total cost of infrastructure roll-out for the building. This is then added to the cost of the new status that the building becomes once the basic infrastructure is placed. In addition, the number of addresses associated with each building was used to compute a per-address cost metric. This enables easier comparison between large and small buildings. The per-address cost is given by:

$$C_i^{\text{per\_addr}} = \frac{C_i}{n_i}$$

where n_i is the number of addresses in building i.

Due to the routing package depending on nodes and edges for routing it sometimes shows inconsistencies. For this reason, the Euclidean distance between the building and the selected DP, as well as the geographic coordinates of both endpoints are included in the end result for checking. These features help identify unusual cases or bugs in the osmnx package.

All in all, the Fiber Routing Cost Model does not attempt to produce exact project-level cost forecasts. It is designed to differentiate buildings based on their relative financial costs between one another. The use of routing-based heuristics makes scalable and interpretable prioritization possible in the absence of detailed engineering estimates. This approach makes sure that the scoring model accounts for connection effort and infrastructure proximity without taking into account specific extra costs made in a project due to engineering difficulties.

4.3.3 Feasibility model

The third component of the selection model evaluates whether a building can feasibly be connected to the fiber-optic network under current operational and technical constraints. Unlike the connectivity and cost components that are combined in the composite score, the feasibility model has a separate as a separate evaluation mechanism. It uses a categorical label rather than a numeric score and is not included in the composite scoring formula. The reason for this is that the financial viability of the scoring of the buildings should be seen separately as it is a quantitative measurement while the feasibility is mostly qualitative.

Each building is assigned a feasibility label — **green**, **orange**, or **red** — based on a rule-based evaluation of known blockers and technical hindrances. These labels reflect increasing levels of roll-out risk:

- **Green:** No known feasibility constraints.
- **Orange:** Moderate or soft hindrances. For example competitor presence.
- **Red:** Hard blockers. For example planned demolition of the building.

The feasibility classification is based on deterministic business logic. It is derived from prior roll-out experiences. For example, buildings with delivery status codes that signify legal or technical exclusion are automatically marked as red. Similarly, the presence of soft constraints such as competitive overbuild or uncertain routing potential may result in an orange label.

If multiple addresses within a building have conflicting feasibility indications, the most conservative label is applied to the entire building to ensure operational robustness, with a warning that something in the building's reporting is wrong.. The full set of rules and logic used to assign these feasibility labels is included in Appendix F.

This color-coded feasibility system is intentionally designed to prioritize transparency and alignment with internal KPN operational logic. It enables stakeholders to quickly assess whether a building should be considered for roll-out. Therefore, Feasibility acts as a precondition for execution, not as a predictive input to the scoring model presented in Section 4.3.4.

4.3.4 Composite Scoring Model

To support data-driven roll-out planning, the two predictive model components—connectivity potential and connection cost—are integrated into a single composite score S_i for each building i. This prioritization score ranks all technically feasible buildings based on a combination of customer acquisition likelihood and connection cost efficiency.

The components are defined as follows:

- $P_i \in [0, 1]$: the predicted probability that a customer (KPN or wholesale) will be acquired from an address i, as estimated by the final LightGBM model trained on historical fiber adoption data.
- $T_i \in \mathbb{R}_{\geq 0}$: the estimated cost of connecting building i, calculated as the sum of the expected contractor price and a heuristic routing distance to the nearest fiber distribution point.

To ensure comparability, the cost component is normalized to a common scale using min-max normalization. Since lower costs are preferable, the cost term is inverted after normalization to yield a normalized cost score $C_i \in [0, 1]$:

$$C_i = 1 - \frac{T_i - \min(T)}{\max(T) - \min(T)}$$

To reduce the influence of extreme values, total cost values T_i are clipped at a predefined percentile before normalization. This prevents high-cost outliers from disproportionately compressing the scoring range for typical cases.

The final composite score $S_i \in [0, 1]$ is calculated as a weighted sum of the two normalized components:

$$S_i = \alpha \cdot P_i + \beta \cdot C_i$$

where $\alpha, \beta \in \mathbb{R}_{\geq 0}$ are weights reflecting the relative importance of customer conversion potential and costs. In this thesis, an equal-weighted formulation is adopted for benchmarking purposes, with $\alpha = \beta = 0.5$.
The feasibility status of each building is not incorporated into the composite score directly. Instead, it is evaluated separately and assigned a categorical label—**green**, **orange**, or **red**—based on operational roll-out constraints. These feasibility labels are a qualitative gating mechanism. Only buildings labeled green or orange are considered eligible for composite scoring and prioritization. Red-labeled buildings are excluded from roll-out consideration.
The resulting score S_i ranks eligible residential buildings within the retrospective connection dataset and serves as a foundation for dashboard-driven project formulation described in Section 4.4.

4.4 Dashboard Integration & Project Formulation

To make the model outputs accessible for the team, a Power BI dashboard is developed. The dashboard can be used to explore, filter, and select high-potential buildings and project areas based on the composite prioritization scores described in Section 4.3.4.

To ensure operational relevance, an address density filter is applied to exclude overly sparse areas from project formulation. This density logic is computed by applying DBSCAN clustering to the projected coordinates (RD New system) of buildings within each project area. Buildings within a radius of x meters and with at least y nearby points are grouped into clusters. For each project, the maximum cluster size and the average intra-cluster distance are calculated. The average distance is determined using k-nearest neighbor (KNN) search within each dense cluster. Only clusters with sufficiently high density are considered viable for project selection. This approach helps ensure that proposed projects have geographically coherent deployment potential. The impact of this density filter will be further evaluated in the model experiments in Chapter 5.

Page 1: Score-Based Overview. The first page provides a general overview of all nationally available buildings, segmented into three score brackets:

- High potential: buildings with a score between 0.7 and 1
- Medium potential: buildings with a score between 0.4 and 0.7
- Low potential: buildings with a score between 0 and 0.4

This segmentation allows users to quickly identify regions with the highest expected return on investment. The dashboard also highlights the best-scoring clusters that could form new roll-out projects. These clusters are generated based on historical project areas and the current scoring outputs.

Page 2: Building-Level Selection. The second page allows detailed filtering and selection of individual buildings. Users can filter by:

- Address or postcode

- Past project region
- Urbanisation level, province, or delivery status
- Score components, for example high connectivity or low cost
- Feasibility label

This page facilitates manual validation of new projects, inspection of model results, and further area selection for operational follow-up.

4.5 Validation Approach

This section discusses how the three components of the proposed selection model are validated to ensure robustness and practical value. Each component requires a different validation approach due to the different method with which they were developed.

4.5.1 Customer Uptake Prediction Model

The Customer Uptake Prediction Model is formulated as a binary classification problem. The goal is to estimate the probability that an address would become a fiber customer (KPN or wholesale) after being connected. The model was trained using addresses that were already technically connected (HC = 1) to reflect real-world adoption behavior.

Validation of this model included both performance metrics and robustness checks. To avoid data leakage and ensure generalizability, **stratified k-fold cross-validation** was applied across multiple random splits. Each fold uses full retraining of the model and probability calibration.

The following performance metrics are used:

- Area Under the ROC Curve (AUC)
- Macro-averaged F1-score (used for threshold optimization)
- Confusion matrix analysis
- Precision, recall, and k-precision (top-k conversion accuracy)

To evaluate feature robustness, two feature importance methods were used:

- Permutation importance, which measured the drop in validation performance when each feature was randomly shuffled
- Importance gain, based on LightGBM's internal feature gain score during tree construction

Together, these methods validated not only the predictive performance but also the interpretability of the model.

4.5.2 Cost Estimation Heuristic

The cost component is a deterministic heuristic. It estimates the total connection cost per building by combining contractual connection prices (based on delivery status) and trenching costs computed using Dijkstra's shortest path algorithm on the OpenStreetMap-based national street network [2]. The shortest path from each building to the nearest fiber distribution point (DP) was computed using a KD-tree to identify candidate DPs, with a Euclidean fallback for edge cases[2].

Validation was performed through:

- Checking costs predictions compared to previous projects
- Manual inspection of routing results for extreme-cost buildings
- Distributional checks on estimated costs per meter and cost per address

Additionally, costs were clipped at a $^{\text{th}}$ percentile before normalization to reduce the influence of outliers and to ensure meaningful scoring for the majority of buildings.

4.5.3 Feasibility Score

The feasibility labels are validated by using rule logic checks rather than statistical metrics. The scoring logic was based on KPN's internal business rules and roll-out experience. These roll-out checks have been confirmed within the team and can easily be changed if it matters. The feasibility score is used more as an extra feature than a true part of the composite score. Validation steps included:

- Sanity checks: verifying that RNA codes and delivery statuses flagged as hard blockers resulted in a red score
- Internal consistency: ensuring that feasibility scores were logically consistent across addresses within the same building
- Visual inspection: reviewing feasibility score distributions across multiple contractor zones and building types

4.5.4 Composite Score Validation

The composite score S_i was calculated as a weighted average of two scaled components:

$$S_i = \alpha \cdot P_i + \beta \cdot C_i$$

where both components were first min-max scaled to the $[0, 1]$ range, and costs were inverted so that a lower cost would reflect a higher score. To ensure interpretability and testing stability, the weights were set to $\alpha = \beta = 0.5$ in this study. These can be varied in future applications to reflect strategic priorities.

Composite score validation included:

- Visual inspection of score distributions across the dataset
- Manual review of top-ranked and bottom-ranked buildings

- Dashboard exploration of spatial clustering

To further validate the robustness of the composite score S_i, a sensitivity analysis was performed. This analysis explored how changes in the weights α and β influence the final building rankings. The goal was to assess whether the model consistently identifies high-priority buildings under different strategic emphases. The results of this analysis are presented in Chapter 5.

4.6 Chapter Conclusion

This chapter covers the design of a data-driven methodology to support KPN in prioritizing buildings for retrospective fiber-optic connections. The proposed solution integrates three key components into a scoring framework that can be used to identify good roll-out choices at the address level.

Firstly, the relevant datasets were preprocessed and combined to create a clean and enriched dataset, suitable for model development. A predictive model was then developed to estimate the likelihood that a connected address will convert into a paying customer. Using LightGBM, categorical features and domain-informed interactions were incorporated to capture underlying behavioral and demographic patterns.

Next, a cost estimation heuristic was introduced to approximate connection costs in the absence of detailed engineering calculations. This heuristic combines delivery-based pricing and routing distances based on a national street network. A feasibility labeling score was also developed, capturing technical and contextual blockers derived from internal business rules.

These components were integrated through a composite scoring model and labeling scoring system. For testing purposes, equal weights were applied. The composite score is intended to support prioritization and project generation by ranking buildings based on their strategic value for roll-out. At the same time, the feasibility labels help determine which projects are easy to handle.

Finally, the scoring model was prepared for integration into a Power BI dashboard to facilitate spatial filtering, visualization, and decision-making. The next chapter presents the results of model implementation, score distributions, dashboard outputs and a discussion of roll-out implications.

5 Results

This chapter addresses the fourth research question:

What are the results of the developed data-driven selection methodology, and how do they impact the performance of the fiber-optic connection planning process?

This chapter presents the results of the developed model components and evaluates the effectiveness of the full data-driven selection methodology. Building on the solution design outlined in Chapter 4, this chapter reports the performance of the individual model elements, investigates their sensitivity and interaction effects, and assesses their combined impact on rollout prioritization at the building level.

To answer the main research question, the following sub-questions are addressed:

- *How does the predictive model perform in estimating customer acquisition potential at the address level, and what insights can be derived from model experiments?*
- *What are the outcomes of the cost estimation and feasibility scoring components in identifying financially viable and feasible buildings?*
- *How does the composite score support the prioritization of buildings, and what patterns emerge from experimental ranking results?*
- *How sensitive is the output of the scoring methodology to changes in input parameters and component weightings?*
- *How does the developed selection methodology enable the formulation of viable project areas and support operational roll-out decisions?*

Section 5.1 presents the experiments and results for the Customer Uptake Prediction Model, including algorithm comparison, feature engineering, and performance evaluation. Section 5.2 addresses the cost estimation logic, covering routing heuristics, fallback strategies, and the comparison to historical project data. Section 5.3 introduces the feasibility model results, including rule activation and distribution analysis. Section 5.4 outlines the integration of these components into a composite score, with sensitivity testing and ranking behavior. Section 5.5 describes how the score is used to formulate project areas, select addresses, and support decision-making through the Power BI dashboard. Finally, Section 5.6 concludes the chapter with a synthesis of findings and critical reflection on the model's implications.

5.1 Customer Uptake Prediction Model

For the Customer Uptake Prediction Model several experiments have been performed with the goal to develop a model capable of predicting the chance that a household becomes a KPN or wholesale customer once connected to KPN's fiber optic network. Afterwards, the best found model results are evaluated and validated.

5.1.1 Experimental Setup

In the section, several experiments will be performed to determine which machine learning models and modeling techniques could be effective for this specific case. The goal of the experiments is to find a suitable machine learning algorithm and configure it as good as possible.

Machine Learning Algorithm Performance

Firstly, from the literature it was found that ensemble machine learning algorithms perform adequately well on telecom customer acquisition prediction. The starting train/test dataset has 35 features and approximately 4.4 million rows. The train/test split is 0.8/0.2 divided to allow for both sizable training and testing sets. Due to the used dataset being very large, an important KPI besides the pure performance of the model is the running time and RAM usage. The reason for this is that there are limitations to the computer with 32 GB RAM and an Intel Core i7-1370P. For this reason, the different algorithms will first be tested on various samples of the dataset to determine their performance and running times to determine which algorithm will be used for the entire dataset. Below a table with the performance of the algorithms can be found over 100000, 500000 and 1 million samples. For all samples, the same random state is used for fair comparison.

Table 5.1: *Model performance comparison across algorithms and sample sizes*

Model (Sample Size)	AUC	F1 Class 0	F1 Class 1	Accuracy	P@Top1%	P@Top5%	P@Top10%
Random Forest (100K)	0.624	0.390	0.742	0.635	0.7700	0.7500	0.7345
Random Forest (500K)	0.634	0.407	0.739	0.640	0.6940	0.7522	0.7503
Random Forest (1M)	0.640	0.407	0.742	0.640	0.7055	0.7420	0.7477
XGBoost (100K)	0.648	0.334	0.756	0.643	0.8900	0.8110	0.7855
XGBoost (500K)	0.645	0.316	0.753	0.637	0.8820	0.8264	0.7877
XGBoost (1M)	0.650	0.309	0.742	0.640	0.8920	0.8189	0.7935
LightGBM (100K)	0.641	0.530	0.668	0.611	0.8955	0.7990	0.7700
LightGBM (500K)	0.658	0.541	0.672	0.617	0.9180	0.8438	0.8012
LightGBM (1M)	0.665	0.549	0.677	0.624	0.9155	0.8423	0.8091

As can be seen in Table 5.1, the Random Forest model performs significantly worse compared to both XGBoost and LightGBM across nearly all evaluation metrics. This is particularly evident in the AUC and precision at top-k percentiles, which are crucial for reliably identifying the most promising customer acquisition opportunities. Additionally, Random Forest exhibits the longest run times and highest memory consumption, making it less practical for large-scale deployment. While XGBoost and LightGBM yield similar overall predictive performance, their internal learning mechanisms lead to different classification characteristics. Notably, the F1-score for class 0 (correctly identifying addresses unlikely to become fiber customers) is substantially higher in LightGBM. This indicates a more balanced classification performance, which is especially relevant given the inherent class imbalance of approximately 60% class 1 and 40% class 0. A model overly biased toward the dominant class could achieve deceptively high overall accuracy, yet fail to generalize to underrepresented outcomes. This would risk incorrectly assigning high uptake probabilities to addresses unlikely to convert, undermining the model's practical value. It is also important to note that the overall classification accuracy is only marginally better than random chance. However, accuracy is not a meaningful performance metric in this context, as it does not capture the quality of probability estimates nor the cost of misclassification across classes. Instead, the AUC and macro F1-scores provide more relevant indicators of the model's effectiveness. Given that the final prioritization relies on calibrated probabilities rather than hard 0/1 decisions, the ability to rank addresses correctly (as reflected by AUC) and to maintain balanced sensitivity across classes (captured by macro F1) are substantially more critical for supporting informed investment decisions.

Furthermore, LightGBM demonstrates superior training speed and computational efficiency, with a training time of only 3.75 seconds on 1 million rows. This is significantly faster than XGBoost's 253 seconds for the same sample size. In environments with limited computational resources, such as this study's setup with 32 GB RAM and a normal laptop CPU this is important. Furthermore, it is also preferable if the algorithm is light to run as it needs to be rerun frequently at KPN to adjust for changes in the data.

Precision at top percentiles is another key evaluation metric used to compare the models, as the primary objective is to prioritize the most promising buildings for fiber-optic network expansion. Precision@1% specifically measures the proportion of actual customer conversions among the top 1% of buildings that the model ranks as having the highest probability. In this case, the LightGBM model achieves a precision@1% of 91.55%, meaning that among the top 1% of buildings identified by the model as most likely to convert, more than 9 out of 10 indeed become customers. This is a crucial performance indicator, as the business use case involves targeting a small subset of buildings with the highest expected return on investment. LightGBM not only performs well at the top 1% but also maintains strong precision scores at the 5% and 10% thresholds. This demonstrates that the model is not just balanced and computationally efficient, but also highly effective at accurately identifying high-probability customer conversion cases.

Given the combination of the best AUC between the models , balanced class performance, good top-k precision, and the lowest training time, LightGBM is selected as the model to continue experiments with. XGBoost is a strong alternative, but the higher computational cost and more pronounced bias toward the dominant class make it less suitable for this prediction task.

Pipeline with one-hot encoding vs Native LightGBM

LightGBM is natively capable of handling categorical variables by treating them as category-type features. This allows the model to learn optimal splits without the need for data transformation. However, the mostly commonly used method in many machine learning pipelines is to use one-hot encoding combined with standardized numeric features. This is typically implemented through preprocessing pipelines such as sklearn's ColumnTransformer.

An experiment was conducted to test if native categorical LightGBM gives the same performance benefits as using a preprocessing pipeline. The first configuration uses LightGBM in its native form, converting all object-type columns to categorical data types and passing them directly to the model. The second configuration uses a pipeline that performs one-hot encoding for categorical variables and standardization for numeric features, followed by training with LightGBM.

The goal of this experiment is to evaluate if pre-processing via a pipeline improves model performance metrics such as AUC, F1-score, and precision at top-k percentiles. The advantage of one-hot encoding is that it sometimes offers advantages in linear models or algorithms without native categorical support. However, it may increase feature dimensionality and reduce efficiency in tree-based models. Thus, this comparison helps validate whether LightGBM's native categorical feature handling is preferable in terms of both accuracy and computational performance in the context of large-scale telecommunications datasets.

Table 5.2: *Performance Comparison: LightGBM with and without Pipeline*

Method	**AUC**	**F1 Class 0**	**F1 Class 1**	**P@Top1%**
Native categorical (no pipeline)	0.6652	0.552	0.679	0.9275
One-hot encoded (pipeline)	0.6576	0.547	0.665	0.9089

As can be seen in the table 5.2, the native LightGBM model outperforms the pipeline-based variant across all relevant performance indicators. The AUC of the native model is higher, and it also has better F1-scores for both class 0 and class 1, suggesting better generalization in a class-imbalanced setting. Moreover, top-k precision is consistently higher without the pipeline: for instance, precision@1% is 92.75% natively, compared to 90.89% using one-hot encoding. This trend can be seen across the top 2%, 5%, 10%, and 20% segments as well.

This difference in performance can be attributed to LightGBM's histogram-based tree building and leaf-wise growth, because it handles categorical variables more efficiently. One-hot encoding expands feature space and introduces sparsity. This increases computational overhead and can dilute signal strength. This is especially the case when high-cardinality features are involved. Additionally, the native approach trains in less time (12.98s vs. 19.87s), showing better computational efficiency.

In conclusion, LightGBM's native categorical handling not only simplifies preprocessing but also improves both prediction quality and training speed. Therefore, for the final model configuration, the native LightGBM variant is preferred over the one-hot encoded pipeline approach.

Ordinal Encoding & Interaction Features

The next experiment is to see if interaction features could improve predictive performance. The interaction features were engineered by combining selected ordinal demographic and behavioral attributes. These features are taken from the dataset used for the prediction task. They are chosen based on domain knowledge and feature importance rankings. To prepare for this experiment, categorical variables are converted to ordinal integers to be able to perform mathematical operations such as multiplication. This transformation itself did not significantly affect model performance.

Each interaction feature was added individually to the baseline LightGBM model to assess its effect on the Area Under the Receiver Operating Characteristic Curve (AUC). This variable was used as it gives a good overview of whether the complete model performs better or worse.

The results are summarized in Table 5.3. After this, all interaction features were tested together as an extension on the normal dataset used for the training and testing.

Table 5.3: *Effect of Individual Interaction Features on Model AUC*

Interaction Feature	**AUC**
Baseline (no interaction features)	0.6686
Income × Purchasing Power	0.6684
Income x House Value	0.6686
Urbanisation × Volume of House	0.6686
Life Phase × Number of Persons	0.6684
Purchasing Power × Type of House	0.6685
Volume × Relocation Activity	0.6684
Life Phase × Switch Sensitivity	0.6683
WOZ Value × Type of House	0.6684
Income × Main Breadwinner Function	0.6685
Energy Label × Streaming Services	0.6685
Age × Purchasing Power	0.6685
All interaction features combined	0.6687

The results show that the inclusion of individual interaction features led to negligible performance changes. The AUC values remained stable around the baseline of 0.6686, with variations no larger than 0.0003. Even when all engineered interaction features were added simultaneously, the overall AUC only increased marginally to 0.6687.

These results suggest that the existing base features already capture the majority of relevant variance, and that simple multiplicative combinations of ordinal features offer limited incremental predictive power. For this reason, the interaction features are not included in the final model as they did not notably increase the performance of the model.

5.1.2 Results

In this subsection, the final model found after all experiments is evaluated. It covers calibration of the model, threshold optimization, hyperparameter tuning and a review of the output of the model to determine if data leakage is prevalent.

Model Performance

To further refine the predictive power of the connectivity model and improve its practical usability for ranking purposes, calibration and threshold optimization were applied to the LightGBM model. Whereas traditional classification tasks rely on hard thresholds such as 0.5, this study is primarily interested in accurate probability estimates to make prioritization of addresses possible. Therefore, well-calibrated probabilities and balanced classification performance across classes are crucial.

The base LightGBM model achieved an AUC of 0.6686 and a macro F1-score of 0.609. To enhance these outcomes, the model was calibrated using Platt scaling via the CalibratedClassifierCV method with a sigmoid function. This step aligns predicted probabilities more closely with observed frequencies. This makes sure that comparisons across addresses is done fairly.

Subsequently, threshold optimization was performed to maximize the macro F1-score. By iteratively testing thresholds between 0.01 and 0.99, the best-performing threshold was found at 0.574, which resulted in an improved macro F1-score of 0.618. The model's final AUC after calibration was 0.6692, showing a marginal yet consistent improvement.

Table 5.4: *Performance Comparison: Baseline vs. Calibrated Model*

Metric	Baseline Model	Calibrated Model
AUC Score	0.6686	0.6692
Macro F1-Score	0.609	0.618
Accuracy	0.631	0.636
Precision@Top1%	0.8955	0.9304
Precision@Top5%	0.7990	0.8522
Precision@Top10%	0.7700	0.8147

The results in Table 5.4 show that the calibrated model slightly improves overall classification metrics such as AUC and accuracy. It also shows slightly higher performance in top-k precision. Especially in a business context where only the most promising addresses are connected first, high precision in the top-ranked predictions is important. The reason for this is that this means that the addresses that are connected based on their actually also have the highest probability of becoming a customer.

Figure 5.1 displays the confusion matrix using the macro F1-optimized threshold. The model achieves a balanced distribution of true positives and true negatives, which is particularly important given the class imbalance in the dataset. It is better at identifying class 1 in absolute numbers due to the total number of class 1 in the dataset being higher. This means that even when the class balancing is used, the algorithm still favors to predict class 1 due to it naturally having the higher chance to result in a good prediction.

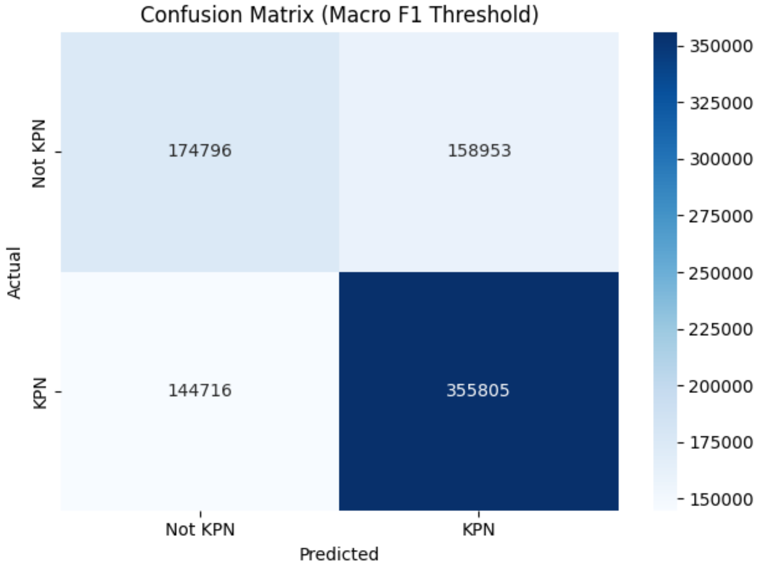


Figure 5.1: *Confusion matrix of the final calibrated LightGBM model at the macro F1-optimized threshold. It demonstrates balanced classification performance, crucial for managing both likely and unlikely adopters in a class-imbalanced dataset.*

The ROC curve in Figure 5.2 shows a stable improvement in discrimination capacity, while the calibration curve in Figure 5.3 confirms that the predicted probabilities are well-aligned with the actual observed outcomes.

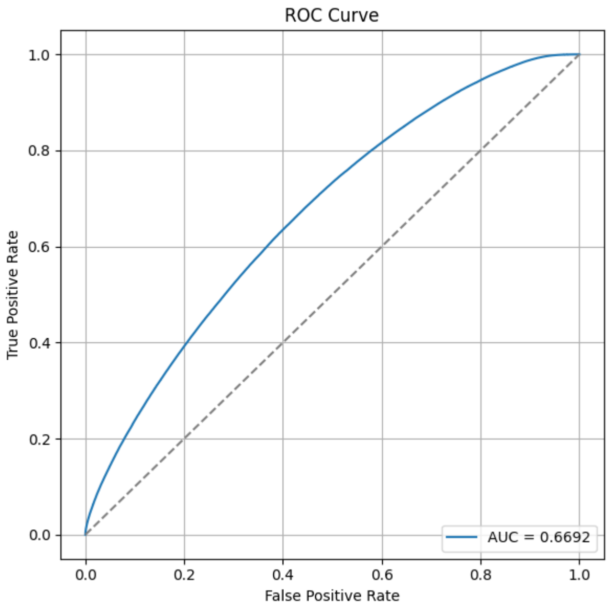


Figure 5.2: *ROC curve of the final model, yielding an AUC of 0.669. This indicates moderate ability to discriminate between buildings with high versus low probability of new customer uptake, supporting probability-based ranking.*

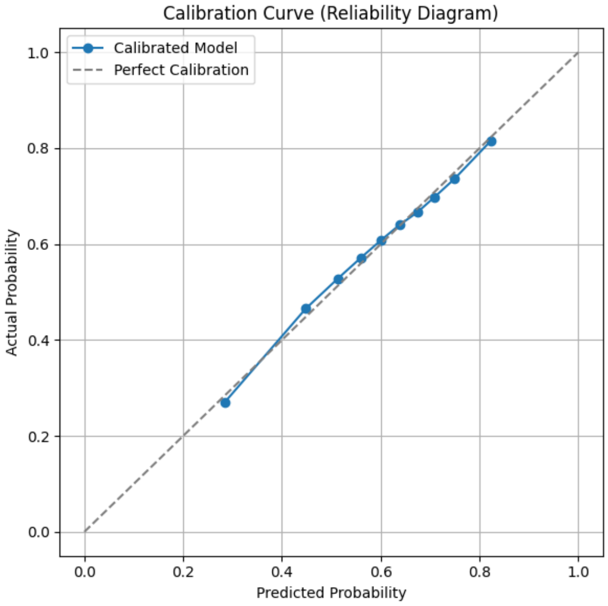


Figure 5.3: *Calibration curve of predicted probabilities against observed outcomes. The close alignment confirms the model's probability estimates are reliable for prioritization decisions*

In conclusion, The calibration and threshold optimization steps result in a slightly better model. The increased macro F1-score and precision at top-k levels show that the calibrated model is more suitable for prioritization-based fiber-optic network decisions. This calibrated probability model is therefore used as the base model for the connectivity potential score. To validate its robustness, stratified 10-fold cross-validation was conducted, retraining the model on each fold. The resulting average AUC was 0.6686 ± 0.0007, indicating consistent performance across folds. A detailed overview of the AUC scores per fold can be found in Appendix E.

To get the final bit of performance out, hyperparameter tuning using Optuna is used to optimize the Macro F1 score. This is done to tune the parameters of the LightGBM to fit the best possible with the current features. This gives the following final results for the model:

Table 5.5: *Final Model Performance After Hyperparameter Tuning (Macro F1 Optimized)*

Metric	**Class 0**	**Class 1**	**Macro Avg.**
Precision	0.553	0.695	0.624
Recall	0.529	0.715	0.622
F1-Score	0.541	0.705	0.623
Accuracy	0.641 (threshold = 0.583)		

To evaluate the added value of the calibrated model, a comparison was made against a random baseline that assigns uniform probabilities to each address. As shown in Table 5.6, the calibrated

LightGBM model outperforms the random classifier in both AUC and Precision@TopK metrics. The AUC result demonstrates that the model has meaningful discriminative power, while the Precision@TopK scores confirm its ability to correctly rank high-potential addresses. The calibrated model consistently identifies more true positives in the top-ranked percentiles. The importance gain and permutation importance of the features of the final model can be found in Appendix E.

Table 5.6: *AUC and Precision@TopK Comparison: Calibrated Model vs. Random Model*

Metric	Calibrated Model	Random Model
AUC Score	0.6753	0.5000
Precision@Top1%	0.9357	0.6070
Precision@Top2%	0.9009	0.6047
Precision@Top5%	0.8568	0.6015
Precision@Top10%	0.8193	0.6006
Precision@Top20%	0.7788	0.5996

Finally, using this model the predictions are performed for the possible retrospective connection addresses. This gives the following results graph in terms of probabilities:

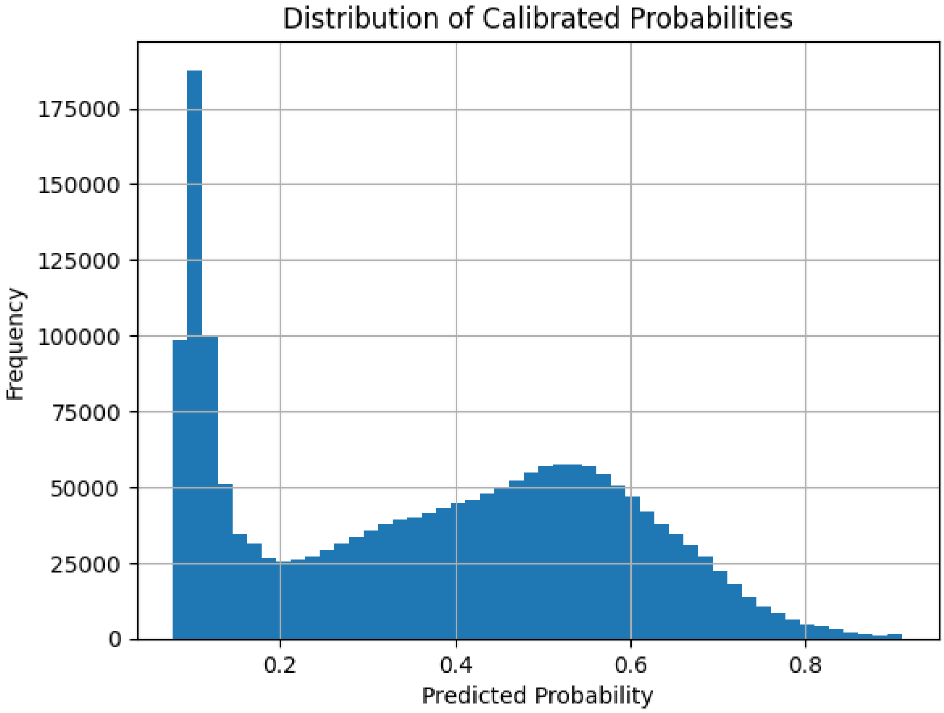


Figure 5.4: *Distribution of predicted probabilities for uptake across all addresses. Most buildings fall between 0.2 and 0.7, reflecting the conservative probability calibration and inherent variability in customer behavior.*

Observations & Limitations

The final calibrated LightGBM model achieves consistent performance across folds and demonstrates strong prioritization capabilities. This is particularly clear at the top of the prediction spectrum. As shown in Table 5.6, the model achieves a Precision@Top1% of 93.6%. This confirms its effectiveness in identifying the highest-potential addresses for fiber roll-out. This makes the model highly suitable for its primary purpose which is supporting in financially viable network expansion. However, the overall AUC score of 0.6686 indicates that the model performs moderately when considering all predictions across the full range. This is mostly due to it being hard to determine whether someone becomes a user of the network. The reason for this is that there are factors at play that cannot be easily captured such as current contract length, competitor offers, or how kind and reliable the contractor was during the roll-out.

This difficulty of predicting is further illustrated by the distribution of predicted probabilities in Figure 5.3, which shows a wide spread between 0.2 and 0.7 and a large spike just above 0.1. While the model rarely assigns extreme probabilities close to 1.0, it does assign sufficiently distinct values to enable effective top-k ranking. This distribution reflects a conservative calibration approach that avoids overconfidence. This is likely due to the noisy and overlapping nature of the input data. As a result, the model excels in separating the top-tier candidates but does not provide good confidence margins throughout the entire dataset. This limitation is partly due to the nature of the available features, which do not capture all relevant behavioral or contextual variables influencing customer conversion. Finally, it should be noted that the model's strength lies in ranking, not classification, and it should therefore be interpreted and applied as a decision-support tool rather than as a strict predictor of customer acquisition.

5.2 Fiber Routing Cost Model

For the Fiber Routing Cost Model several experiments have been performed with the goal to develop a model capable of predicting the cost of connecting a residential property to the fiber-optic network. Afterwards, the best found model results are evaluated and validated.

5.2.1 Experimental Setup

To validate the logic and behavior of the Fiber Routing Cost Model as described in Chapter 4, several experiments were conducted. These are used to verify the correctness of routing computations, the robustness of the clipping threshold, and the realism of the cost output in both synthetic and real-world contexts.

Routing Mechanics & Fallback Strategy

To estimate realistic trenching and connection costs, routing distances between each building and its nearest distribution point (DP) were calculated using OpenStreetMap (OSM)-based shortest-path algorithms. This routing heuristic is more realistic for true life distance rather than straight-line approximations. An example of this routing heuristic can be seen in Figure 5.5.

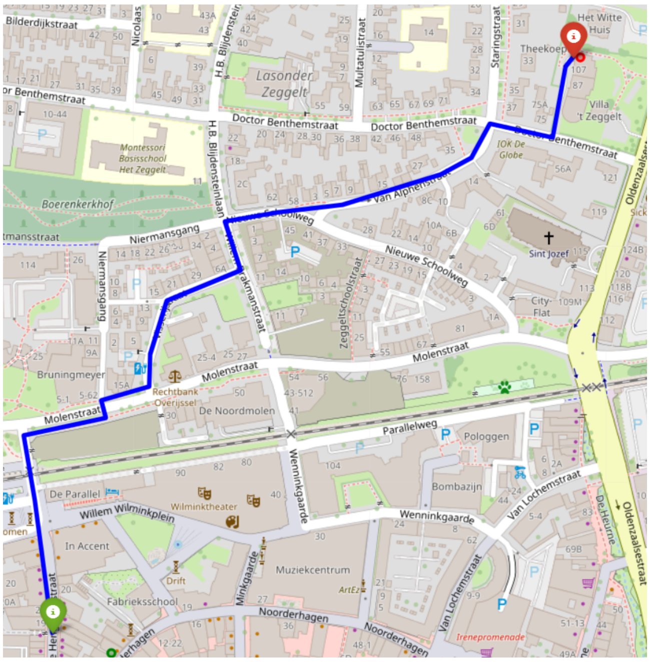


Figure 5.5: *Example of shortest-path routing between a building and its nearest distribution point using an OSM-based street network. This approach provides realistic trenching estimates compared to straight-line distances.*

Figure 5.6 shows the distribution of the ratio between routing and Euclidean distance. The average routing distance was 27% longer than the Euclidean equivalent, with some routes up to three times longer. These findings confirm that route-based distance estimation is needed to avoid underestimation of the trenching costs.

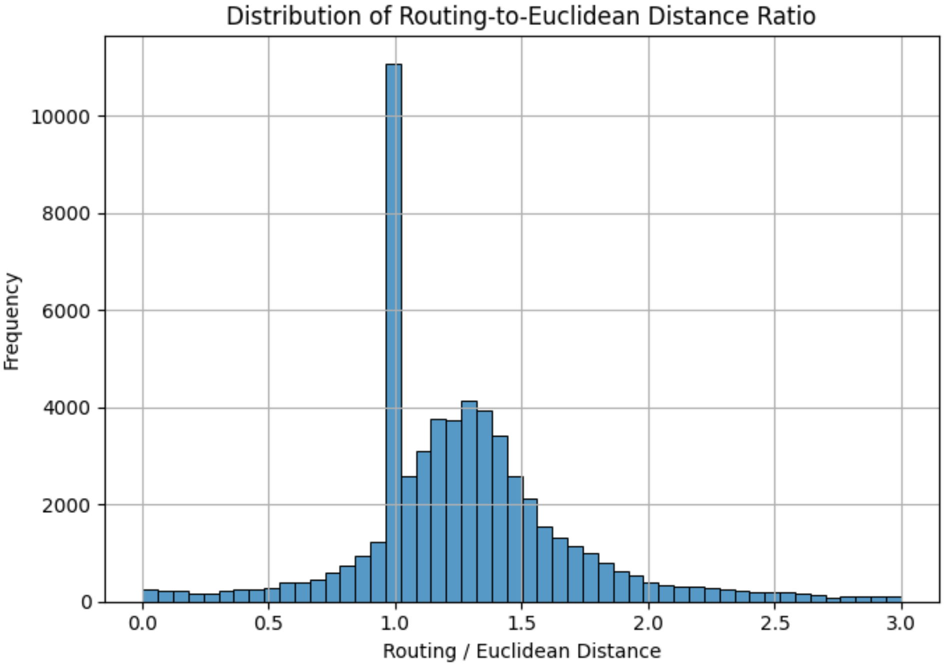


Figure 5.6: *Ratio of routing distances to Euclidean distances, showing trench paths average 27% longer. This highlights the necessity of route-based heuristics for credible cost estimation.*

However, certain edge cases trigger a fallback mechanism. This is needed due to the python package used having certain limitations mentioned in Chapter 4. The fallback mechanism is applied when the computed routing distance is either zero or 3000 meters. Zero is often caused by snapping both building and DP to the same OSM node. The other option is that the closest distribution point is more than 3,000 meters away which indicates disconnected infrastructure, infeasible roll-out or a lack of nodes in the area. In these cases, the Euclidean distance is substituted to ensure continuity in the estimation logic. Table 5.7 summarizes the number and average cost of fallback versus routing-based estimates.

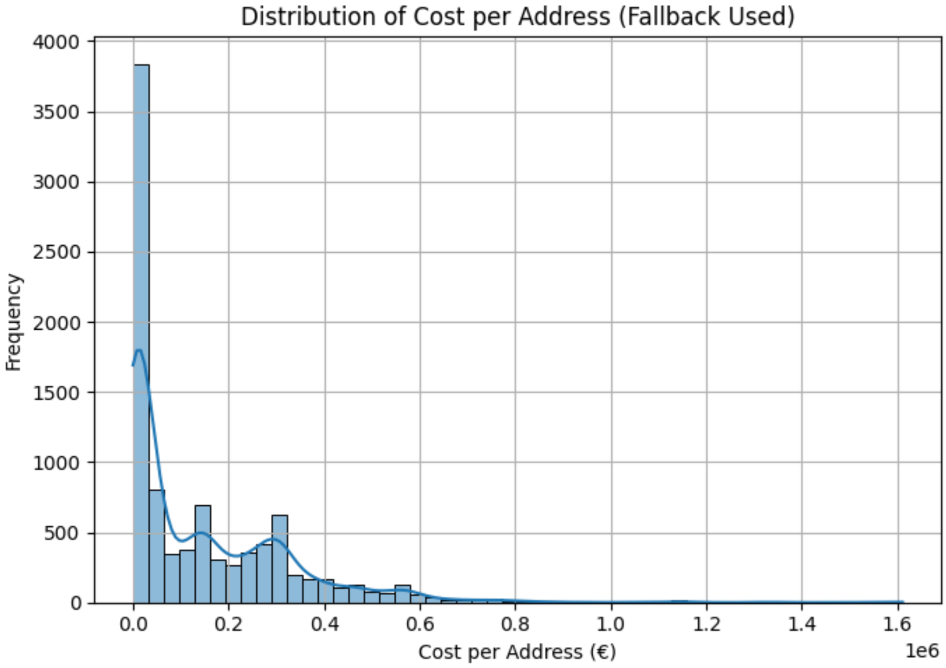


Figure 5.7: *Frequency of fallback to Euclidean distances in cases where routing fails or exceeds 3,000 meters. This ensures cost estimates remain available across the full dataset*

Clipping Strategy

Due to the presence of extreme outliers which are addresses located disproportionately far from the fiber-optic network,the average connection cost per address is heavily skewed. This distorts the cost normalization in the composite scoring model. This results in inflated influence from a small number of unfeasible cases and mapping most cost numbers close to the maximum score for costs. To mitigate this issue, a percentile-based clipping strategy is applied. With this strategy, cost values above a defined threshold are capped at the corresponding percentile value. This makes sure that addresses with extraordinarily high costs do not dominate the scoring function.

Table 5.7: *PUBLIC EXAMPLE Clipping Strategy: Impact of Percentile Thresholds on Cost Metrics*

Percentile	**Clip Value (€)**	**Avg Cost (€)**	**% Clipped**
75th	73,303	33,062	25
85th	114,058	40,822	15
90th	164,691	47,134	10
95th	242,781	52,826	5
99th	461,782	57,752	1

Table 5.7 summarizes the effect of various clipping thresholds. For example, applying a 95th percentile cap results in an average cost of €52,825.85 per address and affects only 5% of the dataset. The more conservative 75th percentile threshold clips 25% of the addresses and reduces the average cost to €33,062.34. In the context of the composite scoring model, all clipped addresses will receive a cost score of zero. This makes the cost component robust against infeasible extremes while preserving discrimination in the majority of realistic cases. The impact of the various settings will also be tested in the composite scoring sensitivity analysis.

Comparison with Real Project Costs

Currently, only a couple buildings have needed trenching work from the DP to get connected. For these buildings the costs were 81% accurate. However, due to the limited sample size, it is hard to validate if this is truly accurate. Also the delivery statuses were the routing from the DP is required and thus require significant trenching work are generally not good options to do first. This is because all the other statuses are often cheaper than the ones needing trenching work.

5.2.2 Results

For the total model results these are the average costs derived from both the contractual defined ones as well as the residential buildings that require trenching. Figure shows it can be seen that all the statuses that require trenching are on average significantly more expensive than the delivery statuses that do not require trenching.

The trenching cost is always divided over all the addresses in an entire building. For this reason, buildings with a lot of addresses in them often high-rise buildings can still have a good comparative price. This is evaluated in Figure 5.8.

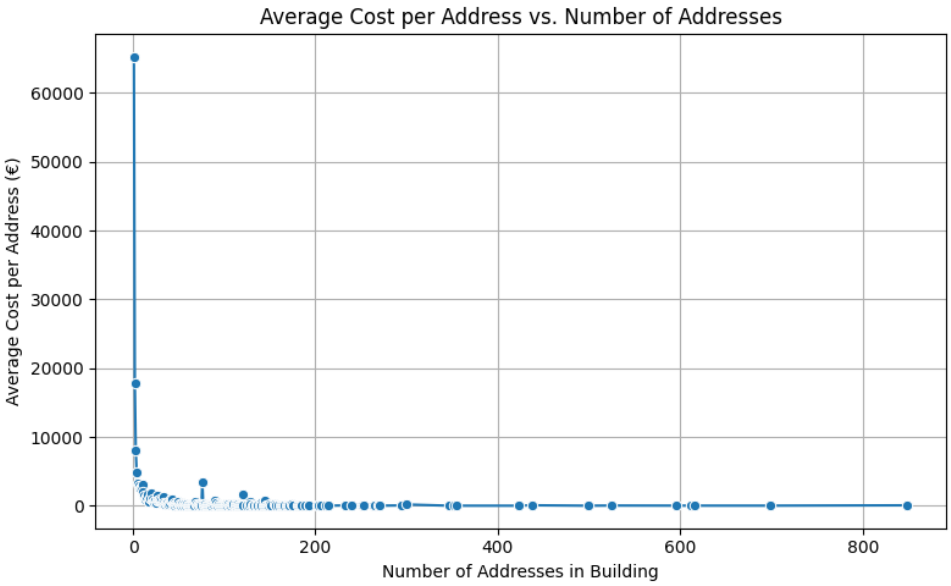


Figure 5.8: *PUBLIC EXAMPLE Relationship between trenching costs and the number of addresses in a building. Larger multi-unit buildings dilute per-address costs, favoring high-rise connections for cost efficiency*

As can be seen in Figure 5.8, the trenching costs for buildings with a lot of addresses are quite marginal. For this reason, it can be concluded that even buildings that have trenching costs can be attractive options to connect. At the same time, connecting an individual address with one of these statuses is extremely expensive using the current shortest path heuristic to determine costs.

5.3 Feasibility Model

Below the results of the feasibility model can be found. For this no experiments were performed due to this being a set of determined rules that only support the business as a notifier of a possible operational infeasibility. This function is build-in to support the team by having an easy overview of what needs to be done.

The model was applied to all unconnected addresses with a valid Reason Not Connected (RNA) and a residential function. As shown in Figure 5.9, the majority of buildings were assigned a green label. This suggests that there are no technical objections for this address. Approximately 41% of the addresses were assigned an orange label, indicating uncertainty due to specific flags or delivery status. A small minority received a red label, indicating that connection is infeasible under current conditions.

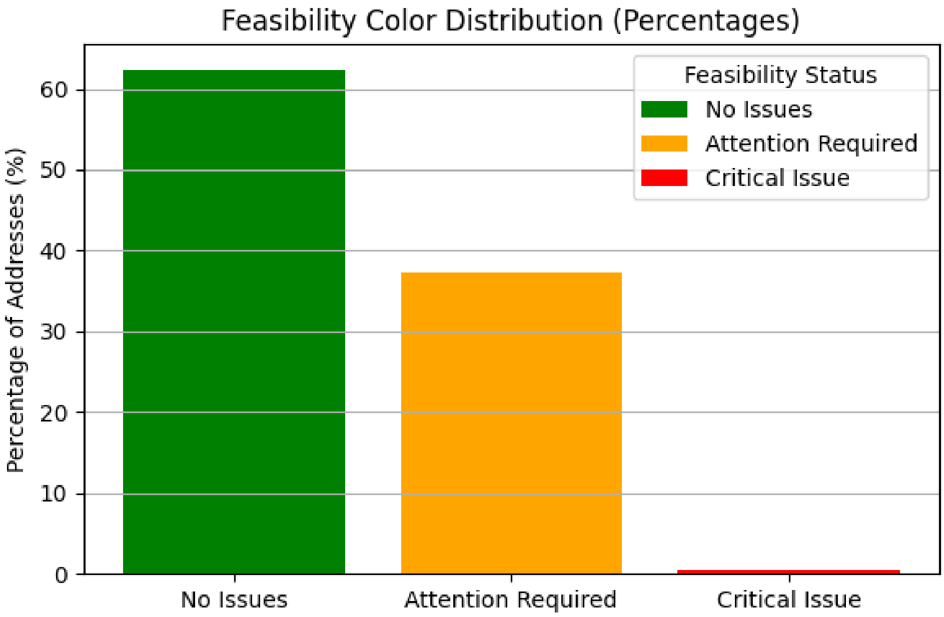


Figure 5.9: *Distribution of feasibility labels across addresses. Most are classified as green (directly viable), with orange indicating moderate operational uncertainties and a small fraction flagged red*

To explain which conditions most frequently triggered a specific label, Figure 5.10 displays the absolute occurrence of each contributing factor, with bar color reflecting the assigned feasibility outcome. The most common trigger for the orange label was overbuild. This was followed by delivery statuses 1, 4, or 7, which typically indicate pre-inspection or construction phases. These conditions do not prohibit connection but show implementation risk. Ambiguous RNA codes such as R11 and R24 were also frequent contributors, suggesting limitations in data clarity.

Green labels were often assigned in the absence of any triggering condition or if the address was previously served via copper by KPN. Red labels were primarily linked to RNA R15, which denotes a non-residential or demolished property. Other red conditions were rare.

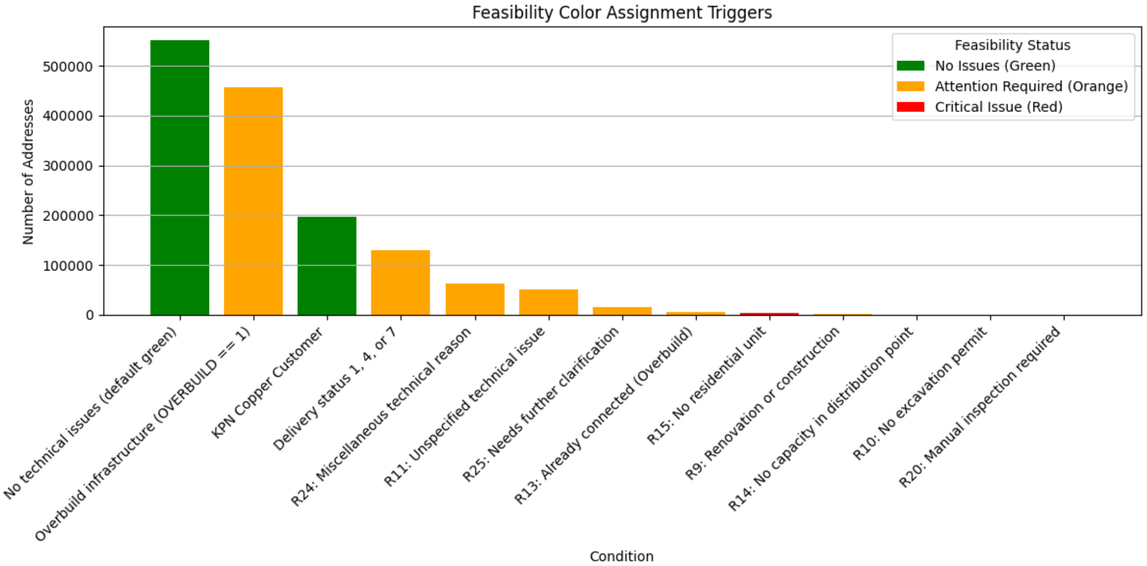


Figure 5.10: *PUBLIC EXAMPLE Frequency of conditions triggering feasibility labels, with most orange labels arising from competitor overbuild or ambiguous RNA codes. This informs where manual reviews may be needed.*

The feasibility label provides a binary indication of whether a building should be included in roll-out evaluation. While green-labeled buildings are assumed directly viable, orange-labeled addresses are flagged for manual review. Red-labeled addresses are excluded entirely from further consideration.

The model enables operational teams to differentiate between high-potential and high-risk areas in large-scale datasets. It is implemented as an interactive filter in the Power BI dashboard, where feasibility color can be used to highlight or exclude addresses during manual project formulation. This approach allows users to balance automation with expert judgment, and supports scalable but responsible roll-out planning.

5.4 Composite Scoring Model

The composite scoring model is based on both the Customer Uptake Prediction Model and costs model. As a weighted scoring method is used for the calculated score between the two a sensitivity analysis will be done for the entire model to determine the best weight settings for the composite scoring as well as the clipping strategy.

5.4.1 Sensitivity Analysis

To test the robustness and sensitivity of the composite scoring model, a sensitivity analysis is performed on the two model components: the clipping strategy for cost normalization and the weight distribution between connectivity potential and cost. The purpose is to determine how different parameter settings affect the final composite score and whether these settings influence the results of the composite score.

Clipping strategy

In the composite score, total estimated connection costs are first clipped to reduce the effect of extreme values and then normalized. To evaluate the impact of this clipping step, several percentile thresholds are tested ranging from the 70th to the 99th percentile. For each clipping level, the total cost is capped at the corresponding percentile value and subsequently included in the composite score at a 50/50 weighting.

As shown in Figure 5.11, the mean composite score remains relatively flat between the 70th and 90th percentiles, with limited variation and a stable average around 0.38. From the 92nd percentile onward, the mean score begins to increase more significantly. This trend indicates that mild clipping thresholds suppress too much of the cost variation, making the score less informative for differentiating viable addresses.

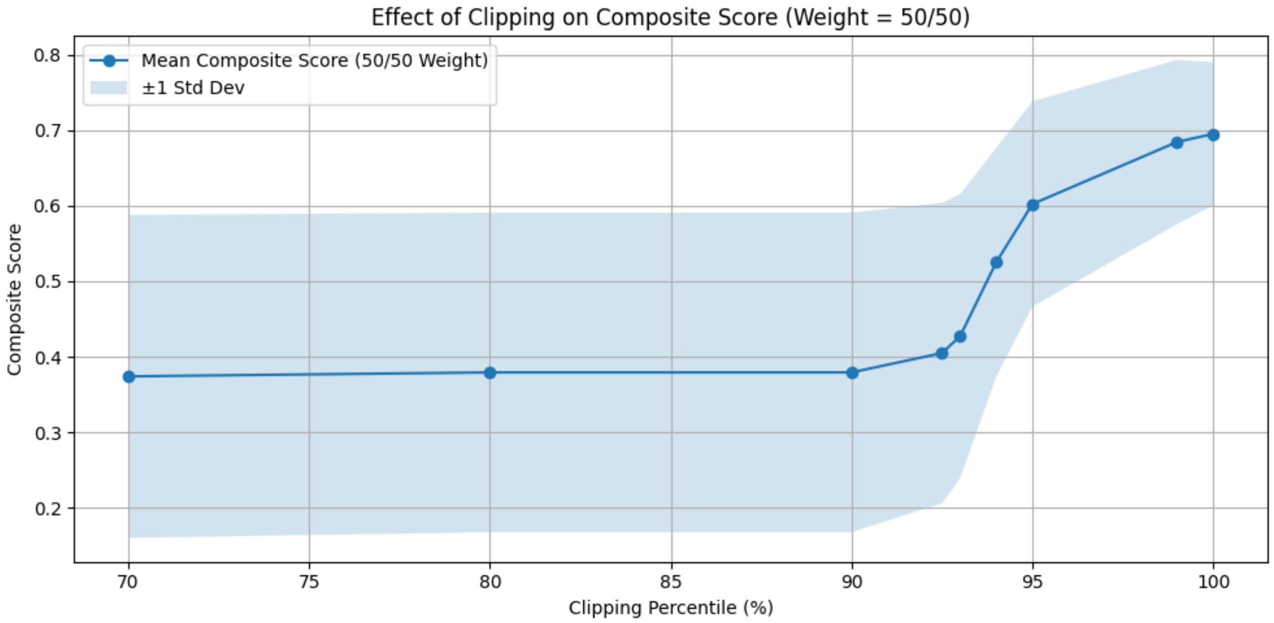


Figure 5.11: *Mean composite score under different clipping thresholds for cost normalization. The 94th percentile balances variation, preventing extreme costs from overly distorting prioritization.*

The 94th percentile represents the optimal trade-off. At this level, the average composite score increases substantially to approximately 0.53, while the standard deviation remains well-contained. This means that most cost outliers are effectively suppressed, but meaningful differentiation between moderately high and low costs is still retained. More lenient clipping levels such as the 95th or 99th percentile further increase the mean score, but also introduce greater variance and potential distortion from extreme cost values. Therefore, the 94th percentile is selected as the final clipping threshold for the composite model.

Weighting strategy

After selecting the clipping threshold, a second sensitivity analysis is done to examine the effect of the relative weight placed on the connectivity potential versus the cost. The composite score is computed for all addresses using weight settings ranging from 0% to 100% on the connectivity component in increments of 10%.

Figure 5.12 shows that the mean composite score declines as more weight is placed on the connectivity potential. This pattern is expected, since the cost score distribution is more favorable on average due to normalized cost scaling and the effects of clipping. Although connectivity potential is a valuable indicator, its distribution is slightly more compressed. This leads to lower average composite scores when it dominates the formula.

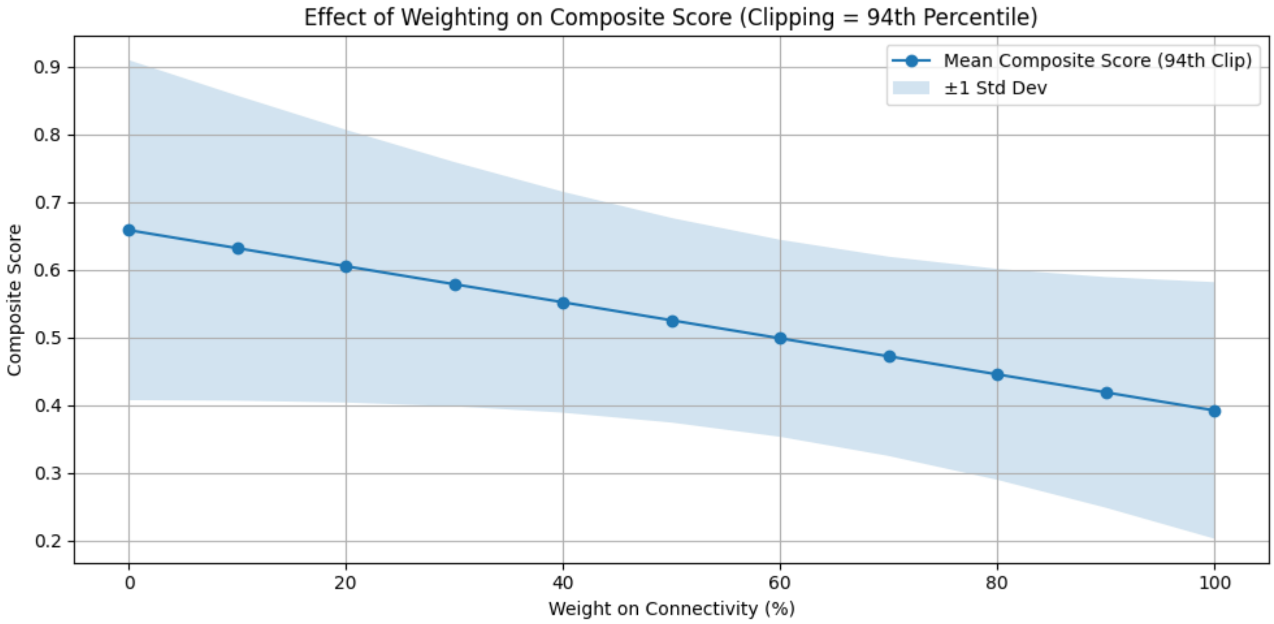


Figure 5.12: *Sensitivity of mean composite scores to changing weight emphasis on connectivity versus cost. Scores decrease and standard deviation rises with heavier weighting on predicted uptake, reflecting conservative probability distributions.*

Importantly, the standard deviation increases as the connectivity component becomes more dominant. This suggests that the standard deviation is higher in the probability model compared to the cost component. This makes a strong case for applying equal weights: the 50/50 configuration offers a well-balanced trade-off between average performance and variance, ensuring that both customer demand and financial viability are taken into account fairly in the final score.

Together, the two sensitivity analyses support the choice of a 94th percentile clipping threshold and a 50/50 weight split between cost and connectivity. This configuration ensures that the composite score remains both informative and interpretable, while avoiding distortion from extreme values or overemphasis on one model component.

5.4.2 Results

The final composite score is based on the normalized connectivity potential and inverted normalized connection cost, combined using a 50/50 weight. As shown in the experiments, the total cost was clipped at the 94th percentile to reduce the impact of extreme values. As a result, the composite score reflects both market potential and financial feasibility.

Figure 5.13 shows the distribution of composite scores across all technically viable addresses. Most scores fall between 0.35 and 0.65, with visible peaks resulting from patterns in the underlying input models. Only a small share of addresses receive scores above 0.75, indicating that highly attractive combinations—low cost and high predicted adoption—are relatively rare. This is also logical as the Customer Uptake Prediction Model can never predict for certain if someone is going to be a customer or not thereby scores close to 0 and 1 do not exist. However, The scoring model still successfully differentiates between more and less attractive connection opportunities as higher scores reflect better business opportunities.

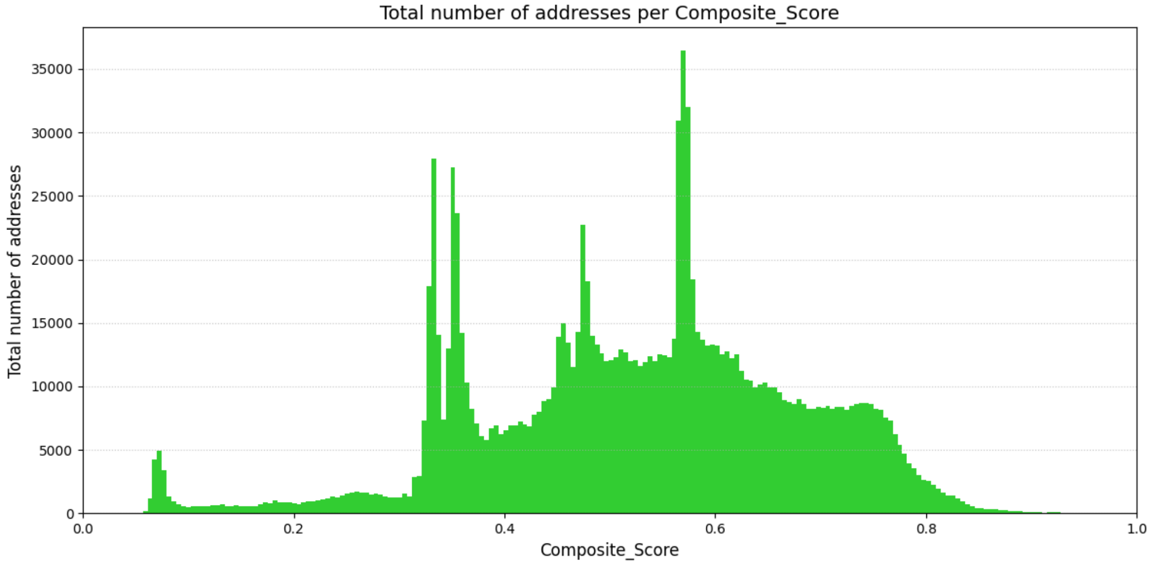


Figure 5.13: *PUBLIC EXAMPLE Distribution of final composite scores, with most buildings clustering between 0.35 and 0.65. This supports nuanced differentiation of rollout opportunities.*

To assess the consistency of the score, the average composite score is plotted per delivery status in Figure 5.14. The results align with expectations: addresses with delivery statuses 9, 4, and 2 receive the highest scores. This is logical as they have the most existing infrastructure. These are followed by addresses in a preparatory stage, such as status 1 and 11. Lower scores are assigned to addresses that require major construction work like status 14 and 15 as these have higher expected costs due to trenching and lower feasibility. The score thus accurately represents existing logic within KPN.

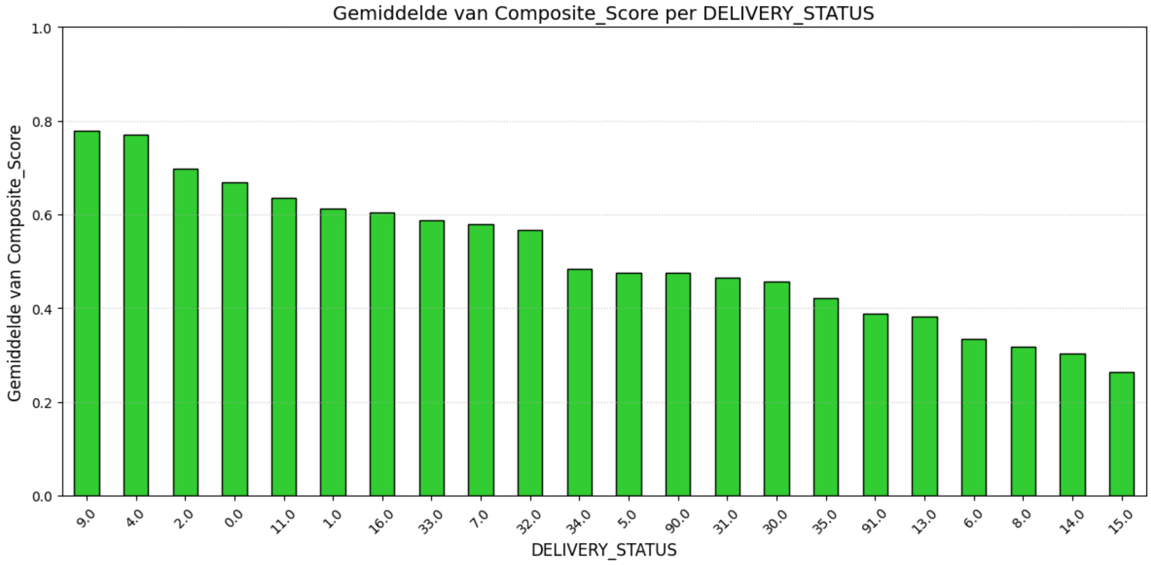


Figure 5.14: *PUBLIC EXAMPLE Average composite scores by delivery status. Buildings with more pre-existing infrastructure rank higher, validating the model's alignment with practical rollout economics.*

In sum, the composite score provides a transparent and scalable method for comparing connection opportunities. The score balances predicted adoption potential with normalized cost levels and aligns with operational insights. In the next section, the score is used to rank addresses and define potential roll-out projects.

5.5 Project Creation

For the project formulation based on the single project scores the address density was used. The address density is defined as the distance in meters to the next house. If the next house falls within this distance, it is included in the project. This distance will be tested in the experiments to see if it functions correctly.

5.5.1 Sensitivity analysis

In this subsection, it is tested if the project formulation formula functions correctly. This is mostly useful for the dashboard itself as it should function as intended if KPN wants to see which projects might be possible for certain settings. As they are able to manually use these switches themselves, the intention of these experiments is mostly to determine if the project formulation code functions as intended. An overview of the dashboard can be found in Appendix E.

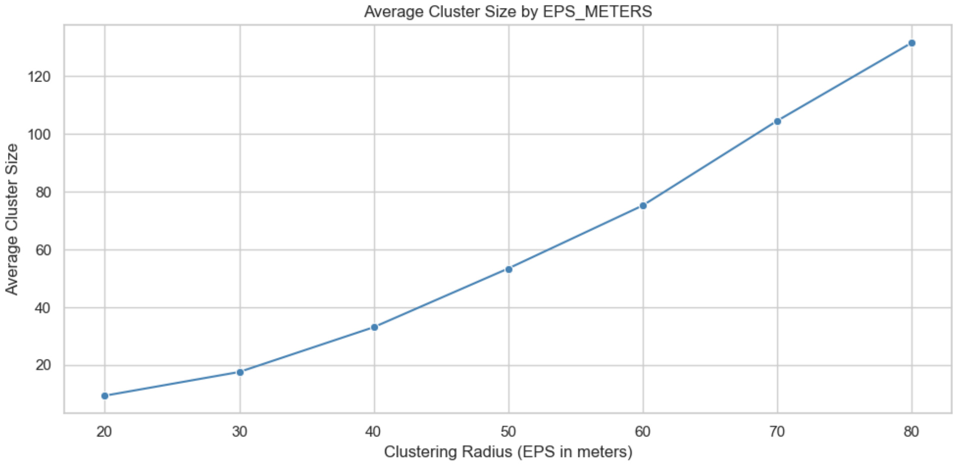


Figure 5.15: *Average project cluster size formed by DBSCAN at different epsilon distance thresholds. Increasing epsilon allows larger, more geographically dispersed clusters, aiding scalable project formation.*

As can be seen in the Figure 5.15, the density function behaves as expected, as more meters between buildings mean that the average cluster size increases. This means that when the distance between buildings is allowed to be larger, the average project size also increases automatically.

5.6 Chapter Conclusion

This chapter addressed the fourth research question: *What are the results of the developed data-driven selection methodology, and how do they impact the performance of the fiber-optic connection planning process?* The results confirm that the proposed solution effectively supports KPN in identifying and prioritizing financially viable buildings for fiber-optic connection at the address level.

The first model component, the Customer Uptake Prediction Model, achieves consistent predictive performance, with an AUC of 0.6753 and a precision@top1% exceeding 93%, indicating that the model is highly effective at flagging the very best candidates. Despite the inherent uncertainty in predicting individual customer behavior, this AUC means the model correctly ranks a future customer above a non-customer in approximately 68% of random address pairs, representing a 35% improvement over random selection. This confirms its value in reliably identifying high-priority buildings for fiber connection. LightGBM with native categorical handling and calibrated probabilities was ultimately chosen as the most suitable algorithm. While individual interaction features did not substantially enhance predictive power, the final calibrated model is robust and well-suited for prioritization purposes, particularly excelling at isolating the most commercially valuable addresses.

The Fiber Routing Cost Model aims to provide realistic connection cost approximations by combining contractual delivery status pricing with a routing-based heuristic. Routing distances were on average 27% longer than Euclidean distances. This shows the importance of realistic trenching estimates. A percentile-based clipping strategy (94th percentile) was applied to mitigate the influence of cost outliers. This results in a robust and interpretable cost score that can be integrated into the composite score.

The Feasibility Model assigns color-coded feasibility labels to reflect operational constraints. Approximately 41% of addresses are labeled orange. This indicates uncertainty in its operational feasibility, so should be evaluated carefully by decision-makers. Only a minority are classified as red and excluded from roll-out consideration. The rule-based feasibility component ensures that technical and contextual risks are transparently captured and factored into roll-out planning.

The Composite Scoring Model integrates the connectivity potential scores and the normalized cost scores into a single composite score. A 50/50 weight setting combined with the 94th percentile clipping threshold offers a balanced trade-off between financial viability and customer acquisition potential. As indicated in the chapter, the score effectively differentiates between buildings based on financial viability with high-ranking buildings aligning well with the known delivery status advantages and infrastructure availability.

Lastly, the dashboard-supported Project Formulation logic was shown to function as intended, with sensitivity analysis confirming that denser address clusters could be formed by relaxing spatial constraints. The integrated scoring and filtering system enables dynamic project generation and supports strategic decision-making in retrospective roll-out planning.

In conclusion, the results validate the effectiveness of the data-driven selection methodology in improving KPN's planning process. By integrating predictive analytics, cost heuristics, and operational feasibility into a composite prioritization model, the approach improves both the efficiency and quality of decision-making for fiber-optic network expansion. The next chapter discusses these findings in a broader context and formulates theoretical and practical recommendations.

6 Conclusion, Discussion and Recommendations

This chapter summarizes the main findings of the research, reflects on the strengths and limitations of the developed selection methodology, and discusses its broader implications. Additionally, it presents concrete recommendations for practice and outlines directions for further research. Lastly, it reflects on the practical and theoretical contribution of this thesis.

6.1 Conclusion

The goal of this thesis is to solve the core problem that KPN's current selection method for determining which buildings to connect to its fiber-optic network is not suited for individual property scoring and does not explicitly take financial viability into account. To address this, the central research question was formulated as follows:

How can KPN develop a data-driven selection method to identify and prioritize financially viable buildings to connect to its fiber-optic network?

To answer this question, the thesis first analyzed the current situation at KPN. This analysis revealed that the existing selection method, although valuable for evaluating area-level returns, lacks the building-level detail, transparency, and cost precision needed for today's brownfield environment. As a result, the current process often depends on manual interpretation and subjective judgment, which increases the risk of inefficient capital allocation and inconsistent project selection.

Based on these insights, a new data-driven selection methodology was developed. This method integrates three essential components. First, a machine learning algorithm, called Customer Uptake Prediction Model, estimates the likelihood that a building will become a KPN fiber customer once connected. This achieved an AUC of approximately 0.675 and a precision@top1% above 93%, resulting in a roughly 35% improvement over random selection in identifying the most commercially attractive buildings. Second, a routing-based cost estimation was introduced that calculates expected trenching distances, showing that actual paths are on average 27% longer than simple straight-line assumptions, thereby providing more accurate financial insights. Third, a feasibility assessment was implemented that uses business rules to filter out buildings that are technically unfeasible or strategically undesirable to connect. By combining these components into a normalized composite score, all eligible buildings can be prioritized based on their combined commercial potential, expected connection costs, and operational feasibility.

The results presented in Chapter 5 demonstrate that the developed selection method enables KPN to make decisions at the individual building level with significantly greater granularity, transparency, and financial accuracy than was previously possible. By combining predictions of customer conversion likelihood, realistic routing-based cost estimates, and feasibility checks into a single composite framework, planners can clearly identify which buildings offer the most favorable balance of commercial potential, expected capital expenditure, and operational viability. The integration of this method into a Power BI dashboard ensures that the results are directly actionable for planners, providing a structured, data-driven tool that aligns closely with expert expectations.

In conclusion, this thesis successfully developed and validated a data-driven selection method that fundamentally changes how KPN can approach its fiber-optic network expansion. Where previously it was not possible to accurately identify attractive individual buildings or evaluate the quality of entire projects, the new approach now makes it possible to quickly find the most promising buildings, group them into optimal projects within existing project IDs, and ensure that capital expenditure is directed toward the highest-value opportunities. As a result, KPN can reduce reliance on manual and subjective decision-making, better align operational decisions with strategic objectives, and connect more homes (HCs) within the same budget. This maximizes the impact of investments and accelerates progress toward KPN's long-term strategic goals.

6.2 Discussion

This section reflects on the findings of the developed selection methodology. It discusses its broader implications, underlying assumptions, and limitations. The goal is to critically assess the robustness and applicability of the approach, and to identify factors influencing its real-world adoption and long-term value.

One of the main strengths of the developed methodology is its ability to translate large volumes of heterogeneous data into a single actionable score. By integrating predictive modeling, cost estimation, and feasibility logic, the approach provides a structured and actionable alternative to the previously used selection method. The composite score allows the Fiber Connect Planning & Capacity team to make decisions based on the predicted connection potential and cost of residential buildings.

However, the approach also introduces several assumptions and limitations. First, the Customer Uptake Prediction Model relies on historical data that reflects past behavior under specific market conditions. While the model generalizes well on the current dataset this could change over time. The reason for this is that shifts in customer behavior, competitive intensity, or promotional strategies could change its predictive power over time as its predictive power is solid but not incredibly strong. As a result, retraining and updating the model with features periodically will be important to maintain its performance over time.

Additionally, the Fiber Routing Cost Model uses a heuristic routing algorithm combined with contractor-specific delivery prices. While this method enables cost differentiation across delivery statuses and accounts for routing complexity, it still abstracts from contractor negotiation dynamics, terrain conditions, and address-level variation in trenching difficulty. As such, the estimated costs should be interpreted as relative rather than absolute indicators of financial feasibility. Furthermore, when using this method for the project formulation some routes may be taken into account multiple times indicating more costs than actually need to be made for the trenching as these could be combined. However, in general the logic holds as the buildings that require extensive trenching work will almost always remain bad choices for connection.

Furthermore, the feasibility logic, while based on expert rules and validated RNA codes, depends on the completeness and accuracy of the underlying administrative data. In practice, misclassifications or outdated contractor inputs could affect the reliability of feasibility labels. This highlights the importance of integrating expert validation or crowd-sourced updates into future versions of the model.

Another consideration is the assumption of independence between the model components. In reality, connectivity potential, cost, and feasibility may be correlated. For instance, a building with poor feasibility may also have low customer acquisition potential or high cost. While the composite score implicitly balances these factors, future research could explore joint modeling approaches to capture these dependencies in a better way.

Lastly, the approach assumes that decision-makers value cost and potential equally, as reflected in the default 50/50 weighting scheme. While sensitivity analysis confirmed the robustness of this assumption, the optimal weighting may vary depending on KPN's strategic priorities, budget constraints, or regional goals. A dynamic weighting mechanism, possibly integrated into the dashboard, would allow for more flexible adaptation to business needs.

Despite these limitations, the method represents a significant step forward in supporting data-driven infrastructure decisions at the address level. This is substantiated by the results in Chapter 5, which show that the method systematically ranks buildings based on their predicted customer potential, realistic connection costs, and feasibility constraints, producing outcomes that align closely with expert evaluations and historical project benchmarks. Its modular design and interpretability further facilitate internal adoption, ensuring that technical feasibility and commercial priorities are transparently incorporated into the selection process. Nonetheless, successful implementation will require ongoing maintenance, robust data governance, and alignment with evolving business processes.

6.3 Recommendations and Further Research

Based on the development and evaluation of the proposed selection methodology, several recommendations are formulated for both practical implementation within KPN and potential avenues for further academic or organizational research.

Practical Recommendations

Firstly, it is recommended that the developed model be integrated into KPN's operational roll-out process through a monthly updated dashboard. The current Power BI prototype provides address-level prioritization and project-level grouping but should be further developed into a standardized internal tool. It is also recommended to yearly re-optimize the model to improve performance when new data features become available.

Secondly, in order to maintain predictive relevance the Customer Uptake Prediction Model should be retrained on a monthly basis. This allows the model to adapt to shifting customer preferences, competitor strategies, and promotional effects. Regular performance monitoring should be conducted using updated labeled data, with key metrics like the AUC, F1-score, precision@k scores being tracked over time.

Thirdly, it is advised that KPN refines its cost estimation practices by incorporating additional geographic, technical, or contractor-specific variables. Currently, cost predictions rely on heuristics derived from routing distance and delivery status pricing. Including this kind of data on ground conditions, contractor performance, and project-level overheads could improve absolute cost accuracy and help inform CAPEX budgeting. Additionally, a different cost estimation heuristic for the trenching could work better for the project formulation part in the dashboard

Finally, it is recommended that KPN formalizes a feedback loop to capture feasibility feedback from field engineers and contractors. This would help to improve the quality of RNA data and ensure that feasibility scoring reflects real-world installation constraints. Such feedback could be incorporated into the dashboard or recorded through structured evaluation forms during site visits.

Further Research

Several directions for further research emerge from the findings and limitations of this thesis. First, future work could explore the use of multi-objective optimization or reinforcement learning to dynamically balance competing objectives rather than relying on static weighted combinations.

Second, the current approach assumes independence between model components. Future research could develop a joint probabilistic model that integrates cost, feasibility, and customer likelihood into a single estimation framework, thereby capturing potential interactions between variables.

Third, it would be valuable to investigate the effect of incorporating real-time external data, such as competitor roll-out activity, neighborhood demographics, or marketing engagement data. These could significantly improve the model's accuracy in predicting migration behavior and help identify high-risk areas for customer churn.

Fourth, the clustering logic used for project formation could be expanded into a formal project selection algorithm. Incorporating spatial optimization techniques, such as capacitated clustering or route-aware project selection, could yield even more efficient project configurations.

Lastly, a longitudinal study evaluating the actual uptake rates and financial returns from prioritized connections could validate the model's long-term impact. This would not only assess the predictive performance but also quantify the business value of the data-driven approach in practice.

In summary, while the current methodology provides a solid foundation for improved decision-making, future developments should focus on increasing predictive granularity. This will improve interpretability and will help the model with the operational uses of the model.

6.4 Theoretical and Practical Contributions

This thesis offers both theoretical and practical contributions at the domains of machine learning, infrastructure planning, and telecommunications operations.

From a theoretical standpoint, it addresses two underexplored areas in the academic literature. Firstly, machine learning has been widely applied in marketing and churn prediction. However, there is almost no literature available of studies on predictive customer acquisition modeling within the telecommunications industry. This is even more pronounced in the context of fiber-optic service uptake as there is literally no literature on this. Secondly, most literature on network expansion focuses on greenfield planning or technical design aspects, with little attention to brownfield expansion in existing residential environments. By modeling customer conversion likelihood, feasibility constraints, and connection costs at the building level, this thesis contributes a novel, data-driven perspective on how fiber-optic network expansion can be guided by predictive analytics in complex brownfield contexts.

Practically, the research delivers a decision-support system that transforms how KPN evaluates and formulates fiber-optic expansion projects from manual work to a more automated method. The integrated dashboard and composite scoring logic give a prioritization of connection candidates based on commercial potential, technical feasibility, and cost-efficiency. This can all be scaled and expanded automatically if required. Crucially, the solution enables retrospective project formulation for previously disconnected buildings. Until now, retrospective connections were restricted to individual high-rise buildings with specific delivery statuses. This thesis therefore lays the foundation for more systematic and strategic decision-making across KPN's brownfield portfolio, bridging predictive modeling and operational roll-out planning.

Bibliography

[1] Abdelrahim Kasem Ahmad, Assef Jafar, and Kadan Aljoumaa. "Customer churn prediction in telecom using machine learning in big data platform". In: *Journal of Big Data* 6.1 (2019), pp. 1–24.

[2] Ravindra K Ahuja, Thomas L Magnanti, James B Orlin, et al. *Network flows: theory, algorithms, and applications*. Vol. 1. Prentice hall Englewood Cliffs, NJ, 1993.

[3] David Alderson et al. "Understanding internet topology: principles, models, and validation". In: *IEEE/ACM Transactions on networking* 13.6 (2005), pp. 1205–1218.

[4] Johanna Axehill et al. "From Brownfield to Greenfield Development - Understanding and Managing the Transition". In: July 2021.

[5] Johanna Axehill et al. "From Brownfield to Greenfield Development - Understanding and Managing the Transition". In: July 2021.

[6] Candice Bentéjac, Anna Csörgő, and Gonzalo Martínez-Muñoz. "A comparative analysis of gradient boosting algorithms". In: *Artificial Intelligence Review* 54 (2021), pp. 1937–1967.

[7] Jon Louis Bentley. "Multidimensional binary search trees used for associative searching". In: *Communications of the ACM* 18.9 (1975), pp. 509–517.

[8] Christopher M Bishop and Nasser M Nasrabadi. *Pattern recognition and machine learning*. Vol. 4. 4. Springer, 2006.

[9] Leo Breiman. "Random forests". In: *Machine learning* 45 (2001), pp. 5–32.

[10] Tianqi Chen et al. "Xgboost: extreme gradient boosting". In: *R package version 0.4-2* 1.4 (2015), pp. 1–4.

[11] Carlos Mateo Domingo et al. "A reference network model for large-scale distribution planning with automatic street map generation". In: *IEEE Transactions on Power Systems* 26.1 (2010), pp. 190–197.

[12] Martin Geidl and Göran Andersson. "Optimal power flow of multiple energy carriers". In: *IEEE Transactions on power systems* 22.1 (2007), pp. 145–155.

[13] Hans Heerkens and Arnold Van Winden. *Solving managerial problems systematically*. Routledge, 2021.

[14] Alison J Heppenstall et al. *Agent-based models of geographical systems*. Springer Science & Business Media, 2011.

[15] Gareth James et al. *An introduction to statistical learning*. Vol. 112. 1. Springer, 2013.

[16] Guolin Ke et al. "Lightgbm: A highly efficient gradient boosting decision tree". In: *Advances in neural information processing systems* 30 (2017).

[17] M. Khelifi et al. "Hybrid heuristic for Capacitated Network Design Problem". In: *Journal of High Speed Networks* 21 (Nov. 2015), pp. 313–330. DOI: `10.3233/JHS-150528`.

[18] Youngjin Kim, Youngho Lee, and Junghee Han. "A splitter location–allocation problem in designing fiber optic access networks". In: *European Journal of Operational Research* 210.2 (2011), pp. 425–435.

[19] A. Kokangul and A. Ari. "Optimization of passive optical network planning". In: *Applied Mathematical Modelling* 35 (July 2011), pp. 3345–3354. DOI: `10.1016/j.apm.2011.01.017`.

[20] KPN. *Resultaten vierde kwartaal en jaarresultaten 2024 — overons.kpn.* `https://www.overons.kpn/nieuws/resultaten-vierde-kwartaal-en-jaarresultaten-2024/`. [Accessed 12-02-2025].

[21] Ji Li and Gangxiang Shen. "Cost Minimization Planning for Greenfield Passive Optical Networks". In: *Journal of Optical Communications and Networking* 1 (June 2009), pp. 17–29. DOI: `10.1364/JOCN.1.000017`.

[22] Thomas L Magnanti and Richard T Wong. "Network design and transportation planning: Models and algorithms". In: *Transportation science* 18.1 (1984), pp. 1–55.

[23] María Óskarsdóttir et al. "Social network analytics for churn prediction in telco: Model building, evaluation and network architecture". In: *Expert Systems with Applications* 85 (2017), pp. 204–220.

[24] O Ozkan and S Kilic. "A Monte Carlo simulation for reliability estimation of logistics and supply chain networks". In: *IFAC-PapersOnLine* 52.13 (2019), pp. 2080–2085.

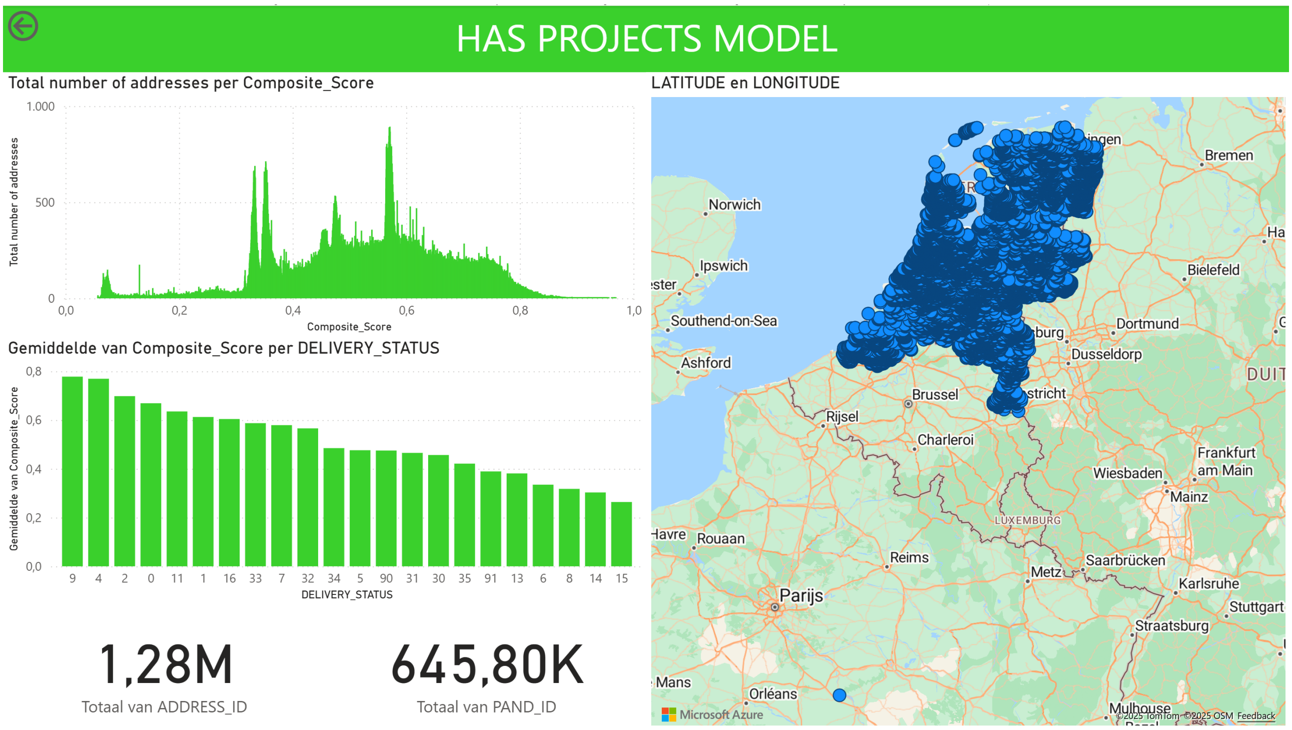
[25] Christos H Papadimitriou and Kenneth Steiglitz. *Combinatorial optimization: algorithms and complexity.* Courier Corporation, 1998.

[26] Parth H Pathak and Rudra Dutta. "A survey of network design problems and joint design approaches in wireless mesh networks". In: *IEEE Communications surveys & tutorials* 13.3 (2010), pp. 396–428.

[27] Iqbal H Sarker. "Machine learning: Algorithms, real-world applications and research directions". In: *SN computer science* 2.3 (2021), p. 160.

[28] Latha Narayanan Valli and N Sujatha. "Predictive Modeling and Decision-Making in Data Science: A Comparative Study". In: *2024 5th International Conference on Recent Trends in Computer Science and Technology (ICRTCST).* IEEE. 2024, pp. 603–608.

A Dashboard example



HAS PROJECTS MODEL

Total number of addresses per Composite_Score

Gemiddelde van Composite_Score per DELIVERY_STATUS

1,28M

Totaal van ADDRESS_ID

645,80K

Totaal van PAND_ID

LATITUDE en LONGITUDE

Figure A.1: *Dashboard front page*

B AI Statement

During the preparation of this master thesis, AI tools such as GitHub Copilot and ChatGPT were used for assistance in coding and refining written text. After using these tools, all text and code were reviewed, evaluated, and rewritten to reflect the author's own understanding, intentions, and academic standards. The final thesis is the result of the author's independent work.