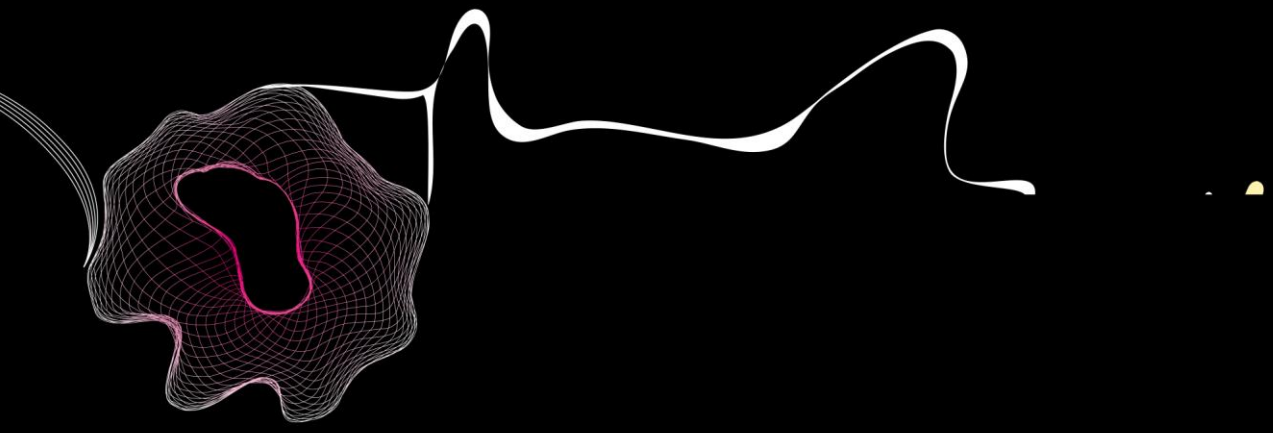
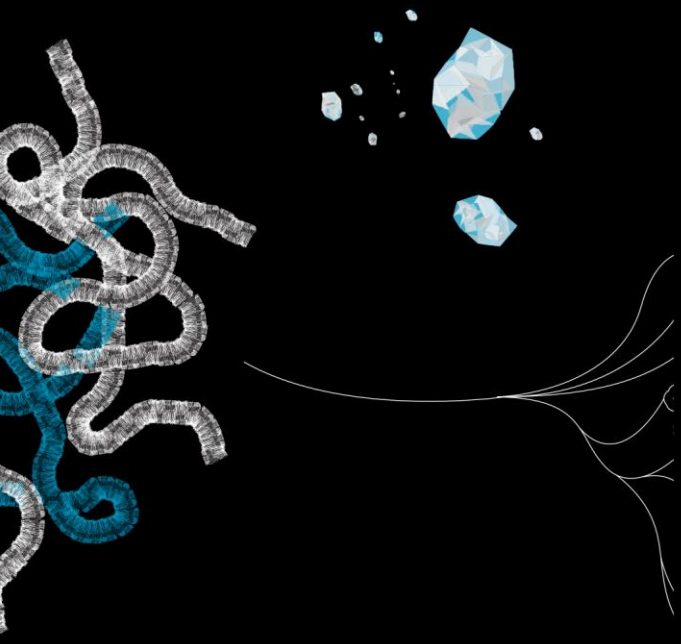
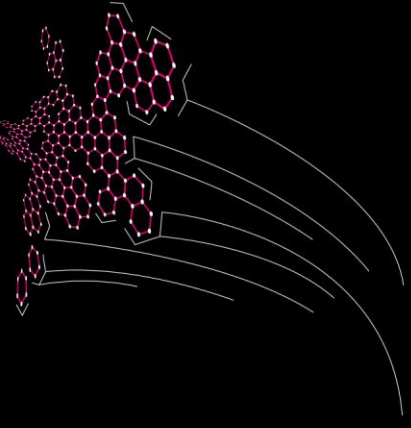


# Predicting mortgage demand using machine learning techniques



Kevin Bonnes
kevin.bonnes@topicus.nl



topicus
Fintech Software

UNIVERSITY OF TWENTE.

# MASTER THESIS

August 2017

Master Business Information Technology, University of Twente
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
Faculty of Behavioural, Management and Social sciences (BMS)

Parts of this thesis have been redacted due to confidentiality

# AUTHOR

Kevin Bonnes
kevin.bonnes@topicus.nl
Master Business Information Technology

# GRADUATION COMMITTEE

Dr. Ir. Maurice van Keulen
m.vankeulen@utwente.nl
Faculty EEMCS, University of Twente

Dr. Chintan Amrit
c.amrit@utwente.nl
Faculty BMS, University of Twente

Dennis Spangenberg
dennis.spangenberg@topicus.nl
Topicus

Stefan Hessels
stefan.hessels@topicus.nl
Topicus

# PREFACE

This thesis marks the end of my study period at the University of Twente. After completing my Bachelor Business & IT in 2014, this thesis is written as a final assignment for my Master Business Information Technology. It is a perfect example of combining both the Business and IT fields. By specializing in Business Analytics, I learned a lot about data science and its related topics. This thesis contains a practical example of applying this knowledge in a business context.

I would like to thank my supervisors at the University of Twente, Maurice van Keulen and Chintan Amrit, for their support, guidance and feedback during the project. Especially the feedback on my thesis provided me with useful directions on how to improve the quality of the thesis, and maintain a clear structure.

I would also like to thank Topicus, for providing me the opportunity and resources to work on my thesis. Thanks to my supervisors at Topicus, Dennis Spangenberg and Stefan Hessels, for their support and guidance during the project. In particular the standups every other day, and the retrospectives every two weeks, helped a lot during the project. Furthermore, thanks to Lennart Boot and Michel Brinkhuis for their feedback and help during the project.

Kevin Bonnes

August 2017

# ABSTRACT

In the Dutch mortgage market, it is difficult for the financial institutions to determine the amount of personnel needed to handle the mortgage applications coming in. There are multiple factors influencing the amount of mortgage applications, such as the mortgage interest rates, which cause the amount of mortgage applications to differ day by day. In this research we aim to provide more insight in the amount of personnel needed by developing a machine learning model that predicts the amount of mortgage applications coming in per day for the next week, using the CRISP-DM framework. After conducting a literature study and interviews, multiple features are generated using historical data from a Dutch financial institution and external data. A number of machine learning models are developed and validated using cross-validation. The predictions of our best model differ on average --- NUMBER REDACTED --- mortgage applications per day compared to the actual amount of mortgage applications. A dynamic dashboard solution is proposed to visualize the predictions, in which mortgage interest rate changes can be manually entered in the dashboard, and recommendations have been given for the deployment of the model at the financial institutions.

# MANAGEMENT SUMMARY


## Context

At Topicus, a software product called FORCE is developed, which is used at multiple financial institutions to manage the mortgage application process. A mortgage application coming in at the financial institution has to pass a large number of actions and checks before it can be turned into an offer. A large part of this mortgage application process is performed automatically by the system, but some of the actions have to be performed manually by employees of the financial institution, depending on the type of mortgage application and the quality of the data.

## Research Problem

In order to process the mortgage applications within the designated time standards, it is important for the financial institutions to have enough mortgage acceptors available to handle the continuous inflow of mortgage applications. The amount of mortgage applications coming in at a certain moment fluctuates significantly, and is influenced by multiple factors. Hence, it is difficult for the financial institutions to determine the optimal amount of personnel needed at any day, to balance the tradeoff between throughput time of the mortgage applications and personnel costs. In this research, a predictive model is developed using machine learning techniques that provides a prediction of the amount of mortgage applications per day coming in for the next week, in order to provide assistance with the personnel planning. Furthermore, an overview is given of the most important factors that influence the amount of mortgage applications, and recommendations are given on how to use the model at the financial institutions for personnel planning. The research question of our research is as follows:

*How can domain data be used to predict the amount of mortgage applications per day for the next week?*

## Methodology

In order to develop a machine learning model that can predict the amount of mortgage applications per day, for the next week, the CRISP-DM process model is used to structure the project. CRISP-DM is a process model that is used to ensure the quality of data mining projects. It describes the most common steps used in a data mining project and helps structuring the project. A literature study and multiple interviews were conducted in order to get an understanding of the context of the research problem, and to develop a list of possible predictors. Historical event log data and publicly available data were used as input for our predictive model, and five machine learning techniques (Decision Tree, Random Forest, Gradient Boosting Machines, Support Vector Regression and Neural Networks) were applied to create the predictions. The models are validated using repeated cross-validation, and evaluated using several evaluation criteria.

## Results

The results of our models can be found in the table below. The Random Forest model gave the best result on each of the four evaluation criteria used to evaluate the models. The Mean Absolute Error of the Random Forest model is --- NUMBER REDACTED --- mortgage applications per day. The Gradient Boosting Machines model scored slightly worse, on a second place. The percentual error of the Random

Forest model is around --- NUMBER REDACTED --- of the actual amount of mortgage applications per day.

| Model | RMSE | MAE | MAE/Mean | $R^2$ |
|---|---|---|---|---|
| Random Forests (RF) | --- RESULTS REDACTED DUE TO CONFIDENTIALITY --- | | | |
| Gradient Boosting Machines (GBM) | | | | |
| Support Vector Regression (SVR) | | | | |
| Neural Networks (NN) | | | | |
| Classification and Regression Tree (CART) | | | | |

The most important features of our model are as follows:

- The financial institution's interest rate
- Changes in the financial institution's interest rates
- The amount of mortgage applications on the previous day
- Holidays
- The day of the year

By analyzing the results of our model, we can see that in particular the 'outliers' (i.e. the days with an extremely high amount of mortgage applications) are consistently under-predicted. These outliers are often influenced by changes in mortgage interest rates, which implies that there is still room for improvement in our model. As the changes in interest rate are one of the most important features of our model and are influenced by many factors, and hence hard to predict, a dynamic dashboard is proposed. In this dashboard, interest rate changes can be entered manually, so that their impact on the amount of mortgage applications is shown real-time.

## Conclusions

A predictive model was created using the Random Forest technique, which predicts the amount of mortgage applications per day with a mean absolute error of --- NUMBER REDACTED --- mortgage applications per day. This can directly be converted to the amount of personnel needed at the mortgage application department of the financial institutions, by dividing it by the amount of mortgage applications handled per person per day.

The mortgage interest rates have the biggest impact on our model, but are difficult to predict. Hence, a dynamic dashboard solution was proposed, and a prototype was developed. This dashboard is yet to be validated, in order to see if it will be accepted by the stakeholders and provides significant value.

Several features can be added to the model in order to improve its predictive power. Amongst others Open Source Intelligence, relative interest rates to the financial institution's competitors and data regarding the marketing budget may provide additional value to our model. Furthermore, there is still improvement in the feature regarding mortgage interest rate changes, as a significant part of the error of our model is caused by under-prediction of the outliers.

# LIST OF FIGURES

# LIST OF TABLES

# GLOSSARY

| | |
|---|---|
| API | Application Programming Interface |
| ARM | Adjustable Rate Mortgage |
| BKR | Bureau Krediet Registratie |
| BVAR | Bayesian Vector Autoregressive model |
| CART | Classification and Regression Trees |
| CBS | Centraal Bureau Statistiek |
| CCC | Correctheid en Compleetheid Controle |
| CRISP-DM | Cross Industry Standard Process for Data Mining |
| CRM | Customer Relationship Management |
| DNB | De Nederlandsche Bank |
| ECB | European Central Bank |
| Euribor | Euro Interbank Offered Rate |
| FinTech | Financial Technology |
| FRM | Fixed Rate Mortgage |
| GBM | Gradient Boosting Machines |
| IDE | Integrated Development Environment |
| LTV | Loan-To-Value |
| MAE | Mean Absolute Error |
| MAE/Mean | Mean Absolute Error divided by the mean |
| MLR | Multiple Linear Regression |
| NHG | Nationale Hypotheek Garantie |
| NN | Neural Networks |
| OSINT | Open Source Intelligence |
| RF | Random Forest |
| RFE | Recursive Feature Elimination |
| RMSE | Root Mean Square Error |
| SaaS | Software-as-a-Service |
| STP | Straight-Through Processing |
| SVR | Support Vector Regression |

# TABLE OF CONTENTS

# 1 INTRODUCTION

*In this chapter, the research domain is introduced. An introduction on Topicus is given as well as the software product called FORCE, which can be used to manage the mortgage application process for financial institutions. A brief introduction is given on the Dutch mortgage market, and the position of the* financial institutions *in the Dutch mortgage market, and finally an introduction is given on the mortgage application process in FORCE.*

## 1.1 TOPICUS

Topicus is an IT service provider that offers software solutions for different industries. The company is located in Deventer, but has multiple offices throughout The Netherlands. Topicus was founded in 2001 by five employees, but has currently grown to over 650 employees and is still growing.

The company specializes in chain integration and Software-as-a-Service (SaaS) solutions. Chain integration is the concept of integrating different systems within a chain in a business process and facilitating information exchange between them. This is often done using Application Programming Interfaces (APIs). An API is a set of standards that define how one can communicate with a specific software system.

Topicus offers software solutions in different sectors: finance, healthcare, education, government and legal. Each of these sectors consists of different business units, and each business unit consists of one or multiple teams. At Topicus.Finance a software product called FORCE is developed, which can be used to manage mortgage requests and quotations. Multiple financial institutions use FORCE. Each of these financial institutions has their own implementation of the software product, with customized functionality, and sometimes their own teams within Topicus that are dedicated to their implementation of the product. The implementations of FORCE used by these financial institutions are different from the standard FORCE product because they need to provide integration with a number of external systems used at these financial institutions.

In FORCE, the financial institution's employees can process and manage mortgage applications. A mortgage is a loan given by a financial institution to a house owner, in which the borrower's property functions as a security for the loan. Since most individuals do not have the funds to buy a house straight away, they can apply for a mortgage so that the mortgage lender helps them providing funding. The mortgage borrower then pays a monthly payment to the lender, until the debt is paid off. The lender also expects an interest rate to be paid as a compensation. There are different types of mortgages. The two basic mortgage types are the fixed rate mortgage (FRM) and the adjustable rate mortgage (ARM). The difference between these types is the interest rate, which either remains fixed for a certain period of time or is adjustable depending on interest market index. Next to this there are differences in the payment schemes, the two main types are linear mortgages and annuity mortgages. The difference between these types is the repayment scheme. In an annuity mortgage, the monthly payment increases gradually, whereas it remains constant in a linear mortgage.

## 1.2 DUTCH MORTGAGE MARKET

Topicus provides FinTech solutions for multiple companies in the Dutch mortgage market. FinTech or "Financial technology" is the term that refers to the use of technology to deliver financial solutions (Arner, Barberis, & Buckley, 2015). It describes the digitalization of the financial industry, and is often seen as the intersection between financial services and information technology. It is a term that has

gained a lot of popularity lately, and has attracted interest from both industry participants and consumers. It aims to provide automated financial services that reduce the throughput time of transactions and transaction costs (Dapp, Slomka, AG, & Hoffmann, 2014).

In the Dutch mortgage market there are numerous different mortgage lenders, which can roughly be categorized in three different categories: banks, insurance companies and other financial institutions. The banks are generally the most influential type of mortgage lenders, and have a combined market share of about 60% (van Dalen, 2016). Insurance companies and other financial institutions both have a market share of about 20% of the total mortgage market in The Netherlands.

Recently, the competitive position of the banks in the Dutch mortgage market has received some pressure. Due to the entry of new players in the Dutch mortgage market, that can offer mortgages at a lower interest rate, the market share of the big banks has decreased by about 20% over the last four years (van Dalen, 2016). This can be seen in Figure 1. These new entrants have lower costs than the traditional players, and can thus provide 'cheaper' mortgages in terms of interest rates, which can often save the consumers a significant amount of money. Due to this, they have taken a considerable amount of market share and are slowly pressuring the competitive position of the big banks, which are having difficulties to remain competitive in terms of interest rates. This has led to an increased interest in FinTech from the banks, hoping that it allows them to regain market share by increasing their competitiveness in terms of interest rates and reduced costs.

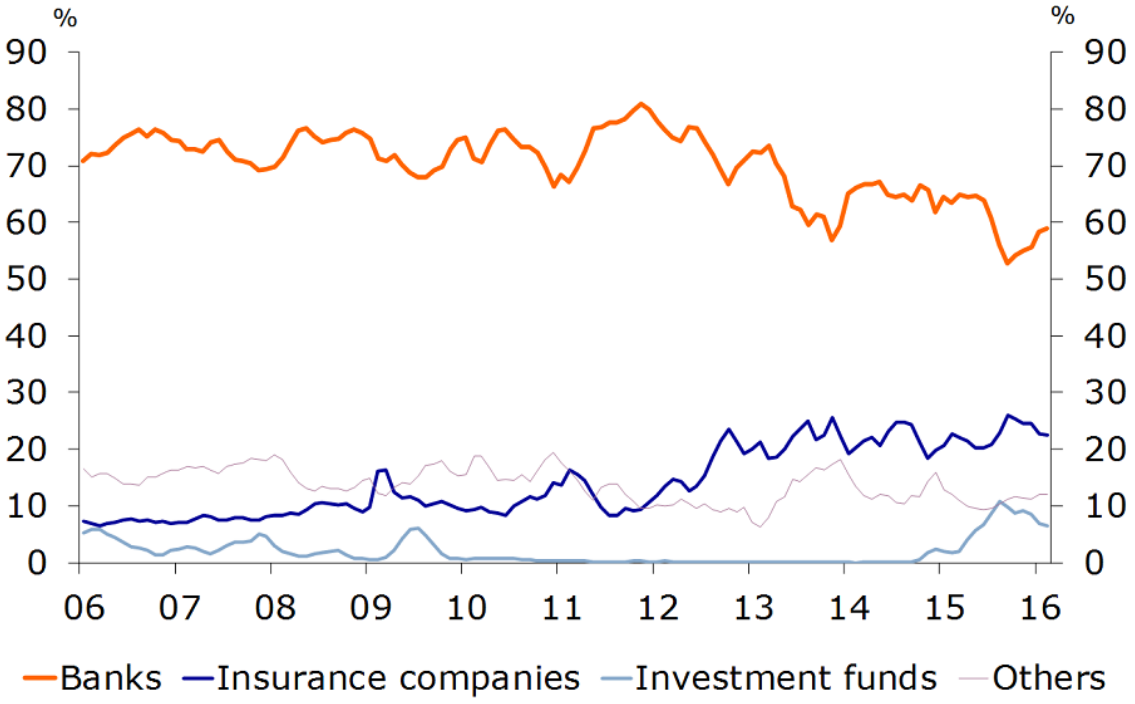



*Figure 1 – Competitive position of financial institutions in the mortgage market. Retrieved from (van Dalen, 2016).*

## 1.3 Mortgage application process

FORCE is a software product developed by Topicus, used to handle the mortgage application process. With FORCE, a large part of the mortgage application process can be automated. It is a mid- and backoffice product, which offers functionality for processing and managing mortgage applications. The financial institutions that work with FORCE can have their own frontend systems, in which a mortgage advisor or a user can apply for a mortgage. Once the application is completed, the frontoffice systems

can send it to FORCE. When a mortgage application enters the system in FORCE, several main steps will be conducted. First, the information in the application is extracted and connected to the CRM-system of the financial institution. Afterwards, a Correctheid en Compleetheid Controle (CCC) is performed, to ensure the completeness and correctness of the information. Afterwards, multiple other steps are conducted, such as reviewing the information and documents in the mortgage application and conducting a credit check at Bureau Krediet Registratie (BKR), in order to see if the applicant is creditworthy. Finally, the application is either accepted, rejected, or sent to another department for further checks. If the application is accepted, a quotation will be made by one of the employees. This entire process of handling a mortgage application is too complex to be shown in detail, as there are dozens of other steps and statuses that a mortgage application will undertake, up until the quotation process.

In the best possible case, when all of the checks are positive and the mortgage application gets accepted, the process up until the quotation is completely automated. This is called Straight-Through Processing (STP). For all other requests, manual input will be required from the financial institution's employees.

All of these steps that are performed by the system or by any of the employees are logged in a database. This log data contains multiple variables, amongst others the id of a mortgage application, the status of the mortgage application, the time and date of the status change and the employee that performed the action. This date and timestamp can be extracted from the database and can be used to train a predictive model.

Since the amount of applications and the nature of these applications differ from day-to-day and hour-to-hour, it is difficult to predict how many employees will be needed in order to handle these applications within a certain time limit. There are a number of factors that influence this, for example the interest rates (e.g. when the interest rate is low, more mortgages are sold) and the time of the year (e.g. a higher number of mortgage applications at the end of the year, a lower amount of mortgage applications during the summer holidays). Once we can make an accurate prediction of this, we will be able to determine the amount of personnel needed more accurately.

# 2 RESEARCH PROBLEM

*This chapter covers the research problem of our research, and describes how this research will provide value to Topicus and its customers. First, the motivation of the research is discussed, which provides the reasoning behind our research problem. Second, the research question and its subquestions are proposed.*

## 2.1 MOTIVATION

The mortgage application process consists of multiple actions and subprocesses, of which a large part is performed automatically, without human interference. However, some of the actions have to be performed manually by employees of the financial institutions. This entirely depends on the type of application. Each application has different characteristics so the processing time is never the same. For example, applications submitted by entrepreneurs require manual processing and thus have a longer processing time, due to the fact that they have to add extra documents in their application (Geertsma, 2016).

Due to the variability in the amount of mortgage applications and the different natures of them, it is difficult for the financial institutions to determine how much personnel is needed at any given time to process these applications. Currently they have to keep on scaling up and down in terms of personnel real-time manually, in order to keep the processing time of these applications within the designated time standards, which is not cost-efficient. Providing insight into the amount of personnel needed at any given time can significantly decrease the personnel overhead by preventing personnel over-allocation, and it lowers the average throughput time of mortgage applications and helps preventing outliers in the throughput times.

The entire process of getting insight into the amount of personnel needed at any given time is a too complex research problem to be handled at once, but it can be split up in multiple smaller research problems, that can each be handled within a master thesis. For example, by providing a more accurate prediction of the amount of mortgage applications and their expected processing time, the amount of personnel needed at any given time could be more accurately predicted. In this research we will focus on one of these research problems: predicting the amount of mortgage applications coming in at any specific time within FORCE, in order to get more insight in the amount of personnel needed at the dedicated departments of the financial institutions. This will be done by creating a predictive model. Not all of the mortgage applications will lead to an actual mortgage, but the amount of mortgage applications generally does give a decent indication of the expected activity on the mortgage market (Boumeester, 2016). This study will contribute to this by conducting a subpart of this research problem, and providing recommendations for future research, for example for a subsequent master thesis.

In terms of business value, Topicus is currently working on a management information dashboard for its customers, in which they can incorporate this information in order to improve the value of the dashboard. Next to this, Topicus may want to use this information for future purposes, for example in one of their start-ups called Jungo. Jungo is a mortgage lender that makes use of 'crowdlending', it allows third-party lenders to crowdfund a part of the mortgage, in order to lower the interest rates ("Jungo," 2016).

For Topicus' customers, the goal is to reduce the costs involved with scaling their personnel up and down due to an unexpected increase or decrease in demand. Currently, domain experts estimate the amount of personnel needed by looking at several factors. The predictive model can assist these experts

in decision making, and provide a better estimation. Predictions will be made for a timeframe of one week.

## 2.2 Research Questions

Based on the problem mentioned above, the following research question was defined.

RQ. How can domain data be used to predict the amount of mortgage applications per day for the next week?

In order to answer this research question, the following subquestions were defined.

SQ1. What are the variables in the domain data that influence the amount of mortgage applications?
SQ2. Which techniques and algorithms can be used to create a model that is able to predict the amount of mortgage applications?
SQ3. Which technique performs best on our dataset?
SQ4. How can we use this model to determine the amount of personnel needed during the next week?

# 3 RESEARCH METHODOLOGY

*In this chapter the research methodology is discussed. An overview is given of the CRISP-DM model, and each of its stages is discussed briefly and applied on our research project. Finally, an explanation is given of the tool selection and the remaining structure of this report is discussed.*

## 3.1 CRISP-DM

In order to complete the project and develop a valid predictive model several steps will need to be conducted. To model these steps, the CRISP-DM process model will be used (Chapman et al., 2000). An overview of the CRISP-DM model can be found in Figure 2 (Alnoukari & El Sheikh, 2012).

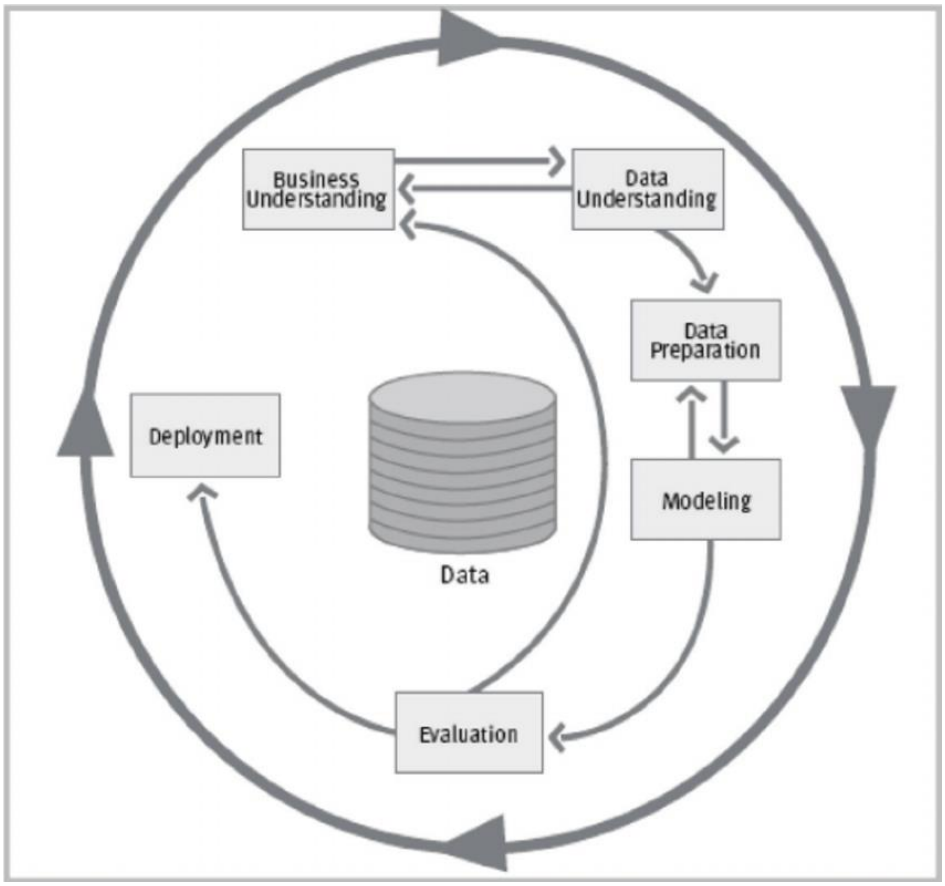



*Figure 2 – CRISP-DM Process Model. Retrieved from (Alnoukari & El Sheikh, 2012)*

The CRISP-DM model is a process model to ensure the quality of knowledge discovery project results (Chapman et al., 1999). It describes the common steps used in a data mining process and helps structuring the project (Wirth & Hipp, 2000). It is considered as the leading methodology for data mining and knowledge discovery projects (Kurgan & Musilek, 2006; Marbán, Mariscal, & Segovia, 2009; Piatetsky, 2014).

The CRISP-DM model consists out of six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation and Deployment. The sequence of these stages is not strict, the arrows only indicate the most important flows, but in any particular project the sequence of these steps

may vary (Wirth & Hipp, 2000). In general, this is an iterative process in which new features are continuously created and evaluated.

In our research, the CRISP-DM model will be used as a framework for our predictive model. Each of these six stages contain a number of activities, but not all of these are applicable to our research project. In order to customize the framework to our research project, we have mapped the different stages of the CRISP-DM framework onto the activities required in our research project. An overview of this can be found in Figure 3. Below, the six stages as well as the activities in these stages are explained in more detail and are applied to our project.

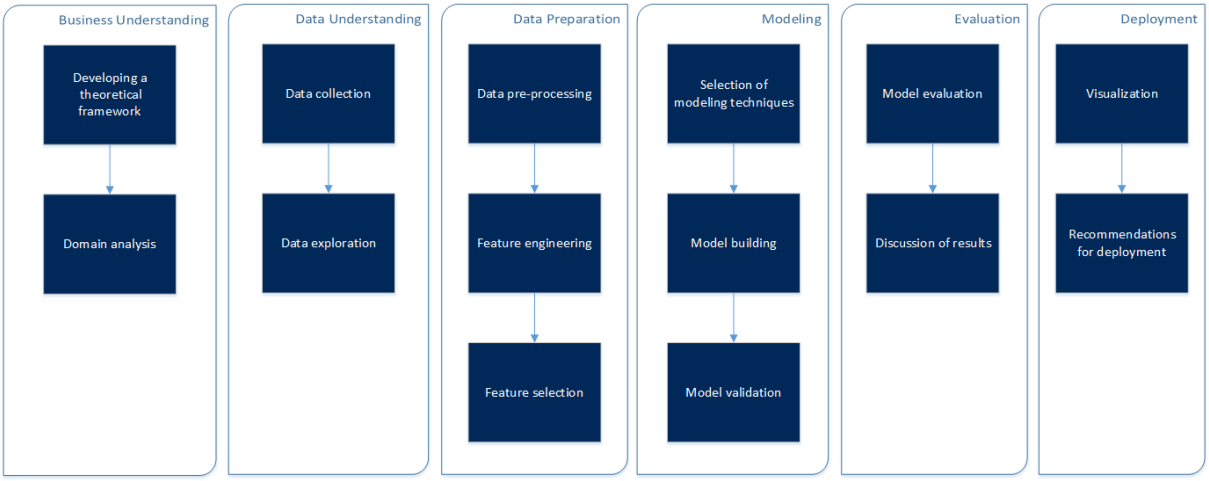



*Figure 3 – Research Method using the CRISP-DM framework*

### 3.1.1 Business Understanding

In the first stage, a theoretical framework is developed and background information is collected on the domain. This is done in order to get an overview of the objectives and requirements of the final solution from a business perspective. A theoretical framework is developed by conducting a literature review. The theoretical framework contains an overview of the related research in this research area, and an overview of the concepts in predictive analytics and the different models that are feasible for our research project. The theoretical framework covers the most important terms and theories from the different topics that are involved in our research.

For the domain analysis, background information on the mortgage application process and the mortgage domain is collected in order to get a better understanding of the different topics. Also a list of possible predictors is developed by using literature and conducting interviews. An initial list of predictors is formulated using literature, and interviews with several domain experts within Topicus are used to validate that list.

### 3.1.2 Data Understanding

In the Data Understanding stage the raw data is collected from the database using an SQL query and its characteristics and distributions are explored. Event logs are kept in the database and can be used for predictive modeling. Once the data is collected, the data is explored and visualizations are made of the different variables to get an understanding of the data. With these visualizations we can already see some of the relationships in the data and identify possible features. The data exploration activity is important for becoming familiar with the data and identifying data quality problems (Chapman et al., 2000).

### 3.1.3 Data Preparation

The goal of the Data Preparation stage is to transform and enrich the dataset so that it can be fed into the models. After the data is collected and explored, it can be pre-processed so that it can be used directly in our predictive model. With the pre-processed data one can perform feature engineering. Using historical data and external data, different features can be generated. For some of these features the data has to be collected first, from publicly available sources (e.g. Centraal Bureau voor de Statistiek). After the feature engineering activity, a subset of features will be selected that provide predictive value for our models.

### 3.1.4 Modeling

In this stage, several models are developed based on the dataset. First, a selection of predictive models is made (e.g. Neural Networks, Random Forests). These models are trained on the dataset and used to make predictions. The models are validated using a test set and repeated 10-fold cross-validation (Friedman, Hastie, & Tibshirani, 2001). For each of the models, hyperparameters are optimized and data pre-processing is done if needed (e.g. centering, scaling, multi-collinearity checks). Some of the models have specific requirements on the form of the data, which require specific pre-processing activities (Chapman et al., 2000).

### 3.1.5 Evaluation

In the Evaluation stage, the results of the different models in the previous stage are discussed and evaluated, and a final model is selected. Model evaluation will be done using several criteria, amongst others the performance of the model and the model interpretability. For ensuring the validity of our research, cross-validation is used. The selection of a final model is done by a t-test.

### 3.1.6 Deployment

In the last stage, our model is fed with real-time data and predictions are made for the upcoming period. These predictions are visualized in a dashboard. Furthermore recommendations for deployment are written to give the financial institutions some guidance on the deployment of our model, and to use it within their decision making processes.

## 3.2 TOOL SELECTION

There are a number of tools available for performing predictive analytics, both open source tools and commercial tools. Some of the most popular open source tools are RStudio, Weka, RapidMiner and KNIME. According to multiple studies, the differences between these tools are minor, and the tool selection is mostly based on personal preferences (Al-Odan & Al-Daraiseh, 2015; Meka & Patil, 2015). Because of its ease of use and abilities to visualize the data, RStudio will be used for our project ("RStudio," 2016). RStudio is an integrated development environment (IDE) for R, a programming language for data analysis and statistics ("R: The R Project for Statistical Computing," 2016). It contains a wide range of predictive models in different libraries.

## 3.3 STRUCTURE OF THIS REPORT

In the remainder of this report we will discuss the six stages of CRISP-DM and its activities in more detail, as described in Figure 3. Each of the stages will be discussed in its designated chapter. Chapter 4 discusses the Business Understanding stage, including a domain analysis and an overview of the possible predictors. In Chapter 5, the Data Understanding stage is discussed, including the data collection and data exploration activities. Chapter 6 discusses the Data Preparation stage, in which a list of features is generated that will be used for our models. In Chapter 7, the modeling activities are discussed. A selection of modeling techniques is made and several models are built and validated. In Chapter 8, a

final model is selected and the results are evaluated. Chapter 9 discusses the deployment stage of the CRISP-DM framework, and recommendations are given for the financial institutions. Finally, in Chapter 10, the conclusions of our research project are presented, as well as limitations and recommendations for further research.

# 4 BUSINESS UNDERSTANDING

*In this chapter a domain analysis is conducted using interviews and a literature study, and an overview of predictors is presented. Also an introduction is given on predictive analytics. In the domain analysis, interviews and a literature study are conducted in order to get an overview of related research and provide a list of possible predictors. Also the dynamics involved with the mortgage interest rates and mortgage interest rate changes at the* financial institutions *are discussed, as well as changes in rules and regulations. In the last subchapter, several types of problems in predictive analytics are discussed and an overview of appropriate models is presented.*

## 4.1 DOMAIN ANALYSIS

An important aspect of a data analysis project is obtaining domain information, by conducting a domain analysis. A domain analysis can be used to get an understanding of the different factors that influence the amount of mortgage applications and to formulate a list of possible predictors that can be used for our model. There are multiple studies that confirm the importance of conducting a domain analysis (Kopanas, Avouris, & Daskalaki, 2002; Wu, Zhu, Wu, & Ding, 2014). In our research project, the domain analysis will consist out of a literature study and several expert interviews. A part of the domain analysis has already been discussed in Chapter 1, the remaining information will be presented in the following subchapters.

### 4.1.1 Related work

Most of the research done in the mortgage prediction domain is kept internally at the financial institutions, due to the classified nature of the data and the importance of the results. Financial institutions that provide their data for data analysis generally use the results of this analysis for own use. Their goal is to improve their internal processes and to provide competitive advantage to their company, and generally do not want their competitors to have access to the analysis. This conflicts with the goal of scientific research, to provide publicly accessible information that can be used for further research. This is one of the reasons that there is a lack of scientific research publicly available. However, albeit the scarcity of literature on predicting mortgage applications, there are a number of papers that provide similar research in related domains, for example the prediction of home sales.

In a study performed on the Swiss mortgage market, the authors show the relationship between house prices and mortgage demand (Basten & Koch, 2015). The authors investigate the different causal relationships between house prices and mortgage demand, and use regression analysis to calculate the correlations between these two variables. They found that house prices do not only influence mortgage demand, but that mortgage demand in return also influences the house prices. Both correlations are positive, the variables affect each other in a positive way.

A study performed by Dua & Smyth (1995) uses a Bayesian Vector Autoregressive model (BVAR) to predict home sales, using variables such as housing prices, mortgage interest rates, unemployment rates and real disposable income as input for the model. Another study by Dua, Miller & Smyth (1999) reaches similar conclusions. They forecast US home sales based on a combination of different variables. They found that variables with a longer lead seem to perform better in their model than variables with a shorter lead.

Baghestani, Kaya & Kherfi (2013) have shown that changes in consumer's assessment of house prices and mortgage interest rates have impact on home sales. They have conducted surveys to measure these assessments. The study has shown that changes in these assessments are able to predict the direction of change in home sales 3 months later.

In a study conducted by Gupta, Tipoy & Das (2010), the authors have used several univariate and multivariate models to predict home sales in the US, using amongst others variables such as home prices, mortgage interest rates and unemployment rates. The results have shown that Bayesian models seem to outperform the other models used in this study.

There are also a number of papers predicting the probability of default of a mortgage loan. The probability of default denotes the chance that a mortgage borrower is unable to make its payments. Galindo & Tamayo (2000) have used a number of machine learning models to predict the probability of default. The results show that the Classification and Regression Trees (CART) algorithm performed the best, with an average error rate of 8.31%. In another paper, Feldman & Gross (2005) also use CART to predict the probability of default. Their results show that the borrowers' features, rather than the mortgage contract features, are the best predictors of the probability of default.

The dynamics around mortgage interest rates are also interesting for our research. There are a number of papers that explain parts of these dynamics for the Dutch mortgage market (De Haan & Sterken, 2011; Toolsema & Jacobs, 2007). Amongst others, these papers show that the mortgage interest rates seem to respond asymmetrically to cost changes in the Dutch mortgage market, rising faster than falling. In general, there can be many reasons that influence the mortgage interest rates, amongst others the degree of competition, the costs of lending, the risks financial institutions are facing and regulatory measures on the financial institutions (Mulder & Lengton, 2011).

### 4.1.2 Interviews

In order to get an overview of the different aspects involved in the mortgage application domain, multiple interviews were conducted with domain experts within Topicus. Since not all of the domain knowledge is available via scientific literature, and a part of the domain knowledge is client-specific and may contain sensitive data that can provide a competitive advantage in the mortgage market, we use interviews to fill this gap.

In general, three types of interviews can be categorized: unstructured interviews, structured interviews and semi-structured interviews (DiCicco-Bloom & Crabtree, 2006). The main difference between these categories is the degree to which questions are formulated upfront and whether one can divert from these questions or not. In our research project we use unstructured interviews. Unstructured interviews provide the benefit that the interviewer does not need to restrict the questions that can be asked (Doody & Noonan, 2013). Unstructured interviews are particularly useful when little is known about a certain topic, or for collecting background data (Ryan, Coughlan, & Cronin, 2009). It offers the benefit of asking in-depth questions on a certain topic.

For our interviews, even though we did not use a predefined set of questions, a list of topics was formulated to form the 'skeleton' of the conversation. Publicly available information was used to develop an initial list of predictors, and the interviews were used to validate and extend this list. Furthermore, the interviews were used to provide insight in the different aspects involved in the mortgage domain, and the context of our research.

A total of five interviews were held with different business experts at Topicus. Four of the interviewees are product owners at different teams within the mortgages business line. The fifth interviewee is a mortgage advisor. The results of the interviews are presented in the next subchapters, a summary of the individual interviews (in Dutch) can be found in Appendix A.

## 4.2 PREDICTORS OF THE AMOUNT OF MORTGAGE APPLICATIONS

"Garbage in, garbage out" is a widely used term in the field of computer science **("Garbage in, garbage out," 2005; Hand, 1999)**. The term refers to the fact that software systems will only provide sensible output if they are provided with the right input. Even the most intelligent software systems will produce incorrect output as long as the input data is nonsensical. Basically, the performance of our model is mostly determined by the quality of our input data. If the different predictors we choose to include in our model do not have enough predictive power, the model will never be able to provide any accurate predictions. In order to prevent this risk we will combine our domain knowledge with the domain knowledge of several experts within Topicus in order to create a list of possible predictors.

### 4.2.1 Overview of predictors

An overview of the predictors can be found in Table 1. In the next subchapters, the predictors will be discussed in more detail.

*Table 1 – List of possible predictors*

| Category | Predictor | Mentioned in literature | Mentioned in interview(s) |
|---|---|---|---|
| **Seasonality** | Date | X | 1, 5 |
| | Vacations | X | 1, 4, 5 |
| | Holidays | | 1 |
| | Historical amount of mortgage applications | X | |
| **Mortgage interest rates** | Mortgage interest rates | X | 1, 2, 3, 4, 5 |
| | Changes in mortgage interest rates | X | 1, 2, 3, 4, 5 |
| | Relative interest rates | | 3 |
| **Changes in regulations** | Changes in regulations | X | 1, 2, 3, 4, 5 |
| **State of the housing market** | House prices | X | 4, 5 |
| | Rental prices | X | 5 |
| | Amount of houses available | X | 4, 5 |
| **State of the economy** | Economic growth | X | 4, 5 |
| | Income growth | X | |
| | Affordability | X | |
| | Consumer confidence | X | |
| | Propensity to buy | X | |
| | Unemployment rates | X | |
| **Influence of the media** | Influence of the media | | 3, 4, 5 |

### 4.2.2 Mortgage interest rates

One of the most discussed factors in the literature and the interviews is the mortgage interest rate **(Basten & Koch, 2015; Boumeester & Lamain, 2016; Deira, 2015; Pettinger, 2009, 2013)**. Mortgage interest rates have a significant impact on the amount of mortgage applications. If the interest rates are low, the mortgages are relatively cheaper for the borrower as they have to pay less interest, which leads to an increased amount of mortgage applications. A high mortgage interest rate means the mortgage borrower pays a high amount of interest to the lender, which makes the mortgage less attractive for the borrower. Interest rate changes have a significant impact on mortgage applications, as was seen in November of last year, where a sudden increase in interest rates led to a large peak in mortgage applications **(Mebius & Haegens, 2016; "Piek in hypotheekaanvragen," 2016)**. The impact of mortgage interest rates on the amount of mortgage applications was confirmed in the interviews.

As mentioned in Chapter 1.2, there is a stiff competition between banks, insurance companies and other competitors in the mortgage market in The Netherlands. The main difference between the mortgages offered by these types of companies lies in the mortgage interest rates. Even a small difference in mortgage interest rates can often save or cost the borrower a vast amount of money, due to the large sum of a mortgage. In Figure 4, an overview can be found of the historical interest rates over the last 15 years (Vrieselaar et al., 2017). From the figure, it can be seen that the mortgage interest rates are at its lowest point since 2003.

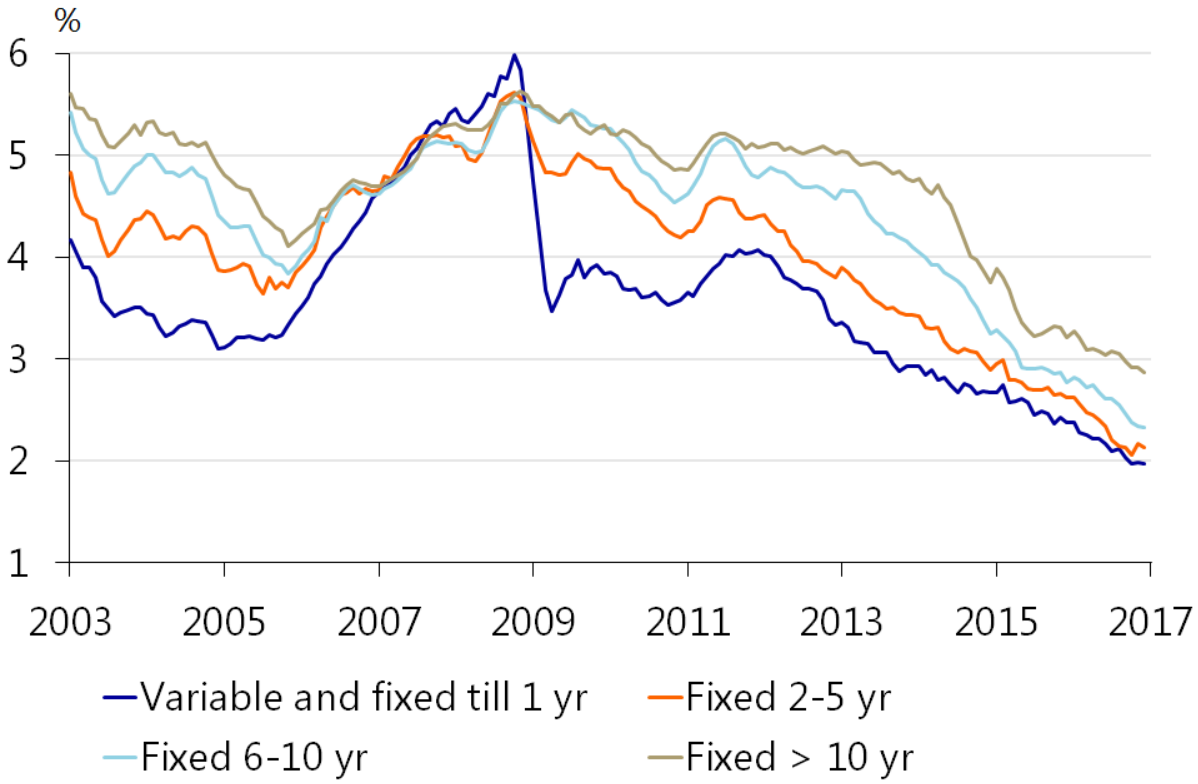



*Figure 4 – Mortgage interest rates are currently at its lowest point. Retrieved from (Vrieselaar et al., 2017)*

In general, there are two types of mortgage interest rate: variable rates and fixed rates. Variable interest rates are generally lower than fixed interest rates, but can change every month. Fixed interest rates are slightly higher, but are fixed for a certain period of time. When this period of time increases, the mortgage interest rates also go up. E.g., the mortgage interest rate for a fixed period of 30 years is higher than the mortgage interest rate for a fixed period of 5 years. A fixed interest rate is generally preferred when the mortgage interest rates are expected to rise, or when the borrower wants to know its monthly expenses upfront. A variable interest rate is preferred when interest rates are expected to decrease. As can be seen in Figure 5, the long-term fixed interest period have received increased popularity the last two years, due to the low mortgage interest rates (Vrieselaar et al., 2017). Next to the type of interest rates, there are multiple other factors that affect the interest rates, such as the default risk of the borrower (i.e. with a higher default risk, the lender requires a higher interest rate) and interest rate discounts.

Furthermore, interest rates are also influenced by the cost of lending for the financial institutions itself. By far the biggest part of a mortgage is funded by debt, which the financial institutions lend from the ECB or from other financial institutions. The mortgage interest rate depends heavily on the cost of which the financial institution can get its funding, and the fixed interest period of the lending. For the short-term interest rates (i.e. the interest rates with a fixed interest period of up to one year), the Euribor can be used as an indicator for the mortgage interest rate. For the long-term interest rates (i.e. interest

rates with a fixed interest period of 5 to 30 years), the capital market interest rates can be used as an indicator.

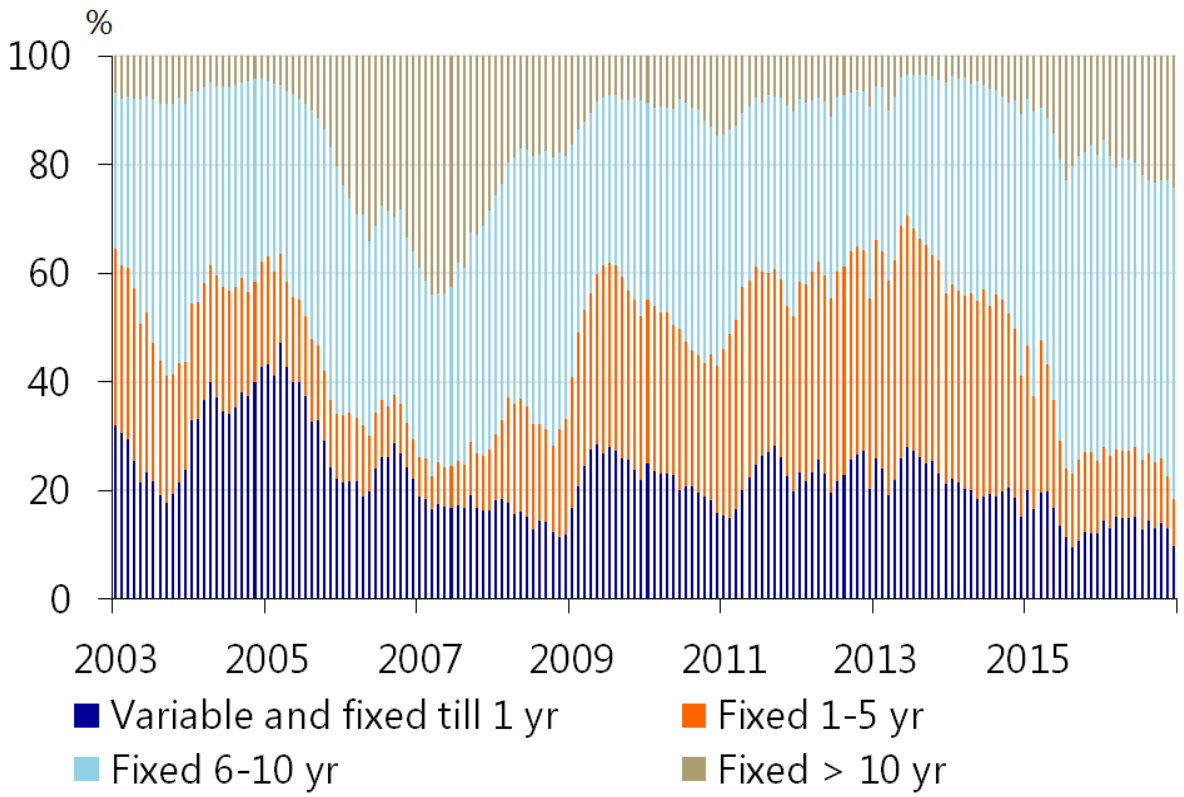



*Figure 5 – Long-term fixed interest rate periods are receiving increased popularity. Retrieved from (Vrieselaar et al., 2017).*

As mentioned in the interviews, for the financial institutions, the relative height of its mortgage interest rates compared to their competitors is important, as it has a large impact on the amount mortgage applications coming in. If a financial institution has a significantly higher interest rate than its competitors, it will generally receive fewer mortgage applications as the independent mortgage advisors will forward its customers to a different mortgage lender.

--- PARAGRAPH REDACTED DUE TO CONFIDENTIALITY ---

Besides the mortgage interest rate itself, interest rate changes also have a big impact on the amount of mortgage applications. Whenever the interest rate will decrease, one can expect a sudden drop in the amount of mortgage applications right before the interest rate decrease, and a peak in the amount of mortgage applications right after the interest rate decrease. With an interest rate increase, this relationship goes the other way around: right before an increase in interest rate the amount of mortgage applications peak, as consumers generally want to submit their mortgage application against the lowest rate, and right after an increase in interest rate the amount of mortgage applications drops.

As mentioned in the interviews, changes in mortgage interest rates are often announced somewhere between 1 and 2 days before the actual change. The financial institutions deliberately announce these as late as possible, so that they do not provide their competitors with useful information. Mortgage advisors then have a few days to submit their mortgage applications, in case they want to make use of the old interest rate. Advisors often have multiple mortgage applications ready to be submitted, and are waiting for the best moment in terms of interest rates for the actual submission.

One of the factors influencing these decisions are the predictions for the mortgage interest rate for the upcoming period (i.e. if the mortgage interest rate is expected to decrease, it can be beneficial to wait with submission). There are multiple financial institutions in the Dutch mortgage market that provide mortgage interest rate predictions on a monthly or quarterly basis (Bokeloh, 2017; Vrieselaar et al., 2017). These predictions give an indication of what the interest rate might do in the upcoming period, but by no means give an accurate prediction. Even the domain experts seem to disagree every now and then, and the interest rates are dependent on so many factors that it is often difficult to provide an accurate prediction.

For the financial institutions, there can be a number of reasons to change its mortgage interest rate, as mentioned in the interviews. First of all, the mortgage interest rate is based on the cost of lending for the financial institutions itself. If the cost of debt is higher, the financial institutions will compensate this by charging higher interest rates for its mortgages, in order to keep a profitable margin on their products. This cost of lending is mainly based on the capital market interest rate, for the long-term loans, and the Euro Interbank Offered Rate (Euribor), for the short-term loans. If either of these changes significantly, one can expect the financial institutions to respond by changing their own mortgage interest rates. This usually happens after a few days.

Second, financial institutions generally work with a budget for their mortgages. Based on the amount of funding they can get, and on the interest rates and the duration of the funding, they determine a budget for their mortgages for the upcoming period. Ideally, financial institutions want to match the duration of the fixed interest period of a mortgage with the duration of the lending of debt for that mortgage. Once a financial institution is almost out of budget for a specific fixed interest period, it may choose to increase the interest rate for mortgages with that fixed interest period. This way, borrowers will apply for mortgages with a different fixed interest period, or may choose to go to another mortgage lender.

Finally, financial institutions sometimes increase their interest rates during the summer months, and at the end of the year, as there is less personnel available to handle the requests due to vacations and holidays. With less personnel available they can handle less mortgage requests, so in order to keep the processing time the same they choose to reduce the input, by increasing the interest rates. Financial institutions may also specifically keep interest rates low for mortgages with a certain fixed interest period. For example, in order to increase its market share, a financial institution can offer a low interest rate for mortgages with a fixed interest period of 20 years, whereas the interest rates for other mortgages are in line with the market. Interest rate changes are not always directly influenced by changes in the cost of lending, but can have numerous reasons.

### 4.2.3 Changes in regulations

Another factor that impacts the amount of mortgage applications is changes in regulations (Van der Laan, 2015). Depending on the type of regulations change and the impact of the change, there is generally an increase or decrease in mortgage applications before and after the regulations change. Over the last few years the regulations have become stricter quite a few times, which has led to sudden peaks in mortgage applications. These changes in regulations often happen on the 1st of January or the 1st of July (Boumeester, 2016). An example of a recent change in regulations was the change in the Nationale Hypotheek Garantie (NHG) regulations at the 1st of July 2015 (Boon, 2015; "Uitstekend half jaar voor hypotheekaanvragen," 2016). As can be seen in the data, there was a huge peak in June 2015 due to a decrease of the maximum mortgage (i.e. the maximum rentable amount) at the 1st of July, and a decrease in mortgage applications right after the change. The interviews confirm the importance of a change in regulations.

Generally there are two types of changes in regulations directed by the government: changes in the mortgage loan regulations and changes in Nationale Hypotheek Garantie (NHG). Changes in the mortgage loan regulations include amongst others changes in the maximum mortgage (i.e. the maximum rentable amount), also called the Loan-To-Value (LTV) ratio, and changes in the mortgage interest deduction. The LTV ratio is a financial term that indicates the ratio of the mortgage loan to the value of the property. The maximum LTV ratio is set by the Dutch government, and is currently capped at 101%. This means a mortgage borrower cannot lend more than 101% of the value of the property. This ratio was decreased by 1% each year, for the past 5 years, and is expected to decrease even further in the upcoming years, as the government wants the mortgage borrowers to bring in more equity in order to reduce risks. Changes in the mortgage loan regulations generally come into force the 1st of January each year.

Changes in NHG often occur at the 1st of July. NHG is a guarantee system for mortgage borrowers that buy a house, which serves as a safety net in case a mortgage borrower is unable to pay its mortgage costs due to circumstances such as unemployment. If a mortgage borrower goes default (i.e. is unable to pay its mortgage), NHG will offer possibilities to temporarily resolve the problem. A mortgage borrower can qualify for a mortgage with NHG under several conditions, the most important one being that it is only available for mortgages where the maximum mortgage is capped at a certain amount. The main advantage of having a mortgage with NHG is that the default risk is much lower for the mortgage lender, which results in a lower mortgage interest rate for the borrower and thus saves costs.

As mentioned in the interviews, these changes in regulations generally come into force at the 1st of January or the 1st of July, and are announced at Prinsjesdag, in which the government introduces its regulations for the next year. Changes in regulations can have different effects on the amount of mortgage applications, depending on the impact of the change. If a change in regulations has a negative impact on the consumer, one can generally expect an increase in the amount of mortgage applications before the change in regulations takes place, and a decrease in mortgage applications after the change, as the consumers generally want to submit their mortgage applications before the regulation change affects them. The other way around, if a change in regulations has a positive impact on the consumers, one can expect a decrease in the amount of mortgage applications before the change, and an increase in mortgage applications after the change.

In the last few years, the changes in regulations generally had a negative impact on the consumer. The NHG maximum rentable amount has decreased multiple times, and the LTV has gone down from 106% to 101% over the past five years. It is expected to negatively affect the consumer for at least a few more years in the upcoming future.

### 4.2.4 Other predictors

Besides the mortgage interest rates and changes in regulations, several other predictors were mentioned in the literature and interviews. These will be discussed briefly below. Furthermore, besides looking at the predictors of the amount of mortgage applications, we will also look at factors that affect the housing market. Since there is a strong relationship between the housing market and the mortgage market (i.e. the amount of houses sold and the amount of mortgage applications), we can assume that the factors that influence the housing market may affect the amount of mortgage applications (Boumeester, 2016).

The first obvious pattern in the data is related to *seasonality*. As can be seen in historical data, there is a clear seasonal pattern in the amount of mortgage applications per month with a peak at the end of the year and a bottom at the beginning of the year ("Terugblik 2015 en vooruitblik 2016," 2016). During the summer months, the amount of mortgage applications is also lower due to the vacations

("Zomerdipje in hypotheekaanvragen," 2016). According to Boumeester (2016), the amount of mortgage applications are low in January and February and during the summer months, and there are peaks in May, June, November and December. These findings were also confirmed in the interviews. We also expect a drop in the amount of mortgage applications during the weekends, as most of the mortgage advisors are not working during these days. Furthermore, holidays (e.g. Easter, Pentecost, and Christmas) are expected to have a negative impact on the amount of mortgage applications. Finally, even though there seems to be no autocorrelation between the amount of mortgage applications over time, we still want to include the historical amount of mortgage applications in our model (Brockwell & Davis, 2016). Even though the amount of mortgage applications are not dependent on the amount of mortgage applications for the previous day, week or month, there is still a time component present and there may be a correlation between these factors.

Another aspect that influences the amount of mortgages is the *state of the housing market* (Basten & Koch, 2015; Boumeester & Lamain, 2016; Deira, 2015; Pettinger, 2009, 2013; Van der Laan, 2015). In the housing market there are several factors that are of importance. First of all, the average house prices and expected changes in house prices have an effect on the amount of mortgage applications. If house prices are expected to increase, people might think it is a good moment to buy a house, and thus more mortgage applications may be coming in at the financial institutions. As mentioned in Chapter 4.1.1, this relationship between the two variables also works the other way around. House prices do not only influence the amount of mortgage applications, but mortgage applications in return also influence house prices. Second, rental prices are important. If the cost of renting is high, buying becomes more attractive compared to renting. Third, the amount of houses available has an effect on the amount of mortgages sold. If the supply of houses is large, the average house prices are going to drop and it will become more attractive to buy a house. All of these factors influence the amount of houses sold and thus the amount of mortgage applications. In the interviews each of these factors was identified as a possible predictor.

There are also a number of factors mentioned that are related to the *state of the economy* and the financial position of the consumers. In case of economic growth, or a growth in income, the amount of mortgage applications may increase as people are more likely to buy a new house (Boumeester & Lamain, 2016; Pettinger, 2013). The affordability of a house (i.e. the ratio of house prices to income) also plays an important factor here. If the incomes increase, but the house prices lag behind, the demand in houses should rise and thus the amount of houses and mortgages sold (Pettinger, 2009). Also an increase in consumer confidence has a positive effect on the housing market, and may thus increase the amount of mortgage applications (Deira, 2015). Finally, economic factors such as the propensity to buy and unemployment rates may also influence the financial position of a potential house buyer. In the interviews the importance of the economic situation was confirmed, albeit they only influence the amount of mortgage applications on the long term.

In the short term the economic features may also impact the amount of mortgage applications, but this is generally caused by the media. If there are a relatively large number of news stories about economic growth or an economic crisis within a short period of time, then this will have a certain impact on the consumers. This effect was mentioned multiple times during the interviews. Even though the financial position of the consumer does not directly change, it indirectly has a psychological effect on the consumer, which could have an effect on the number of houses sold and thus mortgage applications. The same effect also applies to interest rates. If a number of news stories appear about an expected increase in interest rates for the next month, it will indirectly have an effect on the consumer. If at the same time the European Central Bank (ECB) announces an increase in interest rate, one can expect to see impact on the amount of mortgage applications. We will refer to this type of predictor as the

*influence of the media*. A technique to collect information from news websites and other kinds of publicly available sources is called Open Source Intelligence (OSINT) (Stalder & Hirsh, 2002).

## 4.3 PREDICTIVE ANALYTICS

Predictive analytics is a field in data mining that encompasses different statistical and machine learning techniques that are aimed at making empirical predictions (Finlay, 2014). These predictions are based on empirical data, rather than predictions that are based on theory only (Shmueli & Koppius, 2010). In predictive analytics, several statistical and machine learning techniques can be used to create predictive models. These models are used to exploit patterns in historical data, and make use of these patterns in order to predict future events. These models can be validated using different methods to determine the quality of such a model, in order to see which model performs best. The quality of such a model is also called predictive power (Shmueli & Koppius, 2010).

Predictive analytics can be applied in many different fields. Some of these include marketing, financial services and retail. According to a survey regarding the applications of predictive analytics (Eckerson, 2007), some of the most used applications of predictive analytics are cross-selling/upselling, campaign management and customer acquisition. However, in practice it can be applied to almost any field (Gandomi & Haider, 2015).

There are generally two types of problems predictive analytics is used for: classification problems and regression problems. The main difference between these two problems is the dependent variable, the target variable that is being predicted. In classification problems, the dependent variable is categorical (e.g. credit status). In regression problems, the dependent variable is continuous (e.g. pricing) (Gandomi & Haider, 2015).

The techniques that are used in predictive analytics to create a model depend heavily on the type of problem. For classification problems, classification techniques are used such as Naïve Bayes and decision trees. These techniques often consist out of one or multiple algorithms that can be used to construct a model. For decision trees, some of the algorithms are Classification and Regression Trees (CART), C4.5 and Conditional Inference Trees. Especially C4.5 and CART are some of the most influential data mining algorithms (Chen, Chiang, & Storey, 2012), because they are generally easy to use and simple to understand and interpret.

For regression problems, regression techniques such as multiple linear regression, support vector machines or time series are used. These techniques focus on providing a mathematical equation in order to represent the interdependencies between the independent variables and the dependent variable, and use these to make predictions (Gandomi & Haider, 2015). One of the most popular regression techniques is linear regression. Linear regression has been studied for decades, and is used extensively in practice (Finlay, 2014; Yan, 2009). When applied correctly, regression is a powerful technique to show the relationships between the independent and the dependent variables. However, linear regression requires some assumptions in the dataset (Armstrong, 2011). One of these assumptions is that there has to be a linear interdependency between the independent variables and the dependent variable. A pitfall of linear regression is that the regression line contains no information about the distribution of the data. It needs to be combined with a visualization of the regression line in order to draw conclusions. In Figure 6, it can be seen that four different datasets that have the same means, variances, correlation and linear fit, still have a completely different distribution, even though their regression lines are the same (Anscombe, 1973; Owen, 2015). Hence, a regression line always needs to be combined with a visualization in order to draw conclusions about the distribution of the data.

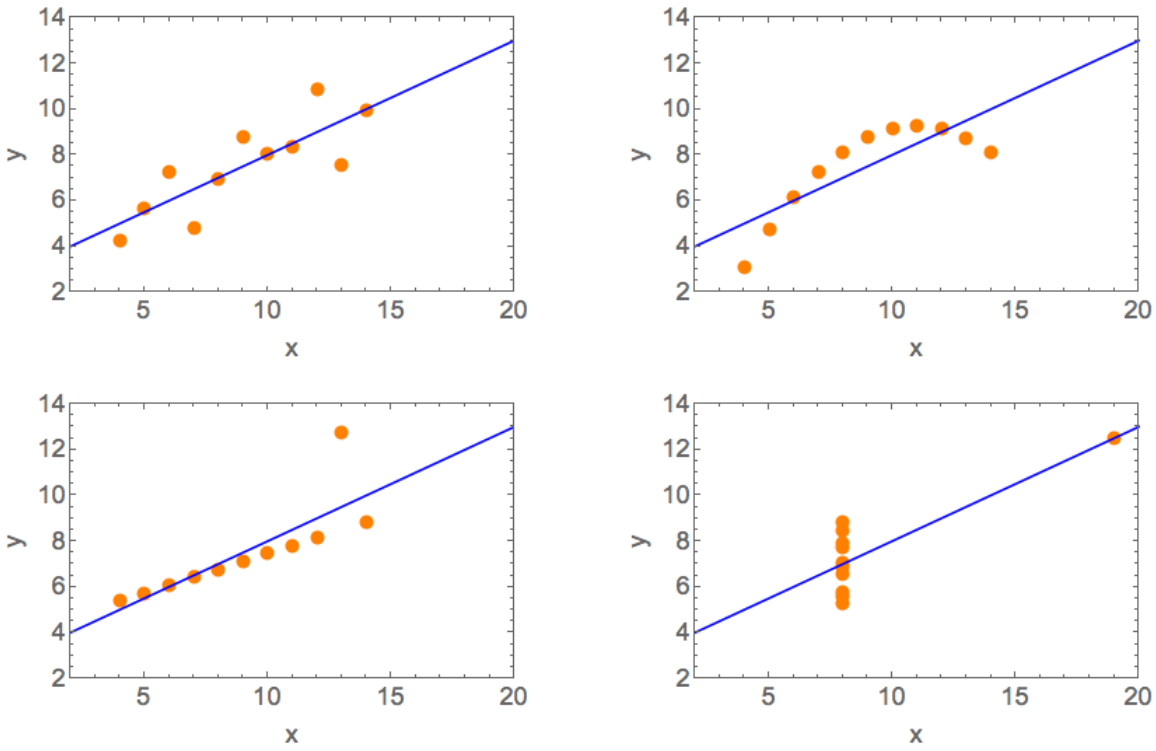



*Figure 6 – Anscombe's quartet. Retrieved from (Owen, 2015).*

In order to compensate for the disadvantages of the individual models, ensemble models can be used. An ensemble model is a set of individually trained models, which predictors are combined to increase the predictive performance. Ensemble models are generally more accurate than any of the individual models that make up the ensemble model **(Opitz & Maclin, 1999)**. Examples of techniques used for creating an ensemble model are bagging and boosting. With bagging, multiple versions of a predictor are used to create an aggregated predictor, in order to increase the accuracy of the model **(Breiman, 1996)**. An example of a bagging algorithm is random forests, which combines a set of decision trees to increase the model performance **(Ho, 1995)**.

A combination of the machine learning techniques mentioned above can be used to create predictive models. These models can then be validated and compared based on predictive power, which can be calculated using a set of statistical measures, such as Mean Absolute Error (MAE), Root Mean Square Error (RMSE) and $R^2$ **(Willmott & Matsuura, 2005)**.

# 5 DATA UNDERSTANDING

*This chapter discusses the Data Understanding stage of the CRISP-DM framework. The Data Understanding stage has been split up in two parts: data collection and data exploration. In the data collection subchapter we will discuss characteristics of the event log data, and how the data has been extracted from the database. In the data exploration subchapter, the data and its characteristics are explored to extract useful information for our models.*

## 5.1 DATA COLLECTION

For our research we are working with a dataset from one of the clients of the FORCE product. Since the structure of the event log data is the same for each of the clients of the product FORCE, our solution can easily be generalized to other clients of the system. We will be using an anonymized version of the FORCE database of that specific financial institution to extract the data from. This is an SQL database that contains a number of tables that are used within FORCE. Since we are only interested in the event log data we will only be using one of the tables. This table contains data about the status changes in FORCE, for every mortgage application. Every action performed by the system or by a user on a mortgage application is logged, and the status before and after that specific action is logged. The table contains amongst others information about these statuses, the object of which the status was changed and the date and time of the status change. An example of the table structure and the data can be found in Table 2. For our analysis we are mainly interested in the date and time at which each of the mortgage applications have entered the system.

In this table, the Id is the unique identifier of the row. The columns called fromStatus and toStatus indicate the status before and after the status change. In case the fromStatus column is NULL, there is no previous status, which means the object has just entered the system. The statusChangedBy column indicates the id of the employee that performed the status. If the statusChangedBy column is NULL then this action was performed automatically by the system. The statusChangedDate column contains the date and time of the status change. The streamingObject column contains the id of the object. This object is not always a mortgage application, it can also be for example a contract, so a further selection on the type of object has to be made.

*Table 2 – Example overview of the database table*

| Id | fromStatus | toStatus | statusChangedBy | statusChangedDate | streaming Object |
|---|---|---|---|---|---|
| **54783854** | NULL | 1 | NULL | 2016-10-19 11:25:29.000 | 279024 |
| **54783855** | 1 | 4 | 49 | 2016-10-19 11:26:45.000 | 279024 |
| **54783856** | 92 | 103 | 115 | 2016-10-19 11:26:58.000 | 278978 |
| **54783857** | NULL | 1 | NULL | 2016-10-19 11:27:06.000 | 279025 |
| **54783858** | 4 | 5 | 49 | 2016-10-19 11:27:33.000 | 279024 |

Using an SQL query we select all of the mortgage applications with their application date and time. Since we are only interested in the amount of mortgage applications we exclude the objects from different processes in the system, and exclude the status changes that were already in the system. This query was validated with domain experts at Topicus to ensure its correctness. With this query we can output a list containing the timestamps of the mortgage applications and load it in R to perform further analysis.

## 5.2 DATA EXPLORATION

Since the objective is to make a prediction per day, our dataset can be grouped per day to create meaningful visualizations. The dataset contains data from January 2013 until May 2017. In order to get a feel of the amount of mortgage applications per day and the distribution of the mortgage applications, different visualizations can be made using R. Two graphs have been created, which can be found in Figure 7 and Figure 8. Both of these graphs only contain the amount of mortgage applications on the weekdays. As there are almost no applications coming in on the weekends they have been excluded from the graphs.

--- FIGURE REDACTED DUE TO CONFIDENTIALITY ---

*Figure 7 – Daily amount of mortgage applications over time*

--- FIGURE REDACTED DUE TO CONFIDENTIALITY ---

*Figure 8 – Density plot of the amount of mortgage applications per day*

As can be seen from the graphs, there seems to be a seasonal pattern on a monthly level, but from these graphs it is not very clear. It also seems like there are some outliers, so these data points will have to be investigated to see if they will have to be included in our model, as there can be multiple underlying reasons for outliers in our dataset. It also seems there is an increase in mortgage applications during the last few months of each year, which confirms our findings in Chapter 4.2.4. The amount of applications per day during these months is almost doubled compared to the other months. This can have multiple explanations so this will have to be accounted for in the model.

The density plot shows the distribution of the amount of mortgage applications. It seems the distribution of the amount of mortgage applications is normally distributed, slightly skewed to the right with a long tail. This is due to the outliers mentioned before. The median seems to be at around --- NUMBER REDACTED --- mortgage applications per day.

Next to these plots, it is also interesting to look at the amount of daily mortgage applications per month. An overview of this can be found in Figure 9. In this plot we can confirm some of the findings from the literature, as discussed in Chapter 4.2. The amount of applications is slightly lower during the beginning of the year, and lower during the summer months. At the end of the year it rises again.

--- FIGURE REDACTED DUE TO CONFIDENTIALITY ---

*Figure 9 – Total amount of mortgage applications per month*

# 6 DATA PREPARATION

*The Data Preparation stage of the CRISP-DM framework contains three elements: data pre-processing, feature engineering and feature selection. In the data pre-processing subchapter we pre-process the data so that it contains the right format for our models. In the feature engineering subchapter we create a list of features, using the predictors mentioned in Chapter 4.2.1. In the feature selection subchapter we select a subset of features that are useful for our model.*

## 6.1 DATA PRE-PROCESSING

In the data pre-processing phase we pre-process our data to a suitable format for our predictive model. This phase consists out of the following steps. First, the data is grouped by day. Since we want to get insight in the amount of personnel needed on a daily level, we want to group the amount of mortgage applications per working day. In our definition of a working day at the financial institution, a working day ends at 17:00 so any application that enters the system after 17:00 will be processed the next day. The time component of a mortgage application is ignored afterwards.

Second, as there are minimal applications coming in during the weekends, we choose to transfer these applications to Monday. Third, mortgage applications that entered the system before the 1st of January 2013 have been removed from our dataset. From the dataset we can see that the log data starts at the --- DATE REDACTED ---. However, due to instability in the system and different changes in FORCE that affected the amount of mortgage applications coming in, we choose to exclude the data from before the 1st of January 2013. As this data would not be useful for our model, it is excluded. Finally, missing dates (i.e. dates without any mortgage applications) have been added to the model, as these need to be predicted as well.

## 6.2 FEATURE ENGINEERING

Feature engineering is the process of encoding the predictors in a way that they can be useful for prediction **(Kuhn & Johnson, 2013)**. Often this is done by applying domain knowledge, for example using a combination of two predictors (e.g. calculating a ratio) can sometimes be more effective than using the individual predictors. In our research, we generate features for our model that may have predictive power based on our domain knowledge of the mortgage application domain. As mentioned in Chapter 4.2, there are a number of predictors that influence the amount of mortgage applications. In order to use these predictors in our model, we can encode them into features that can directly be fed into our model. For each of the predictors mentioned, one or multiple features are created. For example, for the predictors related to seasonality, we generate features related to the time of the year, vacations and holidays. We can incorporate external holiday data, to see if the holidays have an effect on the amount of mortgage applications. For the other predictors, we will need to include external data in our model. There are a number of institutions that offer publicly available data related to the economy and housing market (e.g. Kadaster, CBS, Rijksoverheid). This data can often be extracted using a web API, and incorporated into our model. There is also data already available in the financial institution's database. For example the financial institution's interest rates can easily be extracted from the database.

A single predictor sometimes translates into multiple features. For example for the mortgage interest rates, we can include the financial institution's own interest rates but also the Euribor rate, which is the rate used between financial institutions in Europe. An increase in Euribor rates often leads to an increase in the financial institution's rate as well, as mentioned in Chapter 4.2.2. Another factor that we have to take into account is the announcements of interest rate changes. As soon as an increase in interest rate

is announced, we may already expect an increase in mortgage applications. This increase in mortgage applications happens before the actual interest rate change itself.

In Table 3, an overview of the features can be found, as well as their corresponding predictors as mentioned in Chapter 4.2.1 and listed in Table 1. As can be seen in the table, we created a total of 27 features using different datasets. For four predictors (relative interest rates, rental prices, income growth and affordability), there was no suitable external data available. These have therefore been excluded from our research project.

*Table 3 – Overview of features used in our models*

| Predictor | Id | Feature |
|---|---|---|
| Date | F1 | Year |
| | F2 | Month |
| | F3 | Day |
| | F4 | Week |
| | F5 | Weekday |
| | F6 | Yearday |
| Vacations | F7 | Vacations |
| Holidays | F8 | Holidays |
| Historical amount of mortgage applications | F9 | Previous day |
| | F10 | Same day last week |
| | F11 | Average of last week |
| | F12 | Average of last month |
| Mortgage interest rates | F13 | Financial institution's mortgage interest rate |
| | F14 | Euribor |
| | F15 | Capital market interest rate |
| Changes in mortgage interest rates | F16 | Changes in financial institution's mortgage interest rate |
| Relative interest rates | - | No data available |
| Changes in regulations | F17 | Changes in regulations January |
| | F18 | Changes in regulations July |
| House prices | F19 | House prices |
| Rental prices | - | No data available |
| Amount of houses available | F20 | Amount of houses available |
| | F21 | Amount of houses sold |
| Economic growth | F22 | Economic growth |
| Income growth | - | No data available |
| Affordability | - | No data available |
| Consumer confidence | F23 | Consumer confidence |
| Propensity to buy | F24 | Propensity to buy |
| Unemployment rates | F25 | Unemployment rates |
| Influence of the media | F26 | Google Trends "Hypotheekrente financiele instelling" |
| | F27 | Google Trends "Hypotheekrente" |

For the predictors related to seasonality, multiple features were generated. A total of six features were derived from the timestamp of the mortgage applications (F1 – F6). Using external data, vacations and

holidays (F7, F8) were extracted (Kalender 365, 2017; Schoolvakanties Nederland, 2017). Also aggregations are made of the historical amount of mortgage applications (F9 – F12).

For the mortgage interest rates, the financial institution's mortgage interest rates were derived from their database (F13). Also two indicator of the financial institution's mortgage interest rate were used in our model: the Euribor rate (F14) and the capital market interest rate (F15) using external data from De Nederlandsche Bank (DNB) (De Nederlandsche Bank, 2017). Also changes in the financial institution's mortgage interest rate were used as a feature (F16), explicitly making a separation between a decrease in mortgage interest rate and an increase in mortgage interest rate. Both of these have a different impact on the amount of mortgage applications, as explained in Chapter 4.2.2. These changes in mortgage interest rate almost impact the days before and after the actual change. Data regarding relative interest rates could unfortunately not be found. Current interest rates of all financial institutions in the Dutch mortgage market are available, but no historical data could be found.

For the changes in regulations two features have been created. A separation has been made between changes in the mortgage loan regulations (F17), which generally come into force at the first of January, and changes in NHG regulations (F18), which generally come into force at the first of July. External data was obtained from NHG and Rijksoverheid (Ministerie van Algemene Zaken, 2016; NHG, 2016).

Three features have been derived to incorporate the changes on the housing market into our model. A feature was derived using housing prices on the Dutch housing market (F19), and two other features indicate the amount of houses available and the amount of houses sold given a specific timestamp (F20, F21). External data was extracted using Kadaster and Centraal Bureau voor de Statistiek (CBS) (CBS StatLine, 2017b; Kadaster, 2017). For the rental prices, no external data source was found, so this predictor was excluded from our research project.

Regarding the state of the economy, four features were created using data from CBS: economic growth, consumer confidence, propensity to buy and unemployment rates (F22 – F25) (CBS StatLine, 2017a). For the income growth and affordability, the data was aggregated on a yearly level, and hence not useful for our model as our data is aggregated on a daily level. The variance between the observations would be too low to contain enough predictive power for our models.

Finally, the influence of the media was added by analyzing Google Trends data (Google Trends, 2017). Two features were derived using the search terms "Hypotheekrente" and "Hypotheekrente financiele instelling" (F26, F27), in which the term "financiele instelling" is replaced by the name of the actual financial institution. There are several other sources of external data that could have been used to collect OSINT, but these have been excluded from our analysis due to time restrictions.

## 6.3 Feature selection

All of the features created in the previous step can be used to train our models. However, not all of the features may be relevant in terms of predictive power. The process of selecting an optimal subset of the features is called feature selection. Feature selection helps understanding the data, reducing computational power and it may increase predictive power (Chandrashekar & Sahin, 2014). Some of the predictive models have built-in feature selection, for example decision tree-based models. Others, such as Linear Regression and Neural Networks, require manual feature selection. Figure 19.1 of the book Applied Predictive Modeling confirms this; the performance of Neural Networks and Support Vector Regression clearly decrease after adding multiple non-predictive features, whereas the performance of tree-based models stay about the same (Kuhn & Johnson, 2013, p. 489).

There are multiple solutions for the feature selection. In general, one can distinguish between filter methods and wrapper methods **(John, Kohavi, & Pfleger, 1994)**. Filter methods evaluate the relevance of the predictors according to one or multiple criteria, after which a selection of predictors can be made that pass the criteria. An example of such a criterion is the correlation coefficient. Predictors that pass a certain level of correlation with the target variable could be included in the model, and the remaining predictors could be excluded. Wrapper methods evaluate multiple models on subsets of features, in order to find the optimal subset that maximizes the performance of the models. An example of a wrapper method is Recursive Feature Elimination, in which features with low weights are removed recursively until an optimal subset of features is found.

For our models, variable selection is done using Recursive Feature Elimination (RFE), using the Caret package in R **(Fernandez-Lozano et al., 2015; Kuhn, 2012)**. An optimal subset of features is selected from the feature list mentioned in Table 3. A total of 12 features were selected from the total of 27 features. The other 15 features are rejected from the final model. An overview of the selected features can be found in Table 4.

*Table 4 – Overview of features selected for final model (included variables are denoted by 'X', excluded variables are denoted by '-').*

| Id | Feature | Included |
|---|---|---|
| F1 | Year | X |
| F2 | Month | X |
| F3 | Day | X |
| F4 | Week | - |
| F5 | Weekday | - |
| F6 | Yearday | X |
| F7 | Vacations | X |
| F8 | Holidays | X |
| F9 | Previous day | X |
| F10 | Same day last week | - |
| F11 | Average of last week | - |
| F12 | Average of last month | - |
| F13 | Financial institution's mortgage interest rate | X |
| F14 | Euribor | X |
| F15 | Capital market interest rate | - |
| F16 | Changes in financial institution's mortgage interest rate | X |
| F17 | Changes in regulations January | X |
| F18 | Changes in regulations July | X |
| F19 | House prices | - |
| F20 | Amount of houses available | - |
| F21 | Amount of houses sold | - |
| F22 | Economic growth | - |
| F23 | Consumer confidence | - |
| F24 | Propensity to buy | - |
| F25 | Unemployment rates | - |
| F26 | Google Trends "Hypotheekrente financiele instelling" | - |
| F27 | Google Trends "Hypotheekrente" | - |

As can be seen from Table 4, all of the features regarding the state of the economy and the state of the housing market were rejected by RFE. The Google Trends features were also rejected. The features that

were included in our models are related to seasonality, interest rates and changes in regulations. Reasons for excluding a feature can be a lack of predictive power, or strong multi-collinearity with other features. For example, the financial institution's mortgage interest rate, the Euribor and the capital market interest rate are all strongly correlated with each other. They may not all be of significance to the predictive power of our model, as the predictor's effect on the amount of mortgage applications may have already been captured by adding the other features. In this case, RFE has decided to drop the capital market interest rate. The predictors related to the state of the economy and the state of the housing market are measured on a monthly basis, so may not add enough predictive power on a daily level. Furthermore, they generally have influence on the amount of mortgage applications on the long term, rather than the short term. These kind of features do not capture differences in the amount of mortgage applications on a daily level, so are considered irrelevant for our models.

Besides the feature selection, some models require additional pre-processing. For example, Support Vector Regression and Neural Networks require the data to be scaled, normalized and centered (Kuhn & Johnson, 2013, p. 550).

# 7 MODELING

*In the Modeling stage of the CRISP-DM framework we discuss the activities related to the model building part of our research. A selection of five modeling techniques is made that are applicable to our research project. From each of these five modeling techniques, a model is built with the featureset provided in the previous chapter, and the models are validated using repeated cross-validation.*

## 7.1 SELECTION OF MODELING TECHNIQUES

For our modeling we use a combination of predictive techniques. Multiple techniques are selected and applied on the data. For our research, we are only interested in regression techniques that can identify non-linear relationships, as discussed in Chapter 4.3. This way, classification techniques and linear regression techniques such as Multiple Linear Regression (MLR) are not suitable for our problem. Since there is only a limited amount of literature available that discuss a few basic modeling techniques, we do not know yet if there are already any more advanced techniques implemented successfully. We will therefore manually select a set of techniques that are applicable to our case and are expected to give promising results.

For the non-linear regression techniques, we use *Support Vector Regression (SVR)* and *Neural Networks (NN)*. SVR has shown to obtain excellent performances in regression and time series applications (Basak, Pal, & Patranabis, 2007). Neural Networks are a widely used method for time series data that generally gives mixed results (Zhang & Qi, 2005).

Another technique we use is *Classification and Regression Trees (CART),* which is a simple technique that is easy to visualize (Kuhn & Johnson, 2013). Also two ensemble techniques are included, in order to improve the performance of the CART (Varian, 2014). These ensemble techniques are *Gradient Boosting Machines (GBM)* and *Random Forests (RF)*. These techniques create a multitude of regression trees and select a combination of them in order to maximize the performance.

## 7.2 MODEL BUILDING

Using these five techniques (SVR, CART, RF, GBM and NN) we can create five models. We use the list of features mentioned in Chapter 6.3 as input for our models. A total of 12 features are included, the remaining features were excluded after performing feature selection. For each of the five models hyperparameters were tuned, using grid search (Hsu, Chang, & Lin, 2003). Hyperparameters are the model-specific parameters that are used for optimizing the model. They generally have to be tuned in order to optimize the model's performance, and reduce the variance and bias of the model. For example, the Neural Network model requires two hyperparameters: the size of the Neural Network (i.e. the number of units in the hidden layer) and the decay (i.e. the regularization parameter to avoid overfitting). Grid search is used to determine possible values for the hyperparameters. By training the model with different values of the size and the decay, and evaluating its performance, we can select the hyperparameters that result in the best performing model in terms of predictive power.

## 7.3 MODEL VALIDATION

Since our models are trained on historical data, we are not sure how these models will perform during the forecasting. In general, the explanatory power of the model on historical data is almost always higher than then predictive power of the model on new data (Shmueli & Koppius, 2010). If this difference is too big, we talk about overfitting. This happens when the model is too complex and starts

capturing noise as well. In order to prevent overfitting and accurately estimate the predictive power of our model, we use *repeated 10-fold cross-validation*.

10-fold cross-validation is one of the most widely used methods for estimating the prediction error (Friedman et al., 2001). Using 10-fold cross-validation, we are able to split the dataset in *out-of-sample data* and *in-sample data*. The dataset is randomly split in 10 equally sized subsets, and the model is run 10 times where each of the 10 subsets is used as the validation set once. The other 9 sets are used as training sets. By changing the validation set at every run, every data point has been used for validation exactly once. After 10 observations, the results of the observations are averaged. This entire process is repeated five times, in order to account for outliers affecting the cross-validation outcome (Kim, 2009). Afterwards, the performance of the five repeats is averaged to calculate the final performance of the models. Using this method, we can get a feeling of how the model will perform on 'unknown' data. This is still not a guarantee that the model will perform the same on live data, but it will give a decent approximation.

# 8 EVALUATION

*In the Evaluation stage, the models are evaluated using several evaluation criteria, and the results of the models are discussed. A final selection of the best model is made, using an independent t-test to check if the best model performs significantly better than the other models. Also the most important features for our models are selected.*

## 8.1 MODEL EVALUATION

In order to measure the model's performance we use four different metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), MAE/Mean and $R^2$. These are some of the most common metrics used for the evaluation of regression models (Kuhn & Johnson, 2013). Each of these metrics use the residuals (i.e. the differences between the observed values and the predicted values) in order to measure the performance of the model.

RMSE uses a squared value of the absolute error, hence giving more emphasis on higher residuals. This metric is particularly useful because of this last characteristic. One of the key elements of the model is that it should be able to predict 'outliers', i.e. the non-standard days, with an unusually high amount of mortgage applications. RMSE takes this into account by giving more emphasis on the higher residuals. The formula of RMSE can be found below (Kuhn & Johnson, 2013).

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}$$

MAE and MAE/Mean use the mean absolute differences between the predictions and the actual values for the model evaluation. MAE denotes the absolute difference between the predictions and the actual values, whereas MAE/Mean denotes the relative differences of the residuals to the mean of the actual values (Hoover, 2009; Kolassa & Schütz, 2007). These evaluation criteria are primarily used for explaining the model's performance to the customer. Formulas for MAE and MAE/Mean can be found below (Rieg, 2010). In these formulas, $n$ denotes the amount of predictions, $x_i$ denotes the actual values at period $i$, and $\hat{x}_i$ denotes the predicted values at period $i$.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i|$$

$$MAE/Mean = \frac{\frac{1}{n}\sum_{i=1}^{n}|x_i - \hat{x}_i|}{\frac{1}{n}\sum_{i=1}^{n}x_i}$$

The $R^2$ explains the proportion of the variance in the data that is explained by the model. The value of $R^2$ is always between 0 and 1. When the value is close to 1, the model has a great fit on the data. Values close to 0 denote a bad fit. The formula of $R^2$ can be found below (Kuhn & Johnson, 2013). In this formula, $\bar{x}$ denotes the mean of the actual values.

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_i - \hat{x}_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

For evaluating the model's performance we will primarily be using RMSE, in which a lower RMSE implies a better model.

Besides the four metrics mentioned above, the interpretability of the model is another factor that will be taken into account. By interpretability we refer to the degree by which the model's predictions can be explained to the clients **(Lipton, 2016)**. For example, a simple decision tree or a linear model is more interpretable than a deep neural network. Since the interpretability of the model can only be judged subjectively, it will not be used as the primary evaluation criterion **(Browne, Cudeck, Bollen, & Long, 1993)**. However, in case the predictive power of multiple models are statistically insignificant, the interpretability of a model can be used to decide the outcome.

In Table 5, an overview of the five different models and their results is provided. The results indicate the average repeated cross-validation performance of the different models, for the four metrics mentioned above. Predictions are made for the next day, based on the final list of predictors mentioned in Chapter 6.3.

*Table 5 – Model performance of the five different models*

| Model | RMSE | MAE | MAE/Mean | $R^2$ |
|---|---|---|---|---|
| Random Forests (RF) | --- RESULTS REDACTED DUE TO CONFIDENTIALITY --- | | | |
| Gradient Boosting Machines (GBM) | | | | |
| Support Vector Regression (SVR) | | | | |
| Neural Networks (NN) | | | | |
| Classification and Regression Tree (CART) | | | | |

## 8.2 Discussion of results

Using these metrics to evaluate the performance of the models, we can already see differences between the models. In Table 5 we can see that the Random Forest model performed best, not only in terms of RMSE, but also in MAE, MAE/Mean and $R^2$. The Gradient Boosting Machines model seems to perform only slightly worse. We can also see that the CART model performs worst, which was expected. This is a basic decision tree, whereas RF and GBM are ensemble methods of multiple decision trees. In order to test whether the difference in RMSE between RF and GBM are statistically significant, an independent t-test can be conducted ($p < 0.05$) **(Rice, 2006)**. For the t-test, the out-of-fold performances (i.e. the predictions on the 10 different folds that were used as validation set in our cross-validation) of the five repeats of the RF and GBM models are compared using the characteristics described in Table 6. Using a 95% significance level, there appears to be a significant difference between the average RMSE of the RF and GBM models. This way we have enough statistical evidence to conclude that the Random Forest performs better than the Gradient Boosting Machines.

*Table 6 – Characteristics of the RF and GBM models*

| Model | Mean RMSE | St. Dev. RMSE |
|---|---|---|
| Random Forests (RF) | --- RESULTS REDACTED DUE TO CONFIDENTIALITY --- | |
| Gradient Boosting Machines (GBM) | | |

In Figure 10 we can see a graph of the actual amount of mortgage applications versus the predicted amount of mortgage applications for the RF model. The diagonal black line is the ideal situation, where the actual values and the predicted values are the same. In general, the closer to the black line the better, but getting too close means the model is possibly overfitted as there is always an amount of random variation involved. At first glance, the model seems to 'over-predict' the days with a low amount of mortgage applications, and 'under-predict' the days with a high amount of mortgage applications. The outliers in particular seem to be under-predicted by quite a lot, as they are quite far from the diagonal black line.

--- FIGURE REDACTED DUE TO CONFIDENTIALITY ---

*Figure 10 – Random Forest: Actual vs. Predicted*

In order to analyze these predictions in more detail, the differences between the predicted values and the actual values, also called the residuals, are plotted in Figure 11. We ideally want the residuals to be symmetrically distributed along the horizontal black line, but as we can see from the figure there are some patterns visible. Our model seems to do quite well at predicting the 'average' days, i.e. the days in which the amount of mortgage applications seems to lie between --- NUMBER REDACTED --- and --- NUMBER REDACTED ---, which is by far the main part of our dataset. For the lower end of the graph, i.e. the days in which the amount of mortgage applications seems to be between --- NUMBER REDACTED --- and --- NUMBER REDACTED ---, the residuals seem to be negative, which means the predicted values are higher than the actual values, and our model is over-predicting.

--- FIGURE REDACTED DUE TO CONFIDENTIALITY ---

*Figure 11 – Random Forest: Residual plot*

In the higher end of the residual graph, the residuals seem to be positive. The predicted values are structurally lower than the actual values, so the model is under-predicting. In particular the predictions for the outliers are significantly lower than the actual values. This explains the big difference between RMSE and MAE. Since these days have the highest number of mortgage applications, they will also require the highest amount of personnel and thus have the biggest impact on personnel planning. Our model does not handle this well yet.

In general, it seems that our model's predictions are skewed towards the mean. This is a common effect observed at the predictions of random forests, due to the fact that there is a certain amount of noise in the data. Our random forest is unable to capture the entire variance of the target variable. In our case the $R^2$ is about --- NUMBER REDACTED ---, which means the remaining --- NUMBER REDACTED --- of the variance is not captured by the model, and hence will be correlated with the mean of the amount of mortgage applications. This effect is particularly visible at the lower and higher end of the predictions, i.e. the outliers. Almost all of the outliers visible in the graph are explained by interest changes. In Figure 12, an overview of the predictions of our best model can be found. Weekends are not predicted, hence their predictions are set to zero.

--- FIGURE REDACTED DUE TO CONFIDENTIALITY ---

*Figure 12 – Predictions of the amount of mortgage applications per day: June & July*

As can be seen from the graph, most of the days in June were predicted quite accurately. In July, there seem to be two values in particular that are heavily under-predicted. By looking at the data, we can see that these values are most likely caused by interest rate changes. For future research, it would be interesting to see if there are any underlying reasons that can explain the differences between the peaks. Perhaps there are other factors involved with the interest rate changes, which have impact on the amount of mortgage applications, that are not captured by our model.

Besides using performance measures such as MAE and RMSE, the interpretability of the models is also an important factor when implementing the model at the financial institutions. Not only the prediction itself is important, also the most important variables in the model and the reason why the model has come to the predictions. For example, when the domain experts at the financial institutions predict that there will be --- NUMBER REDACTED --- new mortgage applications tomorrow, whereas the model predicts --- NUMBER REDACTED --- new mortgage applications, there is a gap between these

predictions. In these cases, it is important that the model can 'explain' why it has come to that specific prediction. Although tree ensembles such as RF and GBM are renowned for their predictive power, their interpretability is generally considered limited. There are however various methods mentioned in literature to increase the interpretability of ensemble trees (Hara & Hayashi, 2016; Ribeiro, Singh, & Guestrin, 2016). Another way to increase the interpretability is to run a simple decision tree alongside the RF, and use the decision tree for the interpretation.

In terms of feature importance, the five most predictive features are the following:

- The financial institution's interest rate
- Changes in the financial institution's interest rates
- The amount of mortgage applications on the previous day
- Holidays
- The day of the year

These are all related to seasonality and interest rates, which shows the importance of these two categories on the prediction of the amount of mortgage applications. Changes in regulations, the state of the economy, the state of the housing market and the influence of the media are all less important, or not even included in our final model.

By using the model to predict the amount of mortgage applications a week upfront, the accuracy of our model will most likely decline. As the model uses the amount of mortgage applications on the previous day as a feature, for predicting the amount of mortgage applications on the next day, the accuracy of the predictions multiple days upfront will decrease, as this value is simply not known for these days. There are multiple ways of handling this. In our current model, the historical amount of mortgage applications for the days upfront are filled with the predictions of the day before. The model is run for one day upfront, and then this value is used to predict the value of the upcoming day, and so on. This means that for making a prediction for the next week, the predict-function of our model will have to be ran 5 times. Another option would be to replace the value of the historical amount of mortgage applications with the last known value, or with an average of the previous week. Either way, the model's predictions will most likely lose accuracy, as predictions are made further upfront. A recommendation would be to update the predictions every night, to keep the predictions as accurate as possible.

Using the results of the models and the t-test, we saw that the Random Forest model performs the best in terms of RMSE, with an average RMSE of --- NUMBER REDACTED ---. Since RMSE is difficult to interpret in a business context, we will be using MAE to explain the model's performance to the customers. The RF model has an MAE of --- NUMBER REDACTED ---, which means that the model's predictions of the amount of mortgage applications differ on average --- NUMBER REDACTED --- of the actual amount of mortgage applications for that day, excluding the weekends. The difference between these predicted values and the actual observations can be either negative or positive.

This difference can directly be related to the personnel planning, by dividing it by the number of mortgage applications handled per person per day at the relevant departments at the financial institutions. As we can see from the database, there are approximately --- NUMBER REDACTED --- employees working at the mortgage application department at the financial institution, per day, over the last months. By dividing the amount of new mortgage applications per day by the amount of personnel working at the financial institution each day, we can see that every employee at the mortgage application department handles around --- NUMBER REDACTED --- applications per day. Even though these numbers are not accurate, as there are a lot of other factors that influence this number, it can still give us an indication of the error of our model in terms of personnel allocation. However, in reality this

number is most likely larger, as not all of the employees may work fulltime, not all of the employees at the mortgage application department may work as a mortgage acceptor, the 'type' of mortgage application is not taken into account, and the mortgage acceptors work in teams, each on their own specialization. If we make the assumption that the amount of mortgage applications handled per person per day lies around --- NUMBER REDACTED ---, we can say that the MAE of our model in terms of personnel is around --- NUMBER REDACTED --- employees per day. However, once again, multiple assumptions have been made during this calculation. Hence, we would recommend to further analyze this in more detail, to make a more accurate calculation.

# 9 DEPLOYMENT

*The Deployment stage of the CRISP-DM model contains two activities: visualization and recommendations for deployment. In the visualization subchapter we discuss how the model can be used in practice, and proposes an option for visualizing the information in a dynamic dashboard. In the recommendations for deployment subchapter, we give recommendations on a number of important aspects that are involved with deploying the dashboard at the financial institutions.*

## 9.1 VISUALIZATION

In order to deploy the model in practice it can be incorporated in a dashboard, in which an overview of the predictions for the next week can be shown. The model can be re-trained every night, to incorporate the new mortgage applications that have entered the system on that day. One issue that arises however during the transition from an explanatory model to a predictive model is the availability of data for making the predictions. As we want to predict the amount of mortgage applications a week upfront, we will also need the data for our model a week upfront. Some of the features needed may already be known (e.g. features based on seasonality, holidays, vacations, changes in regulations) but other data may not be available a week upfront, such as mortgage interest rates.

For the mortgage interest rates, there are three aspects that are required in order to make an accurate prediction: the 'moment' of interest rate change (i.e. the day on which the mortgage interest rate changes), the percentage of increase/decrease and the impact on the days before and after the interest rate change. The third aspect is 'known', this can be predicted by using previous interest rate changes. The first and second aspect however are unknown for the upcoming week. There are multiple options for solving this issue:

1. The unknown data can be predicted using time series
2. The currently available mortgage interest rates can be used for our predictions for the next week
3. Changes in mortgage interest rates could be incorporated in our model as a feature, so that one can manually enter new mortgage interest rates at the moment of announcement.

Option 1 is not feasible for our model, due to the nature of the changes in mortgage interest rates. As described in Chapter 4.2.2, changes in mortgage interest rates are only announced one or two days upfront, and there appears to be no structural pattern in the moments of the changes in interest rates. Besides, the changes in interest rates only have impact on a small amount of days around the actual change itself. Predicting the changes in interest rates would only increase the model's error, as a wrong prediction would totally change the distribution of mortgage applications over the next week.

Option 2 would be the easiest solution, but by choosing option 2 changes in interest rates would be ignored, which would increase the model's error, as they do have significant impact on the model's predictions. Hence, option 3 would be the most viable option, a *dynamic dashboard* in which interest rate changes can be entered manually, which shows the result of a change in interest rate real-time. Even though this means that interest changes have to be entered manually in the model, it gives the best results in terms of error.

In this dynamic dashboard, interest rates are added as a *feature* of the model. The dashboard's users are able to manually enter new interest rates, which are directly fed into the model. This way the impact of a change in interest rate on the amount of mortgage applications is visible real-time. Besides the date

of the interest rate change, the percentage of change and whether the interest rate increases or decreases are also configurable.

In order to give a short demo of what the dynamic dashboard could look like, a simple prototype was built using R and its libraries Shiny, Flexdashboard and Rmarkdown. The default layout was used, and there is no integration with other software solutions such as FORCE, so the sole purpose of this prototype is to give an indication of what the dynamic dashboard could look like. A screenshot of the dashboard prototype can be found in Figure 13. In dark blue color, the historical amount of mortgage applications per day can be found. In light blue color, a prediction of the amount of mortgage applications for the next week is shown. To the left of the graph, a dropdown field and a slider field can be seen, in which a change in interest rate can be defined. In this particular screenshot, a change in interest rate was manually set on July 20th, in which the interest rate will increase by 0.1%. As can be seen in the graph, our model expects a peak on the 19th and the 20th. By changing the height of the interest rate change or the moment of change, the graph will automatically create new predictions and update itself. Furthermore, the dashboard contains features such as zooming, panning and resizing.

--- FIGURE REDACTED DUE TO CONFIDENTIALITY ---

*Figure 13 – Screenshot of the dynamic dashboard prototype*

## 9.2 Recommendations for deployment

When deploying the model at the financial institutions, there are a number of aspects that are of importance. First of all, the model is not perfect so we suggest to use it in combination with the domain knowledge already available. This way the financial institutions can manually 'learn' how to use the model's predictions, and how to combine it with their own domain knowledge.

As discussed in Chapter 8.2, the outliers are often under-predicted, so in case the domain experts at the financial institutions expect a relatively large amount of mortgage applications it may be wise to add a safety margin on top of the model's predictions. In case there appears to be a large difference between the predicted amount of mortgage applications and the actual amount of mortgage applications, the financial institutions can adapt to this by planning extra personnel the next day(s).

The *optimization criterion* is important here: there is a tradeoff between personnel costs and mortgage application throughput times. By planning extra personnel, the throughput times of the mortgage applications are going down, but the personnel costs will go up. Planning too little personnel will save personnel costs, but increase the throughput times of the mortgage applications. Finding a 'sweet spot' here is important.

By re-training the model periodically (e.g. every night), one can ensure that the model uses all available data to make its predictions, and the model is optimized for predicting the upcoming week. As one of the variables used in the model contains data about the amount of mortgage applications on the previous working day, new data needs to be fed to the model on a daily basis to keep the data up-to-date.

Initially, the current model can be used to predict the amount of personnel needed, but optimally the financial institutions also would like to incorporate insight in the different types of mortgage applications (e.g. a mortgage application by an entrepreneur takes far more time to process than a regular mortgage application, as mentioned in Chapter 2.1), and insight in the bottlenecks in the mortgage application process (e.g. some stages in the mortgage application process might contain more workload than other stages, and hence need more dedicated personnel).

Finally, we suggest subsequently using a simpler model on the exact same dataset in order to make the predictions more interpretable. In general, the model's interpretability tends to go down as the predictive performance increases (Liu et al., 2014). A random forest generally gives better predictive performance than a decision tree, but is less interpretable. Using a random forest for the predictions and one of the interpretability methods as discussed in Chapter 8.2, or a subsequent model such as a basic decision tree, will make the model more interpretable, and could assist in determining the optimal amount of personnel needed. Another way of increasing the adoptability of the model is by running the model in the background for a period of time, in order to see how it performs in a 'live' environment, so that the personnel at the financial institutions can 'learn' how to use the model.

# 10 Conclusions, Limitations and Further Research

*In this chapter we propose the conclusions of our report, answering the research question and its subquestions. We also discuss the limitations of our research, and options for further research.*

## 10.1 Conclusions

In this report we developed a model using machine learning techniques that predicts the amount of mortgage applications for the upcoming week, using R and the CRISP-DM process model. For the visualization of the model, a dynamic dashboard is proposed. Below, we answer the research question and its subquestions, as formulated in Chapter 2.2.

*SQ1.* *What are the variables that influence the amount of mortgage applications?*

In Chapters 4.2.1 and 6.2, a list of 27 features for our model is developed, for a total of 6 categories. In Chapter 6.3 a subset of this list is made. Eventually, a total of 12 variables is used for our model, which can roughly be grouped in 3 categories: variables related to seasonality (e.g. holidays, vacations, time of the year), mortgage interest rates (e.g. the financial institutions's interest rates, changes in the financial institutions's interest rates, Euribor rates) and changes in regulations (e.g. changes in the mortgage loan regulations each January, changes in NHG regulations each July). These 12 variables are used for making the predictions.

*SQ2.* *Which techniques and algorithms can be used to create a model that is able to predict the amount of mortgage applications?*

As discussed in Chapter 4.3, there are two main type of prediction problems: classification problems and regression problems. Our research contains a regression problem, for which a specific set of machine learning algorithms can be used. In Chapter 7.1, we selected a list of 5 algorithms that can be used in our research: Support Vector Regression, Neural Networks, Classification and Regression Trees, Gradient Boosting Machines and Random Forests.

*SQ3.* *Which technique performs best on our dataset?*

Each of the models mentioned above was trained on the dataset, and using a t-test with a 95% significance level we can conclude that the Random Forest produces the best results in terms of Root Mean Square Error (RMSE), with the Gradient Boosting Machines model on a second place. The Random Forest model scores a Mean Absolute Error (MAE) of --- NUMBER REDACTED ---, which means the average deviation between the predicted amount of mortgage applications and the actual amount of mortgage applications on historical data are --- NUMBER REDACTED --- mortgage applications per day. As could be seen in Chapter 8.2, the error was partly caused by the under-prediction of outliers in the dataset, due to the inability of the model to fully capture the effects of interest rate changes.

*SQ4.* *How can we use this model to determine the amount of personnel needed during the next week?*

By incorporating the Random Forest model into a dashboard, a visualization can be made of the predicted amount of mortgage applications for the next week. Since we do not have data available upfront regarding mortgage interest rate changes, a dynamic

dashboard is proposed. In this dynamic dashboard, mortgage interest rate changes can be entered manually, and the effect of changing the mortgage interest rate on the amount of mortgage applications can be seen real-time. This dashboard can be used to make a prediction of the amount of personnel needed, by linking the amount of mortgage applications with the amount of personnel needed to handle these mortgage applications.

Finally, the main research question can be answered

*RQ.* *How can event log data be used to predict the amount of mortgage applications per day for the next week?*

In our models we used historical event log data to build a machine learning model that predicts the amount of mortgage applications for the next week, using several machine learning techniques discussed above. Our best model has a Mean Absolute Error of --- NUMBER REDACTED --- across a repeated 10-fold cross-validation, which indicates that the model's predictions of the amount of mortgage applications per day are on average --- NUMBER REDACTED --- applications higher or lower than the actual amount. By using this model to predict the amount of personnel needed to handle these mortgage applications, the financial institutions can save personnel costs and reduce the throughput time of the mortgage applications. A dynamic dashboard is proposed to visualize the amount of mortgage applications, in which interest rate changes can be manually entered in the dashboard.

## 10.2 Limitations

In this section we will discuss the limitations of our research. Most of these limitations are related to the dataset and the features that were engineered using external data.

First, not all of the features could be incorporated in our model, due to the limited availability of external data. For example, for the rental prices and income growth the data was either unavailable or aggregated on a yearly level, which makes these predictors useless for our model. Also the historical interest rates of competitors in the Dutch mortgage market could not be included, due to the unavailability of data. Although the actual interest rates of the competitors are known, there was no overview of the historical interest rates. Furthermore, Open Source Intelligence (OSINT) was not included in our model due to time and scope restrictions, as discussed in Chapter 6.2.

Another limitation is the lack of historical data that was available regarding mortgage applications. We have only used data from January 2013, which means we only had a certain amount of data available for events that occur on a yearly level, such as the changes in regulations. By having more historical data we would be able to capture the impact of a change in regulations more accurately, and make trends over the last few years better identifiable.

Furthermore, several non-repeating events at the financial institutions that do not follow a regular pattern but still have impact on the amount of mortgage applications (e.g. downtime of services or system failure, changes in acceptation criteria at the financial institutions, special offers at the financial institutions) were not included in our model due to scope and time restrictions. By including these events, the predictive power of our model should slightly increase, as the model will be better at explaining the outliers in the dataset.

## 10.3 RECOMMENDATIONS FOR FURTHER RESEARCH

Based on the limitations mentioned above, a number of recommendations for further research are defined. These recommendations can roughly be grouped into three categories: follow-up research, improving the predictive model and validating the dynamic dashboard.

Regarding the follow-up research, a next step would be to use the predictions of our model for determining the amount of personnel needed. This can be done in multiple ways, either by using a mathematical formula, or by using the predictions in a new model that predicts the amount of personnel needed for the next week. Other factors can be included in this new model, such as the expected throughput time of mortgage applications. This could be the subject of a follow-up research.

Open Source Intelligence (OSINT) could be used to improve the model's performance. As mentioned in Chapter 4.2.2, news messages regarding the economic situation, the housing market or mortgage interest rates influence the amount of mortgage applications on the short term. By analyzing news messages from multiple news websites, this effect could be captured and implemented in our model. Further research could be focused on extracting information from multiple media sources, to see if the model's predictive power can be improved. Next to this, data regarding the marketing budget of the financial institution could be added in our model, to see if the size of the marketing budget has impact on the amount of mortgage applications and can be used to improve the predictions.

Furthermore, as could be seen in Chapter 8.2, the model was unable to accurately predict the outliers in our dataset. The outliers were generally under-predicted, due to interest rate changes not being captured completely by our model. For further research we would suggest to look into this feature and investigate if there are any other factors that influence the amount of mortgage applications during an interest rate change.

Finally, a dynamic dashboard was developed to include the interest rate changes in our predictive model. Albeit the other options mentioned in Chapter 9.1 were not feasible, the dynamic dashboard solution requires validation, to check that it will be accepted by the stakeholders and provides added functionality. This can be done both by using scientific literature, as well as validating it internally at the customer.

# BIBLIOGRAPHY


Alnoukari, M., & El Sheikh, A. (2012). Knowledge Discovery Process Models: From Traditional to Agile Modeling. *Business Intelligence and Agile Methodologies for Knowledge-Based Organizations: Cross-Disciplinary Applications*, 72–100.

Al-Odan, H. A., & Al-Daraiseh, A. A. (2015). Open Source Data Mining Tools. In *Electrical and Information Technologies (ICEIT), 2015 International Conference on* (pp. 369–374). IEEE.

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Armstrong, J. S. (2011). Illusions in regression analysis. *Available at SSRN 1969740*.

Arner, D. W., Barberis, J. N., & Buckley, R. P. (2015). The Evolution of Fintech: A New Post-Crisis Paradigm?

Baghestani, H., Kaya, I., & Kherfi, S. (2013). Do changes in consumers' home buying attitudes predict directional change in home sales? *Applied Economics Letters*, *20*(5), 411–415.

Basak, D., Pal, S., & Patranabis, D. C. (2007). Support vector regression. *Neural Information Processing-Letters and Reviews*, *11*(10), 203–224.

Basten, C., & Koch, C. (2015). The causal effect of house prices on mortgage demand and mortgage supply: Evidence from Switzerland. *Journal of Housing Economics*, *30*, 1–22.

Bokeloh, P. (2017, February 14). Woningmarktmonitor april 2017 - Gissen naar plannen volgende regering. *Woningmarktmonitor*. Retrieved from https://www.abnamro.nl/nl/prive/hypotheken/wonen/woningmarkt/woningmarktmonitor.html

Boon, P. (2015, June 13). Topdrukte bij hypotheekverstrekker. *Telegraaf*. Retrieved from http://www.telegraaf.nl/dft/geld/huis-hypotheek/24152602/__Topdrukte_bij_hypotheekverstrekker__.html

Boumeester, H. (2016). *Monitor koopwoningmarkt - 3e kwartaal 2016*. Delft, The Netherlands: OTB - Research for the Built Environment.

Boumeester, H., & Lamain, m. m. v. C. (2016). *Eigen Huis Marktindicator - 3e kwartaal 2016*. Delft, The Netherlands: OTB - Research for the Built Environment.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Brockwell, P. J., & Davis, R. A. (2016). *Introduction to time series and forecasting*. springer.

Browne, M. W., Cudeck, R., Bollen, K. A., & Long, J. S. (1993). Alternative ways of assessing model fit. *Sage Focus Editions*, *154*, 136–136.

CBS StatLine. (2017a, March 20). CBS StatLine - Consumentenvertrouwen, economisch klimaat en koopbereidheid; 1972-2016. Retrieved May 29, 2017, from http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=7388pcr&D1=0,12,18&D2=513-515,517-519,521-523,525-527,530-532,534-536,538-540,542-544,547-549,551-553,555-557,559-561,564-566,568-570,572-574,576-578,581-583&HDR=T&STB=G1&VW=T

CBS StatLine. (2017b, May 29). CBS StatLine - Bestaande koopwoningen; verkoopprijzen prijsindex 2010 = 100. Retrieved May 29, 2017, from http://statline.cbs.nl/Statweb/publication/?DM=SLNL&PA=81884ned&D1=a&D2=306-308,310-312,314-316,318-320,323-325,327-329,331-333,335-337,340-342,344-346,348-350,352-354,357-359,361-363,365-367,369-371,374-376,l&HDR=T&STB=G1&VW=T

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (1999). The CRISP-DM user guide. In *4th CRISP-DM SIG Workshop in Brussels in March*.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business Intelligence and Analytics: From Big Data to Big Impact. *MIS Quarterly*, *36*(4), 1165–1188.

Dapp, T. F., Slomka, L., AG, D. B., & Hoffmann, R. (2014). Fintech–The digital (r) evolution in the financial sector. *Deutsche Bank Research", Frankfurt Am Main*.

De Haan, L., & Sterken, E. (2011). Bank-specific daily interest rate adjustment in the Dutch mortgage market. *Journal of Financial Services Research*, *39*(3), 145–159.

De Nederlandsche Bank. (2017). Rentes - De Nederlandsche Bank. Retrieved May 29, 2017, from https://www.dnb.nl/statistiek/statistieken-dnb/financiele-markten/rentes/index.jsp

Deira, S. (2015, May 15). Vier redenen voor aanhoudend herstel huizenmarkt. *Elsevier*. Retrieved from http://www.elsevier.nl/economie/article/2015/05/vier-redenen-voor-aanhoudend-herstel-huizenmarkt-1762383W/

DiCicco-Bloom, B., & Crabtree, B. F. (2006). The qualitative research interview. *Medical Education*, *40*(4), 314–321.

Doody, O., & Noonan, M. (2013). Preparing and conducting interviews to collect data. *Nurse Researcher*, *20*(5), 28–32.

Dua, P., Miller, S. M., & Smyth, D. J. (1999). Using leading indicators to forecast US home sales in a Bayesian vector autoregressive framework. *The Journal of Real Estate Finance and Economics*, *18*(2), 191–205.

Dua, P., & Smyth, D. J. (1995). Forecasting US home sales using BVAR models and survey data on households' buying attitudes for homes. *Journal of Forecasting*, *14*(3), 217–227.

Eckerson, W. W. (2007). Predictive Analytics. *Extending the Value of Your Data Warehousing Investment. TDWI Best Practices Report. Q*, *1*, 2007.

Feldman, D., & Gross, S. (2005). Mortgage default: classification trees analysis. *The Journal of Real Estate Finance and Economics*, *30*(4), 369–396.

Fernandez-Lozano, C., Seoane, J. A., Gestal, M., Gaunt, T. R., Dorado, J., & Campbell, C. (2015). Texture classification using feature selection and kernel-based techniques. *Soft Computing*, *19*(9), 2469–2480.

Finlay, S. (2014). *Predictive Analytics, Data Mining and Big Data: Myths, Misconceptions and Methods*. Springer.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1). Springer series in statistics Springer, Berlin.

Galindo, J., & Tamayo, P. (2000). Credit risk assessment using statistical and machine learning: basic methodology and risk modeling applications. *Computational Economics*, *15*(1), 107–143.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management*, *35*(2), 137–144.

Garbage in, garbage out. (2005). Retrieved April 10, 2017, from http://www.worldwidewords.org/qa/qa-gar1.htm

Geertsma, P. (2016, June 5). Drukte op woningmarkt zorgt voor vertraging hypotheekverstrekking in 2016. *TechnischWerken*. Retrieved from http://www.technischwerken.nl/nieuws/drukte-op-woningmarkt-zorgt-voor-vertraging-hypotheekverstrekking-in-2016/

Google Trends. (2017). Google Trends. Retrieved May 29, 2017, from /trends/explore

Gupta, R., Tipoy, C., & Das, S. (2010). Could We Have Predicted the Recent Downturn in Home Sales in the Four US Census Regions? *Journal of Housing Research*, *19*(2), 111–128.

Hand, D. J. (1999). Statistics and data mining: intersecting disciplines. *ACM SIGKDD Explorations Newsletter*, *1*(1), 16–19.

Hara, S., & Hayashi, K. (2016). Making tree ensembles interpretable. *ArXiv Preprint ArXiv:1606.05390*.

Ho, T. K. (1995). Random decision forests. In *Document Analysis and Recognition, 1995., Proceedings of the Third International Conference on* (Vol. 1, pp. 278–282). IEEE.

Hoover, J. (2009). How to track forecast accuracy to guide forecast process improvement. *Foresight: The International Journal of Applied Forecasting*, *14*, 17–23.

Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A practical guide to support vector classification.

John, G. H., Kohavi, R., & Pfleger, K. (1994). Irrelevant features and the subset selection problem. In *Machine learning: proceedings of the eleventh international conference* (pp. 121–129).

Jungo. (2016). Retrieved April 10, 2017, from http://www.jungo.nl/

Kadaster. (2017). Downloads. Retrieved May 29, 2017, from https://www.kadaster.nl/downloads

Kalender 365. (2017). Feestdagen 2017. Retrieved May 29, 2017, from https://www.kalender-365.nl/feestdagen/2017.html

Kashyap, A. K., & Stein, J. C. (2004). Cyclical implications of the Basel II capital standards. *Economic Perspectives-Federal Reserve Bank Of Chicago*, *28*(1), 18–33.

Kim, J.-H. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, *53*(11), 3735–3745.

Kolassa, S., & Schütz, W. (2007). Advantages of the MAD/MEAN ratio over the MAPE. *Foresight: The International Journal of Applied Forecasting*, (6), 40–43.

Kopanas, I., Avouris, N. M., & Daskalaki, S. (2002). The role of domain knowledge in a large scale data mining project. In *Hellenic Conference on Artificial Intelligence* (pp. 288–299). Springer.

Kuhn, M. (2012). Variable selection using the caret package. *URL {http://Cran. Cermin. Lipi. Go. Id/Web/Packages/Caret/Vignettes/CaretSelection. Pdf}*.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer.

Kurgan, L. A., & Musilek, P. (2006). A survey of Knowledge Discovery and Data Mining process models. *The Knowledge Engineering Review*, *21*(01), 1–24.

Lastra, R. M. (2004). Risk-based capital requirements and their impact upon the banking industry: Basel II and CAD III. *Journal of Financial Regulation and Compliance*, *12*(3), 225–239.

Lipton, Z. C. (2016). The mythos of model interpretability. *ArXiv Preprint ArXiv:1606.03490*.

Liu, S., Dissanayake, S., Patel, S., Dang, X., Mlsna, T., Chen, Y., & Wilkins, D. (2014). Learning accurate and interpretable models based on regularized random forests regression. *BMC Systems Biology*, *8*(3), S5.

Marbán, Ó., Mariscal, G., & Segovia, J. (2009). A data mining & knowledge discovery process model. *Data Mining and Knowledge Discovery in Real Life Applications*, *2009*, 8.

Mebius, D., & Haegens, K. (2016, November 20). Hypotheekverstrekkers platgebeld nu rente langzaam stijgt. *Volkskrant*. Retrieved from http://www.volkskrant.nl/economie/hypotheekverstrekkers-platgebeld-nu-rente-langzaam-stijgt~a4418695/

Meka, R. C. R., & Patil, A. (2015). Performing Predictive Data Analytics in Data Mining Using Various Tools. *IJITR*, *3*(4), 2229–2233.

Ministerie van Algemene Zaken. (2016, October 26). Wijzigingsregeling hypothecair krediet 2017 [regeling]. Retrieved May 29, 2017, from https://www.rijksoverheid.nl/documenten/regelingen/2016/10/26/wijzigingsregeling-hypothecair-krediet-2017

Mulder, M., & Lengton, M. (2011). Competition and interest rates in the Dutch mortgage market: an econometric analysis over 2004-2010.

NHG. (2016, October 31). Nieuwe Voorwaarden & Normen NHG per 1 januari 2017. Retrieved May 29, 2017, from https://www.nhg.nl/Over-NHG/Nieuws/Actueel-detail/ArtMID/833/ArticleID/98/Nieuwe-Voorwaarden-Normen-NHG-per-1-januari-2017

Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, *11*, 169–198.

Owen, H. (2015, April 20). Anscombe's quartet in mathematica. Retrieved from http://hywelowen.org/anscombes-quartet/

Pettinger, T. (2009, February 10). Factors Affecting Demand for Housing. Retrieved January 3, 2017, from http://www.economicshelp.org/blog/1302/economics/factors-affecting-demand-for-housing/

Pettinger, T. (2013, November 26). Factors that affect the housing market. Retrieved January 3, 2017, from http://www.economicshelp.org/blog/377/housing/factors-that-affect-the-housing-market/

Piatetsky, G. (2014). What main methodology are you using for your analytics, data mining, or data science projects? Poll. Retrieved November 29, 2016, from http://www.kdnuggets.com/polls/2014/analytics-data-mining-data-science-methodology.html

Piek in hypotheekaanvragen. (2016, December 5). *Hypotheken Data Netwerk (HDN)*. Retrieved from https://www.hdn.nl/piek-hypotheekaanvragen/

R: The R Project for Statistical Computing. (2016). Retrieved December 1, 2016, from https://www.r-project.org/

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why Should I Trust You?: Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1135–1144). ACM.

Rice, J. (2006). *Mathematical statistics and data analysis*. Nelson Education.

Rieg, R. (2010). Do forecasts improve over time? A case study of the accuracy of sales forecasting at a German car manufacturer. *International Journal of Accounting & Information Management*, *18*(3), 220–236.

RStudio - Open source and enterprise-ready professional software for R. (2016). Retrieved December 1, 2016, from https://www.rstudio.com/

Ryan, F., Coughlan, M., & Cronin, P. (2009). Interviewing in qualitative research: The one-to-one interview. *International Journal of Therapy & Rehabilitation*, *16*(6).

Schoolvakanties Nederland. (2017). Schoolvakanties 2017. Retrieved May 29, 2017, from https://www.schoolvakanties-nederland.nl/schoolvakanties-2017.html

Shmueli, G., & Koppius, O. (2010). Predictive analytics in information systems research. *Robert H. Smith School Research Paper No. RHS*, 06-138.

Stalder, F., & Hirsh, J. (2002). Open source intelligence. *First Monday*, *7*(6).

Terugblik 2015 en vooruitblik 2016. (2016, January 27). *Hypotheken Data Netwerk (HDN)*. Retrieved from https://www.hdn.nl/terugblik-2015-en-vooruitblik-2016/

Toolsema, L. A., & Jacobs, J. P. (2007). Why do prices rise faster than they fall? With an application to mortgage rates. *Managerial and Decision Economics*, 701–712.

Uitstekend half jaar voor hypotheekaanvragen. (2016, July 5). *Hypotheken Data Netwerk (HDN)*. Retrieved from https://www.hdn.nl/uitstekend-half-jaar-voor-hypotheekaanvragen/

van Dalen, P. (2016). Dutch housing market prospects: strong sales and rising prices.

Van der Laan, S. (2015, June 22). Herstel huizenmarkt zet door: prijzen blijven stijgen. *Elsevier*. Retrieved from http://www.elsevier.nl/Economie/achtergrond/2015/6/Herstel-huizenmarkt-zet-door-prijzen-blijven-stijgen-1780096W/

Varian, H. R. (2014). Big data: New tricks for econometrics. *The Journal of Economic Perspectives*, *28*(2), 3–27.

Vrieselaar et al., N. (2017, February 14). Op weg naar historisch hoge krapte op de woningmarkt. *Kwartaalbericht Woningmarkt*. Retrieved from https://economie.rabobank.com/publicaties/2017/februari/naar-hoge-krapte-nederlandse-huizenmarkt/

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79–82.

Wirth, R., & Hipp, J. (2000). CRISP-DM: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining* (pp. 29–39). Citeseer.

Wu, X., Zhu, X., Wu, G.-Q., & Ding, W. (2014). Data mining with big data. *Ieee Transactions on Knowledge and Data Engineering*, *26*(1), 97–107.

Yan, X. (2009). *Linear regression analysis: theory and computing*. World Scientific.

Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research*, *160*(2), 501–514.

Zomerdipje in hypotheekaanvragen. (2016, August 8). *Hypotheken Data Netwerk (HDN)*. Retrieved from https://www.hdn.nl/zomerdipje-in-hypotheekaanvragen/

# APPENDIX A – INTERVIEW SUMMARIES (IN DUTCH)

--- APPENDIX REDACTED DUE TO CONFIDENTIALITY ---