Hybrid Machine Learning (ML) Models in Banking: An Approach for the B2B Sector

by

Sonakashi Dhawan

A Master's thesis submitted to the
Faculty of Electrical Engineering, Mathematics and Computer Science (EEMCS)
in partial fulfillment of the requirements for the degree of

MSc in Business Information Technology- Data Science & Business

Faculty of Electrical Engineering, Mathematics
& Computer Science (EEMCS)

University of Twente

Enschede, Overijssel, The Netherlands

November 2023

Abstract

The advent of big data has revolutionized decision-making processes within the Business-to Business (B2B) financial sector, primarily by leveraging the predictive power of Machine Learning (ML) models. This study investigates the development of innovative Hybrid Model (HM)s tailored for predicting future investments in the B2B banking sector. By comparing HMs with the traditional models such as Extreme Gradient Boosting (XGBoost) regressor, the study highlights the superiority of HMs, for example, the ones employing k-means clustering, in terms of performance metrics. Furthermore, it uses Explainable artificial intelligence (XAI) techniques such as SHapley Additive exPlanations (SHAP) to increase the transparency and explainability of ML decisions, enhancing trust in automated financial forecasting. A comprehensive analysis reveals the effectiveness of HMs models over traditional ML methods, underscoring the potential of such HMs in reshaping the future of financial services. This research bridges a critical gap by providing empirical evidence on the efficacy of HMs, contributing to academic literature and offering a practical blueprint for financial institutions aiming to adopt advanced analytics in their operational strategies.

Keywords: ML, HM, Prediction, Banking, XAI, B2B

AUTHOR'S DECLARATION

I hereby declare that this thesis consists of original work of which I have authored. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I authorize the University of Twente to lend this thesis to other institutions or individuals for the purpose of scholarly research. I further authorize University of Twente to reproduce this thesis by photocopying or by other means, in total or in part, at the request of other institutions or individuals for the purpose of scholarly research. I understand that my thesis will be made electronically available to the public.

Sonakashi Dhawan

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my academic mentors, Marcos and Faiza, for their unwavering support and wise guidance, which have been instrumental in shaping the course of my research journey. Their mentorship has not only enriched my scholarly endeavours but also broadened my horizons of knowledge. I am particularly indebted to Marcos for his invaluable support throughout my project, for providing insightful feedback in a timely and consistent manner, and for his words of encouragement during the development of my thesis. I would also like to thank Martijn Bosma and Evert van Steen for giving me the opportunity to complete my graduation internship under their wing. Working alongside Jeremy Sabelis, Martijn Bosma and Evert van Steen has been immensely rewarding, providing me with practical perspectives that seamlessly blend academic theories with their application in the professional sphere.

The unwavering support of my family has been my backbone. I am deeply indebted to my father, whose encouragement gave me the freedom to pursue my passions and who has always been a pillar of support. My mother's belief in me and her nurturing support have been my guiding light. My brother's gentle push towards excellence, coupled with reminders to savor life's moments, has been a source of strength. To Vansh, whose encouragement, understanding, and love have been my shelter, especially in moments of stress. I am immensely grateful to my newfound family in Enschede, Rakshitha, Saran and Sai Ganesh, with a special mention to Sai Ganesh. His company has been invaluable, encompassing everything from collaborative academic endeavors to the joy of cooking together and engaging in endless conversation. To my old friends back home, who have left an indelible mark on my heart, and to the new friendships forged in the Netherlands, your camaraderie has been a source of happiness and support.

Thank you all for your integral role in this academic venture. Your collective wisdom has brought me to this significant academic milestone. I sincerely hope that you find the following pages as rewarding to read as I have found them to research and write.

CONTENTS

List of Figures

List of Tables

Abbreviations

$\mathbf{R^2}$ R-sqaured. ix, 15, 20, 23, 30, 40, 44, 46–50, 58, 59

AUC Area under the curve. 12–14, 18, 23

B2B Business-to Business. i, 1, 2, 5, 8, 9, 11–13, 15, 16, 18, 19, 32, 58–61

B2C Business-to Customer. 8, 9, 12–16, 18, 19

BO Bayesian Optimization. 4, 14, 43, 44, 47–50, 58, 59

CART Classification and Regression Tree. vii, 10, 12

CRM Customer Relationship Management. 24

DT Decision Tree. vii, 10, 12–15, 17, 18, 20, 25, 26, 38–40, 46, 47

EDA Exploratory data analysis. 4, 21, 32

EMA Exponential Moving Average. 34

FDR False Discovery Rate. 14, 18

FNR False Negative Rate. 14, 18

FPR False Positive Rate. 14, 18

HM Hybrid Model. i, v, vii, ix, 2–7, 10–13, 15, 18, 19, 36, 40–42, 46, 47, 50, 51, 53, 56–62

HML Hybrid Machine Learning. 17

KNN K-Nearest Neighbour. 13, 14

LDA Latent Dirichlet allocation. 14

LIME Local Interpretable Model agnostic Explanations. 5, 13, 14

LLM Logit Leaf Model. 12, 13, 17

LR Linear regression. vii, 10, 12–15, 17, 19, 20, 25, 38, 40, 46, 47

MAE Mean Absolute Error. ix, 13, 15, 20, 23, 29, 40, 44, 46–50, 58, 59

MI Mutual Information. vii, 37, 38

ML Machine Learning. i, v, vii, 1–6, 9–20, 23–26, 30, 38, 41–43, 49, 51, 58–61

MSE Mean Squared Error. 13, 15, 29

NN Neural Network. vii, 10, 12–14, 18

NPV Negative Prediction Value. 14, 18

OOB out-of-bag. 26

PCA Principal Component Analysis. 23, 63

PDP Partial Dependence Plots. 18

PR Polynomial Regressor. 4, 43, 44, 47–50, 59, 62

RF Random Forest. vii, 3, 10, 12–15, 17–20, 26, 38–40, 43, 46, 47, 57, 59, 62, 63

RFM Recency, Frequency, Monetary. 12, 15, 20, 24, 29

RFMTC Recency, Frequency, Monetary value, Time since first purchase, and Churn probability. 17

RMSE Root Mean Squared Error. ix, 13, 15, 20, 23, 29, 40, 44, 46–50, 59

ROC Receiver Operating Characteristic. 12–14, 18, 23

SEMMA (Sample, Explore, Modify, Model, and Asses. vii, 4, 20–22, 24, 38, 58

SHAP SHapley Additive exPlanations. i, vii, viii, 3–5, 13–15, 18, 20, 30, 31, 45, 46, 51–58, 60, 62, 77–81

SLR Systematic Literature Review. 2–4, 19, 25, 38, 58, 61

SVC Support Vector Classifier. 14

SVM Support Vector Machine. vii, 10, 12–15, 18

XAI Explainable artificial intelligence. i, v, 2, 3, 5, 18, 19, 45, 51, 58, 60–62

XGBoost Extreme Gradient Boosting. i, vii, ix, 3, 4, 13–15, 20, 27, 38–40, 42–44, 46–49, 51–54, 57–60, 63, 71

1
INTRODUCTION

1.1. INTRODUCTION

Big data has brought about a revolutionary transformation in the financial sector. This dynamic force has revolutionized traditional practices and opened up new horizons, reshaping how financial institutions operate, analyze risk, make informed decisions, and cater to customer needs in an interconnected world.

The demand for financial services has significantly increased, leading to the production of a vast amount of data in terms of volume, veracity, and variety [1]. The banking industry is actively gathering substantial volumes of information from both individual customer transactions and business clients i.e., in B2B interactions. This data encompasses a range of elements, including details about the businesses (like industry and size), behavioral insights (such as website interactions and campaign details), metrics regarding product usage (like how often certain actions are taken with the product), and information about the customer-bank relationship [2]. This enormous amount of data is being used by the banking sector to its advantage by adopting a data-driven strategy through utilizing analytics. There is a shift from descriptive analytics to predictive and prescriptive analytics. This shift empowers banks to gain deeper insights into their customers, thus enabling them to create tailored services, proactively manage risk and efficacy of operational frameworks. As a consequence, the profitability and customer satisfaction is increased. Bolívar et al. [3] has shown that implementation of ML models instead of traditional quantitative models in several areas of credit risk management could save 12.4% to 17% in terms of regulatory capital requirements.

Currently, customer churn poses a significant challenge for banks due to increased competition [4], dissatisfaction with the services provided, or enticing offerings from other financial institutions and, hence retaining customers has become a pressing mission for banks in the face of these emerging factors [5]. The costs of acquiring new customers are five to six times more than retaining existing ones [6]. Recognizing the significance of this matter, efforts have been directed toward predicting churn in order to mitigate its impact. Though churn prediction is a useful strategy, it cannot solve the problem entirely on its own. As a result, banks are attempting to adopt new ML models to segment consumer data and produce more accurate forecasts about customers' next investments [7–10]. Predicting investments can inform about churn risk by identifying changes in their behavior or disengagement. If the predicted investments are significantly reduced or diverge from past patterns, it may signal dissatisfaction and a higher likelihood of churn, prompting businesses to take proactive retention measures. Forecasting a customer's future investments to detect churn has emerged as a pivotal factor in organizational marketing decision-making. It is instrumental in elevating business value through the enhancement of operational, social, environmental, and financial performance [11].

1.2. RESEARCH QUESTIONS

This study aims to identify any existing gaps in predicting customer's next investments and propose potential directions for future research through a Systematic Literature Review (SLR). The research presented also provides useful insights and inspiration for future studies. In particular, the application of XAI to HM frameworks within the banking sector for B2B transactions. To achieve the stated research goal, various research questions have been formulated. These questions, detailed below, provide a basis for investigating relevant studies and developing approaches to address the research questions presented. These questions are outlined in the section below help form the guidelines for the review of existing literature and the formulation of strategies to tackle the presented research queries.

MAIN RESEARCH QUESTION

1. How can the application of HMs improve the prediction of future investments by B2B customers in the banking industry?

SUB-RESEARCH QUESTIONS

1. What are the current ML models and techniques used in the banking industry to predict the customers' next investments?
2. What are the different HMs that can be used to predict a customer's next investment?
3. How does the integration of model-agnostic methods with a HM in place of traditional ML models affect the level of explainability?

1.3. RESEARCH GOAL

The primary objective of this study is to discover how the use of HMs can be beneficial in predicting the future investments of B2B customers in the banking sector. The research also aims to shed light on how XAI can be applied to these HMs to increase the level of explainability. The study begins with a detailed review of the current landscape of predictive-modelling across different sectors, highlighting the role of ML methods and HMs. Through an extensive review of the academic literature, this thesis aims to identify existing gaps and research opportunities in the area of ML. It also intends to provide insights and motivation for further research in the development of HMs. Also, XAI is used to explain the predictions of HMs and individual ML models.

To achieve the research objective, the study is conducted in three key steps. The first step is to investigate different individual baseline ML models for predicting the customer's next investment on the entire dataset. Then, the performance of the two best-performing baseline models is improved by employing distinct optimization techniques. In the second step, two different types of HMs are designed, both using different techniques, to predict the customer's next investment. Lastly, a model agnostic approach is applied to obtain and compare the effectiveness of both methods in terms of explainability.

The insights derived from addressing the above-mentioned research questions can be used by organizations to obtain a deeper understanding of their customers. This knowledge can help them make well-informed decisions. Moreover, comprehending the limitations of traditional ML models can pave the way for developing new models with improved explainability. These advanced models can effectively identify the underlying factors influencing customer investment decisions, enabling banks to customize their strategies and offerings accordingly. This tailored approach fosters stronger customer engagement and loyalty, which can result in heightened retention rates and a competitive edge in the market.

1.4. METHODOLOGY

To address the research question, a comprehensive SLR was carried out. Subsequently, three experimental designs were formulated to help in predicting the customer's next investment. A pictorial view of experiments has been presented in Figure 1.1 methodology. The initial experimental setup involved the use of four baseline ML models to predict the target variable. Following this, the two models that demonstrated superior performance i.e., the XGBoost and RF were fine-tuned by means of an optimization process. Additionally, XAI using SHAP was employed to provide insight into the decision-making processes of these models.

The second experimental phase employed a HM approach, utilizing domain knowledge to segment the data into four distinct clusters. An optimized XGBoost regressor was then individually applied to each cluster to make predictions. Then, the SHAP method was used to explain the model's prediction, which offered a deeper understanding of the influence of different features on the model's outcomes.

The third experiment was done in the same way as the second experiment, with the difference that instead of domain knowledge, the data segmentation was performed using ML algorithm called k-means clustering. This resulted in the formation of three clusters, on which the optimized XGBoost regressor was applied to make predictions. The SHAP method was again used as the explanatory tool, to provide a transparent framework for understanding the prediction.

The overarching focus of these experiments was to leverage a range of ML models, with a special emphasis on the XGBoost regressor, to achieve a robust and interpretable predictive framework. Optimization of these models was achieved through advanced techniques such as Polynomial Regressor (PR) and Bayesian Optimization (BO). The use of the SHAP method across these models underscores a commitment to the explainability of the HMs, ensuring that the predictions are not only accurate but also understandable in the context of their contributing factors.

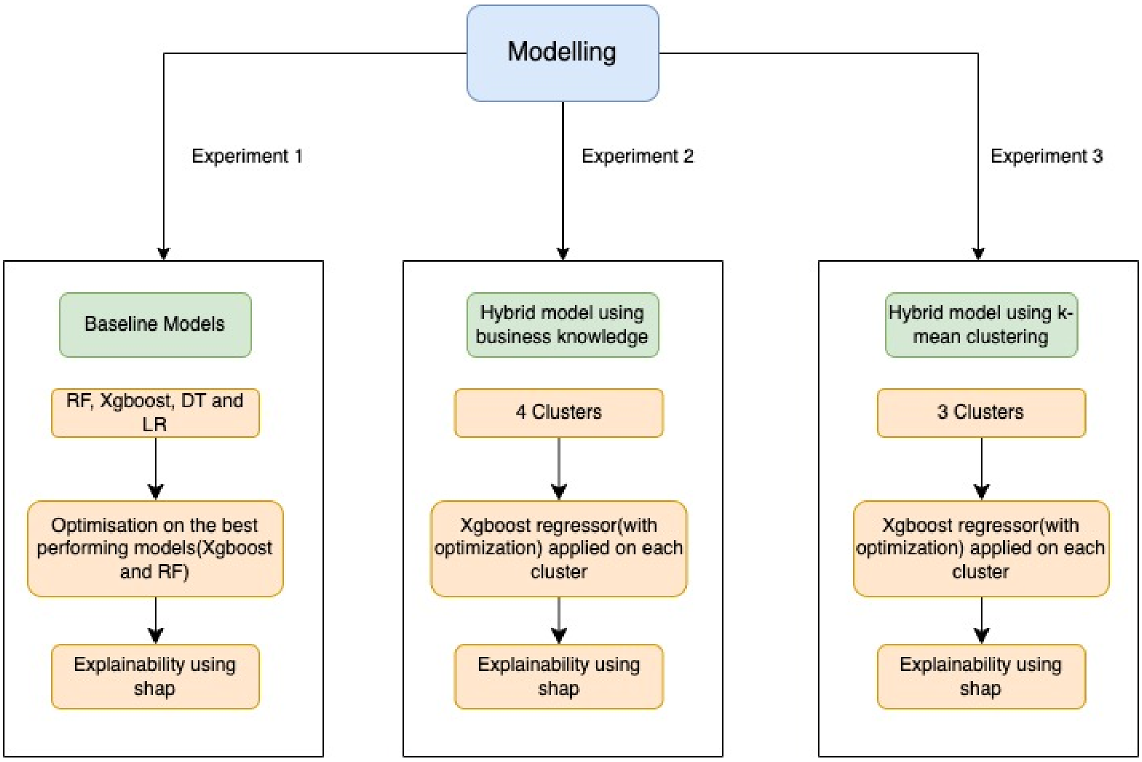

Modelling
Experiment 1
Experiment 2
Experiment 3
Baseline Models
RF, Xgboost, DT and LR
Optimisation on the best performing models(Xgboost and RF)
Explainability using shap
Hybrid model using business knowledge
4 Clusters
Xgboost regressor(with optimization) applied on each cluster
Explainability using shap
Hybrid model using k-mean clustering
3 Clusters
Xgboost regressor(with optimization) applied on each cluster
Explainability using shap

Figure 1.1: Flowchart summarizing the three experimental approaches: Baseline model evaluation, HM optimization with business knowledge, and HM made using k-means clustering.

1.5. THESIS STRUCTURE

This thesis is organized into six chapters, which unfold in the following manner. Chapter 1 provides background information, proposing research questions, and outlining the objectives of the study. Chapter 2 presents a SLR aimed at addressing the proposed research questions, compiling relevant theoretical concepts, and identifying gaps in the current research landscape. Following this, Chapter 3 describes the main research methodology, the SEMMA approach, and examines various analytical models applied in the study, including model-agnostic methods like SHAP for explainability. Chapter 4 delves into the data acquisition process and the preliminary steps of Exploratory data analysis (EDA), which encompasses data pre-processing, feature engineering, and the exploration of various models for data modeling including HMs and individual predictive models, as well as their evaluation. Subsequently, Chapter 5 discusses the results obtained by the application of various HMs and individual ML methods in forecasting customer investment decisions. It also presents the results of SHAP analysis for the best baseline model and of the HMs. Lastly, Chapter 6, provides the conclusions of this thesis by answering the research questions and underscoring the academic and practical significance of the findings. It also provides a critical assessment of the study's limitations and provides an outlook for future research.

1.6. CHAPTER SUMMARY

This chapter introduces the research study on the role of big data in the financial sector, focusing on how banks can utilize data to enhance customer understanding and retention through predictive analytics. It emphasizes the significance of ML and HM in predicting B2B customer investments and the integration of model-agnostic methods like SHAP and Local Interpretable Model agnostic Explanations (LIME) for improved interpretability. The chapter outlines the research questions aimed at uncovering the potential of HM in investment forecasting and the application of XAI to these models. The primary goal is to bridge research gaps and advance the understanding of predictive models in banking. The chapter sets the stage for a thorough literature review, methodology exposition, and exploration of individual and hybrid predictive models, culminating in a comparative analysis of their explanatory power. The insights gained will guide future strategy development for customer retention and engagement in banking.

2

LITERATURE REVIEW

2.1. SYSTEMATIC LITERATURE REVIEW

In this literature review, two citation databases, Scopus[1] and Google Scholar[2], were used. Both databases provide an extensive range of literature related to the research topic. Primarily, the Scopus database was preferred due to its diverse features, which facilitate a more streamlined search process. The platform offers a convenient search bar, enabling the use of keywords to access and download research literature. The research keywords used for downloading relevant material encompassed "HM", "ML", "explainability", "prediction", and "banking". These keywords were combined using AND operator in search queries to address the research questions effectively. By using these search keywords, it is ensured that the results obtained after the research will contain at least one of these specific keywords.

Another citation database tool utilized for finding relevant literature in this research was Google Scholar. This platform offers a user-friendly approach to conducting broad searches for scholarly publications. Users can explore a wide array of disciplines and sources, including articles, theses, books, abstracts, and court opinions, originating from academic publishers, professional societies, online repositories, universities, and other websites. With Google Scholar, researchers can directly input a brief sentence or a few keywords into the search bar, generating a substantial output that requires subsequent filtering to extract the most pertinent results.

Following are the search queries we used while searching for the relevant literature.

Scopus

1. Hybrid AND models AND in AND bank
2. Explainable AND Hybrid AND models

[1] https://www.scopus.com/home.uri

[2] https://scholar.google.com/

3. Churn AND Prediction

Google Scolar

1. Predicting customer's investment in banking
2. Explainable HMs
3. HMs for prediction

2.1.1. INCLUSION AND EXCLUSION CRITERIA

The inclusion and exclusion criteria have a critical role in systematically identifying literature that is both relevant and of high quality. They facilitate the execution of a comprehensive and precisely focused review of the existing knowledge within a particular field. Table 2.1 presents the inclusion and exclusion criteria, which outline the specific protocols for study selection.

Table 2.1: Criteria for literature selection.

Inclusion Critera	**Exclusion Criteria**
Literature is written in English.	Literature not having open access.
Literature belongs to Computer Science, Data science, Management, Accounting Economics, Econometrics, and Finance.	Literature did not have the required keywords.
Literature was published in the last 10 years.	Based on the abstract, the papers were discarded.
Literature belongs to conference proceedings papers.	Duplicates were removed

The inclusion criteria utilized in this systematic literature review encompass several essential elements. Initially, the chosen literature should be composed in English, guaranteeing its accessibility and comprehensibility. Furthermore, the literature should fall under the domains of Computer Science, Data Science, Management, Accounting, Economics, Econometrics, and Finance. This step ensures that the literature is aligned with the relevant topic. Additionally, only publications from 2012 to 2023 were taken into account to uphold a contemporary viewpoint on the subject matter. A preference was also shown for literature stemming from conference proceedings papers.

To ensure the presence of pertinent and top-quality literature, measures were undertaken to eliminate redundant and inaccessible content. Initially, literature lacking full-text accessibility was excluded, followed by the removal of duplicate entries from the combined results of both databases. Subsequently, literature lacking the specified keywords in their abstracts was also filtered out.

2.1.2. LITERATURE SELECTION PROCESS

The process of selecting literature begins with executing queries derived from specific databases. The inclusion and exclusion criteria in Table 2.1 are then applied during this literature selection process. This serves to ensure the quality of the collected research and also helps streamline the data extraction process.

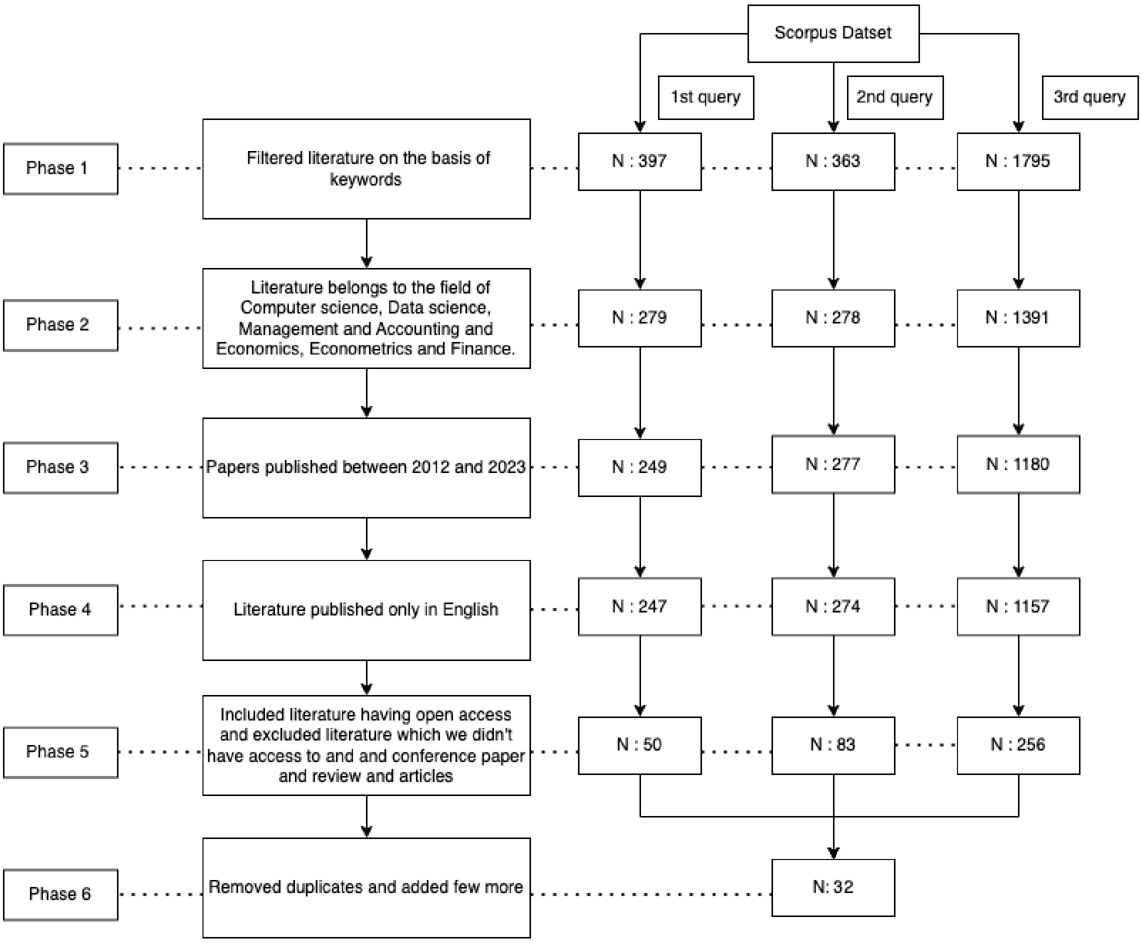

Scorpus Datset
1st query
2nd query
3rd query
Phase 1
Filtered literature on the basis of keywords
N : 397
N : 363
N : 1795
Phase 2
Literature belongs to the field of Computer science, Data science, Management and Accounting and Economics, Econometrics and Finance.
N : 279
N : 278
N : 1391
Phase 3
Papers published between 2012 and 2023
N : 249
N : 277
N : 1180
Phase 4
Literature published only in English
N : 247
N : 274
N : 1157
Phase 5
Included literature having open access and excluded literature which we didn't have access to and and conference paper and review and articles
N : 50
N : 83
N : 256
Phase 6
Removed duplicates and added few more
N: 32

Figure 2.1: Literature Selection Phases.

Figure 2.1 shows the steps followed for the literature selection process. Following the implementation of the above stages, a sum of 50, 83, and 256 articles were identified through each query using the inclusion and exclusion criteria. In the final phase, 32 relevant articles were chosen after eliminating duplicate entries.

A similar analysis was conducted using Google Scholar informally. That is, the exclusion and inclusion criteria were found without employing a step-by-step methodology. Any duplicates found in both Google Scholar and Scopus during the paper review process were subsequently eliminated. The results from both the datasets were then combined.

2.2. RELEVANT TRENDS IN LITERATURE

This section comprises five distinct parts. The initial part will present an analysis of various papers dedicated to two specific research segments: B2B and Business-to Customer (B2C). The second subsection will delve into the year-wise distribution of the published literature spanning the period from 2013 to 2023. Subsequently, the third subsection will provide clear definitions of the diverse ML techniques employed for predicting customer investment in the study literature. Furthermore, the fourth subsection will encompass a comprehensive exploration of different studies conducted across various areas such as telecommunications, the financial sector, and online learning portals. Lastly, a journal-wise distribution and key-wise distribution have been presented in tabular and pictorial form, respectively.

2.2.1. SEGMENT-BASED TRENDS

Figure 2.2 shows the analysis of the literature published between 2012 and 2022, focusing on segment-based trends. The analysis reveals significant trends within the research areas and highlights potential research gaps. After reviewing the literature, it becomes evident that the majority of research concerning customers' upcoming investments and churn predictions in the financial and telecom sectors has predominantly focused on the B2C domain. In contrast, there has been relatively little research done in the B2B field. This observation underscores the existence of notable gaps in the existing literature.

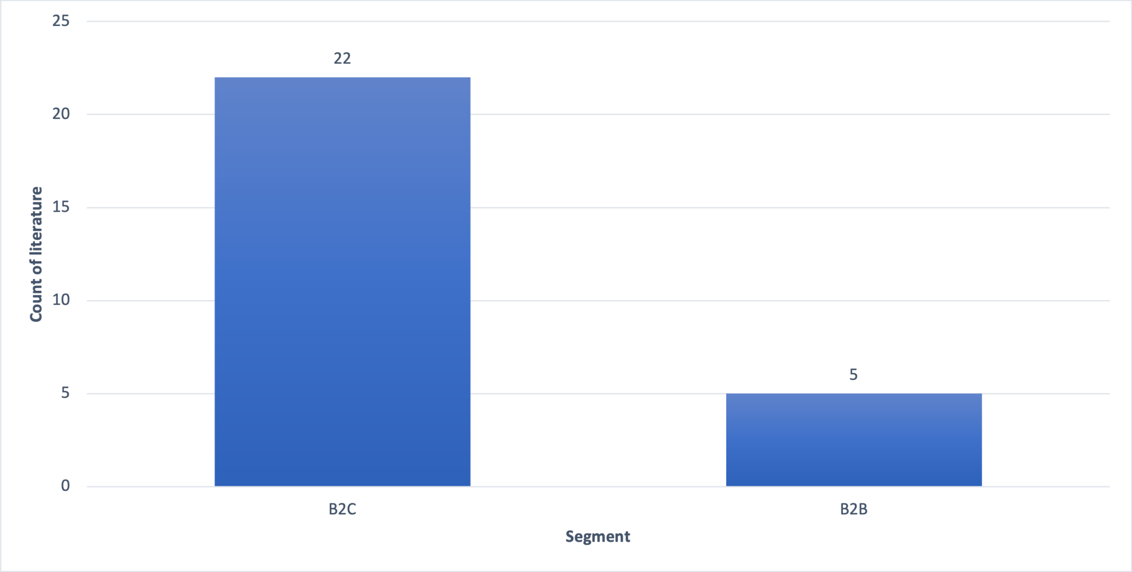


Figure 2.2: Segment-Based Trends.

2.2.2. YEAR WISE DISTRIBUTION

Figure 2.3 shows valuable insights and trends obtained by a year-wise analysis. The study indicates a consistent overall growth in publications from 2013 to 2022, which may signify a rising interest in the subject or the emergence of new sub-fields. Notably, the trend exhibits some fluctuations around the years 2018-2020, followed by a sudden increase in 2022. This significant surge could be attributed to advancements in new technologies and the implementation of regulations, particularly concerning the explainability of ML models. These factors might have contributed to a heightened interest in this field among researchers and practitioners.

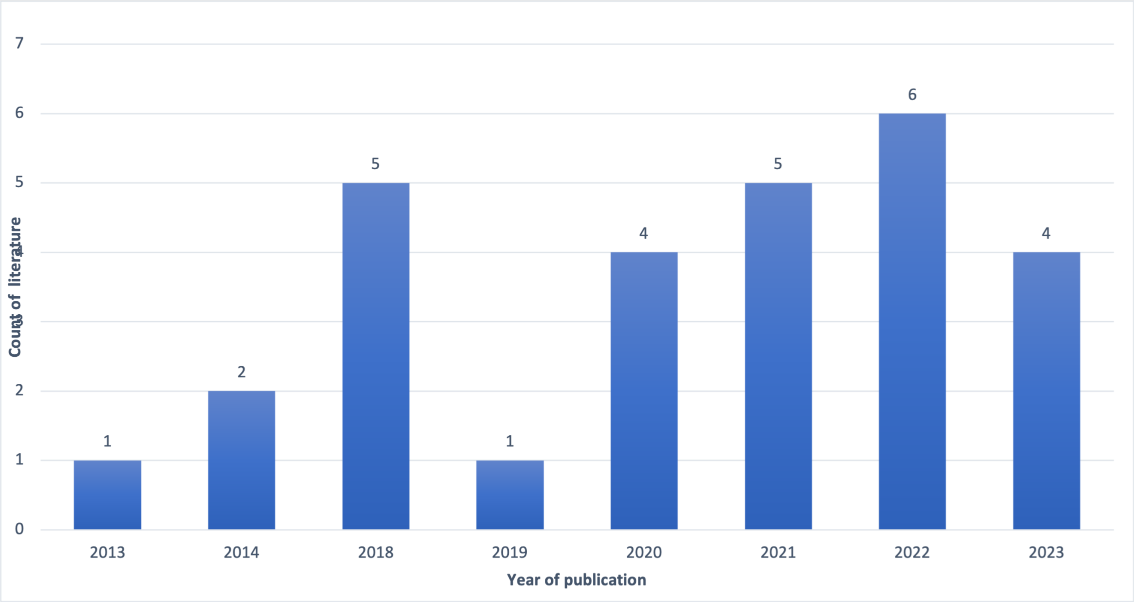


Figure 2.3: Year distribution of research papers.

2.2.3. DIFFERENT ML TECHNIQUES

Figure 2.4 shows different ML methods that have been employed in multiple research articles. The analysis reveals a growing prevalence of HMs as the preferred choice for prediction in the financial industry[3]. Additionally, supervised ML techniques like LR, DT, and RF remain widely used. This diverse array of approaches underscores the ongoing efforts to explore and harness the predictive capabilities offered by various ML techniques.

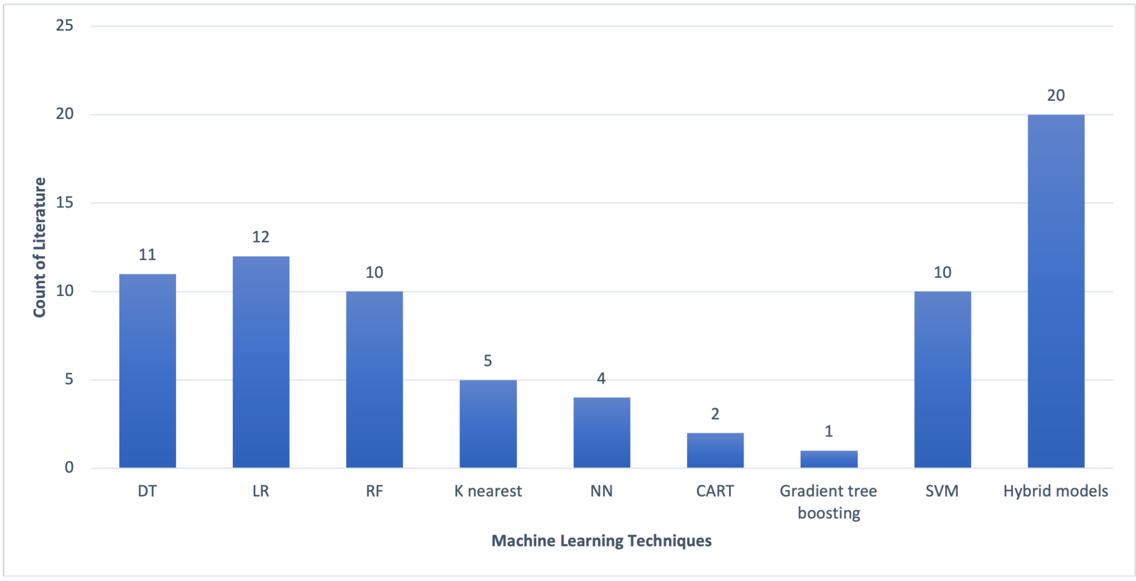


Figure 2.4: The number of times ML techniques such as NN, CART, DT and SVM appeared in the reviewed literature.

[3] This preference could be due to the fact that the term “hybrid model” was used in the search query.

2.2.4. SUBJECT BASED TREND

The trend analysis within the subject area demonstrates the diverse range of research conducted in various fields, while also drawing attention to potential research gaps. Figure 2.5 represents the distribution of the reviewed literature in this research, categorized according to subject areas. The analysis indicates a considerable focus on finance, telecommunications, and healthcare. Within the financial sector, significant research efforts have been dedicated to predicting customers' future investments and anticipating churn behavior. Additionally, other subject areas like online learning portals and manufacturing domain forecasting have also received attention in a few studies.

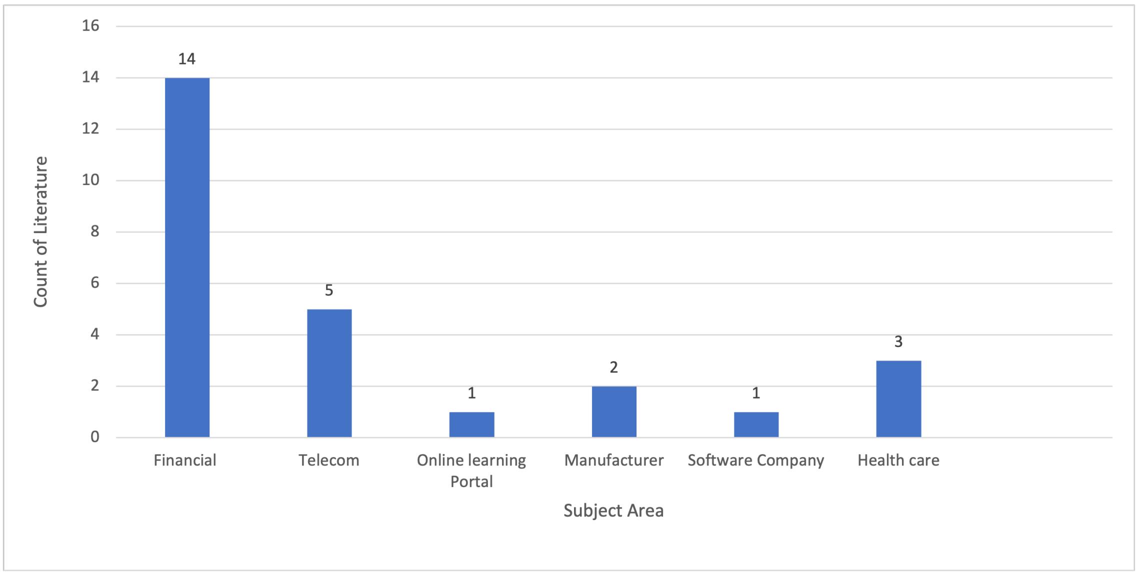


Figure 2.5: Subject distribution.

2.2.5. JOURNAL-WISE DISTRIBUTION

This subsection focuses on examining the distribution of literature gathered from various journals in different subjects such as financial and telecommunication sectors. Table 2.2 comprises a collection of articles relevant to our research topic, indicating the respective journal names where these articles were published, the specific ML techniques used in the research papers, the business domains targeted in the studies, and the subject areas of the research. It can be seen that most of the work has been done in B2B, with only a few research papers using HMs in their work.

Table 2.2: Quantitative Analysis of the Literature.

Author	HM	Subject	ML Techniques	Metrics-evaluation	Dataset	Segment
De Caigny et al. [1]	Yes	Financial	DT, logistic regression, Logit Leaf Model (LLM), RF and logistic model trees	Area under the curve (AUC) and TDL, 5 X 2 cross-validation, missing value imputation, zero, median and modus imputation, dummy encoding, Winsorization, undersampling, fisher score, Test-Holm posthoc test	14 churn data sets from European financial services provider, retailer, DIY supplier, newspaper company, telecom operator, energy company and Duke.	B2C
Moro et al. [12]	No	Financial	NN	Recency, Frequency, Monetary (RFM), AUC and Receiver Operating Characteristic (ROC), lift cumulative curve area (ALIFT), data-based sensitivity analysis (DSA), forward selection method and realistic rolling window scheme	Real data collected from Portuguese bank. 52,944-phone calls	B2B
Martínez et al. [13]	No	Financial	Logistic Lasso regression, Gradient tree boosting	Prediction accuracy, AUC, ROC, 10-fold cross-validation	B2B data of 10000 customers and a total number of 200000 transactions. Set of 274 features	B2C
Tamaddoni et al. [14]	No	Telecom	CART, boosting Logistic regression(benchmarked)	ROC and cumulative lift measures, AUC	–	B2B
Hué et al. [15]	Yes	Financial	Penalised Logistic Tree Regression, RF, logistic regression, linear logistic regression, non-linear regression with an adaptive lasso, and SVM and NN	Monte Carlo simulations, NX2-fold cross-validation, ROC, the percentage of correctly classified(PCC) and partial Gini Index(PGI)	Financial Dataset from Kaggle "Give me some credit"	ROC
De Caigny et al. [16]	Yes	Telecom	BLM, LLM, DT, support leaf model	Fowatd selection- Fisher score, AUC , TDL	–	B2C
Kunchaparthi et al. [17]	Yes	Telecom	SVM, LLM	AUC, TDL, selecting variables: Fisher score	Two customer churn datasets: DS1-WA_Fn-UseC_-Telco-Customer-Churn (no of records- 7044), DS2 - Cell2cell (no of records-71049)	B2C
Coussement et al. [18]	Yes	Online Learning Portal	LLM, LR, SVM, NN, DT, RF, Boost, NB, BN, and Hidden Markov models.	5X2 cross-validation F-test, AUC and TDL	Data set having 10,554 students and 122 variables from global online learning provider	B2C
Pawełek and Pociecha [19]	Yes	Manufacturer	LLM, CART classification, logit model	Sensitivity, specificity, precision, F1, G-mean and AUC.	61 financial ratios from the manufacturing sector in Poland containing 5910 enterprises.	B2C
Vafeiadis et al. [20]	No	Telecom	Ridge classifier, gradient booster, adaptive boosting (AdaBoost), bagging classifier, k-nearest neighbor (kNN), DT, LR, and RF	Feature selection, accuracy, AUC score, precision score, recall score, F1 score	Data set is from the leading telecommunication company in Indonesia consisting of 80,000 customers including both active and inactive.	B2C

De Caigny et al. [2]	Yes	Software Company	Uplift DT, LR, LLM and RF	5-fold cross-validation, Kullback-Leibler divergence splitting criterion, grid search, Qini Coefficient	The data set is from European b2B software company having 6432 observations	B2B
Sayjadah et al. [21]	No	Financial	Logistic regression, Rpart Decision Tree and RF	Correlation-based Feature Selection (CFS), Accuracy, Trye Positive, AUC	The dataset used is generated from credit card operations by the users. It is made up of 30000 instances, 24 attributes	B2C
Bolívar et al. [3]	No	Financial	Logistic Lasso, Tree(CART), RF, XGBoost and deep learning	Mean, standard deviation, k-fold cross-validation, AUC-ROC curve, Brier score	An anonymized dataset from Banco Santander having more than 75,000 credit operations classified as default or not.	B2C
Hudaib et al. [22]	Yes	Telecom	HM (DT and Artificial NN)	encoding, feature selection, Fisher score, AUC	Two datasets from telecommunications sector.	B2C
Thirugnanam [23]	Yes	Health care	SVM, DT, and logistic regression, rule-based algorithm	Removal of missing fields, normalization of data, removal of outliers, purity, Gini coefficient, statistical deviance.k-fold cross-validation, Hosmer-Lemeshow test, classification rule, sensitivity, specificity, and accuracy.	Datasets collected from Cleveland Heart Disease Dataset (CHDD) available on the UCI Repository	B2C
Kaur and Kaur [24]	Yes	Financial	Logistic regression, k-nearest neighbor, DT	Exploratory data analysis, feature selection, AUC-ROC	The dataset is from Kaggle having 28,382 records and 21 features (attributes).	B2C
Chou [25]	Yes	Financial	HM integrating the decision tree with the deep NN	Accuracy, Sensitivity and Specificity, LIME algorithm	The financial statement, corporate governance, and corporate information disclosure of the distressed and stable companies were used as inputs to train the	B2B
De et al. [26]	Yes	Financial	(a) TREPAN decision tree [4] [1] (b) hidden-layer-clustering on a NN	AUC, ROC, LIME	The data (source: UCI ML Repository) consists of 30,000 samples (customers)	B2C
Sheuly et al. [27]	Yes	Manufacturer	A HM of ANN and K-Nearest Neighbour (KNN), SVR and KNN and PLS	SHAP, RMSE, MAE, Mean Squared Error (MSE),	A dataset from the manufacturing companies' logistics containing 155 variables information of 154,029 PTU units	B2B
Jain and Jana [28]	No	–	SGFL algorithm, RF regressor algorithm	SHAP values based on Game theory, RMSE, MAE, MSE, Mean absolute Percentage error	Medical Coimbra dataset, business car evaluation dataset, employee churn problem dataset from Kaggle Human Resource Information System (HRIS	B2B

[4] https://www.sciencedirect.com/science/article/pii/S187705092030394X

Joseph et al. [29]	Yes	Health care	TabNet model, BO-TabNet, FCN, DNN, Gradient Boost, AdaBoost, XGBoost, SVC, KNN, Logistic regression	Median imputation, box plots and interquartile range (IQR),10-fold cross-validation, BO, Grid search (GS),loss-function and accuracy, precision, recall, specificity, F1 score, False Positive Rate (FPR), False Negative Rate (FNR), Negative Prediction Value (NPV), False Discovery Rate (FDR), Cohen's Kappa (κ), and ROC, AUC.	medical Coimbra dataset, business car evaluation dataset, employee churn problem dataset from Kaggle Human Resource Information System (HRIS)	B2C
Desai and Khairnar [30]	Yes	Financial	SVM, RF, KNN and adaptive boost ML	Thiel's U method for correlation, Yeo-Johnson method for transformation, accuracy, precision, recall, f-measure metrics, ROC	Customer data of a Portuguese banking organization.	B2C
Devi and Yalavarthi [31]	No	Financial	Latent Dirichlet allocation (LDA), multivariate discriminant analysis (MDA) and LR, and ML techniques such as artificial NN, SVM and DT, genetic algorithm (GA) and particle swarm optimization (PSO)	Accuracy, precision, sensitivity, and specificity	–	B2C
Koumetio Tekouabou et al. [32]	No	Mathematics	Ensemble methods (RF, B, GB, ET, and AB), individual machine learning-based methods (KNN, Support Vector Classifier (SVC), DT, LR, ANN, and NB)	SMOTE algorithm, feature importance, Shape, and SHAP value, as well as feature importance, 5-fold cross-validation	Dataset, is from Kaggle	B2C
Dias et al. [33]	Yes	Financial	k means was the first clustering approach, XGBoost, RF, AdaBoost, DT, KNN, SVM, LR, ELM	Synthetic Minority Over Sampling Techniques (SMOTE), LIME and SHAP, Intelligent system-based labeling through deep clustering, tree-based learning techniques, recall and F1-score, k-fold cross-validation	Database (Berka) from a Czech bank, dataset from Kaggle	B2C
Choi and Choi [34]	No	Financial	Random Subspace (RS), Multi-Boosting (MB) and Random Subspace-Multi-Boosting (RS-MB)	Partial Dependence Plot (PDP), random subspace rate, 10 cross-validations, RF, min-max Normalization, Synthetic Minority Oversampling Technique (SMOTE), Average Accuracy, F-Measure, FNR, FPR, and AUC, ROC	University of California Irvine Machine Learning Repository (https://archive.ics.uci.edu/ml/), we obtained 41,188 bank telemarketing campaigns for term deposits named 'Bank Marketing Data Set'.	B2C
Uddin et al. [35]	No	Financial	Naïve Bayes, DT and SVM algorithms	The hybrid ML classifier of DT and Naive Bayes algorithms, mean, median, max, and min, var() and std(), recall, sensitivity, specificity, f-measure, accuracy, and precision	Dataset Kaggle website. The dataset consists of 615 records and 13 features of the loan applicant.	B2C

Gastón and Garcia-Viñas [36]	No	–	–	–	–	B2C
Yeh et al. [37]	No	Health care	RFM model to derive a formula	–	Donor database of Blood Transfusion Service Center.748 donors	B2C
Thesis	Yes	Financial	RF, XGBoost, HM	MAE, MSE, RMSE, $\mathbf{R}^2$, SHAP	Data from one of the largest European bank	B2B

2.2.6. KEYWORDS WISE DISTRIBUTION

This subsection explores the trends observed through the analysis of keywords extracted from the reviewed literature. The word cloud visualization in Figure 2.6 presents a graphical representation where words are displayed based on their frequency and significance within the text. This word cloud is generated using the abstract of each article, highlighting key terms that carry significant weight in both research and practical applications within the field. Notably, the word "prediction" is the center of Word Cloud in Figure 2.6, which implies that it has the highest frequency of appearance in the studied research articles. Other important keywords include "B2C", "customer churn", and "ML", which have made significant contributions to the literature review.

Various ML techniques, such as LR, DT, and SVM, have been extensively utilized and contributed significantly to the literature. Additionally, the employment of HMs has played a vital role in enhancing the explainability of the models, particularly in research done in the financial sector dealings with B2B interactions.

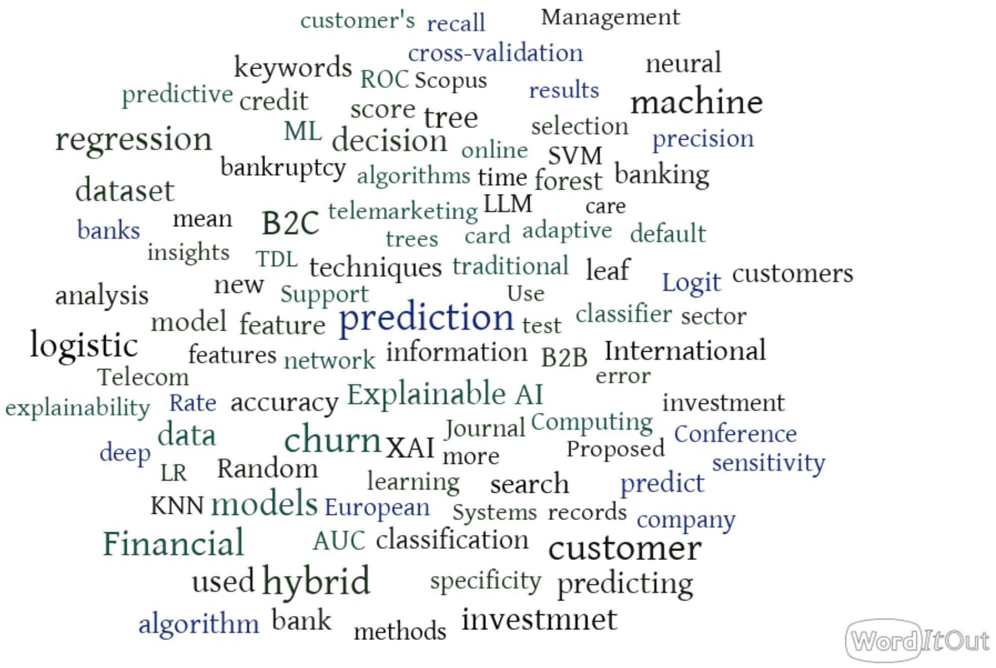


Figure 2.6: Word Cloud depicting the Keywords from the Reviewed Literature.

2.3. DOMINANT THEMES IN LITERATURE

2.3.1. RECURRING THEMES

The first subsection aims to provide an overview of the recurring contexts in which forecasting customer investment has been studied in the literature.

1. In the realm of forecasting customer investment, a significant body of research has concentrated on interactions between businesses and individual customers, commonly known as B2C interactions. In B2B scenarios, where transactions occur between businesses, the dynamics of customer churn can vary from those observed in B2C contexts. This implies the necessity of devising analytical models that are specifically tailored to B2B environments [2]. The requirement for anticipating customer churn becomes particularly pronounced when dealing with B2B contexts, characterized by larger purchase amounts and a higher frequency of transactions [14]. While there exists an extensive body of literature explaining methods for predicting customer churn in B2C scenarios [16] there is a noticeable lack of parallel research addressing the same topic within B2B settings.

2. Numerous ML methods have been utilized in academia to predict customer retention and profitability. In many cases, these approaches involve extracting latent characteristics from a customer's past purchase behavior, operating on the assumption that observed behavior is a manifestation of an underlying stochastic process [38]. This particular approach to predicting customer purchases can be referred to as the 'characteristics approach'.

3. Moreover, three major predictor categories have been employed to forecast customer investments, as evidenced in the articles: past customer behavior, observed customer heterogeneity, and variables associated with intermediaries [39]. Research findings indicate that customer past behavior holds significant importance as a feature in these models. Consistent with this, another study highlights the consideration of features like the number of transactions observed in past time frames, the time of the last transaction, and the relative change in a customer's total spending to develop a model for predicting future customer investments [15].

4. Furthermore, similar analytical approaches have been applied across various sectors, with healthcare and the telecom industry being among the most common. The majority of reviewed papers emphasized the use of customer segmentation techniques in model development.

2.3.2. TECHNIQUE ANALYSIS

Articles suggested the use of the Recency, Frequency, Monetary value, Time since first purchase, and Churn probability (RFMTC) in order to predict customers' next investment [37]. Many papers reviewed in this research focus solely on evaluating the predictive performance of different ML techniques. Notably, DT and LR have emerged as highly popular methods in prediction modeling [40]. Comparative analysis against other models has consistently demonstrated the strong predictive capabilities of DT and RF in forecasting customer investment, and these models have also been successfully applied in other domains, such as healthcare and telecommunications.

Furthermore, a notable observation among most articles is the emphasis on the comprehensibility of the analytical models [1]. Ensuring that users of the models can interpret the results is crucial. For instance, when these models are utilized by management in the financial and telecom sectors or by healthcare professionals, comprehensibility becomes even more critical to align with domain knowledge. It has also been seen that a trade-off exists between the comprehensibility and predictive performance of analytical models. Striking a balance between these two factors is vital for optimizing the model's effectiveness in real-world applications.

In addition, it was observed that several analyzed articles not only concentrated on established traditional models but also explored different Hybrid Machine Learning (HML) models [1, 16, 26]. These HML models combine various data-driven techniques and algorithms to address challenges related to comprehensibility and predictive performance. As an illustration, researchers have employed DT and LR to create a hybrid model known as the LLM [1]. The LLM hybrid approach and enhanced complexity have proven advantageous for building prediction models.

EVALUATION METHOD ANALYSIS

The primary objective of this subsection is to present an analysis of the predominant and emerging evaluation methodologies employed by researchers to validate and assess the effectiveness of their models.

In many analytical models that utilize hybrid approaches for predicting customers' next investments, the data is first segmented using various ML techniques like DT due to the heterogeneity of customer data. Following segmentation, the outputs undergo further processing using different supervised and unsupervised ML techniques, such as DT, SVM, and NN. These hybrid models are evaluated based on their predictive performance and comprehensibility.

The most commonly used evaluation metrics in the research include accuracy, precision, recall, specificity, F1 score, FPR, FNR, NPV, AUC and ROC [1, 3, 13, 26]. Some less frequently used metrics, such as FDR, Qini metric [41], and Cohen's Kappa [28], have also appeared in various studies, see Table 2.2. Additionally, researchers have relied on two major validation techniques to assess the predictors: the train/test split and k-fold cross-validation. The value of k in the latter method is typically chosen at the author's discretion based on dataset characteristics and available computational resources.

Furthermore, in some articles, model-agnostic techniques like SHAP have been employed to enhance the model's explainability [26–28]. These techniques aid in understanding the models better. In a few other papers, the evaluation of ML models has been conducted by selecting variables based on their importance using the RF algorithm and Partial Dependence Plots (PDP) [34].

2.4. GAP ANALYSIS

Existing research articles on predicting the investment behavior of B2B and B2C customers consistently highlight a noticeable gap. While B2C investment behavior has been extensively studied, the lack of research on B2B customers underscores the strong need for comprehensive research. This discrepancy highlights the potential importance and unexplored opportunities for predicting B2B investment.

Moreover, an examination of Table 2.2 reveals that the use of HMs within the financial sector, particularly in the context of B2B customers, is relatively unexplored. Furthermore, very few of the studies have integrated the concept of XAI into HMs when addressing B2B customers in the financial sector. This presents a significant opportunity to address three key gaps simultaneously - the inclusion of HMs, the integration of XAI, and the exploration of B2B customers within the financial sector.

Addressing these gaps not only contributes to a more holistic understanding of investment patterns but also has practical implications for companies and financial institutions serving B2B customers. This thesis seeks to address these existing gaps by shedding light on the unique implications of XAI in the context of HMs for predicting customers' investment. This study also aims to improve our understanding of B2B investment behavior, ultimately benefiting the business community and financial analysts who seek to make informed decisions and optimize investment strategies tailored specifically to B2B clients.

The characteristics of B2C differ significantly from those of B2B e.g the nature of the customers, customer size and their transactions. As a consequence, the development of distinct analytical models tailored specifically to the B2B landscape, with a primary focus on the role of XAI within HMs and its profound impact on the financial sector is required.

2.5. CHAPTER SUMMARY

Chapter 2 provides a SLR focusing on the prediction of B2B and B2C customer investment behavior, highlighting the significant research gap in the B2B domain. It outlines the methodology used for literature search and selection, including inclusion and exclusion criteria. The chapter highlights recurring trends in the literature, such as the distribution of journals by year, the distribution of keywords such as prediction, and hybrid, and the prevalence of ML models such as RF regressor and LR. The evaluation methods and metrics used in the studies are also discussed, with a focus on model comprehensibility and HM. The gap analysis highlights the need for further research in the prediction of B2B investments, the use of HM, and the integration of XAI within the financial sector for B2B customers, setting the context for the focus of this thesis.

3

Methodology

This chapter provides a comprehensive exploration of the SEMMA framework, ML models, evaluation methods, and model explainability. Each stage of SEMMA is discussed in detail, highlighting its importance in guiding the data mining process. Various techniques such as RFM models, ML models (LR, DT, RF regressor, XGBoost regressor, and k-means clustering) are used to predict a customer's next investment. In addition, performance metrics such as MAE, RMSE, and $\mathbf{R^2}$ are used to measure the performance of each of the models. Moreover, model explainability is also explored, wherein insights into feature importance and decision rationale have been provided through techniques such as SHAP, which bridge the gap between complex ML models and human understanding.

3.1. SEMMA

SEMMA is a data mining methodology created by the SAS Institute[1]. It facilitates the comprehension, structuring, construction, and ongoing management of data mining endeavors. It plays a pivotal role in delivering resolutions for business challenges and objectives[42]. SEMMA is closely associated with SAS Enterprise Miner, serving as a structured framework for its functional tools[43]. Figure 3.1 provides a visual representation of the SEMMA methodology. A detailed explanation of each stage is provided below.

3.1.1. Sample

The first stage, "Sample" involves selecting a representative subset of the data from the entire dataset. Sampling is essential when dealing with large datasets, as it reduces computational complexity and speeds up the modeling process. Various sampling techniques, such as random sampling, stratified sampling, or oversampling of rare events, can be used depending on the specific data mining task. The goal is to create a smaller dataset that retains the essential

[1] www.sas.com

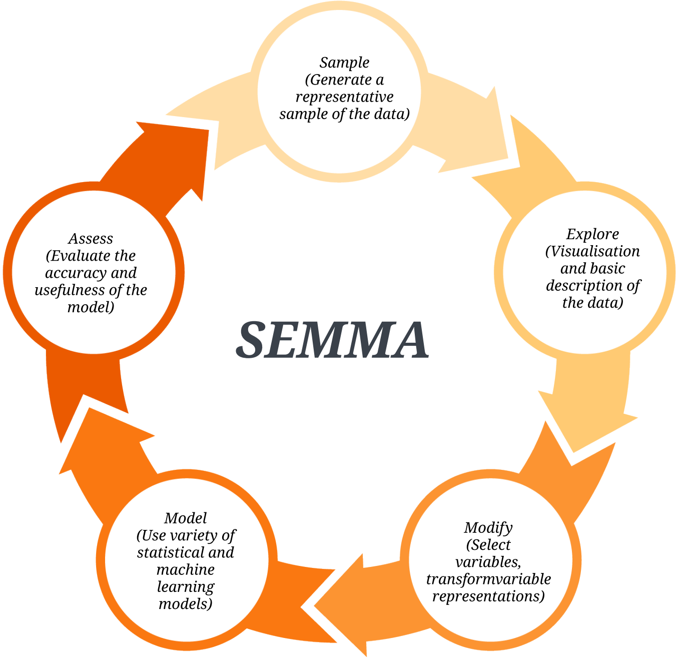

Sample
(Generate a representative sample of the data)
Explore
(Visualisation and basic description of the data)
Modify
(Select variables, transformvariable representations)
Model
(Use variety of statistical and machine learning models)
Assess
(Evaluate the accuracy and usefulness of the model)
SEMMA

Figure 3.1: A visual representation of the SEMMA framework.

characteristics of the original data.

3.1.2. Explore

In the “Explore” stage, the data is thoroughly examined to gain a deeper understanding of its characteristics. EDA techniques are applied to visualize and summarize the data. This includes generating histograms, scatter plots, box plots, and other visualizations to identify patterns, outliers, and relationships among variables. Descriptive statistics, such as mean, median, standard deviation, and correlation coefficients, are computed to quantify key aspects of the data.

This stage involves various EDA techniques and activities:

1. *Data visualization*: Data visualization techniques are used to create graphical representations of the data. This includes various visualization techniques that help in uncovering patterns, trends, and anomalies in the data. Visualization is a powerful tool for identifying relationships between variables, detecting outliers, and understanding the distribution of data.

2. *Summary statistics*: Descriptive statistics, such as mean, median, standard deviation, variance, and percentiles, are computed to summarize key characteristics of the data. These statistics provide a quick overview of central tendencies and dispersion in the dataset, helping data miners understand data distribution.

3. *Data distribution analysis*: Data distributions are examined to assess whether the data follows a normal distribution or exhibits other patterns, such as skewness or kurtosis. Deviations from normality can influence the choice of modeling techniques.

4. *Correlation Analysis*: Data scientists explore correlations and associations between variables using correlation matrices or scatter plots. This analysis helps identify relationships between features, which can be crucial for model development.

5. *Outlier detection*: Outliers, or data points that significantly deviate from the norm, are identified and examined. Outliers can be errors in data or represent important anomalies. Understanding outliers is essential for making informed decisions on whether to retain, transform, or remove them.

6. *Pattern recognition*: Data scientists look for patterns and trends in the data, which may include seasonality, cyclical behavior, or recurring patterns over time. Identifying patterns can guide the choice of appropriate modeling techniques.

7. *Variable identification*: In this phase, data scientists assess the relevance and importance of variables (features) for the data mining task. They identify which variables are likely to have a significant impact on the outcome and which may be less influential.

The “Explore” stage not only helps data scientists understand the data’s underlying structure but also informs subsequent stages of the SEMMA process, such as data modification and model development. It allows data scientists to make informed decisions about data preprocessing, feature engineering, and modeling techniques based on the insights gained during exploration. Additionally, exploring the data often reveals initial hypotheses and patterns that can be further investigated and validated in later stages of the data mining process.

3.1.3. Modify

The “Modify” stage focuses on data preprocessing and cleaning. It aims to prepare the data for modeling by addressing issues such as:

1. *Handling missing data*: Data scientist identifies and address missing data, which can be a common issue in real-world datasets. Strategies for handling missing data include imputation (replacing missing values with estimated values), removal of records with missing values, or using techniques like interpolation.

2. *Dealing with outliers*: Outliers, which are extreme values that deviate significantly from the majority of data points, can impact the performance of models. Data miners decide how to handle outliers, whether by transforming them, removing them, or leaving them unchanged based on the nature of the data and the modeling approach.

3. *Feature engineering*: Feature engineering involves creating new features or transforming existing ones to better represent underlying patterns in the data. This can include encoding categorical variables, scaling or standardizing numerical features, and creating interaction terms. Domain knowledge often plays a critical role in feature engineering, as it helps identify which features are most relevant to the modeling task.

4. *Data reduction*: In cases where the dataset is large or has high dimensionality, data min-

ers may employ dimensionality reduction techniques such as Principal Component Analysis (PCA) or feature selection methods to reduce the number of variables while retaining meaningful information.

5. *Data transformation*: Data transformation techniques like logarithmic transformations, Box-Cox transformations, or normalization may be applied to make the data conform to assumptions required by certain modeling algorithms. These transformations can help improve the model's performance and interpretability
6. *Binning and discretization*: Data binning involves grouping continuous variables into discrete bins or categories. This can simplify the modeling process, especially when dealing with non-linear relationships. Binned variables are treated as categorical features in subsequent modeling stages.
7. *Handling imbalanced data*: In cases where the dataset has imbalanced class distributions (e.g., rare events or minority classes), data miners may apply techniques like oversampling, undersampling, or using different evaluation metrics to address the imbalance and prevent model bias
8. *Data integration*: If the dataset is sourced from multiple origins or systems, data integration is performed to combine and unify the data into a single coherent dataset for modeling.
9. *Data scaling and normalization*: Scaling and normalization techniques are applied to ensure that features have similar scales, preventing certain variables from dominating the modeling process. Common methods include min-max scaling and Z-score normalization.
10. *Data splitting*: The dataset is typically split into training, validation, and test sets to evaluate model performance. Data miners decide on the appropriate split ratios to ensure reliable model evaluation.

3.1.4. Model

The "Model" stage is where predictive models are built using the pre-processed data. Various ML and statistical modeling techniques are applied to create models that can make predictions or classifications. Model selection is an important consideration, and different algorithms may be tested to determine the most suitable one for the task at hand. This stage involves training and fine-tuning models to achieve the best possible performance. The output of this stage is one or more predictive models ready for evaluation.

3.1.5. Assess

The final stage, "Assess" focuses on evaluating the performance of the predictive models generated in the previous stage. Evaluation metrics such as accuracy, precision, F1-score, ROC curves, AUC, RMSE, MAE and $\mathbf{R^2}$ are used to assess how well the models perform on new, unseen data. Cross-validation techniques are often employed to ensure that the model's performance is robust and not overfitting to the training data. Model assessment helps in selecting the best-performing model(s) and provides insights into their strengths and weaknesses.

SEMMA is an iterative process, meaning that after assessing model performance, you may need to go back to earlier stages (e.g., modifying the data or trying different modeling techniques) to improve results. It provides a systematic and flexible approach to data mining, particularly suitable for tasks involving predictive modeling and data exploration.

3.2. RFM MODEL

The RFM model is a customer segmentation and analysis technique used in marketing and Customer Relationship Management (CRM)[44]. This technique quantitatively assesses and categorizes customers based on three critical factors: recency, frequency, and monetary value of their recent transactions [45]. The primary objective is to identify and prioritize the most valuable customers, enabling businesses to tailor targeted marketing campaigns for optimal results. In this method, each customer receives numerical scores based on their transaction history, facilitating an objective and systematic analysis.

RFM analysis ranks each customer on the following factors [46]:

1. *Recency*: This factor assesses how recently a customer made their last purchase. The underlying idea is that customers who have recently interacted with a product or service are more likely to make repeat purchases or engage with the product again. The measurement of recency can vary depending on the nature of the product, spanning from days to weeks, months, or even hours.

2. *Frequency*: Frequency evaluates how often a customer makes purchases within a specific timeframe. Customers who have made multiple purchases are considered more valuable, as they demonstrate a higher level of engagement. First-time customers, in particular, may be targeted for follow-up marketing efforts to encourage repeat business.

3. *Monetary*: The monetary factor gauges the total amount of money a customer has spent during a given period. Customers who consistently spend more are likely to continue doing so in the future, and they typically hold significant value for a business.

By systematically examining these three aspects of customer behavior, RFM attributes can be provided to any ML algorithm, enabling the algorithm to utilize these features and segment customer data based on these attributes.

3.3. MACHINE LEARNING ALGORITHMS

This section offers a comprehensive overview of the ML algorithms that have been strategically employed in the research to effectively address the research questions. The choice of these specific ML algorithms was informed by a comprehensive analysis conducted through the SLR, as depicted in Figure 2.4. These ML methods offer unique strategies for modeling and predicting customers' future investment actions.

3.3.1. LINEAR REGRESSION

LR is a fundamental statistical and ML technique used for modeling the relationship between a dependent variable and one or more independent variables [47]. The principal idea of a LR model is to establish a linear relationship between a dependent variable and one or more independent variables, allowing us to make predictions or infer the impact of changes in the independent variables on the dependent variable [48].
The general form of a LR equation can be expressed as follows:

$$y = \beta_0 + \beta_1 x + \epsilon, \tag{3.1}$$

where

- y is the predicted value of the dependent variable (y) for any given value of the independent variable (x).
- β_0 is the intercept, the predicted value of y when the x is 0.
- β_1 is the regression coefficient – how much y is expected to change under a change in x.
- x is the independent variable i.e., the expected variable that influences y.
- ϵ is the error of the estimate or the variation in the estimate of the regression coefficient.

In the context of forecasting the number of days a customer is inclined to invest, the application of a LR model proves to be a valuable and versatile approach. LR allows us to examine the relationship between various predictor variables, such as customer historical investment patterns, or other relevant attributes, and the target variable—namely, the number of days until an investment is made. By leveraging this model, one can quantify the impact of each predictor on the investment timeframe, providing actionable insights for decision-makers. A significant advantage of using LR is its interpretability; it helps elucidate the relative importance of different factors in influencing the investment timeline.

3.3.2. DECISION TREE REGRESSOR

DT regression is a tree-like structure used to predict the numerical outcomes of a dependent variable. It is sometimes referred to as the M5P algorithm, which is an adaptation of Quinlan's M5 algorithm [49]. M5P are tree-based structures and have trees incorporated in multivariate linear models [50]. The working of DT regressor model is explained below:

The process begins by constructing a tree using a standard DT algorithm. This involves selecting the attribute that leads to the greatest expected reduction in error as the root node, based on a splitting criterion aimed at minimizing the variance within the subsets of data created by each split [51]. The variation is assessed using a standard deviation reduction metric. After the initial tree construction, it undergoes pruning, which simplifies the model by trimming branches from the leaves. To address the abrupt changes that can result from the pruned tree's linear models, a smoothing technique is applied.

The choice of DT regression for the study is informed by its ability to predict numerical outcomes, unlike conventional DT that typically predict categories. Moreover, DT regression is capable of handling high-dimensional datasets effectively. Pseudocode for DT regression:

- Start with a single node.
- For each X, find the fitness function value (S) and choose the split that offers the minimum value of the fitness function.
- In each new node, go back to step 2. If a stopping criterion is reached, exit.

3.3.3. RANDOM FOREST REGRESSOR

The RF regressor is an ensemble ML algorithm used for regression tasks. It operates by constructing a multitude of DTs during the training phase. Each decision tree is grown using a random subset of the training data and a random subset of the available features [52]. These trees work together as a forest, and during prediction, each tree in the forest independently provides an output [53]. In regression tasks, the final prediction is often the average (or sometimes a weighted average) of the individual tree predictions, resulting in an ensemble prediction that tends to be more robust and less prone to over-fitting compared to individual DT.

$$\text{Final Prediction} = \frac{1}{N_{\text{trees}}} \sum_{i=1}^{N_{\text{trees}}} \text{Prediction}_i. \tag{3.2}$$

Here Prediction_i is the prediction from the i-th tree in the forest, and N_{trees} is the total number of trees. This ensemble averaging helps reduce over-fitting and improves the model's ability to generalize to new data. Additionally, the RF uses an out-of-bag (OOB) error estimate [52], which quantifies the prediction error on data points not used in the construction of each tree, aiding in model evaluation. RF regressor leverages the concept of bagging (Bootstrap Aggregating) and random feature selection to reduce variance and improve predictive accuracy, making it a powerful algorithm for handling complex regression problems while maintaining generalization capabilities.

The RF regressor offers numerous advantages. It excels in predictive accuracy by combining multiple DTs, mitigating over-fitting, and accommodating non-linear relationships in data. It ranks feature importance, handles various data types, and provides an OOB error estimate for internal validation. Additionally, it's efficient for large datasets, produces stable results, and can handle missing values. These qualities, along with reduced bias and inherent feature engineering capabilities, make it a versatile and robust choice for regression tasks in diverse domains.

3.3.4. XGBOOST REGRESSOR

Chen and Guestrin [54] presented XGBoost as a novel approach to predicting outcomes based on specific variables. The core concept behind this algorithm is the sequential construction of D-Classification and Regression Trees (CARTs) [55]. Each successive tree is trained on the residuals from the previous one, meaning that each new model refines and addresses the inaccuracies of the previous tree to make its prediction. XGBoost is based on gradient boosting architecture [56], which uses various complement functions to estimate the results using the following equation [57],

$$\overline{y}_i = y_i^0 + \eta \sum_{k=1}^{N} f_k(U_i), \tag{3.3}$$

where $\overline{y}_i$ is the predicted output for i^{th} data, y_i^0 is an initial hypothesis, U_i is the parameter vector and N is the number of estimators associated with independent tree structures corresponding to f_k.

When building XGBoost models, selecting the right hyper-parameters is crucial for establishing accurate correlations. Key parameters considered in this research were [58]:

- max_depth: This determines the maximum depth of the base tree. A higher value indicates a more complex base tree.
- n_estimators: This represents the count of base tree models. A higher number suggests more iterations.
- min_child_weight: This is the least combined weight of child nodes. A higher number leads to more restrained models.
- gamma: This is the minimum loss reduction needed to further split a tree's leaf node. A higher value results in more cautious models.
- subsample: This is the proportion of training samples used.
- colsample_by tree: This is the fraction of columns used when creating new trees.
- reg_lambda: This is the L2 regularization on weights. A higher value makes the model more conservative.

3.3.5. k-MEANS CLUSTERING

k-means clustering is a fundamental unsupervised learning algorithm employed to address clustering problems effectively. This method helps in categorizing a given dataset into a specified number of clusters or groups. It operates by iteratively assigning data points to clusters in such a way that the variance within each cluster is minimized. This optimization process continues until a convergence criterion is met, resulting in well-defined clusters. A detailed flowchart of the k-means algorithm process is illustrated in Figure 3.2.

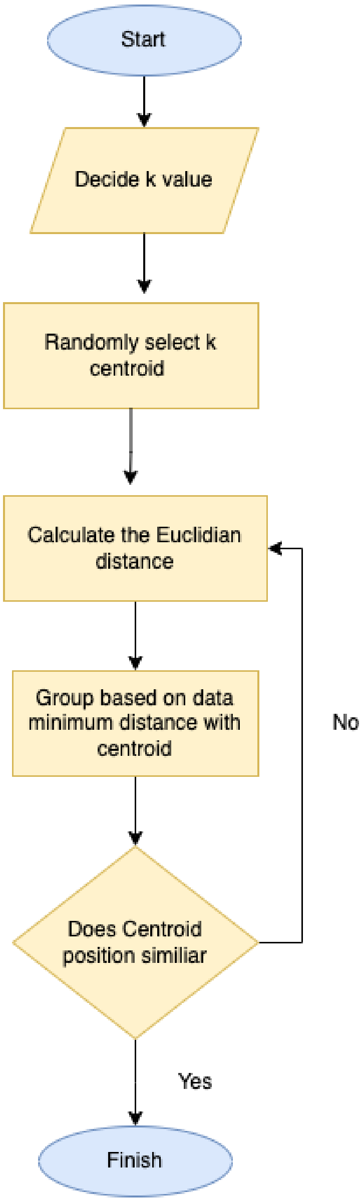

Start
Decide k value
Randomly select k centroid
Calculate the Euclidian distance
Group based on data minimum distance with centroid
Does Centroid position similiar
No
Yes
Finish

Figure 3.2: k-means algorithm flowchart.

There are different methods used to specify the number of clusters on a set of data. One of the most common methods is called the Elbow method. The elbow method is used to produce the best number of clusters by looking at the percentage of the comparison between the number of clusters that will form an elbow at a point [59]. The elbow method examines the percentage of variance that can be explained as a function of the number of clusters. This approach depends on the idea that one should select a sufficient number of clusters such that the data modeling is not substantially enhanced by the addition of another cluster. The percentage of variance explained by the clusters is plotted against the number of clusters. The first clusters will provide a lot of information, but eventually, the marginal gain will decline sharply, giving the graph an angle [60]. The "elbow criterion" refers to the process of selecting the appropriate k, or number of clusters.

The steps used by the Elbow method algorithm to determine the k value in k-means are listed below [60].

1. Initialize the initial value of k;
2. Increase the value of k;
3. Figuring out the results of each value of k's sum of square errors;
4. Analysis of the sum of square error caused by the sharply declining k value;
5. Find the elbow-shaped k value and set it.

K-means clustering can be applied to segment the bank customer data into distinct groups or clusters based on similarities in their RFM profiles. The application of this segmentation technique can enable the formation of significant customer clusters, which might provide a more focused and data-informed insight into customer habits and inclinations [61].

3.4. EVALUATION METHODS

When evaluating the predictive performance of different models used to predict a continuous variable (often referred to as regression models), several evaluation metrics can be used to quantify how well the model's predictions align with the actual values. Here are some common metrics for assessing the performance of regression models [57, 62–64]:

1. MAE measures the average absolute difference between the predicted values and the actual values. It is calculated as:

$$\mathrm{MAE} = \frac{1}{2}\sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{3.4}$$

   where,

   - y_i represents the actual values.
   - $\hat{y}_i$ represents the predicted value.
   - n is the number of data points.

2. MSE measures the average squared difference between the predicted values and the actual values. It is calculated as:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{3.5}$$

3. RMSE is the square root of MSE and provides a measure of the average magnitude of errors in the same units as the target variable:

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} \tag{3.6}$$

4. **R^2** measures the proportion of the variance in the dependent variable that is predictable from the independent variables. It ranges from 0 to 1, with higher values indicating a better fit:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \overline{y}_i)^2} \tag{3.7}$$

   - $\overline{y}_i$ is the mean of actual values.

3.5. EXPLAINABILITY

ML models have transformed decision-making in diverse fields but their inbuilt complexity makes them difficult to understand and trust. One of the reasons is that, in order to achieve high accuracy, ML problems often requires the estimation of a large number of parameters [65]. To tackle this challenge, interpretable techniques such as SHAP have emerged as essential methods for enhancing model explainability. This method provides a bridge between the complex, black-box nature of ML models and the human need for comprehensible explanations of their decisions.

3.5.1. SHAP (SHAPLEY ADDITIVE EXPLANATIONS)

SHAP originated from game theory and has been adapted for the purpose of assessing the significance of each feature that is influencing predictive outcomes. The fundamental idea is that every feature collaboratively impacts the model's prediction, leading it to change in one direction or another [66]. SHAP works to evenly divide these contributions among all possible feature combinations. More particularly, it achieves this equitable distribution through the Shapley value approach, which evenly distributes the difference between the predicted outcome and the average prediction among the feature values of the instance being analyzed.
Shapley values can be used in ML to quantify the contribution of each feature in the model that collectively delivers the prediction [67]. The Shapley value for feature X_j in a model is given by:

$$\text{Shapely}(X_j) = \sum_{(S\subseteq N)\setminus\{j\}} \frac{k!(p-k-1)!}{p!}\big(f(S\cup\{j\}) - f(S)\big) \tag{3.8}$$

where p is the total number of features, $N\setminus(\{j\,j\})$ is a set of all possible combinations of features excluding X_j, S is a feature set in $N\{j\}$, $f(S)$ is the model prediction with features in S, and $f(S\cup\{j\})$ is the model prediction with features in S plus feature X_j. The interpretation of Eq. (3.8) is that the Shapley value of a feature is its marginal contribution to model prediction averaged over all possible models with different combinations of features [65]. There are several types of plots that can be created using SHAP to gain insights into the ML models. Here are some common types of plots:

1. *Summary Plot*: A summary plot provides an overview of feature importance for a specific model. One can determine which features have the biggest impacts on predictions

by examining the Shapley values across all data points. The plot is often displayed as a horizontal bar chart.

2. *Dependence Plot*: A dependence plot illustrates the relationship between a given feature's value and the model's result. It aids in gaining an understanding of the nature and direction of the correlation between a single feature and predictions. Typically, scatter plots used to represent dependence plots have feature values on the x-axis and SHAP values on the y-axis.

3. *Waterfall Plot*: A waterfall plot is similar to a forced plot but it is represented in a waterfall-like format. It illustrates how each feature contributes to the final forecast for a single occurrence, visually breaking down the prediction.

4. *Summary of Importance Plot*: This plot combines a dependence plot with a summary plot. On the left, it displays the importance of the features, and on the right, it displays how the feature values influence the predictions. It gives a thorough overview of the significance of both local and global features.

4

Experimental Set-up

4.1. Exploratory Data Analysis (EDA)

Data exploration is a vital phase in the data mining process, allowing one to gain a deeper understanding of the dataset by uncovering valuable patterns and insights. In this study, real-time data has been obtained from a European B2B banking company that specializes in offering renewable leases to customers globally. This dataset contains information related to these leases and customer data, all stored within the company's database. The data from 2000 to 2023 has been imported for analysis, with a specific focus on the sales department's dataset.

The dataset comprises of two distinct sets of data: one containing customer information, and the other providing contract details. Meaningful connections between these sets are established by utilizing a unique identifier known as the customer ID. The dataset itself consists of 25,260 observations, each with eight features.

Table 4.1, shows all the eight features available in the customer dataset. The customer ID serves as a unique identifier representing each customer's name. The "sector" field indicates the segment to which the customer belongs, such as manufacturing or agriculture. Meanwhile, the "contract ID" is another unique identifier specific to each investment. "start_date" and "end_date" define the contract's beginning and ending dates, based on the lease period. The "money" field denotes the amount invested by the customer for each contract. Lastly, the "asset ID" identifies the asset in which the customer has invested. Additionally, investments are categorized into two types: "Pool X" and "Pool Y". Pool X covers investments up to €50,000, while Pool Y encompasses investments up to €1 million.

4.1.1. Data Preprocessing

The initial phase of EDA revolves around data pre-processing, an important step in the EDA process. This phase involves several key tasks designed to prepare the data for analysis. First,

Table 4.1: Features of the customer Dataset.

Features	**Description**
customer_id	The customer identifier.
sector	Define customers based on their primary economic activity.
contract_ID	The contract identifier.
start_date	Start date of the contract.
end_date	End date of the contract
money	Amount invested per contract.
asset_ID	Information about asset
Type	Range of the amount

a filter is applied to the dataset to include records from 2010 to 2023. This filtering step allows the data to be focused on a specific time period, allowing for more meaningful analysis. The dataset is then carefully checked for missing values and their overall percentage of the total dataset size is calculated. The purpose of this thorough examination is to identify any gaps or discrepancies in the data, ultimately ensuring the quality and reliability of our dataset. During the pre-processing stage, it is found that the percentage of missing values is minimal, less than 3%. To be precise, this corresponds to approximately 101 lines out of a total of 25,260 lines. It is therefore decided to eliminate these rows with missing information from our dataset. This ensures that the data we are dealing with is not only complete but also reliable, providing a robust basis for further analysis.

Second, all the categorial data are transformed into numerical features using technique such as label encoder [47]. This technique creates v–1 dummy variables, where v equals the number of distinct values of the categorical variable.

When visualizing the features, outliers are identified, and their impact on the dataset is analyzed. These outliers are often responsible for abrupt and extreme spikes in the graphs, which can adversely affect the overall data presentation. Figure 4.1 displays the outliers in the money invested per contract over the year. These outliers were identified as investments exceeding €1 billion. Consequently, the dataset should only include customers who have invested less than €1 million, and as a result, these outliers were eliminated. This action is taken to ensure the dataset's representativeness and to minimize the bias introduced by extreme values. Similar outliers were observed when visualizing other features, such as the contract duration and Recency as reflected in Figure A.2. However, it is important to note that not all outliers can be removed, as some of them also signify certain trends in the data.

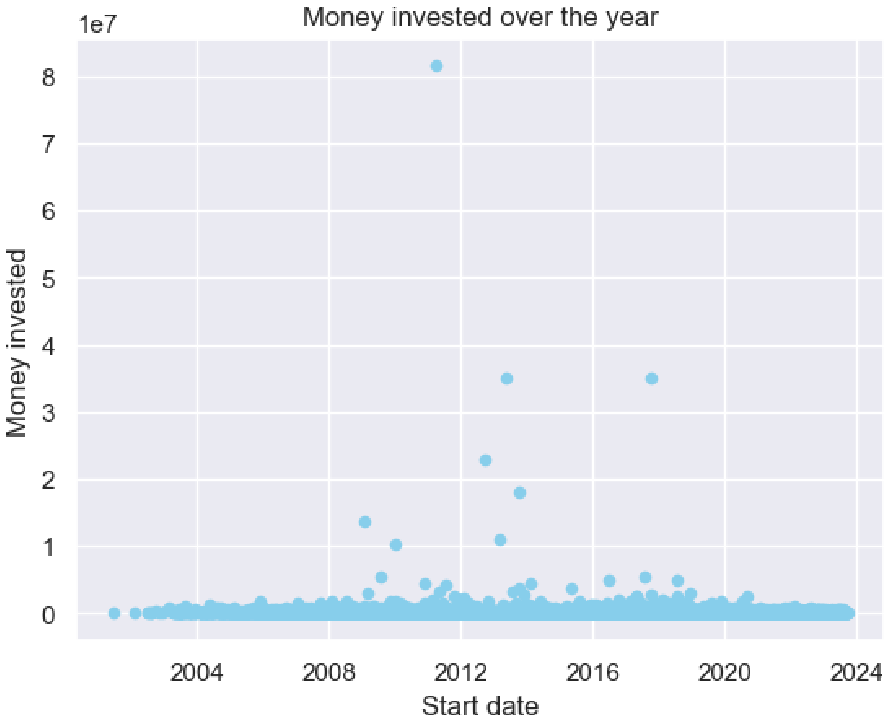


Figure 4.1: Outliners in the total money invested over time.

In addition to outlier removal, the skewness of selected features was evaluated to ensure that the assumptions underlying the modeling techniques were met, recognizing that skewed data could affect model accuracy and interpretation. Skewness was addressed by applying mathematical transformations, such as logarithmic transformation, to make the distribution of skewed features more similar to a normal distribution, thereby improving model performance.

4.1.2. Feature engineering

The next phase of data analysis involves an essential step known as feature engineering. This process is fundamental to improving the dataset for use in our model. In feature engineering, new variables or features are created from the existing dataset that significantly contributes to the performance and accuracy of the model.

Table 4.2 displays the new features that have been generated using the old given features. These newly created features include several variables, such as the no of contracts held by each client, the total money investment for each client, the starting month of their investment journey, the categorization of investments into quarters, the application of an Exponential Moving Average (EMA) for trend analysis, and an assessment of investment frequency. The target variable "days_till_next_investment_in_days" represents the number of days remaining until a customer makes their next investment, essentially forecasting the time interval before their subsequent financial commitment.

Business knowledge has been utilized to calculate the no of contracts and total money invested.

Table 4.2: Features generated after feature engineering.

Features	**Description**
no of contracts	The total no of contracts the customer has over time.
total Money	Total money invested by each customer over time.
start_month	The start month of the contract.
start_quarter	The start quarter of the contract
duration of contract	The contract duration.
investment_frequency	The average days of the last investment
EMA	Exponential moving average
Recency	The customer last contacted days
days_till_next_investment_in_days	The number of days till next investment (Target variable)

It is assumed that when investments are made on a single day or when the interval between investment dates is less than six months, they should be treated as a single investment.

Throughout the feature engineering process, critical attention has been given to preventing data leakage and avoiding the unintentional inclusion of inaccessible data during the prediction process. This precaution ensures the integrity of the model, enabling optimal performance with the current dataset, while also maintaining its reliability when used with novel, unseen data.

The dataset is further analyzed using the newly generated nine new features mentioned in Table 4.2. Using domain-specific knowledge, the data is segmented into four distinct clusters. The customer data is partitioned based on the average time to subsequent investment. This approach results in the creation of four separate clusters. The clustered data is then visualized in terms of the target variable and other customer attributes. Figure 4.2 shows different clusters formed by using domain knowledge. In addition, to gain insight into the distribution of features and their association with the target variable, a number of plots are generated, including scatter plots, box plots, and violin plots.

For example, as shown in Figure 4.5, the scatter plots show the association between the features and the target variable. These visual representations imply that there is no clear linear correlation between “days_till_next_investment_in_days” and the respective characteristics. This suggests that the relationship between “days_until_next_investment_in_days” and these features is non-linear. More such plots highlighting different trends in the data are shown and briefly discussed in Appendix A.3.

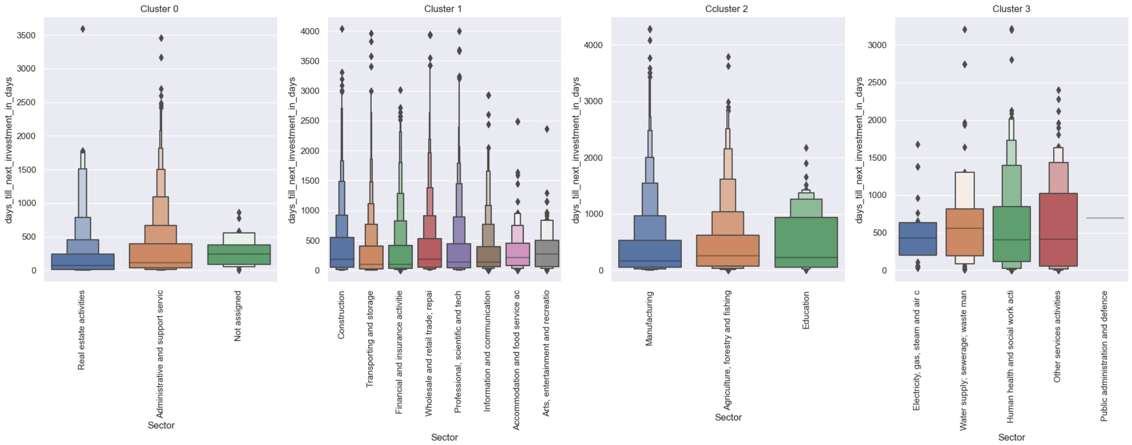


Figure 4.2: Different clusters are formed using HM using domain knowledge.

In addition, feature correlation and mutual information regression analyses are performed on the dataset to assess the degree of correlation between the features and the target variable. By examining feature correlation, insights into the interrelationships among different variables are gained. A strong correlation between two variables indicates that a variation in one variable might correspond with a variation in the other. The correlation heatmap, depicted in Figure 4.3, offers a visual representation of these correlation coefficients. The color scale on the right of the heatmap signifies the strength and direction of the correlation: red represents a positive correlation, blue signifies a negative correlation, and the depth of the color demonstrates the correlation's magnitude. Colors closer to white suggest minimal or no correlation.

Within this analysis, the “days_till_next_investment_in_days” investment feature is designated as the target variable. This target variable exhibits varied correlation magnitudes with other features. For instance, a correlation of -0.16 with “no of contracts" indicates a weak relationship. Similarly, features such as “duration of contract”, “difference_in_days_before_last_investment”, “Investment_frequency”, and “Recency” are found to have weak correlations with the target variable.

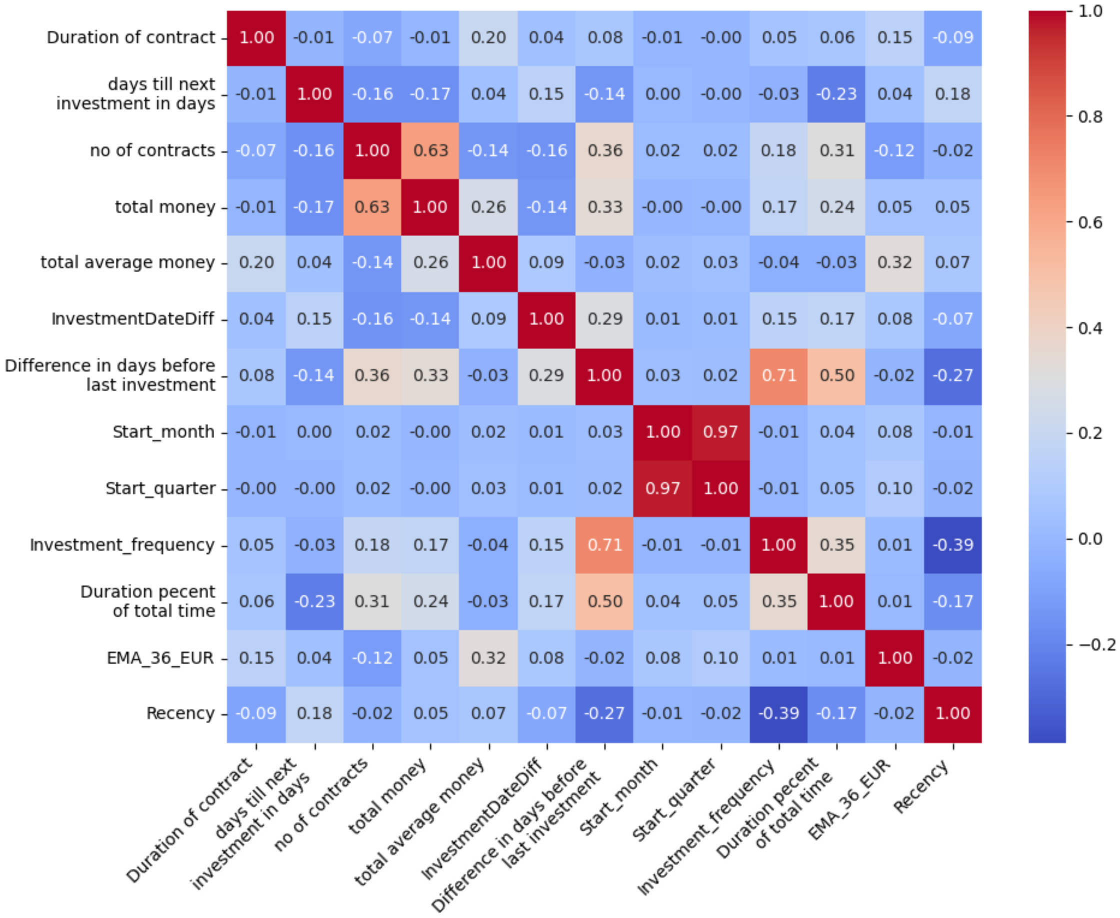


Figure 4.3: Correlation heatmap.

The relationships between various variables are efficiently visualized and understood using the correlation matrix. However, it's emphasized that correlation is not necessarily indicative of causation, highlighting the need for more comprehensive analyses to achieve clear conclusions. As a result, mutual information regression is applied to the dataset. By this method, the information one variable imparts in predicting another is measured, revealing their interdependencies. In Figure 4.4, MI scores are assigned to each feature. Employing such an analysis aids in determining the significance of each feature concerning the target variable and exposes inherent relationships within the dataset. The MI score for feature "investment_frequency", as depicted in Figure 4.4, has a value of 1.2, indicating that despite its low correlation with the target variable as shown in Figure 4.3, it holds significant predictive power. Therefore, the utilization of mutual information regression highlights features that are valuable for training despite not being evidently correlated, aiding in the selection of influential factors for model development.

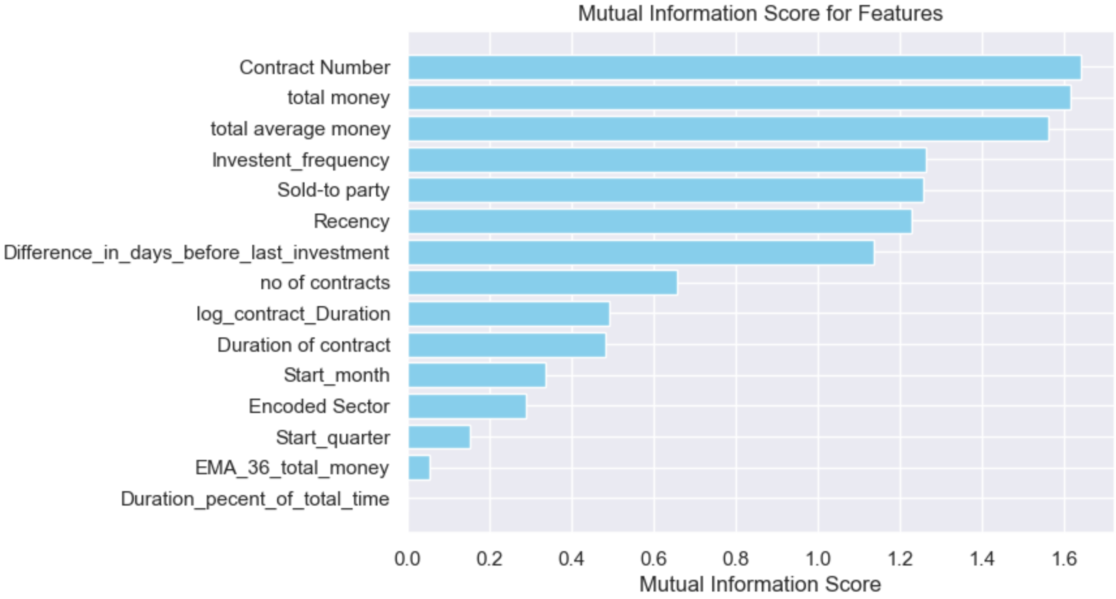


Figure 4.4: MI score for different features.

4.1.3. Data modelling

The ‘Model’ component of the SEMMA framework is effectively demonstrated through the development and evaluation of regression models that are designed to predict the duration of customer investments. Emphasis is placed on the selection of models that are capable of capturing data trends with high accuracy. Two different approaches are employed for data modeling.

An extensive range of regression models has been developed and thoroughly tested to accurately predict the number of days a customer will invest. The primary objective is to select models capable of capturing data trends with a high degree of accuracy. Two different approaches have been used for data modeling.

Baseline Models

Several ML models have been extensively utilized for various predictions in the research articles studied during the SLR, see Figure 2.4 and Table 2.2. Out of these, four models are selected due to their simplicity and ease of explanation, namely, LR, DT, RF, and XGBoost.

The first model considered is LR, recognized for its straightforwardness. It is less computationally demanding compared to other algorithms, making it advantageous in the preliminary phases of analysis. An assumption is made that the relationship between predictors and the target is linear, rendering LR an appropriate choice. Nonetheless, findings from correlation assessments and visualizations, such as the scatter plot illustrated in Figure 4.5, reveal intricacy and non-linearity in the dataset.

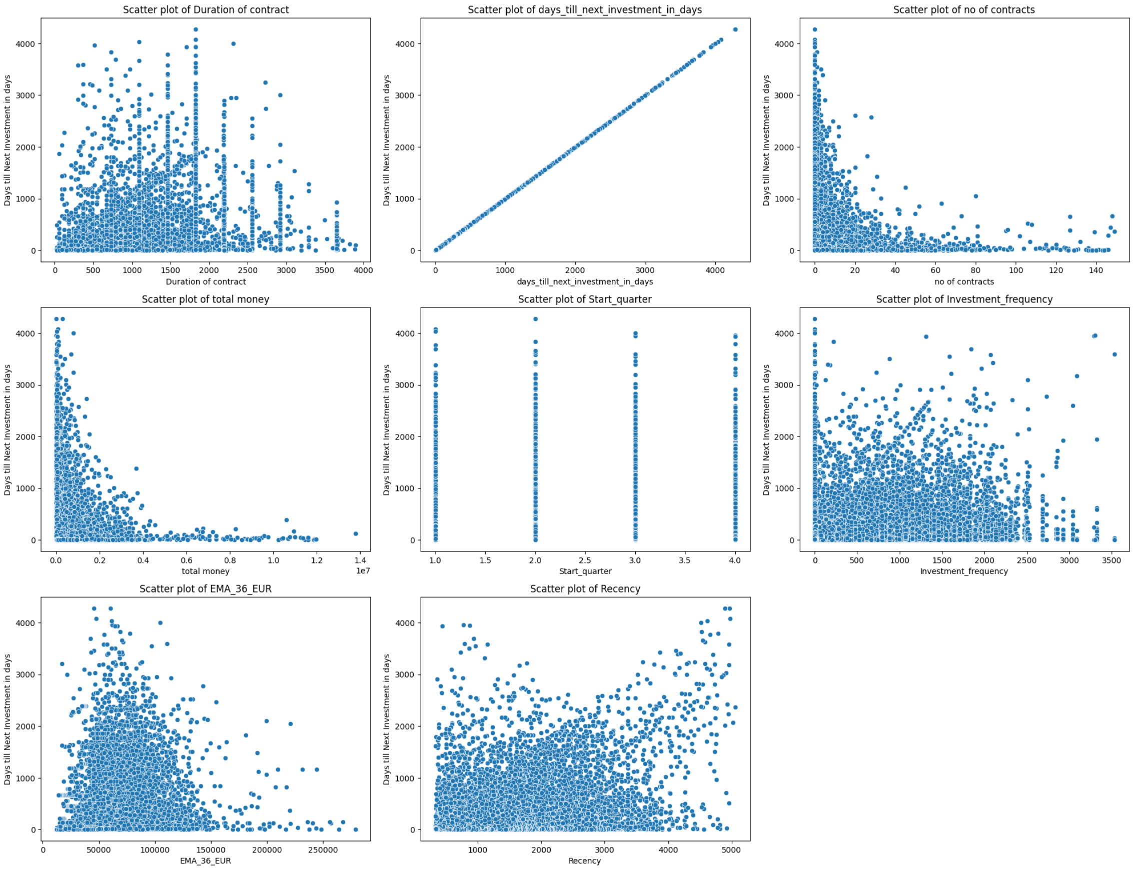


Figure 4.5: Scatter plot illustrating the relationships between features and the target variable.

The second model under consideration is DT, which is valuable for its interpretability and simplicity, making it a good starting point for understanding data patterns. However, DTs have a tendency to overfit data, especially when dealing with complex or noisy datasets, which can result in poor generalization of new data. This limitation often necessitates the use of techniques like pruning, setting minimum samples for node splitting, or combining multiple trees into an ensemble method, like RF or XGBoost to achieve more reliable predictions.

The third model used is RF regressor, esteemed for its proficiency in grasping complex data structures without relying on linear assumptions. Yet, a tendency for overfitting is observed, where an outstanding performance on training data is accompanied by poor performance on the unseen data. This suggests that noise and anomalies are mistakenly identified as patterns, leading to the introduction of a penalty for the magnitude of coefficients.

To address the overfitting issue, an advanced gradient boosting algorithm such as XGBoost regressor is implemented. The success of these models is attributed to several pivotal aspects. Notably, a form of regularization is incorporated, proficiently reducing the over-complexity often seen with RF regressor and enhancing the model's adaptability to unfamiliar data. Their scalability and consistent high performance make them ideal for managing vast datasets with intricate structures, reminiscent of those found in customer investment predictions.

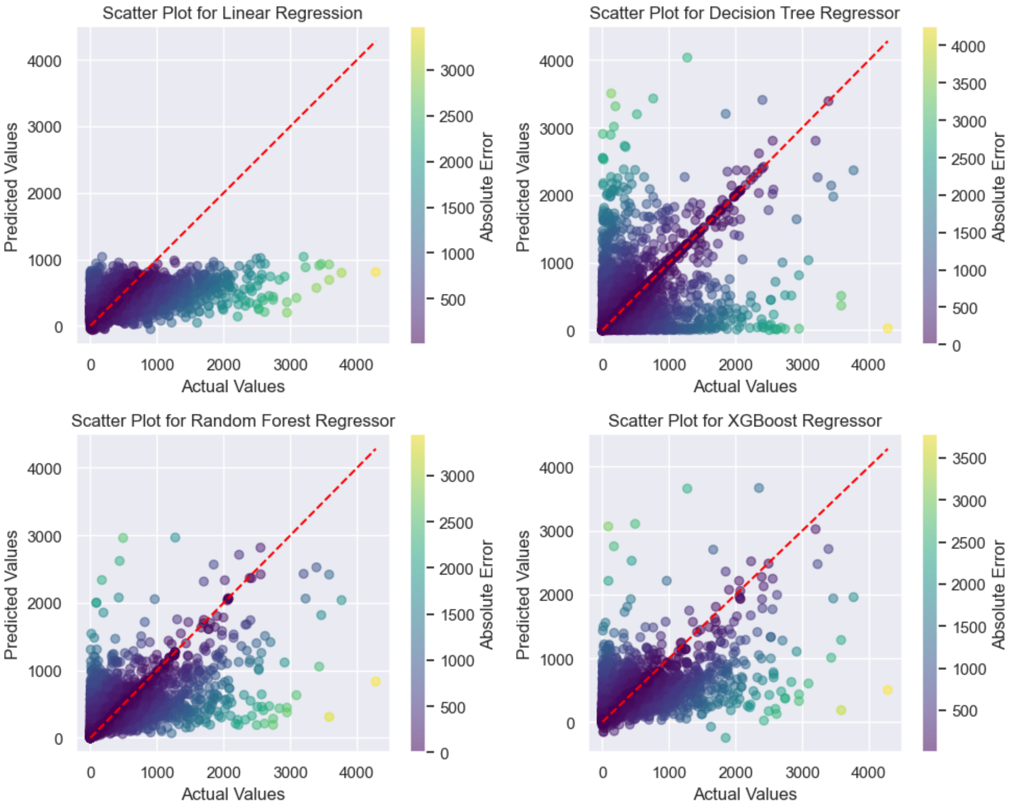


Figure 4.6: Scatter plot showing actual value vs predicted values for LR, DT, RF and XGBoost models.

Figure 4.6 shows a comparative analysis of the four distinct regression models—LR, RF regressor, DT, and XGBoost regressor—in predicting a customer's next investment. Each scatter plot shows the actual investment values against the predicted ones, with a red dashed line serving as the benchmark for perfect prediction. The accompanying color bar signifies the absolute prediction error, with cooler hues denoting accurate predictions and warmer shades indicating larger deviations. The XGBoost regressor demonstrates the closest alignment with the ideal prediction line, especially for lower investment values, suggesting superior accuracy. In contrast, both LR and DT display pronounced scattering for higher investment values, hinting at potential prediction challenges with more substantial amounts. The RF regressor, scatter plot shows a dense clustering of predictions along the entire range of actual values, suggesting a balanced prediction capability across different investment amounts. These visualizations clearly illustrate the strengths and weaknesses of each model, underscoring the critical nature of model selection based on performance indicators. Table 5.1 presents the results of the four baseline models in terms of MAE, RMSE, and $\mathbf{R^2}$.

Hybrid Models

Another approach for predicting a customer's next investment is done by using HMs. By integrating two or more distinct modeling techniques or algorithms, these models are designed to capitalize on the strengths and minimize the limitations of each individual model. This amalgamation frequently results in improved prediction accuracy, reliability, and robustness compared to the use of a single modeling approach. Several strategies are available for crafting HMs. In the analysis, two specific approaches are utilized: the first is informed by domain knowledge for the formation of clusters, while the second employs the k-means clustering technique.

Domain-knowledge clusters

Clusters created using domain knowledge are identified by analyzing the visualizations generated from a boxplot plotting the target variable against the customer's industry. In Figure 4.2, the box plot is shown with sectors on the x-axis and the target variable on the y-axis. Based on average investment days, clusters are created and customers with similar average investment duration are grouped into a single cluster. This methodology results in the formation of four distinct clusters, as shown in Figure 4.7. Basically, clients with similar investment duration are grouped into uniform clusters based on the average number of days invested.

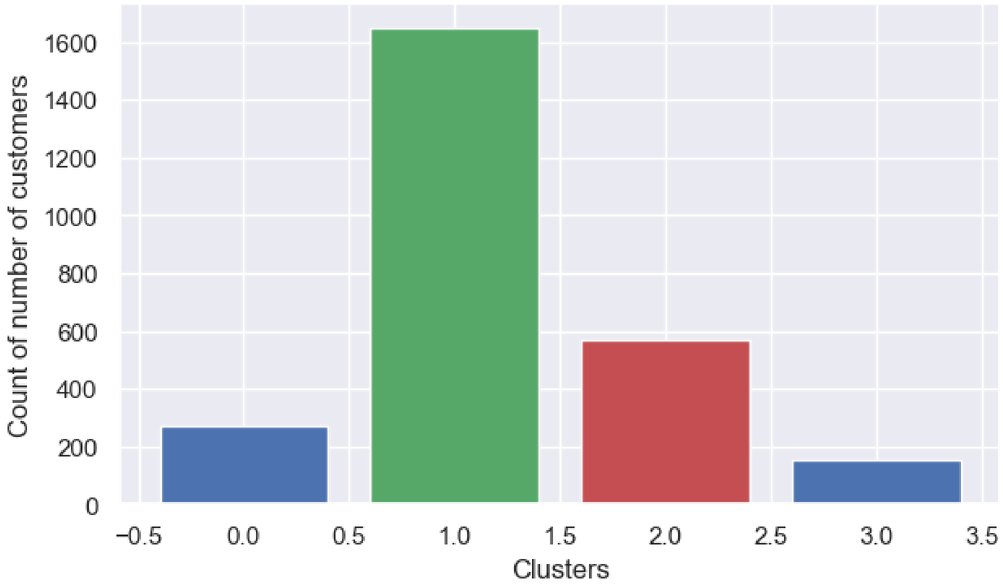


Figure 4.7: Customer segmentation using business knowledge.

k-means clustering

An alternative clustering method uses ML models, called k-means clustering. This technique used features such as the "no of contracts", ”total money" and "Recency" as inputs to the k-means clustering algorithm. This approach produced three distinct clusters and the optimal number of clusters was determined using the elbow method with the inertia function as shown in Figure 4.9. Figure 4.8 illustrates various clusters created through k-means clustering. Prior to implementing k-means clustering, the importance of standardization or normalization of the features was recognized to ensure equal contribution of each feature to the clustering process. The skewness of the data was assessed, and in cases where skewness was present, a method called 'Box-Cox transformation' was used to achieve a more normalized data distribution.

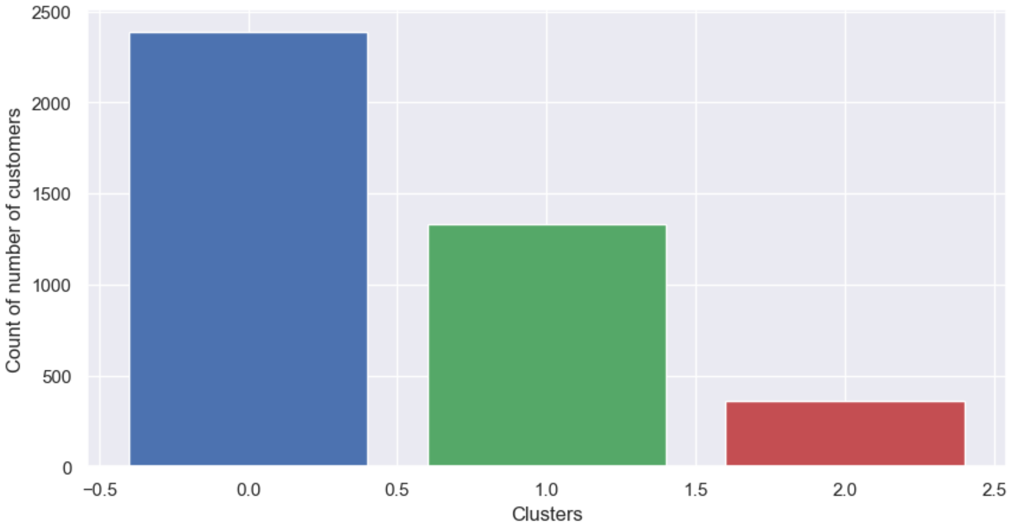


Figure 4.8: Customer segmentation using k-means clustering.

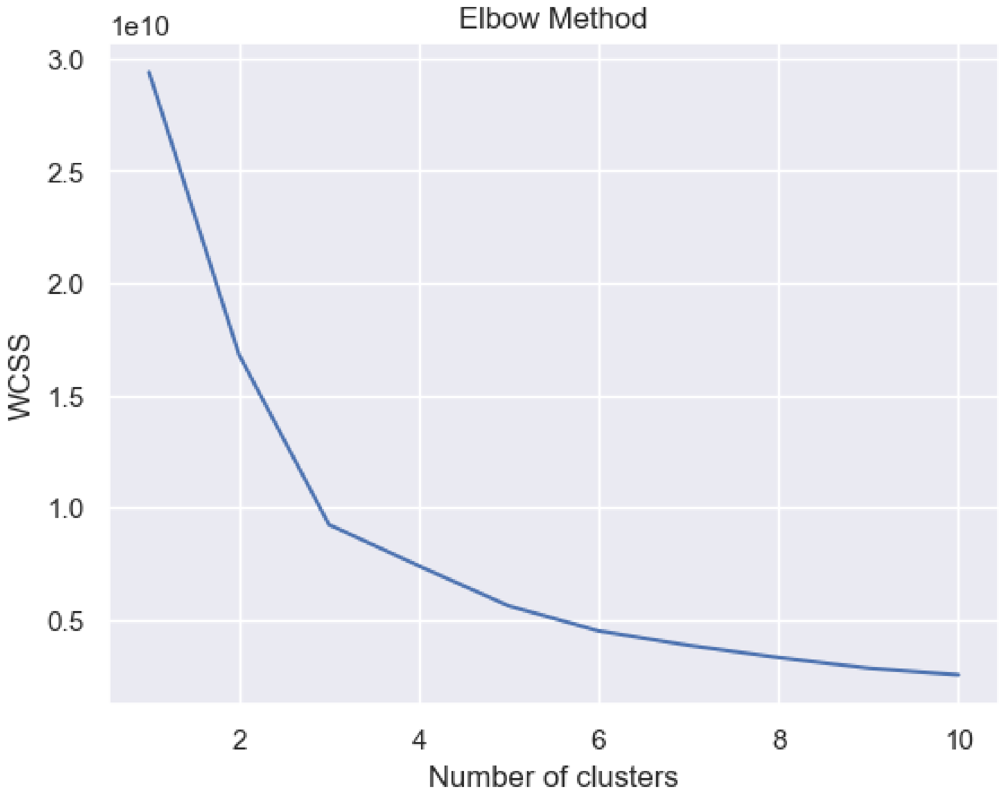


Figure 4.9: Optimal number of clusters using the elbow method.

After segmenting the data, a HM is created. For each cluster, the best baseline models XGBoost regressor, are individually applied for prediction. This hybrid approach leverages the strengths of both clustering and advanced ML models to improve the accuracy and robustness of predicting clients' future investments.

Optimization of the models

Initial evaluations of the baseline models revealed that the RF and XGBoost regressors surpassed other models. Hence, efforts were made towards optimizing these two baseline models for enhanced results. The optimization strategy encompassed two stages; Firstly, applying PR of degree 3 for RF and degree 4 for XGBoost. And secondly, performing hyperparameter tuning along with employing cross-validation techniques.

The selection of the degree of the polynomials was based on iterative testing and the assessment of a model's performance. PR of different degree polynomials were tested with the RF and it was found PR of degree 3 was the most optimal choice in terms of performance and computation time. That is, RF with PR of degree 3 significantly outperformed degree 2 PR but did not turn out to be extremely computationally intensive. Whereas using PR of degree 4 was very computationally demanding, it did not improve the performance of the model substantially. Therefore, making PR of degree 3 was the preferred choice in terms of the accuracy-speed trade-off.

On the other hand, the XGBoost model responded favorably to higher polynomial degrees [68]. Here, a fourth-degree polynomial notably improved the performance, justifying its higher computational cost in light of the significant performance uplift and its alignment with the goals of the research. Thus, a consistent application of the fourth-degree polynomial was maintained for all XGBoost experiments.

Subsequently, hyperparameter tuning was conducted on the two models in conjunction with a 3-fold cross-validation approach [69]. Hyperparameter tuning is essential in the ML workflow as it fine-tunes a model's parameters that are set before training it and are pivotal to the model's success. This diverse parameter space ensured a comprehensive search to improve the model's performance.

To perform hyperparameter tuning for the RF regressor, the random hyperparameter search approach was used with 3-fold cross-validation. This approach provides a search technique, wherein hyperparameter combinations are randomly selected within defined bounds. For the RF model, an array of hyperparameters were adjusted:

- n_estimators: Ranging from 200 to 2000, with 10 equidistant values.
- max_features: Evaluated for both 'auto' and 'sqrt'.
- max_depth: Values spanned from 10 to 110, with an additional 'None' option.
- min_samples_split: Tested for 2, 5, and 10.
- min_samples_leaf: Values of 1, 2, and 4 were considered.
- bootstrap: Both 'True' and 'False' settings were explored.

To perform hyperparameter tuning for XGBoost, BO was used instead of random hyperparameter search as the latter was computationally very expensive in comparison to the former approach. This efficiency is a consequence of its ability to use previous results to deduce subsequent searches and generally fewer required iterations to find optimal settings, making better use of computational resources.

For the XGBoost regressor, hyperparameters are fine-tuned using both PR and BO [70, 71] with 3-fold cross-validation. BO is particularly advantageous for large datasets where optimal utilization of computational resources and execution time are essential [72]. The following are the parameters considered for BO.

- Max Depth: Ranges from 5 to 30.
- Learning Rate: Varies between 0.01 and 0.5 in increments of 0.01.
- Number of Estimators: A choice ranging from 20 to 205, stepping by 5.
- Gamma: Between 0 and 0.50 with a step of 0.01.
- Min Child Weight: Ranges from 1 to 10 in increments of 1.
- Subsample: A range from 0.1 to 1, incrementing by 0.01.
- Colsample by Tree: Between 0.1 and 1.0, stepping by 0.01.
- Alpha: A uniform distribution from 0 to 1.

4.1.4. Evaluation

Performance metrics

Several key metrics such as MAE, RMSE, and $\mathbf{R^2}$ are used to assess the effectiveness of predictive models [57, 62–64]. The evaluation process unfolds in four distinct phases. First, all baseline models are carefully evaluated to determine their RMSE, MAE, and $\mathbf{R^2}$ values. The RMSE is an indicator of the average prediction error of the model, with higher values indicating greater discrepancies. At the same time, MAE measures the average absolute discrepancy between predicted and observed values. $\mathbf{R^2}$ indicates the extent to which the model's predictors account for the variance in the dependent variable, in this case, the expected investment days. A more robust $\mathbf{R^2}$ value indicates that the model can effectively explain variations in the predicted values. In the second phase, top-performing models are fine-tuned using techniques such as PR[64] and hyperparameter optimization along with 3-fold cross validation [73, 74], after which their MAE, RMSE, and $\mathbf{R^2}$ metrics are recalculated. In the third phase, the optimal model, in this instance, XGBoost regressor, is deployed on the four clusters curated using domain knowledge. The MAE, RMSE, and $\mathbf{R^2}$ for each cluster are calculated, and an average score for all clusters is subsequently derived. In the final phase, a similar approach is adopted, but clusters are delineated by the k-means method.

Model performance is often assessed through graphical methods. As depicted in Figure 4.6, a scatter plot plots actual values against their corresponding predicted values. Central to this representation is the identity line, a diagonal line that signifies where data points would fall if the model's predictions were spot-on. The proximity of points to this line serves as an indicator of the model's precision. Points closely aligned suggest high accuracy, while those straying away highlight potential errors. Additionally, the distribution of these points can shed light on the model's reliability across various data ranges. Any noticeable trends or patterns in the data distribution could hint at systematic biases in the model. Points that significantly diverge from the majority might signal outliers, providing clues about potential data irregularities or issues in the model's predictions.

Explainability using SHAP

To understand how predictions are made, it's essential to make the model's decision-making process transparent, especially when dealing with complex models. In such scenarios, post hoc explanation methods are essential. SHAP was integrated into this study to illuminate the underpinnings of the model's predictions, illustrating the individual contribution of each feature and offering in-depth explanations for specific predictive outcomes.

The choice of SHAP was made due to its robust theoretical foundation in cooperative game theory, guaranteeing equitable and consistent determination of feature significance [65]. Its capability to provide explanations for singular predictions caters to the necessity for clarity in intricate model decisions. The accuracy and reliability of SHAP, coupled with its adaptability to different model types, makes it suitable for cross-model comparative analysis [66]. Additionally, SHAP's visualization tools, like summary and dependency plots, aid in intuitively grasping how features impact predictions, which is crucial in domains where the interpretability of models is crucial. Consequently, SHAP distinguishes itself by offering detailed and understandable explanations, thereby meeting the fundamental goals of XAI.

To analyze the predictions of a model, other techniques such as feature importance can also be used. However, it would only provide a broader view of the model's predictive behavior whereas, SHAP provides both a global and a local interpretation, allowing a detailed understanding of individual predictions. This distinction is crucial as SHAP can elucidate the specific reasons behind each prediction, a level of detail not achievable with general feature importance.

Processing time

Processing time is a critical metric for selecting models, especially in real-world scenarios where rapid decision-making is required. It provides insight into the scalability and efficiency of a model, which is crucial when dealing with large datasets or working with limited resources. This metric also highlights an accuracy-speed trade-off that is essential in real-time applications [75]. Processing time is also essential for benchmarking, allowing comparisons to be made between different models or algorithms on the basis of efficiency [76]. This ensures that the model selected is not only accurate but also computationally feasible and cost-effective for its intended purpose. In our analysis, we measured the model's execution time as a metric for assessing processing time.

5

Results

This chapter outlines the outcomes of three experimental setups to forecast a customer's next investment. First, an assessment of the four baseline models is presented. Then the various optimization techniques are applied to the top two best models, which are then evaluated by critical performance indicators such as $\mathbf{R^2}$, MAE, and RMSE. The next section introduces HMs in which the data is segmented using two approaches: using business knowledge and via k-means clustering. The efficacy of these models is also assessed using the specified performance indicators. Lastly, an analysis of the decision processes of the models using SHAP values is presented. Also, a comparison of the results of the SHAP analysis of the baseline model with different clustering techniques is made.

5.1. Comparison of models' performance

5.1.1. Baseline Models

Without optimisation

Table 5.1 presents a comparative evaluation of the four baseline predictive models: LR, RF regressor, DT, and XGBoost regressor based on their performance metrics namely, RMSE, MAE and $\mathbf{R^2}$. These models were applied to the entire dataset without optimization or cross-validation to establish a benchmark for their raw predictive capability.

The RF regressor followed by XGBoost is the two best-performing baseline models in terms of the three performance metrics of interest. Both of these models are sensitive to small variations in the data and are therefore able to capture the complex patterns in the data well. The RMSE values of the two models are extremely close, while the MAE and $\mathbf{R^2}$ values obtained using RF are much better than those obtained using XGBoost. The RF model has the highest $\mathbf{R^2}$ score (0.345), indicating its ability to explain a significant proportion of the variability in the target variable.

On the other hand, LR and DT exhibit poor performance. They both have relatively high RMSE and MAE values and lower $\mathbf{R^2}$ values. These models are vulnerable to overfitting issues, which could explain their lack of performance. Complementing these findings, Figure 4.6 provides a visual comparison of the actual and predicted investment values for each model, allowing for a more intuitive understanding of each of the model's performance.

Table 5.1: Performance metrics evaluation: MAE, RMSE, and $\mathbf{R^2}$ across the four model implementations.

Models	**RMSE**	**MAE**	$\mathbf{R^2}$
LR	488.39	337.04	0.127
RF regressor	**422.96**	**244.37**	**0.345**
DT	555.60	278.67	0.004
XGBoost regressor	**422.94**	272.106	**0.282**

WITH OPTIMISATION

Following the assessment of the four baseline models, the RF regressor and XGBoost regressor emerged as superior in terms of their $\mathbf{R^2}$, RMSE, and MAE metrics. To further enhance the performance of these models, various optimization techniques such as PR and hyperparameter optimization were employed. Table 5.2 illustrates the performance improvements achieved using different optimization techniques on the RF regressor and the XGBoost regressor. For the RF regressor, the use of PR results in slight improved performance, while the random hyperparameter search with 3-fold cross-validation provides notable improvements. On the other hand, the XGBoost regressor exhibits improvement in MAE and $\mathbf{R^2}$ in comparison to the baseline model without optimization.

Although the RF regressor exhibits superior performance with respect to the $\mathbf{R^2}$ value, it has a greater tendency to overfit. This assessment is based on the MAE measured on both the training and testing data. The MAE for the training data was approximately 50.7, while the testing data showed a significantly higher MAE of 380.81. This was evaluated using random hyperparameter search optimization on the RF regressor model.

Table 5.2: Performance metrics evaluation after optimization: MAE, RMSE, and R^2.

Models	**Optimization method**	**RMSE**	**MAE**	$\mathbf{R^2}$
RF regressor	PR with degree 3	426.61	223.40	0.312
RF regressor	Random hyperparameter search with cv=3	380.81	200.5	**0.363**
XGBoost Regressor	PR with degree 4	433.234	265.34	0.33
XGBoost Regressor	PR with degree 4 and BO and cv=3	410.23	256.23	**0.332**

5.1.2. HYBRID MODELS

The use of HMs can yield better results as they can capture the diversity of the data and can make more robust predictions [77]. Therefore, two types of HMs to predict the customer's next investment were designed. The first model uses domain knowledge and the second uses a k-means clustering to segment the customers' data to form distinct clusters. To make predictions for each of the clusters, XGBoost regressor is selected due to its capability to handle complex non-linear patterns (see Figure A.7) and its robustness against overfitting due to regularization.

CUSTOMER SEGMENTATION USING BUSINESS KNOWLEDGE

In this experiment, the dataset is divided into four distinct clusters, guided by domain knowledge, as shown in Figure 4.7. The primary aim here is to enhance the predictive abilities of the models through the utilization of clustering techniques. To achieve this, the XGBoost regressor is used as a predictive model due to its superior performance and its invulnerability to overfitting. In the first phase of the analysis, the XGBoost regressor was applied to all clusters without any optimization. Secondly, PR was introduced into the model in conjunction with the XGBoost regressor. PR increased the adaptability of the models, enabling them to recognize and capture complex, higher-order patterns embedded within each cluster, as a result, the $\mathbf{R^2}$ values are improved, see Table 5.3. Finally, BO was reinforced by 3-fold cross-validation when combined with PR. This strategic combination allowed for the meticulous fine-tuning of hyperparameters for the XGBoost regressor.

Table 5.3 shows the results obtained using the above techniques. It is clear from the results in Table 5.3 that the optimization techniques, in particular, PR and BO, have had a transformative impact on model performance across all clusters. The performance is significantly improved when PR was combined with BO enforced with 3-fold cross-validation, resulting in significantly lower RMSE and MAE values and increased $\mathbf{R^2}$ coefficients for all clusters. Cluster 0,1 and 2, experienced a significant increase in its $\mathbf{R^2}$ coefficient. The $\mathbf{R^2}$ values for the three clusters increased as a result of optimizing the predictive model. This notable increase signals the effectiveness of the modeling techniques in capturing the underlying patterns within this cluster. Meanwhile, the $\mathbf{R^2}$ of Cluster 3 is largely unchanged.

Table 5.3: Performance metrics evaluation for clusters obtained using domain knowledge.

Clusters	**Method**	**RMSE**	**MAE**	$\mathbf{R^2}$
Cluster 0	without optimization	415.0264	213.4121	0.09888
Cluster 1	without optimization	259.3656	258.9068	0.2349
Cluster 2	without optimization	541.1683	336.15554	0.1641
Cluster 3	without optimization	563.6139	362.94009	0.3878
Cluster 0	PR degree 4	347.617	236.607	0.167
Cluster 1	PR degree 4	372.073	246.191	0.3789
Cluster 2	PR degree 4	382.9757	263.46	0.384
Cluster 3	PR degree 4	427.737	344.760	0.3381
Cluster 0	PR degree 4 & BO	314.614	192.143	**0.317**
Cluster 1	PR degree 4 & BO	375.981	248.470	**0.3658**
Cluster 2	PR degree 4 & BO	395.741	279.8962	**0.3429**
Cluster 3	PR degree 4 & BO	419.324	337.0668	0.3638

CUSTOMER SEGMENTATION USING K-CLUSTER ALGORITHM

In this experiment, the k means clustering technique is used to divide the customer data into three distinct clusters, as shown in Figure 4.8. The k means algorithm is known for its effectiveness in partitioning data in the field of ML and data analysis [59]. The resulting clusters generated by the k means algorithm reveal latent patterns within the data set [78]. These patterns can aid the decision-making process by providing valuable insights into the structure and characteristics of the customer data.

The approach employed here is the same as the one adopted to measure the performance of the clusters made using domain knowledge. The clusters are first evaluated using the XGBoost algorithm without any optimization, as presented in Table 5.4. This evaluation is followed by the application of PR of degree 4 to the clusters. Finally, a combination of PR and BO is utilized for assessment. From the presented data, it is inferred that while the application of PR with a degree of 4 does not always enhance performance, the combination of PR and BO appears to improve the performance of the model, especially for Cluster 0.

Table 5.4: Performance metrics evaluation for clusters using k-means clustering.

Clusters	**Method**	**RMSE**	**MAE**	**R^2**
Cluster 0	Without optimization	423.2751	267.7484	0.3007
Cluster 1	Without optimization	460.9731	278.6546	0.4858
Cluster 2	Without optimization	296.0638	180.95336	0.3316
Cluster 0	PR degree	430.7743	266.58611	0.2757
Cluster 1	PR degree	475.7832	285.40447	0.4522
Cluster 2	PR degree	308.1233	181.5793	0.2760
Cluster 0	PR degree & BO (cv=3)	308.1233	249.1143	**0.3388**
Cluster 1	PR degree & BO	408.1233	307.6792	0.4463
Cluster 2	PR degree & BO	293.4911	197.9280	0.3351

AVERAGE ERROR RATE USING HYBRID MODELS

To compare the predictive performance after segmentation using two approaches namely; segmentation using domain knowledge and segmentation using k means clustering, the average performance metrics of the approaches are evaluated. Wherein, the average of a performance metric, implies a mean over the clusters obtained using each of the segmentation approaches. Even though simply using PR, in general, does improve the performance of the predictive method, it is observed that optimization of the methods using PR and BO yields the best results for both of the segmenting methodologies, see Table 5.5. Out of the two segmenting methods (with optimization), k means clustering performs the best, as it yields better results in comparison to clustering using domain knowledge, in terms of their MAE and **R^2** values.

Table 5.5: Average performance metrics for HMs using domain knowledge and $k-$means method for clustering.

Hybrid models	**Methods**	**Avg. RMSE**	**Avg. MAE**	**Avg. R^2**
Cluster using domain knowledge	Without optimisation	489.7935	292.8536	0.2214
Cluster using domain knowledge	Polynomial regression	472.8219	292.5213	0.2704
Cluster using domain knowledge	Polynomial regression with Bayesian optimisation	308.1233	289.2375	0.3155
Cluster using k-means	Without optimisation	393.4373	242.4521	0.3727
Cluster using k-means	Polynomial regression	404.8936	244.5233	0.3346
Cluster using k-means	Polynomial regression with Bayesian optimisation	**308.1233**	**251.5786**	**0.3734**

This suggests that k-means clustering benefits from the algorithm's ability to identify and adapt to the natural groupings in data, which may be more reflective of the underlying patterns than the assumptions inherent in domain knowledge. Moreover, the optimization processes appear to fine-tune the clustering results, leading to k-means clusters having a consistently lower prediction error and a higher explanation of data variability.

5.2. COMPUTATIONAL EFFICIENCY

During the evaluation phase, a significant difference in computational demand was observed between the two clustering techniques. Clusters created using domain knowledge were more computationally demanding, especially when undergoing BO combined with PR with an execution time of approximately 995 minutes (or 16.5 hours). In comparison, the clusters derived from the k-means technique required less computational effort, with the entire k-means clustering process taking approximately 20 minutes to complete. The execution times for each cluster are detailed in Table 5.6. This disparity can be attributed to the inherent complexity of domain knowledge-based clusters. Such clusters, being based on domain-specific expertise, may have intricate relationships and nuances that the optimization process must navigate. On the other hand, k-means derived clusters, being a product of algorithmic processes, may present a more straightforward and structured landscape for optimization. Therefore, when considering computational efficiency alongside performance metrics, the k-means clustering approach not only delivers promising results but does so with less computational overhead, making it a more scalable and efficient choice for large datasets or resource-constrained environments.

Table 5.6: Execution time for both the HMs for each of the clusters.

HM	**Cluster 0**	**Cluster 1**	**Cluster 2**	**Cluster 3**
***k*-means clusters**	6:51 min	7:12 min	6:22 min	—
Domain Knowledge	33.83 min	881.50 min	25.76 min	56.05 min

5.3. FEATURE ANALYSIS

To enhance the understanding of how customers are likely to invest, explainable XAI techniques have been applied in various experiments using SHAP. SHAP values help to clarify how much each input variable contributes to a prediction. This technique is beneficial for stakeholders to gain insights into the decision-making process of ML models by evaluating the significance of each variable [79]. For the analysis, SHAP was used on a baseline XGBoost regressor model, which was selected for its good performance. Subsequently, XAI methods were also applied to two HMs. One model incorporated domain knowledge into its design, and the other used k means clustering to inform its structure.

5.3.1. BASELINE MODEL

Figure 5.1 shows the SHAP summary plot for the baseline model XGBoost regressor, which ranks the variables according to their importance in influencing the target variable. The SHAP summary plot shows that “log_total_money” is the most influential feature in the model, with higher values of this feature generally leading to a decrease in predicted days to next investment. This could mean that customers with more money invested are predicted to invest sooner. The next most significant characteristic is the “no of contracts”. A lower number of contracts has a positive impact on the model. That is, the model predicts that customers who have fewer contracts will take a long time to make their next investment. For the “Recency” feature, higher values have a positive impact on the model’s output i.e., an increase in time since the last investment would result in a longer waiting period for the next investment. This implies that clients who have recently invested are more likely to invest again sooner. The feature “Difference_in_days_before_last _investment” appears to have a mixed effect on the model’s performance, with both high and low values affecting the prediction. This may indicate that the timing of investments is also influenced by other factors and that this characteristic alone does not have a straightforward relationship with the frequency of investment. “Amount_per _contract” also shows a mixed effect on the prediction, suggesting that the amount invested per contract affects the prediction in different ways, possibly depending on how it interacts with other features.

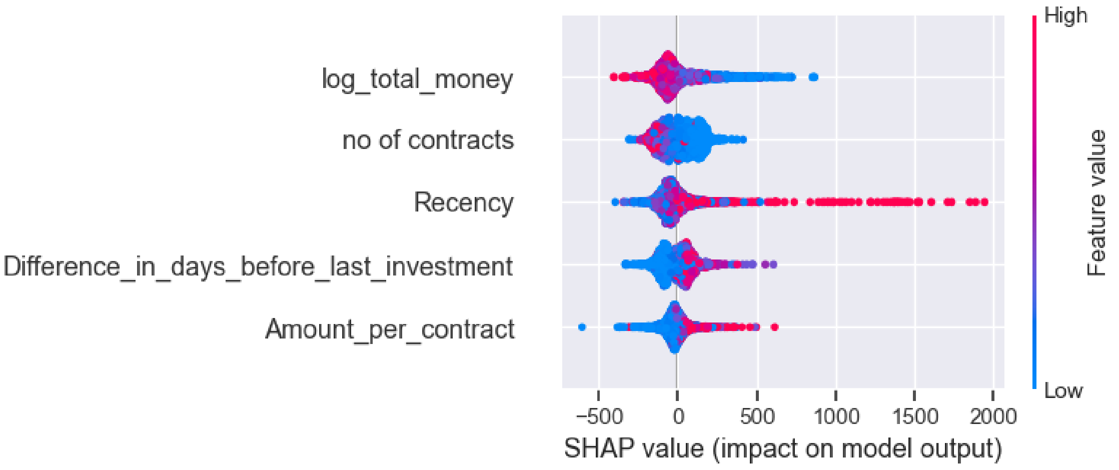


Figure 5.1: SHAP summary plot of the XGBoost model.

Figure 5.2 supports the SHAP summary plot in Figure 5.1, confirming "log_total_money" as the most decisive feature with its highest mean absolute SHAP value, suggesting its strong predictive power on investment timing. The feature importance graph also reinforces the roles of "no of contracts" and "Recency", in line with the summary plot's findings.

Meanwhile, "Amount_per_contract" emerges with a notable negative influence, adding depth to the summary plot's insights. Together, these visualizations validate the model's feature impact and provide a coherent understanding of the factors affecting investment behavior.

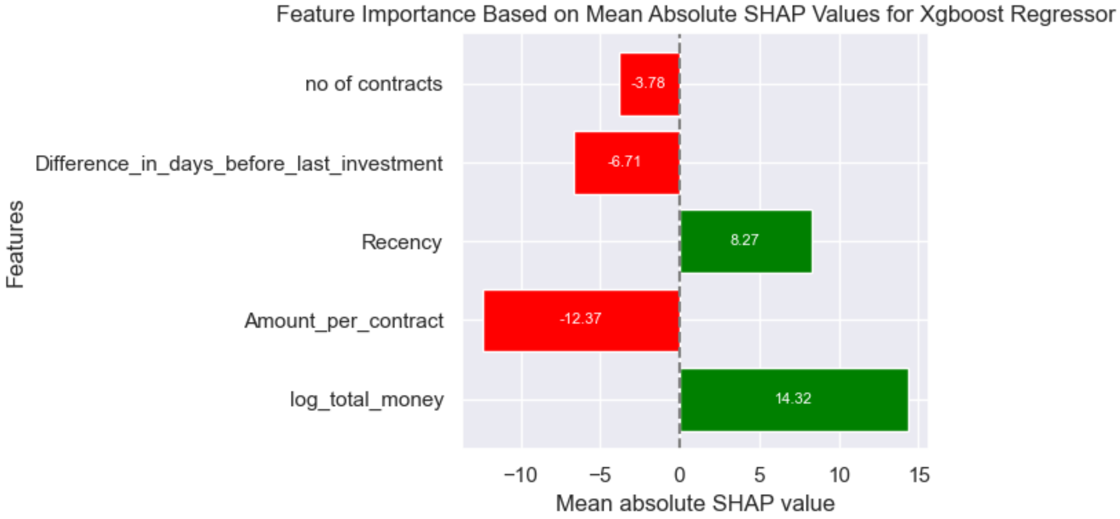


Figure 5.2: Feature importance for XGBoost Regressor.

To delve deeper into how feature values influence predictions, SHAP dependence plots are plotted. These plots illustrate the effect of a single feature on the predicted outcome for each data point. Presented in Figure 5.3, these plots can convey the primary impact of specific predictor variables, as well as their interactions. Through a lens of global interpretability, we observe the overall positive or negative impact of each feature on the prediction score.

In Figure 5.3, the dependency plot illustrates the relationship between a feature's value, shown on the x-axis, and its corresponding SHAP value is plotted on the y-axis, indicating the feature's influence on the model's output i.e., "days_till_customer's_next_investment_in_days_". For example, in Figure 5.3 a) the plot examines how "log_total_money" interacts with "Recency". It shows that larger investment amounts combined with longer periods since the last contact tend to lower the prediction score. This implies that customers who have made substantial investments and have not been recently engaged are more likely to reinvest.

Also in Figure 5.3 b), we can see that a lower "no of contracts" has both a negative and a positive effect on the model predictions. Also, the correlation between "no of contracts" and "Recency" is not very clear. This suggests more complex interactions that may be non-linear or influenced by other factors that are not immediately apparent from the graph.

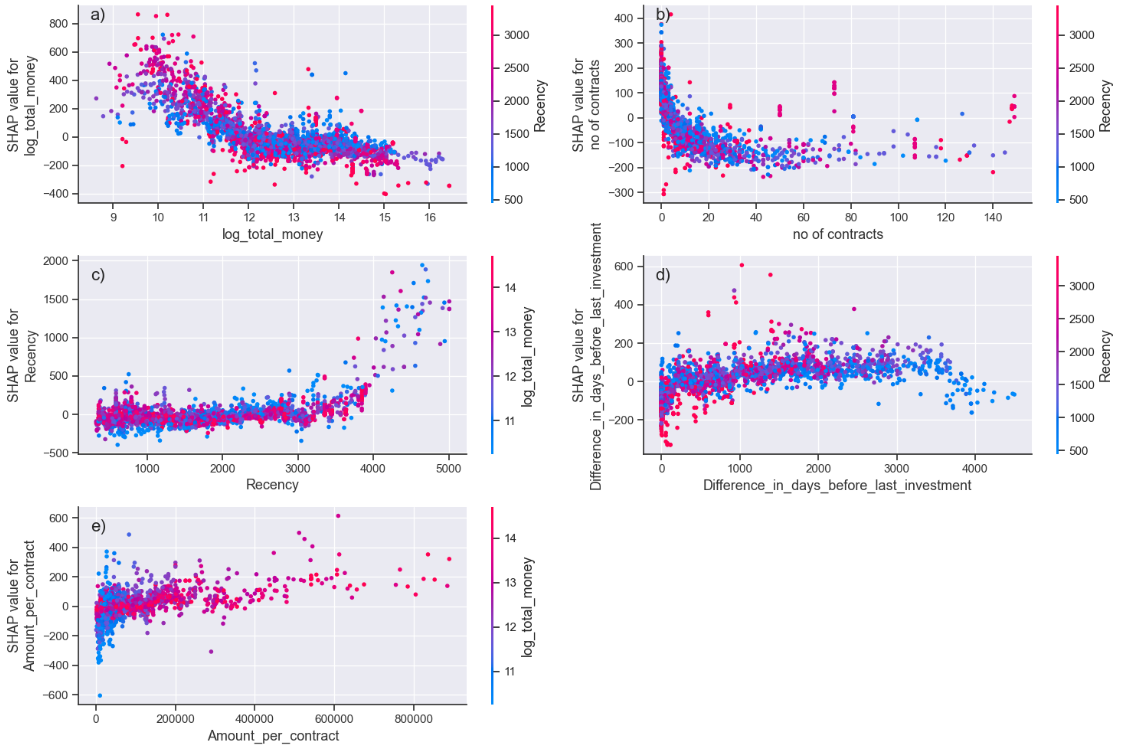
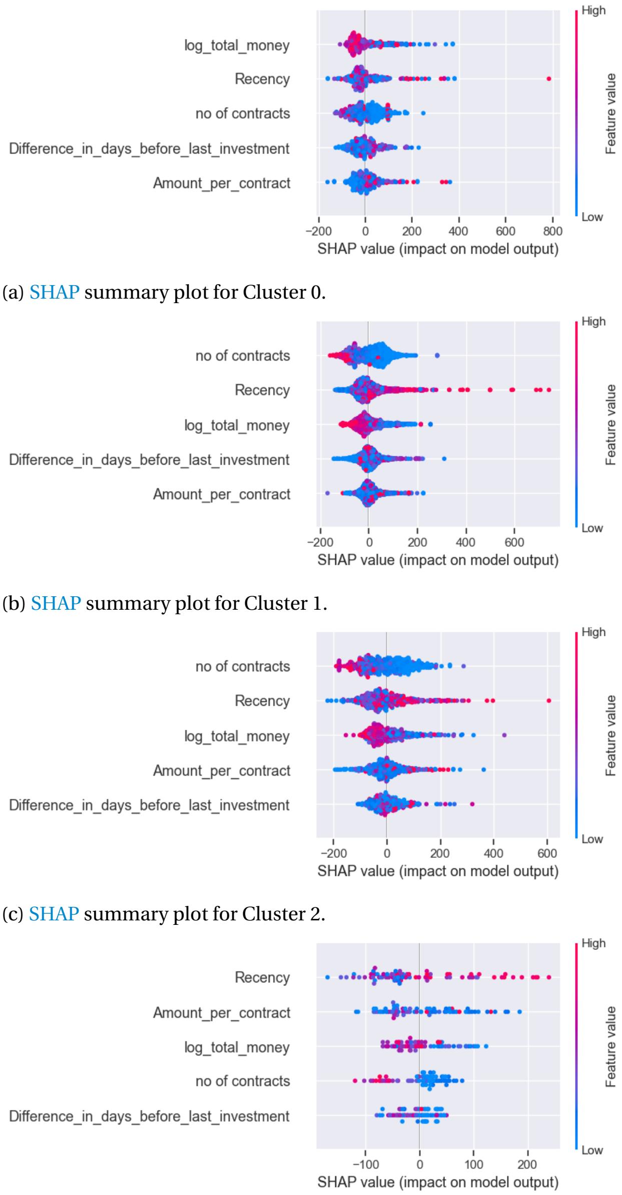


Figure 5.3: SHAP dependence plots for XGBoost model. The x-axis is the value of the feature value and the y-axis is the SHAP value.

5.3.2. CLUSTERS USING DOMAIN KNOWLEDGE

In this analysis, we look at the influence of SHAP values on the predictions of a HM made using the domain knowledge. The visual representations provided by Figure 5.4 highlight the impact of different features, establish a hierarchy of importance, and highlight their influence on the model's predictions. The "no of contracts" feature stands out as a significant predictor across different clusters. It has a positive correlation with the SHAP value, suggesting that a lower number of contracts is associated with a higher likelihood of investment, a pattern that is consistent with the results obtained using XGBoost regressor on the entire dataset. Similarly, the “log_total_money” characteristic reflects a trend consistent with the XGBoost results. A higher value of “log_total_money” tends to negatively influence model predictions. The patterns observed in the domain knowledge clusters are similar to those observed in the XGBoost analysis of the full dataset shown in Figure 5.1. Although each cluster has its own unique dependencies, the overarching trends remain consistent, reinforcing the findings from the full dataset analysis.



(a) SHAP summary plot for Cluster 0.

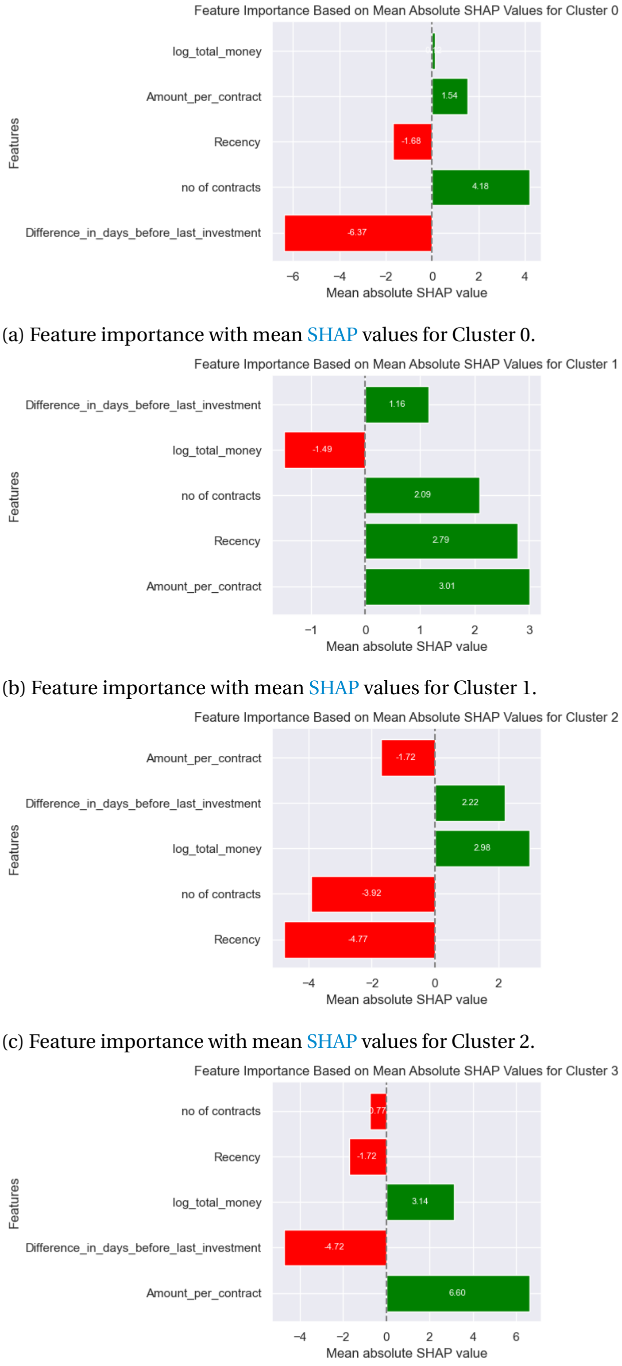
(b) SHAP summary plot for Cluster 1.

(c) SHAP summary plot for Cluster 2.

(d) SHAP summary plot for Cluster 3.

Figure 5.4: SHAP summary plot for all the Clusters made using domain knowledge.

Examination of the importance of the features, as shown in Figure 5.5, reveals a lack of uniformity across the clusters, with each cluster having different SHAP means. In particular, characteristics such as “Amount_per_contract” and “Difference_in_days_before_last_investment” appear to be significant in all clusters, but the extent of their influence varies. This lack of uniformity may suggest that different customer segments are driven by different factors when it comes to their investment behavior. For instance, one cluster might consist of customers who are particularly sensitive to the recency of their investments, while another is influenced more by the average amount invested per contract.



(a) Feature importance with mean SHAP values for Cluster 0.

(b) Feature importance with mean SHAP values for Cluster 1.

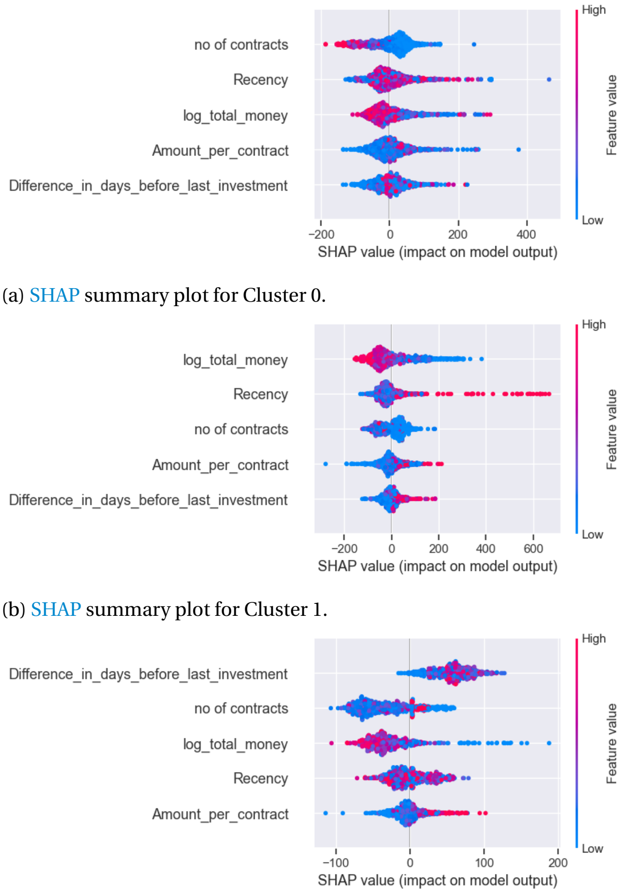
(c) Feature importance with mean SHAP values for Cluster 2.

(d) Feature importance with mean SHAP values for Cluster 3.

Figure 5.5: Feature importance for all the Clusters made using domain knowledge.

5.3.3. *k*-MEANS CLUSTERING

The SHAP method has also been applied to clusters generated by the *k*-means clustering technique. This analysis mirrors that performed on the HM underpinned by *k*-means clustering. As shown in Figure 5.6, the significance and hierarchical order of various features are illuminated by the SHAP summary violin plots. In Figure 5.6a the “no of contracts” emerges as the dominant feature influencing the target variable for Cluster 0. A smaller number of contracts tends to lengthen the predicted interval before the next investment. This suggests that fewer contracts could signal a longer waiting period for subsequent investments according to the model’s predictions. This analysis is also reflected in the feature importance of Cluster 0 shown in Figure 5.7a. Here we can see that a “no of contracts” feature has a value of -1.68, i.e., an increase in the “no of contracts” entails a reduced predicted timeframe for the next investment. In contrast, as shown in Figure 5.6c for Cluster 2, a having lower number of contracts implies a shorter duration until the next investment, indicating an inverse relationship, when compared to the previous cluster. Thus, it can be seen that the characteristics of the different clusters can have a different impact on the outcomes of the model’s prediction.

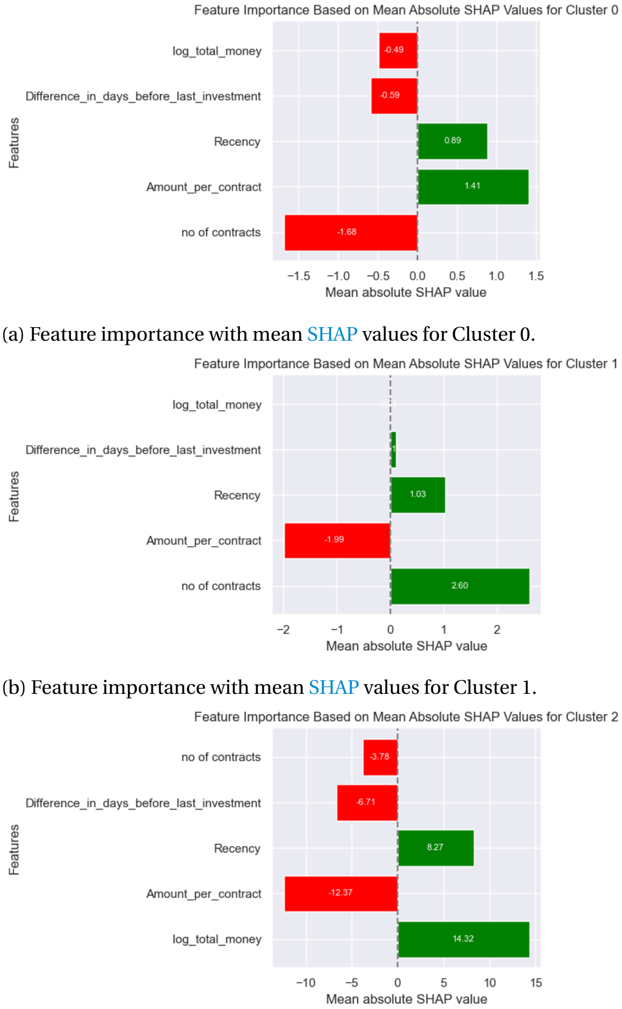


(a) SHAP summary plot for Cluster 0.

(b) SHAP summary plot for Cluster 1.

(c) SHAP summary plot for Cluster 2.

Figure 5.6: SHAP summary plot for all Clusters created using *k* means clustering.



(a) Feature importance with mean SHAP values for Cluster 0.

(b) Feature importance with mean SHAP values for Cluster 1.

(c) Feature importance for with mean SHAP values Cluster 2.

Figure 5.7: Feature importance for all Clusters created using k means clustering.

5.4. CHAPTER SUMMARY

It is found that models like RF and XGBoost outshine others in accuracy as they do not encounter significant overfitting issues. Notably, the application of optimization techniques enhances the performance of the models. Two approaches to HMs were tested, out of which the HM formed using k-means clustering, emerged as a novel approach, yielding more accurate predictions with XGBoost being the model of choice for such complex tasks. The research also delves into computational efficiency, observing that domain knowledge-based clustering demands more resources compared to the k-means approach. Another key aspect of this study is the use of SHAP values for feature analysis, providing insightful revelations about the significant influence of certain features like “total money” and “no of contracts” on investment predictions.

6

Conclusion

The aim of this thesis is to predict the next investment of B2B customers in the financial sector by harnessing the potential of HMs combined with ML techniques. A comprehensive SLR identified the gaps in existing research and highlighted the potential for advanced ML methods in this field. The SEMMA framework guided the study's methodological strategy, beginning with data collection, leading to the iterative cycle of model improvement involving exploratory analysis, model construction, and model evaluation.

The focus of this research is the development of a HM that utilizes both domain knowledge and the ML technique called k-means clustering for customer segmentation, followed by the use of a XGBoost regressor for prediction. HM with k-means clustering has proven superior to baseline and other specialized business knowledge driven HMs in predictive power, as reflected by better $\mathbf{R^2}$ and MAE metrics, while also optimizing computational efficiency. The accuracy of the model is enhanced by a rigorous process of feature engineering and fine-tuning of model parameters, using BO for optimal performance. To increase the transparency of the model's decisions, the SHAP framework has been used to reveal the inner workings of the model's predictive decisions.

The study of B2B investment actions with HMs has provided valuable insights. These HMs are tailored to B2B dynamics, combining multiple ML algorithms to capture the intricate patterns of corporate financial transactions. The integration of domain expertise with customer data clustering has revealed trends that may go unnoticed by traditional analysis. The transparency provided by XAI, particularly SHAP's interpretive power, is essential in the B2B banking landscape, where decision-making processes are meticulously evaluated for integrity and compliance with regulatory requirements. The clarity with which these models can help financial professionals communicate complex investment scenarios to B2B clients with confidence, strengthening relationships and supporting strategic decision-making.

6.1. ANSWERS TO RESEARCH QUESTIONS

6.1.1. COMMON ML METHODS TO PREDICT CUSTOMER'S NEXT INVESTMENT

In addressing the research question of the most common ML methods for predicting a client's next investment, the thesis presents a comprehensive evaluation of ML algorithms, highlighting the superiority of the XGBoost regression model. This model distinguishes itself with accuracy and reliability, outperforming others as per key performance metrics like $\mathbf{R^2}$, RMSE, and MAE. The XGBoost model is specifically adept at modeling the complex, non-linear relationships present in the investment data, with built-in regularization features that minimize the risk of overfitting, making it an optimal choice for this predictive task.

The initial performance metrics for the baseline models, prior to any optimization, set a benchmark and showed strong initial promise with the RF regressor achieving an $\mathbf{R^2}$ value of 0.345 and the XGBoost regressor scoring 0.282. Subsequent optimization using PR and hyperparameter optimization refined these models. For example, the XGBoost model, when optimized with PR and BO with 3 fold cross-validation, showed an improvement, with its $\mathbf{R^2}$ jumping to 0.332, an indication of its improved predictive power.

Through in-depth analysis and refinement, the thesis confirms the effectiveness of RF and XGBoost regressors, especially XGBoost, for predicting B2B customer investment decisions in the financial sector. The study underscores the enhanced performance of these models when finely tuned and optimized, marking them as preferred tools for investment prediction within the complex landscape of B2B finance.

6.1.2. IMPLEMENTATION OF HMS TO PREDICT CUSTOMER'S NEXT INVESTMENT

To address the research question concerning the different HM that can be applied to predict a B2B customer's next investment, the thesis compares two distinctive HMs approaches: one leveraging domain knowledge and the other utilizing k-means clustering for customer segmentation.

The domain knowledge-based approach divides the dataset into four clusters, which enhances the predictive performance of the XGBoost regressor. This technique benefits from deep industry understanding but requires significant computational power due to the complexity of domain-specific clustering when undergoing BO with PR. The results show that, although domain knowledge clusters are effective, they may involve intricate relationships that necessitate intensive computational efforts during optimization. In contrast, the k-means clustering algorithm segments customers into clusters based on data-inherent characteristics, revealing latent investment patterns within the B2B clients that may not be immediately apparent through domain knowledge. This data-driven method resulted in a better average RMSE and $\mathbf{R^2}$ value compared to the domain knowledge approach, indicating more accurate predictions. The k-means approach also demonstrated relatively higher computational efficiency, suggesting a scalable and efficient solution suitable for large or resource-constrained data environments.

This comparative analysis underpins the argument that HM, particularly those utilizing algorithmic clustering like k-means, provide an effective framework for predicting investment behaviors of B2B customers in the banking sector.

6.1.3. INTEGRATION OF XAI WITH THE HMS AND TRADITIONAL ML MODEL

Based on the application of the SHAP methodology to both the HMs and the baseline models, the thesis outlines a clear contrast in the effectiveness of XAI between these two approaches. Evidence from the research suggests that baseline model XGBoost regressor, identified “no of contracts”, “log_total_money” and “recency” as significant factors, which provide a broad prediction of the customer’s next investment. The SHAP plots suggest that customers with a high amount invested and with recent investments are likely to reinvest sooner, a finding that is insightful but provides a generalized view without the granularity of customer segmentation.

Conversely, the HM either informed by domain knowledge or segmented by k-means clustering, reveal a more complex pattern of investment behavior. The “no of contracts” is identified as an important predictor within the HM, revealing its varying influence on the customer’s next investments across different segments. The HM using k-means clustering, defines “no of contracts” as a variable whose impact on investment timing varies significantly across different customer clusters. For example, Cluster 0, obtained using k-means clustering, shows a distinctive trend, where a smaller “no of contracts” correlates with a longer wait for the next investment opportunity. This suggests that within this cluster, customers with fewer contracts might be more cautious or less ready to invest again immediately. In contrast, Cluster 2 of k-means clustering displays an inverse relationship; here customers with fewer contracts tend to reinvest sooner, which may reflect a different investment strategy or financial behavior prevalent in this segment. This differentiation in patterns implies a detailed detection of the investment behaviors that k-means clustering enables within HM, offering a tailored understanding that the baseline models cannot provide.

The application of HMs significantly enhances the prediction of future investments by B2B customers in the banking sector. This enhancement stems from the integration of advanced ML algorithms like XGBoost, which excel in handling complex data. Optimization techniques further improve these models, as evidenced by increased accuracy in predictive metrics. Particularly effective is the use of k-means clustering within HMs for customer segmentation, which uncovers latent investment patterns and offers more accurate, segment-specific predictions. Additionally, the integration of XAI, specifically SHAP methodology, adds clarity to the decision-making process of these models, making them not only powerful in prediction but also transparent and understandable for stakeholders. Overall, HMs provide a comprehensive, efficient, and detailed approach to predicting investment behaviors in the B2B banking context.

6.2. STUDY CONRIBUTIONS

6.2.1. ACADEMIC CONTRIBUTION

This section highlights the contribution of this research to academia. Firstly, it provides a thorough SLR to examine the work on different ML models, optimization techniques, and methods used in predictive modeling. This literature review has revealed certain gaps in the application of predictive modeling within the B2B sector across different industries.

In particular, less work has been done in the area of B2B in the financial sector. There are few studies that have mined financial data in the B2B context to develop analytical models using machine learning, which could improve an organization's understanding of customer behavior, particularly in terms of investment patterns. Furthermore, there is a scarcity of the implementation of HMs in conjunction with XAI in the context of B2B, a gap that this thesis aims to fill.

This study contributes to filling these gaps by exploring, evaluating, and selecting the most suitable HMs for predicting a customer's next investment. The comprehensive evaluation of these HMs created through different approaches enriches the existing body of knowledge in predictive modeling. In addition, the research explores different feature engineering methods, resulting in the creation of novel features that improve model performance.

These academic contributions aim to advance the discussion within the data-driven financial sector, helping institutions to make informed decisions that benefit both the organization and its customers.

6.2.2. PRACTICAL CONTRIBUTION

This study plays a key role in supporting the bank's transition to a data-driven organization by demonstrating the feasibility of implementing predictive modeling. The research involves the use of HMs to predict the future investment activity of clients by analyzing the various data sets available within the bank. Such an approach suggests that the same methodology could be extended to predict other critical aspects within the banking industry, including customers' behavior, fraud detection, and more. The predictions of a HM can be understood and improved through the use of XAI, which highlights the influential features that contribute to the results.

The thorough analysis of the bank's dataset in the study not only uncovers trends and insights but also helps to identify data that can be used for preventative measures. This understanding enables the bank to make more informed decisions and potentially anticipate and mitigate risks before they materialize.

Furthermore, the applicability of the work presented in this thesis extends to other sectors. Two important examples are e-commerce and healthcare. In e-commerce, HMs can significantly improve customer segmentation, allowing companies to categorize consumers into different groups according to their buying patterns and preferences. This segmentation helps to develop more targeted marketing approaches. In addition, in the healthcare industry, the same approach is proving useful in predicting patient outcomes, such as assessing the risk of readmission to the hospital after discharge, thereby contributing to improved patient care protocols. Furthermore, these models are adept at refining treatment strategies by thoroughly examining patient data and their historical treatment outcomes. This is especially crucial in the field of personalized medicine, where such predictive techniques can provide recommendations of specific treatments for conditions like cancer, which can be tailored to individual patient profiles.

In addition, this research can serve as a foundation for the development of other predictive models in various sectors. The insights gained can guide future model development, making them more accurate and efficient, which can help organizations streamline their operations by identifying key trends and patterns that can improve decision-making and operational efficiency.

6.3. LIMITATIONS AND FUTURE RECOMMENDATIONS

While the research suggested the use of HMs to predict customers' next investment, certain limitations of this research should be acknowledged.

One of the main limitations was the availability of the data set. That is, the reliability of the provided data as well as its usability for the models was unknown. Since the dataset of interest came from a financial institution, there were various concerns regarding its confidentiality and privacy. To make more accurate predictions and increase the accuracy of the model, one could include more datasets. The dataset can include information regarding other factors such as repayment behavior of customers, punctuality of repayments, and their frequency, and also consider the influence of wider economic factors such as the impact of the COVID-19 pandemic and the influence of the war on customers' sales, etc. All of these factors can provide critical insights and improve the accuracy and reliability of the predictions.

Additionally, a second key constraint was due to the limited computational resources. This constraint significantly hampered the ability to experiment with different hyperparameters when optimizing models, especially the HM model that used domain knowledge. These limitations also imposed restrictions on the choice of the modeling methods. For instance, PR was limited to a degree of 3 for RF, and cross-validation folds were limited to 3, without the possibility of extending these parameters to potentially more optimal numbers. As a result, the consideration of an accuracy-cost trade-off was essential in the modeling and optimization process.

A limitation in terms of XAI was present when SHAP, which is suitable for supervised learning, was applied to unsupervised k means clustering. In an attempt to interpret the results of k means clustering, the RF classifier was used to determine the customers in each of the clusters. However, using SHAP for this purpose did not provide clear insights, as the explanations provided were more related to the predictions of the RF model than to the underlying logic of the k means algorithm. To better understand the clustering mechanism of k means, it would be beneficial to explore methods designed for unsupervised learning. This could include assessing the characteristics of cluster centroids, calculating mean distances within clusters to measure the influence of features, or performing silhouette analysis to measure cluster separation. The use of dimensionality reduction techniques, such as PCA, could also provide a more understandable visualization of clustering and highlight the role of key features.

In our research, models like RF and XGBoost are used to predict the customer's next investment. These models are adept at handling large datasets and complex modeling scenarios, offering robust predictions across a variety of contexts. However, both RF and XGBoost might not fully capture the intricacies of time-to-event data, a crucial aspect when predicting the timing of customer investments. In future research, exploring advanced statistical models such as survival analysis could be valuable. Survival analysis excels in modeling time-to-event data by explicitly accounting for time-dependent variables, offering a more detailed understanding of when and why customers might make future investments. This approach could significantly enhance the precision of the predictive models, especially in scenarios where the timing of a customer's investment decision is critical.

REFERENCES

[1] A. De Caigny, K. Coussement, K. De Bock. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research 269 (2018). doi:10.1016/j.ejor.2018.02.009.

[2] A. De Caigny, K. Coussement, W. Verbeke, K. Idbenjra, M. Phan. Uplift modeling and its implications for b2b customer churn prediction: A segmentation-based modeling approach. Industrial Marketing Management 99 (2021) 28–39. doi:10.1016/j.indmarman.2021.10.001.

[3] F. Bolívar, M. A. Duran, A. Lozano-Vivas. Business model contributions to bank profit performance: A machine learning approach. Research in International Business and Finance 64 (2023). URL: https://ideas.repec.org/a/eee/riibaf/v64y2023ics0275531922002562.html. doi:10.1016/j.ribaf.2022.1018.

[4] D.-R. Liu, Y.-Y. Shih. Integrating ahp and data mining for product recommendation based on customer lifetime value. Information Management 42 (2005) 387–400. doi:10.1016/j.im.2004.01.008.

[5] K. G. M. Karvana, S. Yazid, A. Syalim, P. Mursanto, in: 2019 International Workshop on Big Data and Information Security (IWBIS), pp. 33–38. doi:10.1109/IWBIS.2019.8935884.

[6] M. Colgate, P. Danaher. Implementing a customer relationship strategy: The asymmetric impact of poor versus excellent execution. Journal of The Academy of Marketing Science - J ACAD MARK SCI 28 (2000) 375–387. doi:10.1177/0092070300283006.

[7] V. Mihova, V. Pavlov, volume 2025, p. 030003. doi:10.1063/1.5064881.

[8] A. Machauer, S. Morgner. Segmentation of bank customers by expected benefits and attitudes. International Journal of Bank Marketing 19 (2001) 6–18. doi:10.1108/02652320110366472.

[9] V. Kumar M. Segmenting the banking market strategy by clustering. International Journal of Computer Applications 45 (2012) 975–8887.

[10] M. Namvar, M. R. Gholamian, S. KhakAbi, in: 2010 International Conference on Intelligent Systems, Modelling and Simulation, pp. 215–219. doi:10.1109/ISMS.2010.48.

[11] P. A. Akhtar, G. Frynas, K. Mellahi, S. Ullah. Big data-savvy teams' skills, big data-driven actions and business performance. British Journal of Management 30 (2019) 252–271. doi:10.1111/1467-8551.12333.

[12] S. Moro, P. Cortez, P. Rita. Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. Neural Computing and Applications 26 (2015). doi:10.1007/s00521-014-1703-0.

[13] A. Martínez, C. Schmuck, S. Pereverzyev, C. Pirker, M. Haltmeier. A machine learning framework for customer purchase prediction in the non-contractual setting. European Journal of Operational Research 281 (2018). doi:10.1016/j.ejor.2018.04.034.

[14] A. Tamaddoni, S. Stakhovych, M. Ewing. Managing b2b customer churn, retention and profitability. Industrial Marketing Management 43 (2014). doi:10.1016/j.indmarman.2014.06.016.

[15] S. Hué, C. Hurlin, S. Tokpavi. Machine learning for credit scoring: Improving logistic regression with non linear decision tree effects. European Journal of Operational Research 297 (2017). doi:10.1016/j.ejor.2021.06.053.

[16] A. De Caigny, K. Coussement, K. De Bock. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research 269 (2018). doi:10.1016/j.ejor.2018.02.009.

[17] J. Kunchaparthi, M. Baburao, K. Chaduvula, K. Chaduvula. A comparative study on logit leaf model (llm) and support leaf model (slm) for predicting the customer churn. INTERNATIONAL JOURNAL OF COMPUTER SCIENCES AND ENGINEERING 7 (2019) 1628–1632. doi:10.26438/ijcse/v7i5.16281632.

[18] K. Coussement, M. Phan, A. De Caigny, D. Benoit, A. Raes. Predicting student dropout in subscription-based online learning environments: The beneficial impact of the logit leaf model. Decision Support Systems 135 (2020) 113325. doi:10.1016/j.dss.2020.113325.

[19] B. Pawełek, J. Pociecha, Corporate Bankruptcy Prediction with the Use of the Logit Leaf Model, 2020, pp. 129–146. doi:10.1007/978-3-030-52348-0_9.

[20] T. Vafeiadis, K. Diamantaras, G. Sarigiannidis, K. Chatzisavvas. A comparison of machine learning techniques for customer churn prediction. Simulation Modelling Practice and Theory 55 (2015). doi:10.1016/j.simpat.2015.03.003.

[21] Y. Sayjadah, I. Hashem, F. Alotaibi, K. Kasmiran, pp. 1–4. doi:10.1109/ICACCAF.2018.8776802.

[22] A. Hudaib, R. Dannoun, O. Harfoushi, R. Obiedat, H. Faris. Hybrid data mining models for predicting customer churn. Int'l J. of Communications, Network and System Sciences 08 (2015) 91–96. URL: https://api.semanticscholar.org/CorpusID:2088003.

[23] M. Thirugnanam. A heart disease prediction model using svm-decision trees-logistic regression (sdl). International Journal of Computer Applications in Technology 68 (2013) 11–15. doi:10.5120/11662-7250.

[24] I. Kaur, J. Kaur, pp. 434–437. doi:10.1109/PDGC50313.2020.9315761.

[25] T.-N. Chou, pp. 122–125. doi:10.1109/ICKII46306.2019.9042639.

[26] T. De, P. Giri, A. Mevawala, R. Nemani, A. Deo. Explainable ai: A hybrid approach to generate human-interpretable explanation for deep learning prediction. Procedia Computer Science 168 (2020) 40–48. URL: https://www.sciencedirect.com/science/article/pii/S187705092030394X. doi:https://doi.org/10.1016/j.procs.2020.02.255, “Complex Adaptive Systems”Malvern, PennsylvaniaNovember 13-15, 2019.

[27] S. Sheuly, M. Ahmed, S. Begum, M. Osbakk, pp. 81–85. doi:10.1109/AI4I51902.2021.00028.

[28] N. Jain, P. K. Jana. Xrrf: An explainable reasonably randomised forest algorithm for classification and regression problems. Information Sciences 613 (2022) 139–160. URL: https://www.sciencedirect.com/science/article/pii/S0020025522010866. doi:https://doi.org/10.1016/j.ins.2022.09.040.

[29] L. Joseph, E. Joseph, R. Prasad. Explainable diabetes classification using hybrid bayesian-optimized tabnet architecture. Computers in Biology and Medicine 151 (2022) 106178. doi:10.1016/j.compbiomed.2022.106178.

[30] R. Desai, V. Khairnar, Hybrid Prediction Model for the Success of Bank Telemarketing, 2022, pp. 693–710. doi:10.1007/978-981-16-2422-3_54.

[31] S. Devi, R. Yalavarthi. A survey on machine learning and statistical techniques in bankruptcy prediction. International Journal of Machine Learning and Computing 8 (2018) 133–139. doi:10.18178/ijmlc.2018.8.2.676.

[32] C. S. Koumetio Tekouabou, C. Gherghina, H. Toulni, P. Mata, J. Martins. Towards explainable machine learning for bank churn prediction using data balancing and ensemble-based methods. Mathematics 10 (2022) 2379. doi:10.3390/math10142379.

[33] J. Dias, P. Godinho, P. Torres, Machine Learning for Customer Churn Prediction in Retail Banking, 2020, pp. 576–589. doi:10.1007/978-3-030-58808-3_42.

[34] Y. Choi, J. Choi, How does machine learning predict the success of bank telemarketing?, 2022. doi:10.21203/rs.3.rs-1695659/v1.

[35] N. Uddin, U. Ahamed, M. A. Uddin, M. Islam, M. A. Talukder, S. Aryal. An ensemble machine learning based bank loan approval predictions system with a smart application. International Journal of Cognitive Computing in Engineering 4 (2023). doi:10.1016/j.ijcce.2023.09.001.

[36] A. Gastón, J. Garcia-Viñas. Modelling species distributions with penalised logistic regressions: A comparison with maximum entropy models. Ecological Modelling 222 (2011) 2037–2041. doi:10.1016/j.ecolmodel.2011.04.015.

[37] I. Yeh, K.-J. Yang, T.-M. Ting. Knowledge discovery on rfm model using bernoulli sequence. Expert Syst. Appl. 36 (2009) 5866–5871. doi:10.1016/j.eswa.2008.07.018.

[38] P. Fader, B. Hardie. Probability models for customer-base analysis. Journal of Interactive Marketing - J INTERACT MARK 23 (2009) 61–69. doi:10.1016/j.intmar.2008.11.003.

[39] B. Lariviere, D. Van den Poel. Predicting customer retention and profitability by using random forests and regression forests techniques. Expert Systems with Applications 29 (2005) 472–484. doi:10.1016/j.eswa.2005.04.043.

[40] W. Verbeke, K. Dejaeger, D. Martens, J. Hur, B. Baesens. New insights into churn prediction in the telecommunication sector: A profit driven data mining approach. European Journal of Operational Research 218 (2012) 211–229. doi:10.1016/j.ejor.2011.09.031.

[41] F. Devriendt, D. Moldovan, W. Verbeke. A literature survey and experimental evaluation of the state-of-the-art in uplift modeling: A stepping stone toward the development of prescriptive analytics. Big Data 6 (2018) 13–41. doi:10.1089/big.2017.0104.

[42] U. Shafique, H. Qaiser. A comparative study of data mining process models (kdd, crisp-dm and semma). International Journal of Innovation and Scientific Research 12 (2014) 2351–8014.

[43] H. I. T. Aziz, A. Sohail, U. Aslam, N. Batcha. Loan default prediction model using sample, explore, modify, model, and assess (semma). Journal of Computational and Theoretical Nanoscience 16 (2019) 3489–3503. doi:10.1166/jctn.2019.8313.

[44] F. Safari, N. Safari, G. A. Montazer. Customer lifetime value determination based on rfm model. Marketing Intelligence Planning 34 (2016). doi:10.1108/MIP-03-2015-0060?journalCode=mip.

[45] H. Roshan, M. Afsharinezhad. The new approach in market segmentation by using rfm model. Journal of Applied Research on Industrial Engineering 4 (2017) 259–267. URL: https://www.journal-aprie.com/article_53422.html. doi:10.22105/jarie.2017.91297.1011. arXiv:https://www.journal-aprie.com/article$_{5}3422_{5}36598f5e6bb29cb85ae20b09930b8ce.pdf$.

[46] D. Nimbalkar, P. Shah, Data mining using rfm analysis, 2013. doi:10.13140/RG.2.2.24229.04328.

[47] J. Hancock, T. Khoshgoftaar. Survey on categorical data for neural networks. Journal of Big Data 7 (2020). doi:10.1186/s40537-020-00305-w.

[48] W. Yao, L. LI. A new regression model: Modal linear regression. Scandinavian Journal of Statistics 41 (2013). doi:10.1111/sjos.12054.

[49] J. R. Quinlan. URL: https://api.semanticscholar.org/CorpusID:1056674.

[50] S. Rathore, S. Kumar. A decision tree regression based approach for the number of software faults prediction. ACM SIGSOFT Software Engineering Notes 41 (2016) 1–6. doi:10.1145/2853073.2853083.

[51] S. Wang, X. Yao. Using class imbalance learning for software defect prediction. Reliability, IEEE Transactions on 62 (2013) 434–443. doi:10.1109/TR.2013.2259203.

[52] V. Rodriguez-Galiano, M. Sánchez Castillo, M. Chica, M. Chica Rivas. Machine learning predictive models for mineral prospectivity: An evaluation of neural networks, random forest, regression trees and support vector machines. Ore Geology Reviews 71 (2015). doi:10.1016/j.oregeorev.2015.01.001.

[53] M. Ahmad, J. Reynolds, Y. Rezgui. Predictive modelling for solar thermal energy systems: A comparison of support vector regression, random forest, extra trees and regression trees. Journal of Cleaner Production 203 (2018) 810–821. doi:10.1016/j.jclepro.2018.08.207.

[54] T. Chen, C. Guestrin, pp. 785–794. doi:10.1145/2939672.2939785.

[55] J. Pesantez-Narvaez, M. Guillen, M. Alcañiz. Predicting motor insurance claims using telematics data—xgboost versus logistic regression. Risks 7 (2019) 70. doi:10.3390/risks7020070.

[56] J. Friedman. Greedy function approximation: A gradient boosting machine. The Annals of Statistics 29 (2000). doi:10.1214/aos/1013203451.

[57] N. M. Shahani, X. Zheng, C. Liu, F. U. Hassan, P. Li. Developing an xgboost regression model for predicting young's modulus of intact sedimentary rocks for the stability of surface and subsurface structures. Frontiers in Earth Science 9 (2021). URL: https://www.frontiersin.org/articles/10.3389/feart.2021.761990. doi:10.3389/feart.2021.761990.

[58] H. Mo, H. Sun, J. Liu, S. Wei. Developing window behavior models for residential buildings using xgboost algorithm. Energy and Buildings 205 (2019) 109564. doi:10.1016/j.enbuild.2019.109564.

[59] R. Nainggolan, R. Perangin-angin, R. Simarmata, A. Tarigan. Improved the performance of the k-means cluster using the sum of squared error (sse) optimized by using the elbow method. Journal of Physics: Conference Series 1361 (2019) 012015. doi:10.1088/1742-6596/1361/1/012015.

[60] E. Umargono, J. Suseno, S. Gunawan. doi:10.2991/assehr.k.201010.019.

[61] O. Dogan, E. Ayçin, Z. Bulut. Customer segmentation by using rfm model and clustering methods: A case study in retail industry. International Journal of Contemporary Economics and Administrative Sciences 8 (2018) 1–19.

[62] C. Willmott, S. Ackleson, R. Davis, J. Feddema, K. Klink, D. Legates, J. O'Donnell, C. Rowe. Statistics for the evaluation and comparison of models. Journal of Geophysical Research (1985). doi:10.1029/JC090iC05p08995.

[63] T. Peterek, P. Dohnalek, P. Gajdo, M. Smondrk, pp. 83–87. doi:10.1109/HIS.2013.6920459.

[64] R. Khan. Performance evaluation of regression models for covid-19: A statistical and predictive perspective. Ain Shams Engineering Journal (2021). doi:10.1016/j.asej.2021.08.016.

[65] Z. Li. Extracting spatial effects from machine learning model using local interpretation method: An example of shap and xgboost. Computers Environment and Urban Systems 96 (2022) 101845. doi:10.1016/j.compenvurbsys.2022.101845.

[66] R. El Shawi, Y. Sherif, M. Al-Mallah, S. Sakr. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. Computational Intelligence 37 (2020). doi:10.1111/coin.12410.

[67] E. Štrumbelj, I. Kononenko. Explaining prediction models and individual predictions with feature contributions. Knowledge and Information Systems 41 (2013) 647–665. doi:10.1007/s10115-013-0679-x.

[68] S. Torrisi, M. Carbone, B. Rohr, J. Montoya, Y. Ha, J. Yano, S. Suram, L. Hung. Random forest machine learning models for interpretable x-ray absorption near-edge structure spectrum-property relationships. npj Computational Materials 6 (2020) 109. doi:10.1038/s41524-020-00376-6.

[69] M. Mitchell. Bias of the random forest out-of-bag (oob) error for certain input parameters. Open Journal of Statistics 01 (2011) 205–211. doi:10.4236/ojs.2011.13024.

[70] R. Shi, X. Xu, pp. 1–6. doi:10.1109/ITSC45102.2020.9294186.

[71] X. Ma, Y. Xu, J. Yang, J. Wu, F. Tong, J. Song. A study of prediction of ground settlement of shield tunnel based on eemd-bo-gru algorithm. Journal of Physics: Conference Series 2450 (2023) 012080. doi:10.1088/1742-6596/2450/1/012080.

[72] J. Zhou, Y. Qiu, S. Zhu, D. Jahed Armaghani, M. Khandelwal, E. Mohamad. Estimating tbm advance rate in hard rock condition using xgboost and bayesian optimization. Underground Space 6 (2020). doi:10.1016/j.undsp.2020.05.008.

[73] J. Bergstra, Y. Bengio. Random search for hyper-parameter optimization. The Journal of Machine Learning Research 13 (2012) 281–305.

[74] A. Camstra, A. Boomsma. Cross-validation in regression and covariance structure analysis: An overview. Sociological Methods Research 21 (1992) 89–115. doi:10.1177/0049124192021001004.

[75] M. Amaris Gonzalez, M. Dyab, D. Trystram, R. Camargo, A. Goldman. doi:10.1109/NCA.2016.7778637.

[76] W. Rühaak, S. Chauhan, F. Khan, FriederEnzmann, P. Mielke, M. Kersten, I. Sass. Rock core microtomography image processing - segmentation using seven different machine learning algorithms. Computers Geosciences 86 (2016) 120–128. doi:10.1016/j.cageo.2015.10.013.

[77] K. Yang, Y. Cai, D. Huang, J. Li, Z. Zhou, X. Lei, in: 2017 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 465–466. doi:10.1109/BIGCOMP.2017.7881759.

[78] C. Ezenkwu, S. Ozuomba, C. Kalu. Application of k-means algorithm for efficient customer segmentation: A strategy for targeted customer services. International Journal of Advanced Research in Artificial Intelligence(IJARAI) 4 (2015). doi:10.14569/IJARAI.2015.041007.

[79] S. Ben Jabeur, S. Mefteh-Wali, J.-L. Viviani. Forecasting gold price with the xgboost algorithm and shap interaction values. Annals of Operations Research (2021). doi:10.1007/s10479-021-04187-w.

A

APPENDICES

A.1. OPTIMIZATUON USING HYPERPARAMETERS

Table A.1: Hyperparameters used for XGBoost regressor which are applied on different clusters formed by both *k*-means clustering and domain knowledge.

Hyparameter	**Range**
max_depth	range(5,30,1)
learning rate	0.01, 0,5, 0.01
n_estimataors	range (20, 205,5)
gamma	0, 0.50, 0.01
min_child_weight	1,10,1
sub_sample	0.1, 1, 0.01
colsample_bytree	0.1, 1.0, 0.01
alpha	0,1

A.2. DESCRIPTIVE STATISTICS

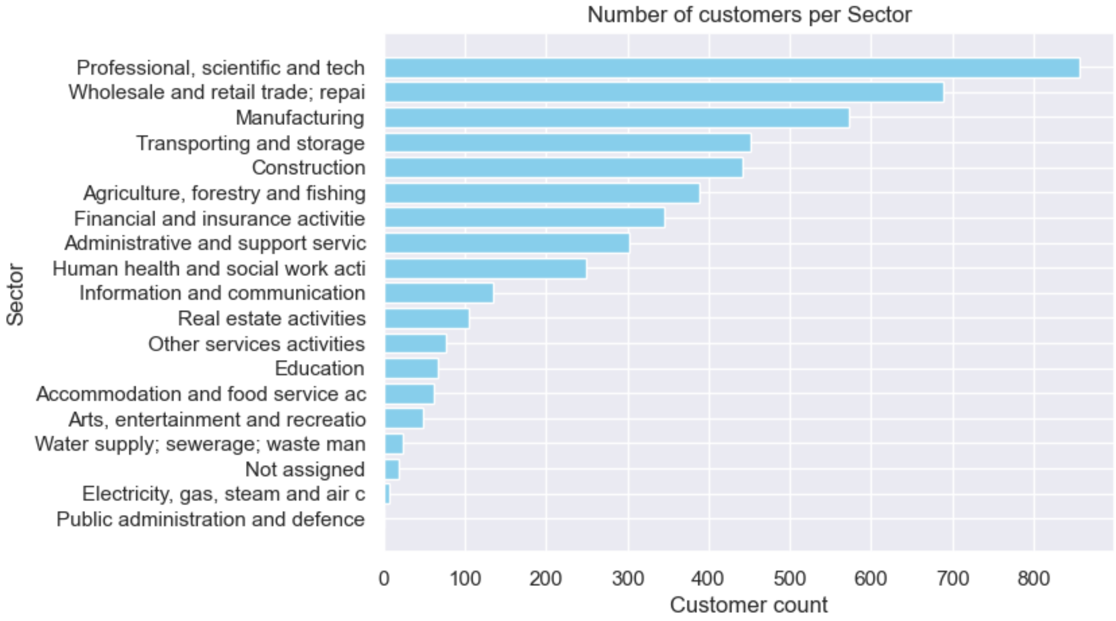
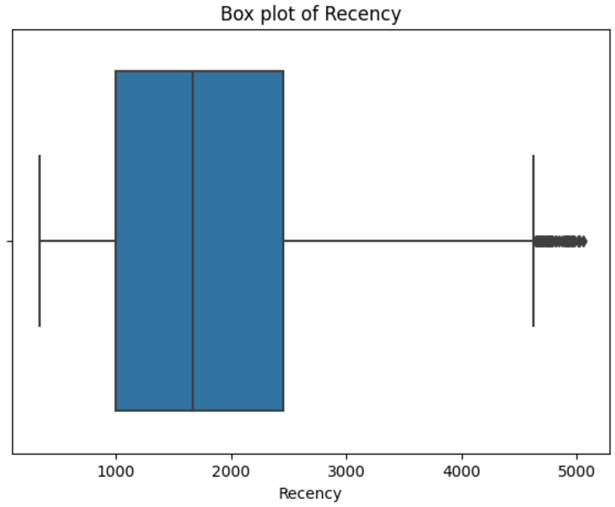
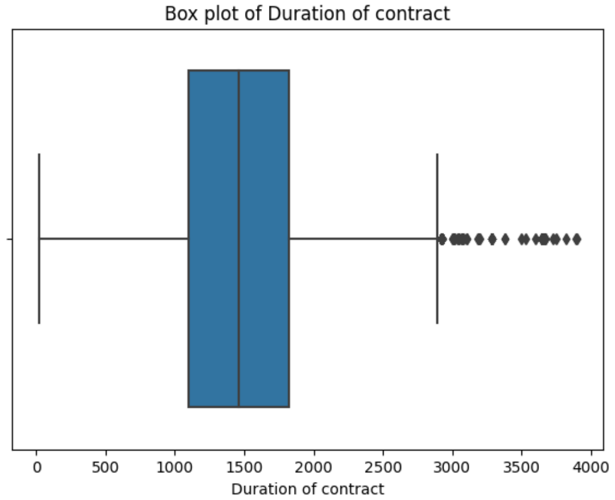


Figure A.1: Count of customers belonging to different sectors.



(a) Box plot showing the outliners for the feature: Recency



(b) Box plot showing outliners in the feature: duration of contract.

Figure A.2: Box plot showing outliners in different features formed after feature engineering.

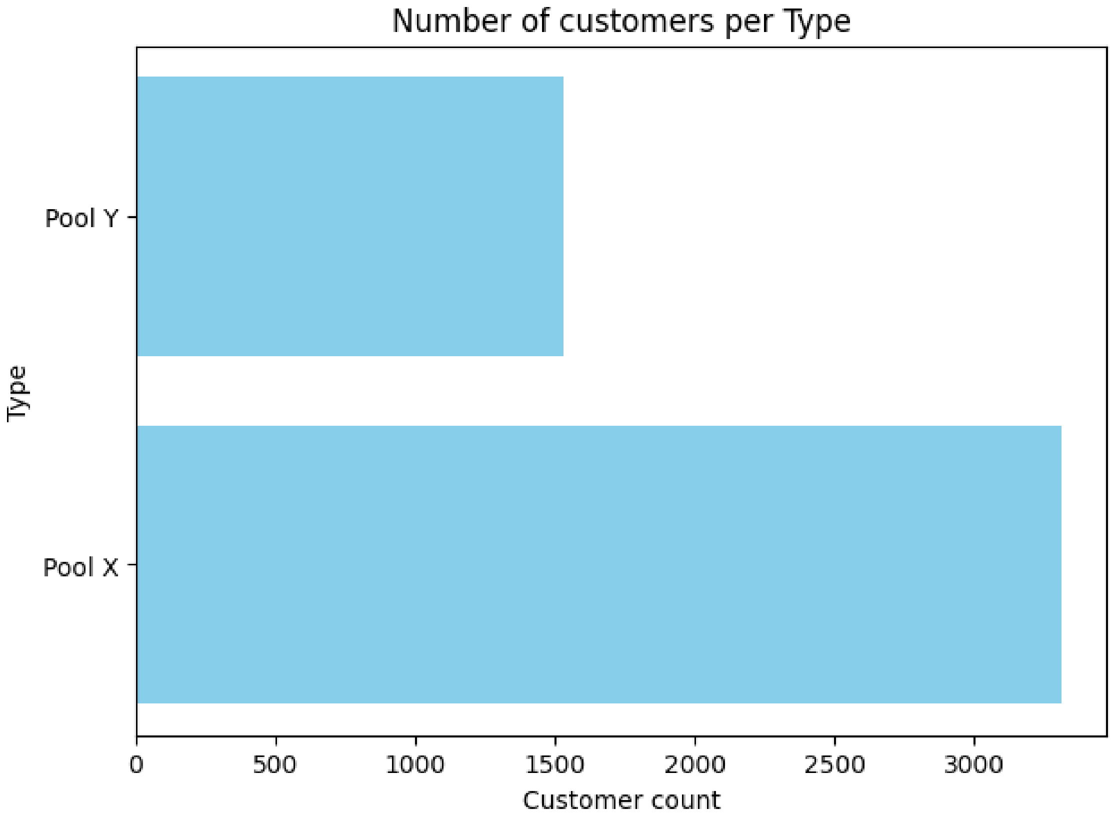


Figure A.3: Total number of customers belonging to the type: Pool X and Pool Y.

A.3. ANALYSIS OF DIFFERENT CLUSTERS

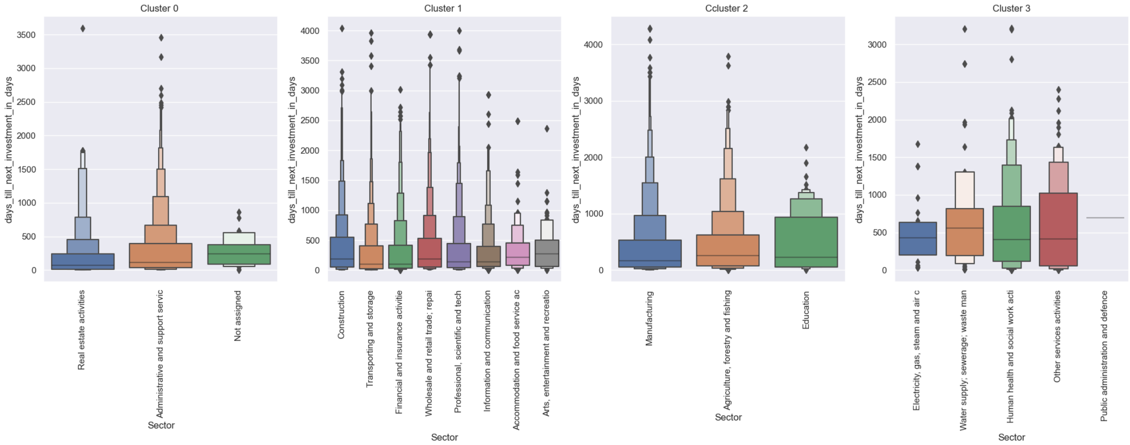


Figure A.4: Customer segmentation using domain knowledge.

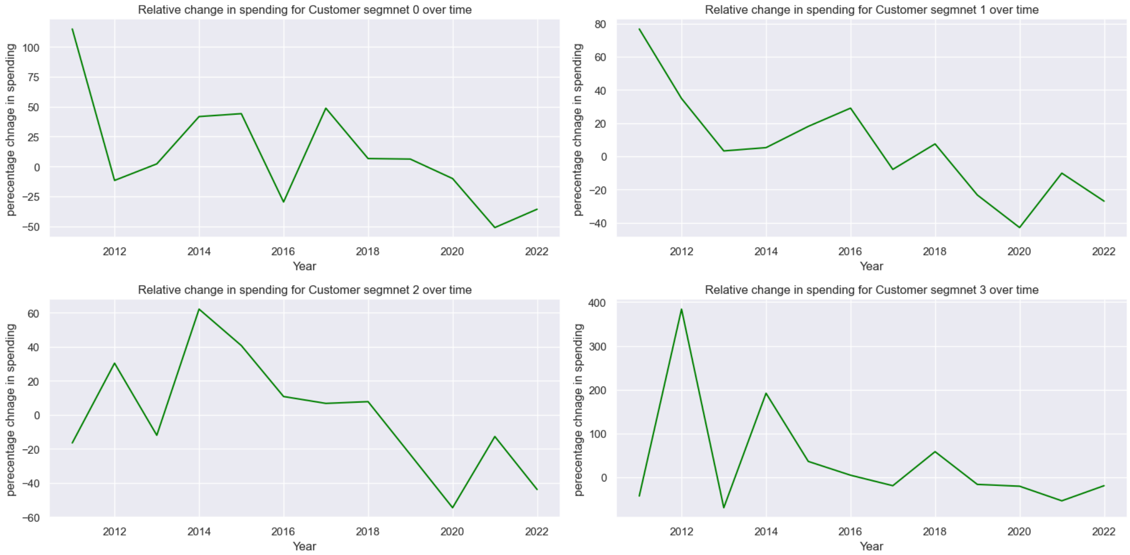


Figure A.5: Comparative analysis of spending trends across four customer segments (2012-2022) using domain knowledge.

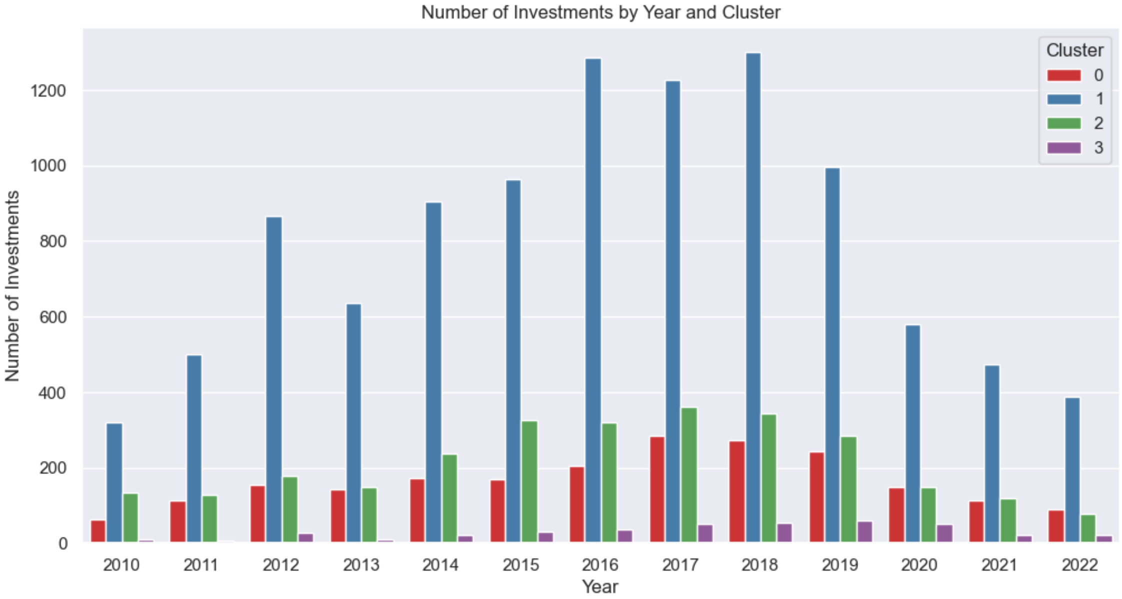


Figure A.6: Annual investment distribution across four clusters (2010-2022) formed using domain knowledge.

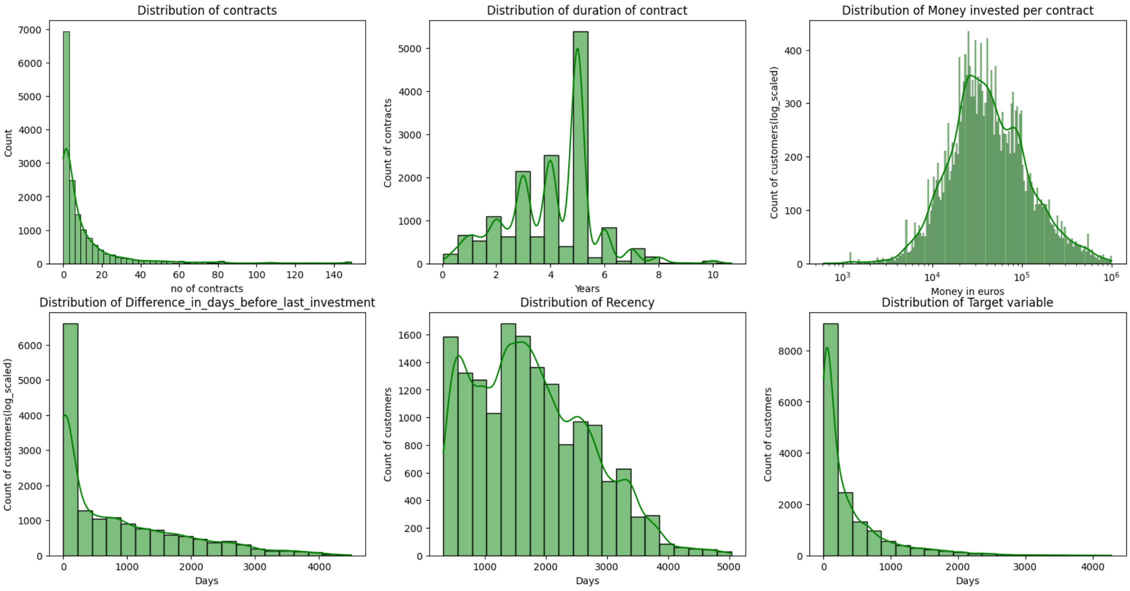


Figure A.7: Distribution of six features formed after feature engineering. This distribution shows that data is highly skewed.

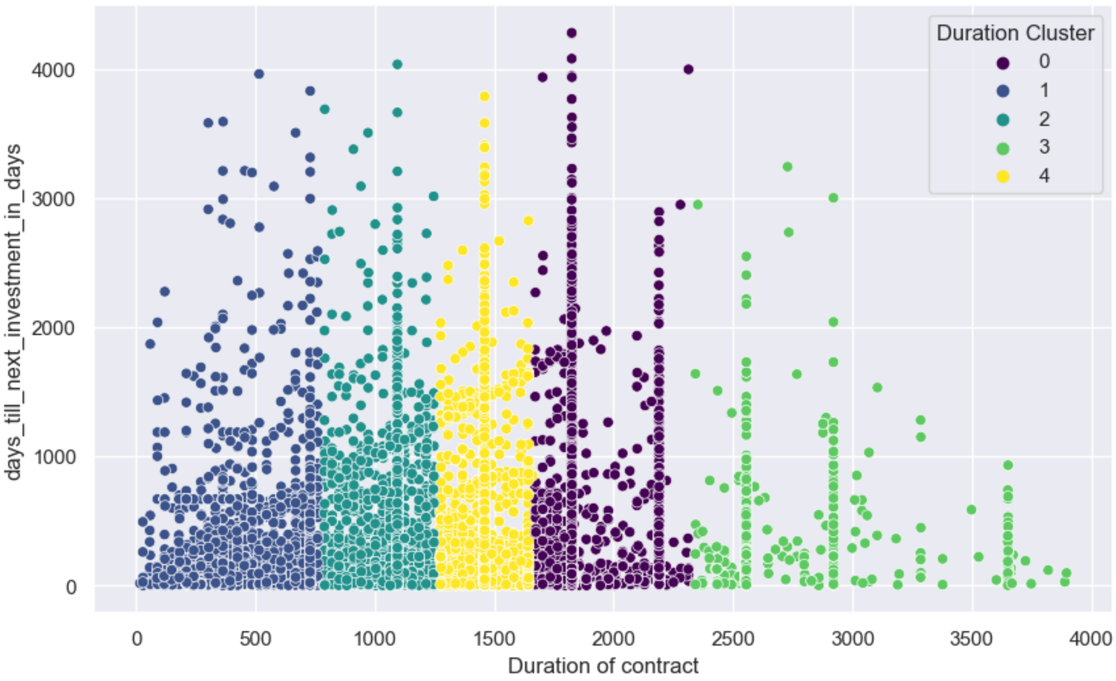


Figure A.8: Distribution of different clusters formed by using k mean clustering where input is the duration of the contract.

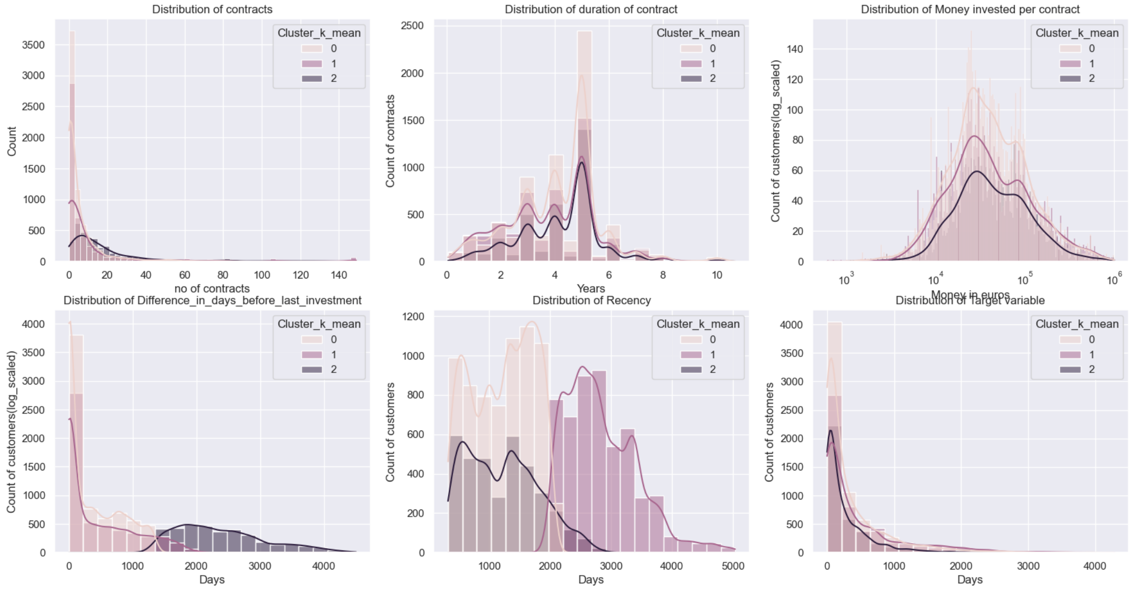


Figure A.9: Distribution of features after clustering using *k*-means clustering. This distribution also reflects the skewness in the data.

A.4. SHAP DEPENDECY PLOT FOR EACH CLUSTER FORMED USING DOMAIN KNOWLEDGE.

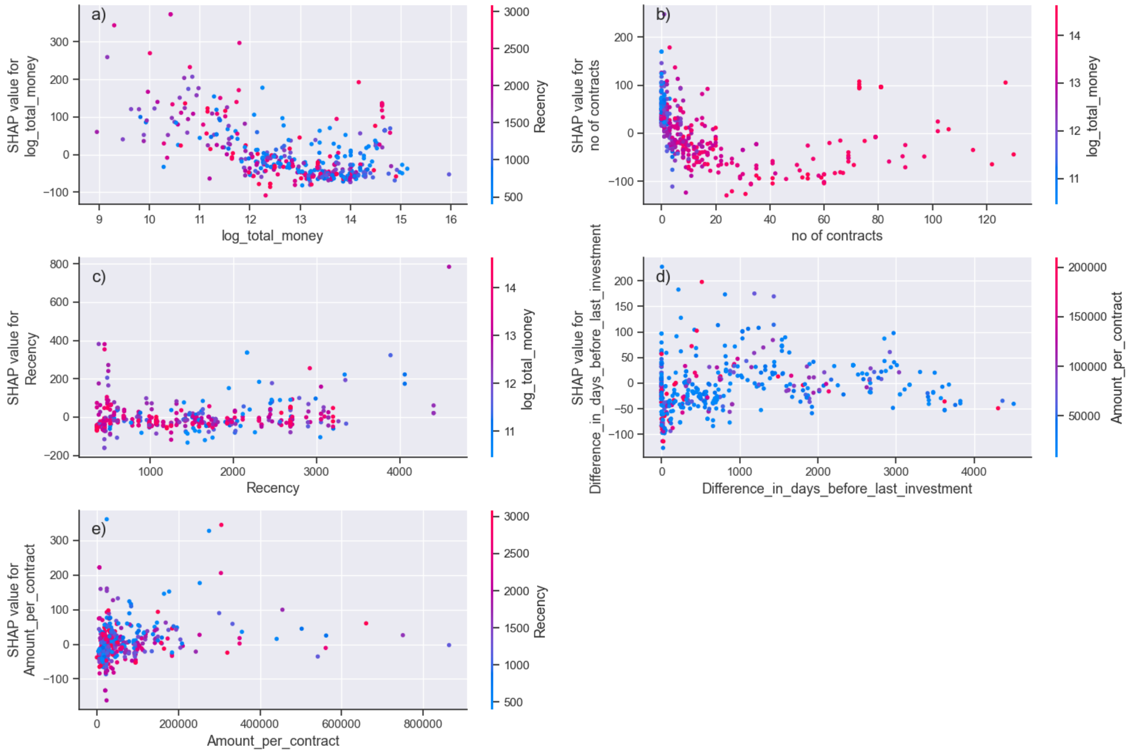


Figure A.10: Dependency plot for the Cluster 0 formed by using business knowledge. The x-axis is the value of the feature value and the y-axis is the SHAP value.

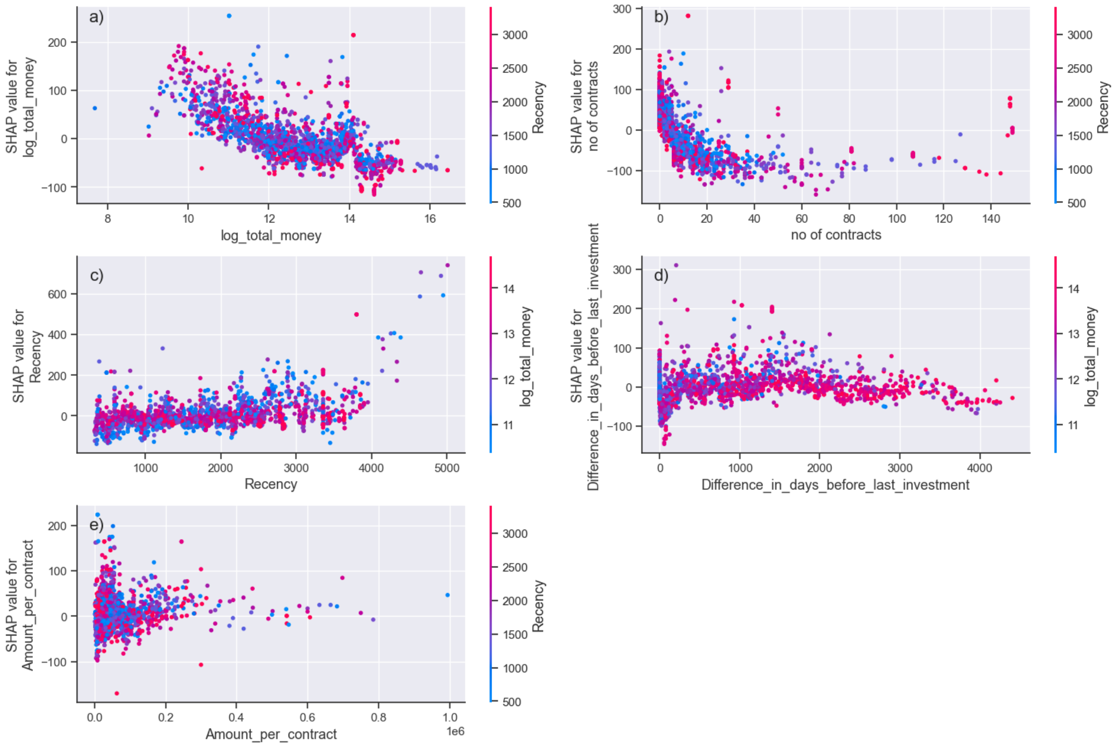


Figure A.11: Dependency plot for Cluster 1 formed by using business knowledge. The x-axis is the value of the feature value and the y-axis is the SHAP value.

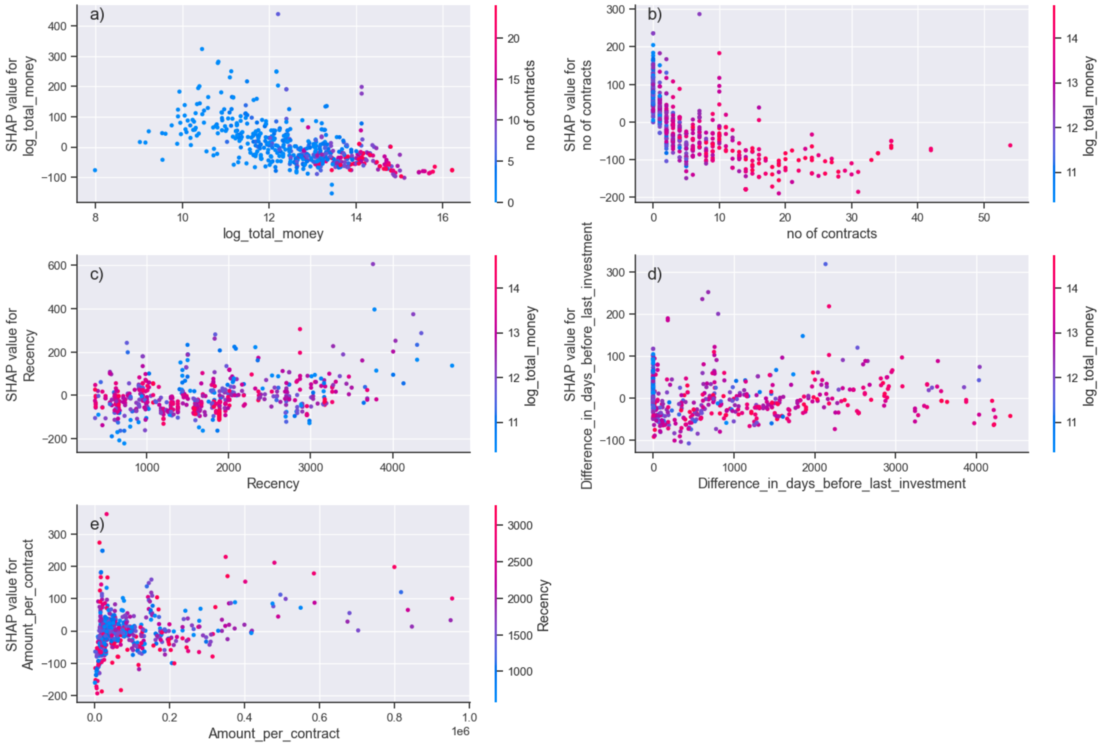


Figure A.12: Dependency plot for Cluster 2 formed by using business knowledge. The x-axis is the value of the feature value and the y-axis is the SHAP value.

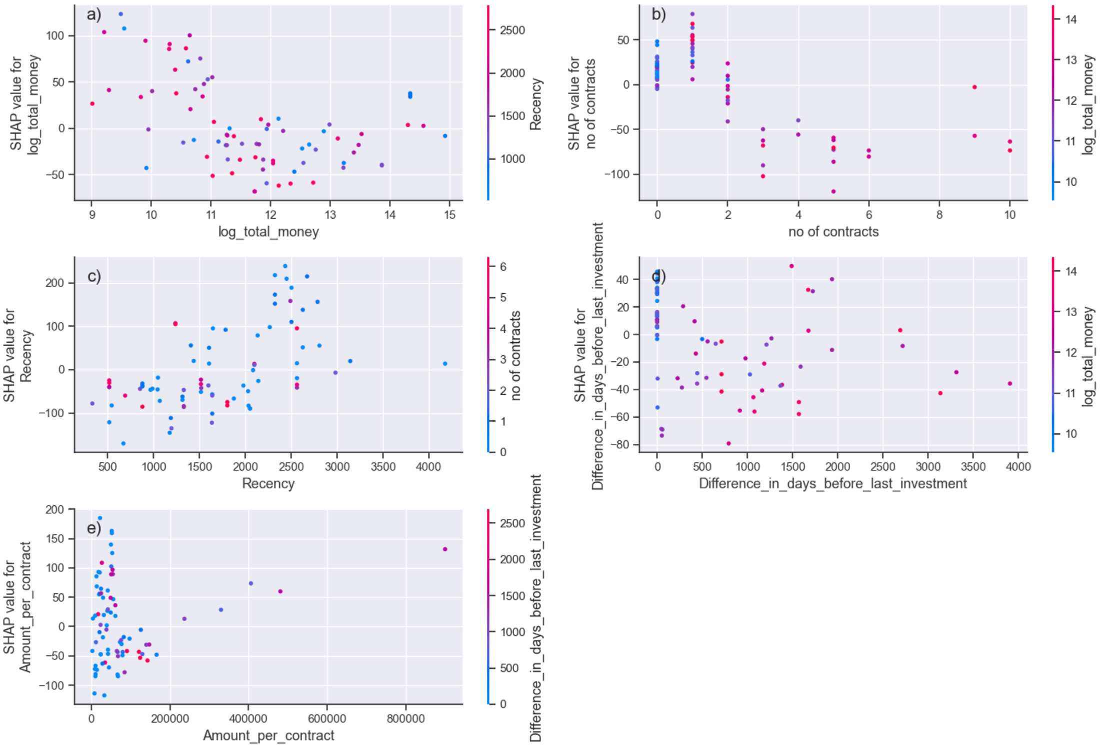


Figure A.13: Dependency plot for the Cluster 3 formed by using business knowledge. The x-axis is the value of the feature value and the y-axis is the SHAP value.

A.5. SHAP DEPENDECY PLOT FOR EACH CLUSTERS OBATINED BY k-MEANS CLUSTERING

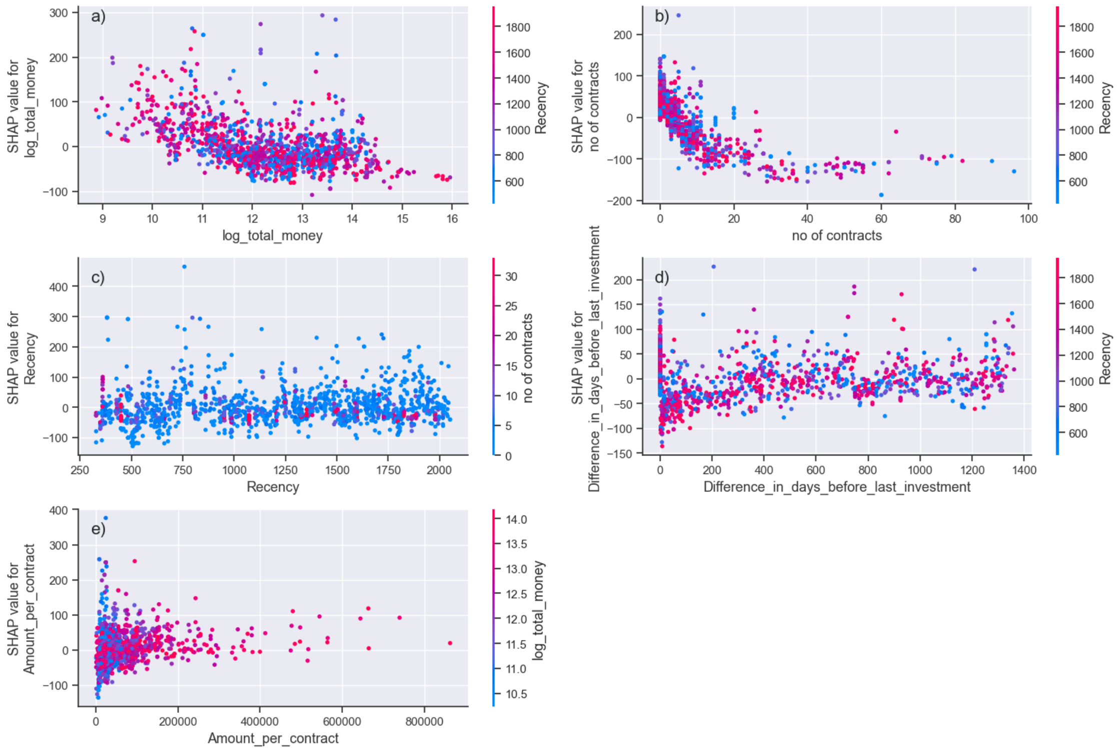


Figure A.14: Dependency plot for the Cluster 0 formed by using k-means clustering. The x-axis is the value of the feature value and the y-axis is the SHAP value.

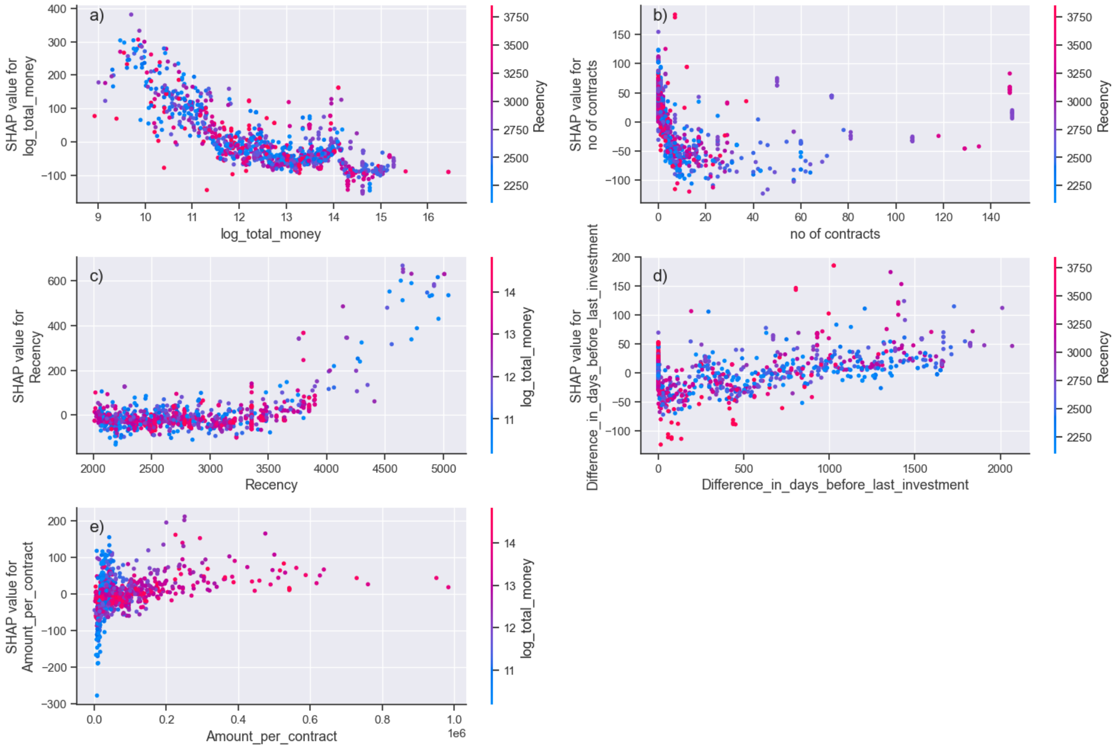


Figure A.15: Dependency plot for Cluster 1 formed by using k-means clustering. The x-axis is the value of the feature value and the y-axis is the SHAP value.

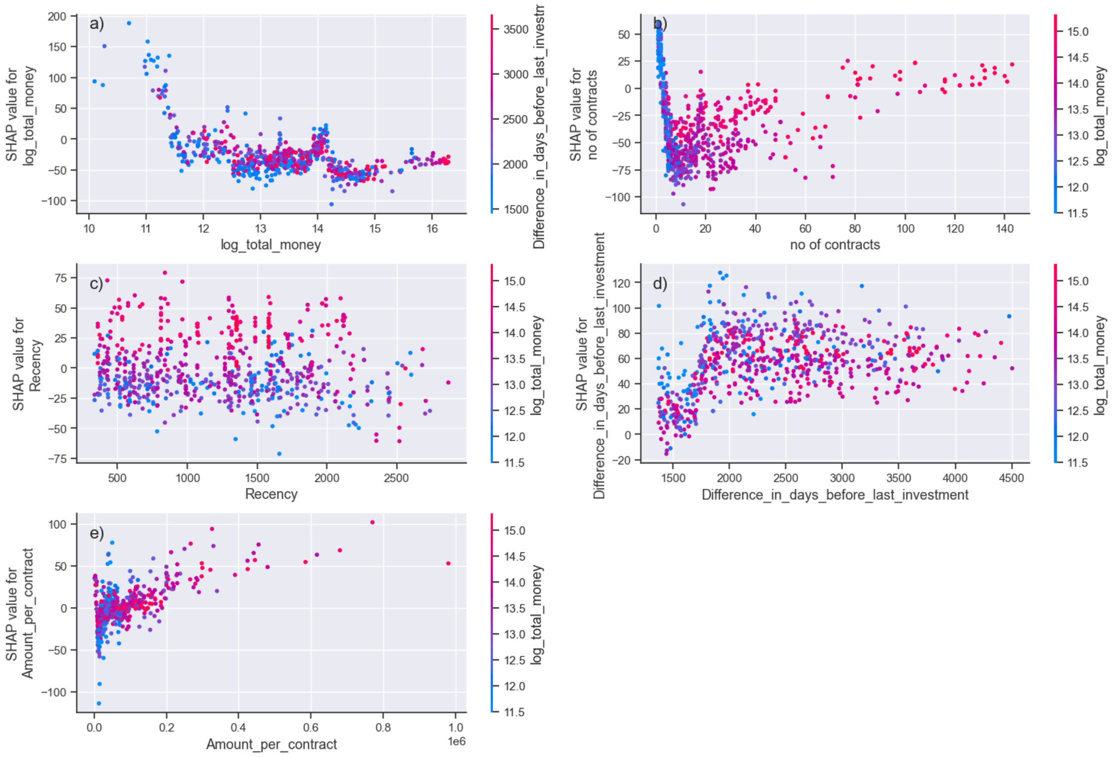


Figure A.16: Dependency plot for Cluster 2 formed by using k-means clustering. The x-axis is the value of the feature value and the y-axis is the SHAP value.