Tilburg University

PREDICTING CROSS-BUYING BEHAVIOR IN THE FINANCIAL SERVICES INDUSTRY

MANON DOUZE

THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE IN DATA SCIENCE & SOCIETY
AT THE SCHOOL OF HUMANITIES AND DIGITAL SCIENCES
OF TILBURG UNIVERSITY

STUDENT NUMBER

1279034

WORD COUNT

8,682

COMMITTEE

dr. M. Postma
dr. M. Klincewicz

LOCATION

Tilburg University
School of Humanities and Digital Sciences
Department of Cognitive Science &
Artificial Intelligence
Tilburg, The Netherlands

DATE

December 3, 2021

PREDICTING CROSS-BUYING BEHAVIOR IN THE FINANCIAL SERVICES INDUSTRY

MANON DOUZE

Abstract

This thesis aims to predict customer cross-buying behavior in the financial services industry. Specifically, customer behaviors and demographics will be investigated in this research. The models logistic regression, k-nearest neighbors and support vector machine are being presented to predict cross-buying actions. For the implementation of these models an external dataset is used to predict cross-buying behavior in a German financial services organization. The dataset contains data for 100,000 customers and covers data for marketing efforts, transaction data, and customer characteristics.

Previous research focused on the effects of customer churn and customer retention, mainly in the retail industry. Increasingly more organizations are pursuing to increase customer value by selling more products or services within the organization. The outcomes of this research could provide an efficient and low-cost manner to identify which customers are most likely to cross-buy. This knowledge will provide an accessible way to identify potential customers, which could be used to steer a firm's marketing efforts accordingly.

The findings of this research suggest that the most important variables predicting cross-buying behavior are related to customer characteristics. Generally, customers who are progressing further on the maturity scale are more likely to cross-buy. Additionally, customer involvement in terms of transaction frequency is an imperative predictor. Customers with higher levels of desktop logins, and cash inflows have shown to be more likely to cross-buy.

DATA ETHICS STATEMENT

Work on this thesis did not involve collecting data from human participants or animals. The original owner of the data and code used in this thesis retains ownership of the data and code during and after the completion of this thesis. The author of this thesis acknowledges that they do not have any legal claim to this data or code. The code used in this thesis is not publicly available. All images used in this thesis were created by the author.

The data used to conduct this research was retrieved from an external source (Harvard Dataverse), which can be downloaded at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/V5NQH7.

CONTENTS

1 INTRODUCTION

Over the past decades, managers have become aware of the fact that solely retaining customers will not be enough to run a successful business (Verhoef, Franses, & Hoekstra, 2001). Since the new millennium, managers have shown significantly more interest in increasing customer value by not merely customer retention and churn prevention. Companies are aiming to increase customer value by selling a wider variety of products and services to a single customer (Verhoef et al., 2001).

However, most previous studies focused on either customer retention or customer behavioral factors such as customer satisfaction and commitment often focused on the retail industry (Estrella-Ramón, 2017; Liu & Wu, 2007). This forms a mismatch between the increasing interest in cross-buying behavior and the currently existing literature regarding this topic.

The phenomenon called cross-buying has been thoroughly researched over the past few decades, however research for the financial services industry specifically is lacking. Regarding the impact of cross-buying behavior a consensus has been reached, supporting the fact that cross-buying behavior has a positive impact on customer lifetime value (Blattberg, Malthouse, & Neslin, 2009). Numerous articles already discuss the effects of cross-buying on increasing customer spending value and purchasing frequency (Dahana, Miwa, Baumann, & Morisada, 2020; Lemon & Wangenheim, 2009; W. Reinartz, Thomas, & Bascoul, 2008).

According to Morisada, Miwa, and Dahana (2018), promotion-induced cross-buying has been addressed frequently, but the effect of different customer demographics has not been investigated. This is a crucial part in examining cross-buying behavior, as it is relevant for managers and society to better understand the results of cross-buying behavior. Creating a better understanding of cross-buying could improve targeting customers for cross-selling efforts by marketers (Li, Sun, & Montgomery, 2011).

In the existing literature, an agreement has been formed concerning the positive impact of cross-buying behavior on customer lifetime value (Blattberg et al., 2009). There are many benefits to cross-buying behavior, which makes it an interesting field of research. Previous studies have shown that cross-buying increases customers' purchasing amount and purchasing frequency (Lemon & Wangenheim, 2009; W. J. Reinartz & Kumar, 2003; Zeithaml, Lemon, & Rust, 2001), customer profitability (Garland, 2004; Hallowell, 1996), and customer retention (W. J. Reinartz & Kumar, 2003).

The goal of this research is to find out what demographics and behaviors would make customers more likely to cross-buy, in order to identify relevant customer characteristics which could be used for marketing purposes. From an ethical perspective, this research provides an efficient way to cluster groups of customers and provides more reliable outcomes in comparison to previous research.

In the current study the goal is to find a more cost-effective and efficient way to target customers by creating models that would identify these customers. Different classification models will be built and trained in order to compare its outcomes to already existing classification models created on a similar dataset in the research conducted by Alves Werb and Schmidberger (2021).

The motivation for this research is to simplify the currently existing models and to create models that could be applied to real-world data. With the successful application of these models, it could cater to the financial organization's needs for analyzing customer data in an efficient and low-cost manner.

The scientific relevance of this research could be found in the simplification of already existing algorithms in the current research (Alves Werb & Schmidberger, 2021). Making the algorithms in this research less complex could reduce action space and could limit the long calculation times (Gentsch, 2018). A less complex outcome could be more valuable to employees who have less knowledge about artificial intelligence and can therefore be easier to interpret.

1.1 *Current Research*

For this research it will be investigated what customer profiles are more likely to cross-buy in a financial services company. A customer profile will be defined based on customer demographics and behavior. The outcome of this research could potentially increase profits within a company by focusing on the right customers, and would therefore be relevant to the research field of business. The importance of acquiring more knowledge about cross-buying and its potential benefits are explained in the research by Dahana et al. (2020).

This research will be conducted using the dataset provided in the article by Alves Werb and Schmidberger (2021). The proposed models in the article will be used as comparison models, in order to analyze to which extent it is possible to alter and improve those models adopting a more practical approach. Ultimately, developing practical models that would analyze cross-buying behavior in a simple, yet accurate way would be most beneficial to the financial services industry and the organization's management teams. The goal of this research is to provide an understandable, yet effective way to analyze customer data and identify potential cross-buying customers.

The general problem statement that will be pursued is as follows:

> *What customer behaviors and demographics affect the customer's cross-buying actions?*

The following research questions follow from the problem statement:

RQ1 *To what extent can cross-buying behavior in a financial services company be predicted by direct mailing campaigns?*

RQ2 *To what extent can customer characteristics such as age, education, and gender be used to predict cross-buying behavior?*

RQ3 *How do external environmental factors as city size and living duration affect cross-buying behavior?*

1.2 *Findings*

The main findings of this research were the positive effects of age and direct mailing. In other words, the higher a customer's age, the more likely they are to cross-buy. The same goes for direct mailing, the more emails a customer received, the more likely they are to cross-buy. Furthermore, a strong positive effect has also been found for the number of desktop logins on cross-buying behavior. On the other hand, the research was not affirmative of a positive effect of the variables city size and living duration. In this research it could therefore not be concluded that there exists a positive effect for these variables.

2 RELATED WORK

In the following section, previous research regarding this topic is being discussed. In subsection 2.1, the research area to which this thesis could be applied is discussed. Afterward in subsection 2.2 previous studies related to this research are being presented. Lastly, in subsection 2.3 it is specified in which ways the gaps in previous research are filled.

2.1 *Research Area*

The area of research to which this thesis contributes is the field of marketing research. Research regarding cross-buying behavior is of major importance to marketing professionals and academics (Mansouri, 2021). Cross-buying is often considered a pivotal predictive modeling problem in the field of marketing research (Alves Werb & Schmidberger, 2021). Previous research in the area mainly focused on customer retention and behavior, besides that the research was most often applied to the retail industry (Verhoef et al., 2001).

Up until now, demographics were not a prominent factor in cross-buying research within the financial services industry. Previous research mainly centered around customer behavioral factors such as customer satisfaction and commitment (Estrella-Ramón, 2017; Li, Sun, & Wilcox, 2005). This forms a gap in the current research, as a general conclusion could not be drawn.

Considering the rise of importance for increasing customer value by cross-buying over the last 20 years, the added value to the already existing marketing research in the financial services industry could be significant (Verhoef et al., 2001). This is because cross-selling continues to be of first concern in various industries, including the financial services industry (Li et al., 2011).

This thesis will add knowledge to this field as it will highlight certain customer behaviors and demographics, in order to indicate which customers would be more likely to participate in cross-buying behavior. This will enable marketers to create customer profiles at which their marketing efforts could be targeted. Moreover, it adds to the current marketing research spectrum as most of the current research regarding customer behavior does not consider cross-buying behavior.

2.2 *Previous Studies*

Over the past few years, cross-buying has been a well-investigated topic of research. The studies by Verhoef et al. (2001) and Kamakura, Ramaswami, and Srivastava (1991) focused on predicting cross-buying behavior for customers. These studies have found positive effects for financial maturity and customer satisfaction, however not for cross-buying specifically.

Another research focused on ensemble methods and the evaluation of model performances by predicting cross-buying behavior (Alves Werb & Schmidberger, 2021). The research uncovered a substantial effect for customers' city size and the living duration on cross-buying behavior, as proposed in research question 3. In the article it is suggested that this would be an interesting topic for future research.

Secondly, it is argued that there exists a positive effect of income, age, employment status, and education on financial maturity (Dahana et al., 2020). Financial maturity increases when a household's funds are allocated to satisfy a higher-order investment, after the basic objectives are met at first. This could lead to cross-buying behavior. This is in accordance with the findings of Verhoef et al. (2001), who stated that customer age also has a significant positive effect on cross-buying behavior. These two papers both relate back to research question 2.

The monetary value of customers is also linked to buying more services or products from the same financial services provider (Estrella-Ramón, 2017). Customers with a higher amount of average monthly assets and liabilities are more likely to acquire other products or services.

According to Li et al. (2005), male households, or households with a higher level of education are more likely to progress further on the financial maturity scale. Older households and higher income households also progress further and quicker on the financial maturity scale. However, for promotion-induced cross-buying the effect of age was negative. Meaning that older people would show a decrease in purchase amount as a result (Morisada et al., 2018). This research showed a negative effect for cross-buying on purchase frequency for males. However, this research does provide a hypothetical answer to research question 2.

Lastly, the adoption of online banking could be a predictor for both loyalty and cross-buying behavior. It is shown that acquiring multiple products or services within the company has a significant effect on adopting online banking, which could suggest a correlation between online banking and cross-buying behavior (Estrella-Ramon, Sánchez-Pérez, & Swinnen, 2016).

Finally, according to Kumar, George, and Pancras (2008), a potential positive effect exists between the effects of direct mailing campaigns and cross-selling behavior. This study investigates the effectiveness of cross-buying behavior by identifying the customer incentives, which corresponds to research question 1.

Generally, most studies have found a significant result for the effects of customer characteristics, customer transaction data and the extent to which a firm executed its marketing strategies (Alves Werb & Schmidberger, 2021).

2.3 *Model Improvement*

The research conducted by Alves Werb and Schmidberger (2021) focuses on creating ensemble methods in order to predict customer's cross-buying behavior. For this thesis the models proposed in their research are being revisited, in order to adjust the models and improve them according to the scope of this research.

The main goal of the research conducted by Alves Werb and Schmidberger (2021) was to investigate the effects of ensemble methods when predicting cross-buying behavior. Additionally, the research pursues the goal of being able to identify customer profiles, rather than the specific effects of a certain combination of models.

The reason for creating new models was to create simple, yet interpretable models, that could be used by managers or marketeers in order to identify which customers to target within a financial services company. The models in previous research established by Alves Werb and Schmidberger (2021) are more difficult to interpret, as four methods are being used to assess the average size, direction, heterogeneity, and variable importance.

The importance of creating well interpretable models is important when establishing a prediction task (Little, 2004). A model that will be utilized by managers and marketing teams needs to be simple, adaptive, easy to interpet, and as complete as possible. The goal of this research is to create a model that would adhere to the needs from the business perspective of this research.

3 METHODS

In this paper three different approaches are considered for predicting cross-buying behavior. The aim of these three models is to predict cross-buying behavior as accurately as possible, in order to provide a model that could be applied to empirical data. The baseline to compare these models with is retrieved from the logistic regression model that can be found in the paper written by Alves Werb and Schmidberger (2021). The dataset used for the research in this thesis originates from an external source (Harvard Dataverse) and was provided by Alves Werb and Schmidberger (2021).

The models created for this research provide an improved performance to the models created in the paper by Alves Werb and Schmidberger (2021), with a more practical and simple application. This practical application means that the created models would better fit to abstract or scarce data. It could be beneficial for models to provide a more accurate prediction with a smaller quantity of data, or data scarcer in its features. The models could be more efficient in training time, data preprocessing and optimization.

The three different approaches are executed by three different techniques, namely logistic regression, k-nearest neighbors (kNN), and support vector machine (SVM).

3.1 *Models*

In this study the following models are being established:

1. A logistic regression model being trained on customer characteristics, marketing efforts, and transaction data. (Model 1)
2. A k-nearest neighbors model being trained on customer characteristics, marketing efforts, and transaction data. (Model 2)
3. A support vector machine being trained on customer characteristics, marketing efforts, and transaction data. (Model 3)

The decision for these three supervised learning methods is based on the fact that kNN and SVM demonstrate certain important trade-offs within the field of machine learning. A balance can be found in the SVM model, which could be considered as computationally less demanding, and provides easier interpretable results. Whereas the kNN model is more exhaustive, however its strength lays in the fact that it is able to find more complex patterns (Bzdok, Krzywinski, & Altman, 2018).

The logistic regression technique was selected because it is an efficient model which is easy to interpret and train. Moreover, it is a model often used in regression-based research and could therefore be a benchmark for future research.

3.2 *Predictors*

The predictor variables originating from the external dataset provided by Alves Werb and Schmidberger (2021) and used to build these models could be divided into three categories. The first category being transaction data, this category provides a broader insight in the customer's interaction within the financial services organization and how many products or services the customer bought. The second category represents the marketing efforts performed by the financial services company. Lastly, the third category exemplifies customer characteristics, which are the demographics belonging to a single customer. In Appendix A on page 47 an overview is being presented with the variables and their definitions that were used in this research.

3.3 *Algorithms*

Logistic Regression

The logistic regression model is selected for this research as it is a model similar to the baseline model created by Alves Werb and Schmidberger (2021). For that reason, this model will serve as a benchmark to compare the performance of the other models to. Moreover, the logistic regression model is a common model in regression-based research studying cross-buying behavior in customers (Alves Werb & Schmidberger, 2021; Larivière & Van den Poel, 2005; Prinzie & Van den Poel, 2008). Therefore, the logistic regression model will not only serve as a benchmark for the other models discussed in this thesis but will also serve as a benchmark for future research.

The logistic regression is used to model the probability of a certain class for classification problems. Therefore, the outcome of the logistic regression model is the probability (or "odds") that the given model input is relevant to a specific class. Whereas for linear regression the relation between the output and the input of the equation is calculated, a logistic regression assumes that the output falls in the range of [0,1].

This could be represented by using the following formula, in which β_0 designates the intercept, β designates the slope, and X designates the predictor variable (Molnar, 2020).

$$P(Y_i = 1) = \frac{1}{1 + exp(-(\beta_0 + \beta_1 X_i + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5))} \quad (1)$$

K-Nearest Neighbors

The k-nearest neighbors algorithm is used in this research to improve the logistic regression model. The kNN algorithm is a suitable model for this, as it has a short training time, it is very well interpretable, and contributes to high model accuracy.

The kNN algorithm can be used as a classification, as well as, a regression algorithm. K-Nearest Neighbors uses instance-based learning, which means the algorithm compares new instances found in the training set to instances which have already been stored in memory. By using the collected data from the training set, the algorithm can make predictions for unseen instances. In this thesis, the kNN output is a class membership, which means an instance is classified based on the majority vote of its neighbors measured by a distance function. The kNN algorithm in this research is used to classify the target variable of cross-buy, by using the predictor variables to assign the most common target variable class to.

The default distance method for continuous variables being used in kNN is Euclidean distance:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2} \quad (2)$$

Support Vector Machine

For the third model, a support vector machine is being used. This model provides a memory-efficient way of classifying an instance that can be separated into classes. The algorithm is suitable to this dataset as the target variable has a clear margin of separation.

Similar to the kNN algorithm, support vector machine is also a supervised machine learning algorithm, that can be used for classification and regression challenges. In a set of training data, each individual instance is being classified. Accordingly, a model is being built in order to assign new instances to one class or the other. A hyperplane is a surface used to separate the two classes. For this research a radial kernel is used. This type of kernel is used when the data is not linearly separable.

The equation used to define a hyperplane, in which w represents the weight vector, x represents the input vector, and b represents the bias, is as follows:

$$w^T x + b = 0 \tag{3}$$

From this equation the separation into two classes could be defined as:

$$w^T x + b \geq 0 \text{ for cross-buy = + 1} \tag{4}$$

$$w^T x + b \leq 0 \text{ for cross-buy = - 1} \tag{5}$$

3.4 *Package References*

The following packages are used to create the models for this research:

- car
- caret
- data.table
- dplyr
- DMwR
- ggplot2
- graphics
- MLmetrics
- pROC
- stats

4 EXPERIMENTAL SETUP

This section will elaborate on the dataset and the experimental procedure used to conduct this research. In section 4.1 the raw data will be described, hereafter in section 4.2, the data preprocessing techniques will be portrayed. In section 4.3 the implementation of the models will be discussed. Lastly, in section 4.4 an overview of the evaluation criteria for the models will be presented.

4.1 *Raw Data Description*

For this thesis the "cross_sell_dataset.csv" is being used. This dataset originates from a paper written by Alves Werb and Schmidberger (2021) on Ensemble Methods for Response Modeling. The dataset contains anonymized empirical data for 100,000 customers of a well-known financial services provider in Germany, containing 34 explanatory variables and one target variable.

The target variable denotes the probability of a customer opening a second checking account, which is recognized as an action of cross-selling. The dataset is formatted in a CSV format. It was uploaded to Harvard Dataverse, an external research repository with data and code available.

The predictors specify the transaction data, the firm's marketing efforts, and customer demographics. The data is collected at one initial time point, therefore a customer's cross-buying decision in the future cannot influence current data. In Appendix A (page 47) a table is provided containing the definitions and interpretations of the used variables. In Appendix B (page 49), a table is displayed describing the data class, score range, mean, and missing data percentage for each variable. Lastly, in Appendix C (page 50), the variable distributions for the categorical and continuous variables are visualized.

Data Imbalance

Next to these raw data characteristics, one remarkable characteristic of the data is its high class imbalance. The dataset is imbalanced with 90% of the target variable being 0 (no cross-buy) and 10% being 1 (yes cross-buy).

A dataset could be considered as imbalanced if the minority class is underrepresented in comparison to the majority class, and the majority class is overrepresented in comparison to the minority class. According to Ramyachitra and Manikandan (2014), there are two main approaches to deal with this kind of data; either preprocess the data in a way that decreases the effect of the data imbalance, or use inventive algorithms that take the data imbalance into consideration.

For this research the data is being preprocessed so that the effect of the data imbalance is diminished. For each of the three models proposed in this study, there is a distinction made in three sampling methods being used for each of the models. This is to show the effect that a different sampling technique could have and to demonstrate which sampling techniques are appropriate for which algorithms. The three sampling techniques used are Synthetic Minority Oversampling Technique (SMOTE), oversampling, and undersampling.

The SMOTE technique is often regarded as an effective, yet simple oversampling method. The SMOTE algorithm tries to synthetically create a new positive sample between two instances (Sun, Lang, Fujita, & Li, 2018). The oversampling and undersampling techniques both simply collect more instances from the minority or majority class, resulting in an equal representation for both classes.

In many studies, it is denoted that class imbalance can affect a model's predictive performance (Prinzie & Van den Poel, 2008). Following from previous research the class imbalance is only altered for the training set to prevent overly optimistic results (Alves Werb & Schmidberger, 2021). In other words, the testing set will still comprise of the 90%/10% split. In this way, we prevent data leaking from the training to the testing set.

The presence of class imbalance in this dataset could be related to the fact that the act of cross-buying in the financial services industry is often done by customers who are further progressed on the financial maturity scale (Knott, Hayes, & Neslin, 2002; Li et al., 2005). As cross-buying is defined as the customer opening an additional checking account, this could simply mean that the vast majority of the organization's customers only possess one checking account.

4.2 *Data Preprocessing*

The data was preprocessed in the same manner for logistic regression, kNN, and SVM. Originally, the dataset consisted of 34 predictor variables and one target variable. The first step for preprocessing the data was to fit our models in order to analyze the variables. By checking the data distribution, the amount of missing data, and the application of each of the variables, the variables could be correctly transformed.

Firstly, the variables representing occupation, share new cars, share new houses, joint account, and marital status were discarded from the dataset. The reason for this is that these variables would not affect the outcome of this research. Moreover, these variables showed high percentages of missing data (more than 10%) and would therefore negatively influence the reliability of our results. The percentages of missing data for the remaining variables can be found in Appendix B (page 49).

As the remaining variables were numeric in the original dataset, the binary variables were converted into factors to prevent them from being treated as continuous variables. This means the variables denoting cross-buy, academic title, gender, get member active, get member passive, and giro mailing were converted.

The segment variables (city size, house size, purchase power, car power, and living duration) were treated as continuous variables for this research. These segment variables represent a scale, as they are all ordered from low to high. As the distributions of these variables are not severely skewed and they all represent a substantial amount of categories, it is recommended to treat these ordinal variables as continuous (Robitzsch, 2020).

For the kNN model specifically, binary predictor variables (academic title, gender, get member active, get member passive, and giro mailing) were one-hot encoded. Since kNN is a distance-based algorithm, binary variables that will be treated as factor variables are assumed to have a natural ordering (0 or 1). Usually, the k-nearest neighbors algorithm does not deal well with categorical variables as it is difficult for most machine learning algorithms to work with labeled data (Sanjar, Bekhzod, Kim, Paul, & Kim, 2020). As there exists no ordinal relationship in these predictor variables, one-hot encoding is being applied to prevent these binary variables from influencing our results in a negative way. One-hot encoded variables are used to represent a category, by creating a new category row and assigning a 1 or 0. This depends on whether this category is present or not. In the end, each new category row represents whether that category type is applicable or not.

The next step in preprocessing the data is creating a train test split. By dividing the data in a training and testing set, this will allow the models to be trained on the training set and to be fitted to the testing set. The testing set represents the unseen data and will form a way to evaluate how the trained algorithm will perform to real-world data. For this research the data is being randomly split in 80% of the observations for training data and 20% for testing data.

As shown in Appendix C on page 50 the majority of the variables are not normally distributed. Most variables such as customer tenure, last account, desktop logins, and brokerage are positively skewed. To normalize the range of these variables feature scaling is applied. As distance-based algorithms are being used in this thesis, it is important to scale the independent variables. If variables fall in the same range, each feature will equally contribute to the final distance function. For this thesis the min-max normalization method is used to perform feature scaling on the predictor variables for the SVM and kNN model, which could both be considered as distance-based algorithms.

According to Patro and Sahu (2015), min-max normalization has an excellent performance in support vector machines. Moreover, it performs well in regards of the cross-validation accuracy, support vector quantity, and it is time-efficient. Min-max normalization transforms the minimum value of a certain feature to 0, and the maximum value to 1. Ultimately, all observations of a certain variable will fall within the range of [0, 1]. For the implementation of the min-max normalization in the preprocessing steps of this study, the values of the training set are first being transformed in the manner that they will fall into the [0,1] range. Afterwards these exact settings are being applied to the testing set, in order to prevent data leakage to the testing set which will bias the results. In other words, the minimum and maximum values of the training set will be used to apply feature scaling on the testing set.

Ultimately, the three sampling methods will be used to establish a 50/50 data balance. For the SMOTE sampling technique, the minority class is oversampled by 100% and the majority class is undersampled by 200%. For the oversampling technique, the minority class is being oversampled. This means that more instances are being selected randomly from the minority class than from the majority class, in order to establish a balanced dataset. Subsequently, in the undersampling technique, the majority class is being undersampled. Namely, more instances are being collected in a random approach from the majority class in comparison to the minority class.

4.3 *Experimental Procedure and Implementation*

In this subsection, the experimental procedure will be further elaborated on. The specific implementation of the algorithms used to conduct this research will be thoroughly discussed, as well as, the specific parameters that were chosen. Furthermore, all of the algorithms discussed in this section were being implemented in the RStudio IDE, using the R programming language. The packages used to run the actual algorithms are mentioned in this section. A list of all packages being used for this research can be found in subsection 3.4.

Logistic Regression

After the data preprocessing, the three algorithms are being trained and then fitted to the testing set. For the logistic regression, the target variable is being run on all predictor variables. To run this model the "glm" function from the "stats" package is being used. The hyperparameter family is denoted as binomial, and the scaled training set is used as its input.

K-Nearest Neighbors

For the k-nearest neighbor algorithm the seed is first set, in order to be able to make the results reproducible. The following step is to set up the "trainControl" function, from the "caret" package. This function is being used to regulate the parameters of the "train" function, which will be used next. For the "trainControl" function used for this algorithm, it is specified which cross-validation method is going to be used.

For the kNN model the repeated k-fold cross-validation is being used. Cross-validation is used to analyze and evaluate algorithms by dividing the data into two sections: the training and testing set. For k-fold cross-validation, the dataset is being split into *k* equal parts, after which multiple iterations are being performed (Refaeilzadeh, Tang, & Liu, 2009). For each iteration a different part of the data is being used for the validation of the training set. For the "trainControl" function the number of iterations is set to ten (number = 10). The number of repeats is set to three (repeats = 3), meaning that three full sets of folds will need to be calculated.

Subsequently, the kNN model is trained on the training set. To execute this step the "caret" package is used. For this classification task, the dependent variable is being run on the independent variables, using the formerly established cross-validation method. Naturally, the algorithm is being trained on the training set. Furthermore, the tune length is being set at 20 (tunelength = 20). This number represents the quantity of granularity in the tuning parameter. The final step after training the model is to fit the model to the testing set. This step is performed with the usage of the "predict" function from the "stats" package. After fitting the model to the unseen data, the model results could be collected.

Support Vector Machine

For the final algorithm, the same method of k-fold cross-validation is used as for the kNN algorithm. The reason for this is to be able to provide more consistent results throughout this research and improve the manner in which both algorithms could be interpreted and compared.

After the "trainControl" step, which conducts the repeated cross-validation with equal hyperparameter settings (number = 10, repeats =

3), the model is being trained on the training set. This step is performed using the "train" function from the "caret" package, a tune length of ten (tunelength = 10), and a radial kernel as our data is non-linear. In the following step the model is being fit to the testing set using the "train" function from the "stats" package, after which the results of the performance of the algorithm could be analyzed.

4.4 *Evaluation Criteria*

The research conducted in this thesis regards a classification task for the cross-buy variable. The error measures being used to evaluate these models are accuracy, specificity, recall or sensitivity, precision, F1 score, and the area under the receiver operator characteristic curve (AUC-ROC). As this research is dealing with an imbalanced dataset, the ROC curve will be one of the more informative assessment criteria (He & Garcia, 2009). However, since the training set is balanced out using SMOTE sampling, undersampling, and oversampling methods, other evaluation metrics are being applied as well. As the models created in the research by Alves Werb and Schmidberger (2021) are evaluated using recall, F1 score, and accuracy, these metrics are included in this research to provide a benchmark in comparison to previous research.

Moreover, a confusion matrix is created for each model individually, to assess the binary prediction task for cross-buying. In Table 1, it is visualized where the above mentioned evaluation metrics originate from. The confusion matrix represents accuracy *((TN + TP)/total))*, specificity *(TN/(TN + FP))*, sensitivity or recall *(TP/(TP + FN)*, and precision *(TP/(TP + FP))* (Alves Werb & Schmidberger, 2021).

	Actual No Cross-Buy (0)	**Actual Yes Cross-Buy (1)**
Predicted No Cross-Buy (0)	True Negative (TN)	False Negative (FN)
Predicted Yes Cross-Buy (1)	False Positive (FP)	True Positive (TP)

Table 1: Confusion Matrix for Predicting the Cross-Buy Variable

Furthermore, the F1 score could be considered as the harmonic mean of recall and precision. It recognizes the ratio between false negatives and false positives (Alves Werb & Schmidberger, 2021). The F1 score could be computed as *(2 * (precision * recall)/(precision + recall))*.

Moreover, the area under the ROC curve is used, as it is a good way to measure a classifier's performance (Bradley, 1997). The ROC curve plots the probability of the extent to which a model is able to separate the two classes. Consequently, the higher the AUC-ROC score, the better the model's performance.

Lastly, the variable importances are provided for the individual results. The importance for each of the features will simplify the extent to which a certain result could be attributed to a feature. Most importantly, this will be helpful to answer the research questions in this thesis. Variable importances could be interpreted as the sum of the decline in the error rate due to a certain variable. For the kNN and SVM model this is represented as a value between 0 and 1, which is the result of dividing the variable importance by the highest variable importance. For the logistic regression model, the values could be interpreted as the absolute value of the t-statistic.

5 RESULTS

In this section, the results of the created models are discussed. In subsection 5.1, the baseline model is being discussed. Secondly, in subsection 5.2, the results from the logistic regression model are being presented. In subsection 5.3, the results of the k-nearest neighbors algorithm are being shown. In subsection 5.4, the results for the support vector machine are discussed. Lastly, in subsection 5.5, an overview of the results for the three models combined is presented.

5.1 *Baseline Model*

The baseline model originates from the research conducted by Alves Werb and Schmidberger (2021). In Table 2, the results for the logistic regression model that is used as the baseline for this research is shown. These results are based on the realistic distribution of the cross-buy variable (90/10 split) (Alves Werb & Schmidberger, 2021). Furthermore, the results are in line with the outcomes of previous studies, which vary from 38.3% to 74.5%. A study conducted by Kumar et al. (2008), found that there was an accuracy of 71% on the holdout data. However, the research conducted by Knott et al. (2002), found an accuracy ranging from 38.3% up until 55.1%. Lastly, Larivière and Van den Poel (2005) were able to obtain an accuracy of 74.5% (Alves Werb & Schmidberger, 2021).

Model	**Accuracy**	**Recall**	**F1 Score**
Logistic Regression	68.1%	66.2%	29.0%

Table 2: Results for the Logistic Regression Model for the testing (Reprinted from Alves Werb and Schmidberger (2021))

5.2 *Logistic Regression*

After training and fitting the logistic regression model, the results were interpreted using a confusion matrix. These confusion matrices are being visualized in each section. Moreover, the variable importances are provided in each section for the five most important variables. In subsection 5.2, the results for the SMOTE sampling technique are shown. Then, on page 22, the results for the undersampling method are provided. Lastly, on page 23, the results for the oversampling technique are given. The evaluation metrics that could be computed from the three sampling methods are shown in Table 9.

SMOTE sampling

In Table 3, the confusion matrix is shown for the SMOTE sampling technique. The adequate evaluation metrics can be found at the end of this section, in Table 9.

	Actual No Cross-Buy (0)	**Actual Yes Cross-Buy (1)**
Predicted No Cross-Buy (0)	12,198	834
Predicted Yes Cross-Buy (1)	3,467	755

Table 3: Confusion Matrix for the Logistic Regression Model Using the SMOTE Sampling Technique on the Testing Set

In Table 4, the five most important variables are displayed. The values given could be interpreted as the absolute value of the t-statistic. The variables age and city size relate back to our research questions regarding demographics and external environmental factors.

Variable	**Importance**
Age	37.3%
House Size	16.7%
Gender (male)	13.2%
City Size	12.9%
Desktop Logins	12.7%

Table 4: Top 5 Variable Importances for the Logistic Regression Using SMOTE Sampling Technique

Undersampling

In the confusion matrix displayed in Table 5, the ratios for the testing set are shown, using the undersampling method. The evaluation metrics that could be computed from this confusion matrix are again shown in Table 9.

	Actual No Cross-Buy (0)	**Actual Yes Cross-Buy (1)**
Predicted No Cross-Buy (0)	9,964	530
Predicted Yes Cross-Buy (1)	5,701	1,059

Table 5: Confusion Matrix for the Logistic Regression Model Using the Undersampling Technique on the Testing Set

In Table 6, the five most important variables are displayed for the undersampling method. The variable age is applicable to our research questions as this is a demographic variable that could affect cross-buying behavior.

Variable	Importance
Age	22.3%
Giro Mailing (yes)	6.7%
Last Account	5.9%
Brokerage	4.5%
Get Member Active (yes)	3.5%

Table 6: Top 5 Variable Importances for the Logistic Regression Using Undersampling Technique

Oversampling

In the confusion matrix displayed in Table 7, the outcomes of the logistic regression model on the testing set are shown, using the oversampling method. The evaluation metrics that could be computed from this confusion matrix are again shown in Table 9.

	Actual No Cross-Buy (0)	Actual Yes Cross-Buy (1)
Predicted No Cross-Buy (0)	11,606	686
Predicted Yes Cross-Buy (1)	4,059	903

Table 7: Confusion Matrix for the Logistic Regression Model Using the Oversampling Technique on the Testing Set

The variable importances for the oversampling technique are shown in Table 8. The variable age is again the most important variable, similar to the SMOTE and undersampling technique, which could be an important predictor for our target variable.

Variable	Importance
Age	72.1%
Last Account	20.2%
Giro Mailing (yes)	13.2%
Brokerage	12.7%
Calls	11.8%

Table 8: Top 5 Variable Importances for the Logistic Regression Using Oversampling Technique

In Table 9, an overview is being shown with all the used evaluation metrics for each of the sampling methods and the baseline model. The evaluation metrics specificity, precision, and AUC-ROC were not included in the research conducted by Alves Werb and Schmidberger (2021).

	SMOTE	**Undersampling**	**Oversampling**	**Baseline**
Accuracy	**75.1%**	63.9%	72.5%	68.1%
Specificity	**77.9%**	63.6%	74.1%	
Sensitivity/Recall	47.5%	**66.7%**	56.8%	66.2%
Precision	17.9%	15.7%	**18.2%**	
F1 Score	26.0%	25.4%	27.6%	**29.0%**
AUC-ROC	68.4%	**71.5%**	71.4%	

Table 9: Results for the Logistic Regression Model on the Testing Set

From Table 9 it could be concluded that the logistic regression model created in this thesis performed better than the baseline model regarding the majority of the evaluation metrics. The best-performing model for each metric is displayed in bold.

For accuracy, the SMOTE model was the best-performing model (75.1%), which performed slightly better than the model implementing the oversampling method (72.5%). Looking at specificity, the model using the SMOTE sampling technique was again the best-performing model (77.9%), with again the model that used the oversampling technique as the second best-performing model (74.1%). For sensitivity or recall, the undersampling method had the highest score (66.7%).

The reason for this could be that recall exemplifies the share of true cross-buyers which are correctly classified by the model (Alves Werb & Schmidberger, 2021). As the undersampling technique selects fewer instances from the majority class in order to establish a 50/50 split, the undersampling technique could be more sensitive to recall.

For precision, the oversampling method is the best-performing model (18.2%). Considering the F1 score, the baseline model was the best-performing one with a score of 29.0%. Lastly, for the AUC-ROC score, the undersampling technique showed to be the best-performing model (71.5%).

5.3 *K-Nearest Neighbors*

After conducting the data preprocessing steps including the one-hot encoding, min-max normalization, training the model on the training set, and fitting the model on the testing set the results are ready to be interpreted. In this section, the results for the kNN model will be presented for the

SMOTE sampling method in subsection 5.3, the undersampling method on page 27, and the oversampling method on page 29. In each of these sections the optimal number of neighbors (k) is being analyzed, a confusion matrix is being presented, and lastly the variable importance is being discussed. At the end, an overview of the results for the three sampling methods is being presented in Table 16.

SMOTE sampling

For the SMOTE sampling technique the model fitted to the training set is plotted in order to get a detailed view of the optimal number of neighbors (k), in order to establish the most optimal accuracy. For SMOTE sampling technique the optimal number of neighbors was set at 27 (k = 27). This is graphically displayed in Figure 1.

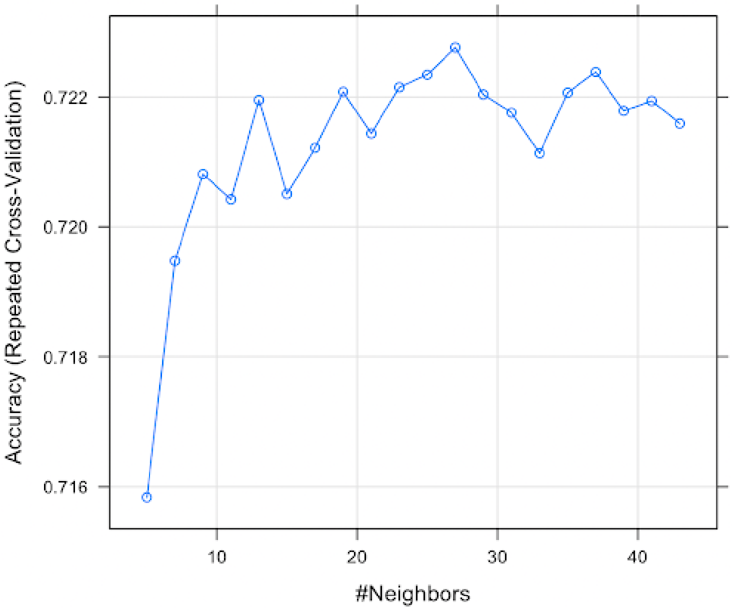

Figure 1: Number of Neighbors Plotted Against Accuracy for SMOTE Sampling

In the confusion matrix in Table 10, the outcomes for the kNN model on the testing set, using the SMOTE sampling technique are being shown. The adequate evaluation metrics that could be computed from this confusion matrix can be found in Table 16.

	Actual No Cross-Buy (0)	Actual Yes Cross-Buy (1)
Predicted No Cross-Buy (0)	12,728	999
Predicted Yes Cross-Buy (1)	1,941	476

Table 10: Confusion Matrix for the kNN Model Using the SMOTE Technique on the Testing Set

In Figure 2, the variable importance for the SMOTE sampling method is shown. The variable importance explains the extent to which the kNN model uses these variables in order to make accurate predictions. In this case, desktop logins are considered to be the most important variable for this model.

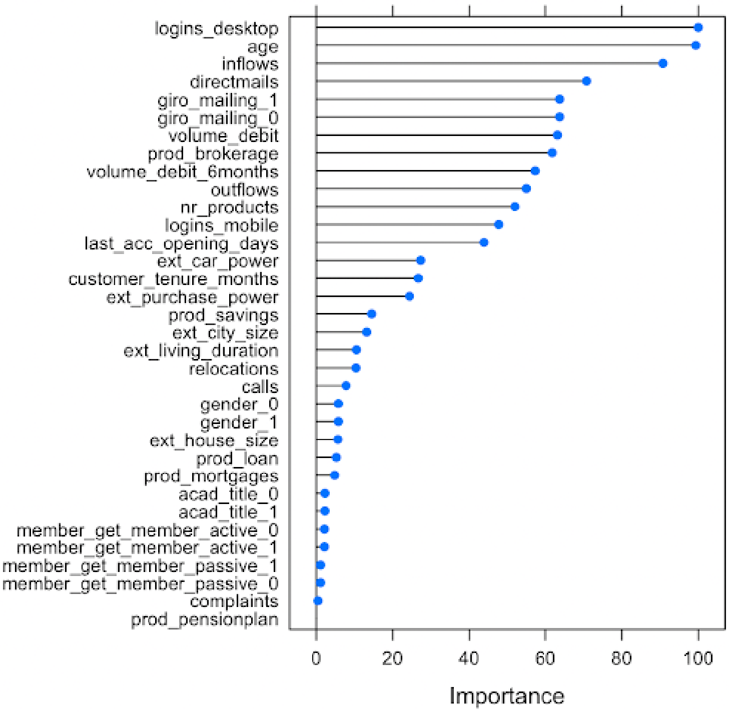

Figure 2: Variable Importance for kNN SMOTE sampling

In Table 11, the five most important variables are being listed, accompanied by the percentage of importance. In this table it is clearly shown which variables are most accurate at predicting the target variable. The variables age, direct mailing, and giro mailing could be affirmative of the research questions proposed in this research.

Variable	Importance
Desktop Logins	100.0%
Age	99.4%
Inflows	90.7%
Direct Mailing	70.8%
Giro Mailing	63.7%

Table 11: Top 5 Variable Importance for the kNN Model using SMOTE Sampling

Undersampling

For the undersampling technique the same procedure regarding the visualization of the optimal number of neighbors. For the undersampling technique the optimal number of neighbors was being established at 43 (k = 43). This is being displayed in Figure 3.

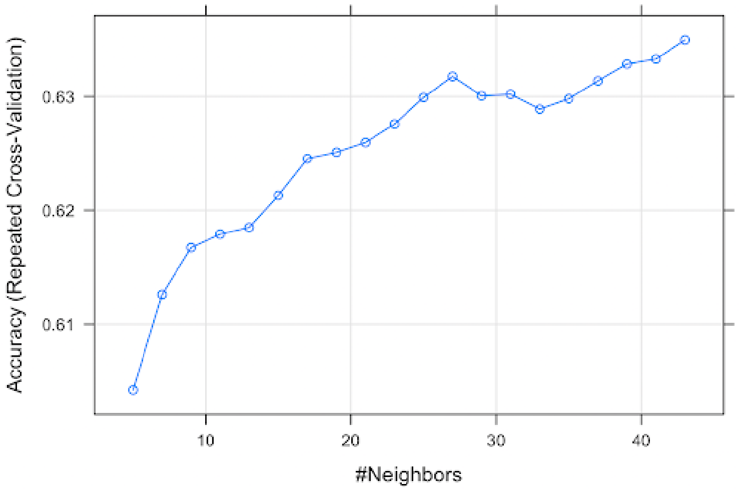

Figure 3: Number of Neighbors Plotted Against Accuracy for Undersampling

The confusion matrix presented in Table 12, shows the predicted outcomes against the actual outcomes for the undersampling technique.

	Actual No Cross-Buy (0)	**Actual Yes Cross-Buy (1)**
Predicted No Cross-Buy (0)	10,819	662
Predicted Yes Cross-Buy (1)	3,850	813

Table 12: Confusion Matrix for the kNN Model Using the Undersampling Technique on the Testing Set

In Figure 4, the variable importance for the undersampling method is displayed. For the undersampling technique used to predict the cross-buy target variable, age could be considered as the most important variable.

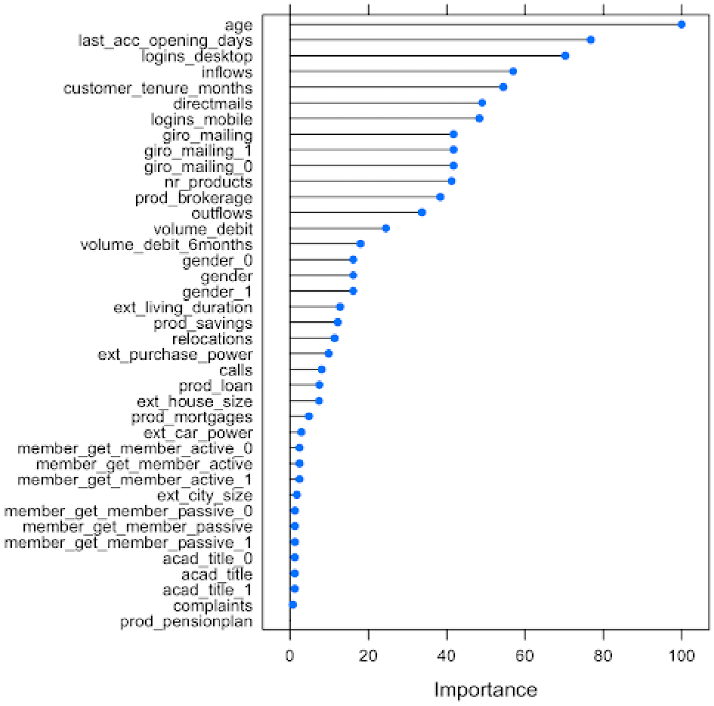

Figure 4: Variable Importance for kNN Undersampling

In Table 13, the five most important predictors for cross-buying behavior for the kNN model using undersampling are being displayed. In this case, it is worth to note that age is the only predictor that relates back to the research questions. Age, desktop logins, and inflows are being considered as important by both the SMOTE sampling, as well as, the undersampling technique.

Variable	**Importance**
Age	100.0%
Last Account	76.8%
Desktop Logins	70.3%
Inflows	57.0%
Customer Tenure	54.4%

Table 13: Top 5 Variable Importance for the kNN Model using Undersampling

Oversampling

Remarkably, for the oversampling technique the optimal number of neighbors was found to be 5 (k=5). This small number of neighbors could imply that the oversampling technique has a higher error rate. Therefore, it could be presumed that this model is overfitting. The evaluation metrics in Table 16 substantiate this effect as well, as the oversampling method generally has the worst performance.

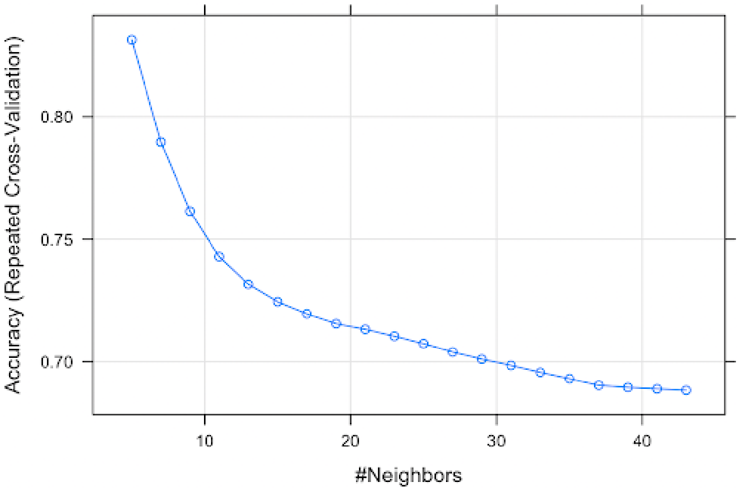

Figure 5: Number of Neighbors Plotted Against Accuracy for Oversampling

Table 14 demonstrates the confusion matrix for the kNN oversampled model.

	Actual No Cross-Buy (0)	Actual Yes Cross-Buy (1)
Predicted No Cross-Buy (0)	10,258	800
Predicted Yes Cross-Buy (1)	4,411	675

Table 14: Confusion Matrix for the kNN Model Using the Oversampling Technique on the Testing Set

In Figure 6, the variable importances for the predictor variables are demonstrated. Similar to the undersampling technique, age is again the most important predictor variable.

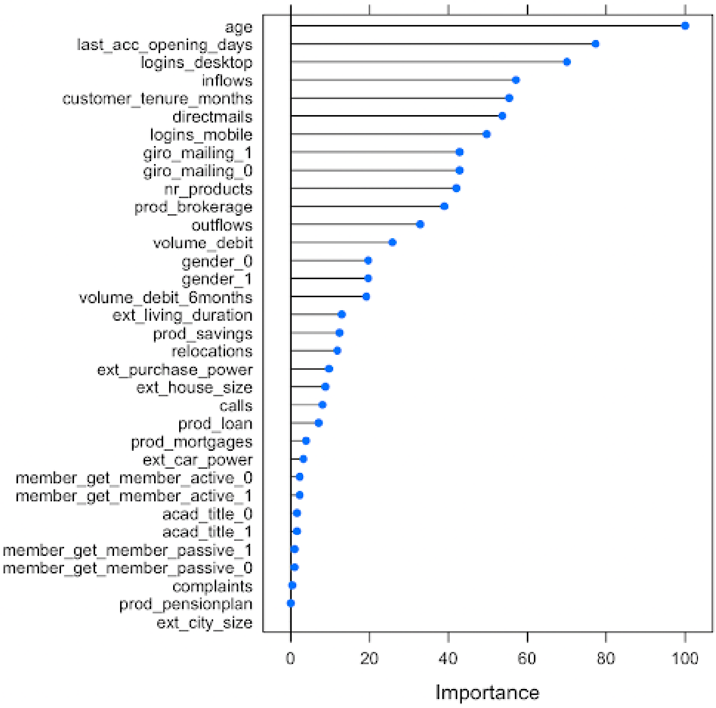

Figure 6: Variable Importance for kNN Using Oversampling

Finally, in Table 15 the five most important predictor variables are being displayed. Again, age, desktop logins, and inflows are considered part of the five most important predictor variables, just as for the SMOTE sampling and oversampling method. For these five variables, only age relates to the proposed research questions.

Variable	Importance
Age	100.0%
Last Account	77.3%
Desktop Logins	70.1%
Inflows	57.1%
Customer Tenure	55.4%

Table 15: Top 5 Variable Importance for the kNN Model using Oversampling

Altogether, the evaluation metrics reflecting the results for all three sampling methods are tabulated in Table 16.

	SMOTE	Undersampling	Oversampling
Accuracy	**81.8%**	72.1%	67.7%
Specificity	32.3%	**55.1%**	45.8%
Sensitivity/Recall	**86.8%**	73.8%	69.9%
Precision	**19.7%**	17.4%	13.3%
F1 Score	24.5%	**26.5%**	20.6%
AUC-ROC	59.5%	**64.4%**	57.8%

Table 16: Results for the kNN Model on the Testing Set

From Table 16 it could be concluded that the SMOTE sampling technique and undersampling technique were the best-performing sampling methods. For each of the evaluation metrics, the best-performing model is being displayed in bold. These two techniques provided for the highest evaluation metric 50% of the time. The SMOTE sampling method has the highest performance outcome looking at accuracy (81.8%), sensitivity or recall (86.8%), and precision (19.7%). The undersampling technique had the best performance for specificity (55.1%), F1 score (26.5%), and AUC-ROC (64.4%).

5.4 *Support Vector Machine*

Lastly, the support vector machine algorithm is being used after the data preprocessing steps, including the min-max normalization for feature scaling, training the model on the training set, and finally fitting the model on the testing set. In this section the results are being proposed for each of the sampling techniques. In subsection 5.4, the results for the SVM model using the SMOTE sampling technique are being discussed. Subsequently, on page 33, the results using the undersampling method in the preprocessing steps for the SVM model is being elaborated on. Lastly, on page 34, the results for the oversampling method are displayed.

In each of these sections a confusion matrix is being presented, followed by the variable importances that could be derived from the model. For the three SVM models a radial kernel has been implemented. Lastly, an overview of the evaluation metrics is being displayed in Table 23.

SMOTE sampling

For the SMOTE sampling technique a confusion matrix has been assembled in Table 17. In the confusion matrix the actual outcomes of a customer's cross-buying tendency is being compared against the predicted outcomes.

	Actual No Cross-Buy (0)	Actual Yes Cross-Buy (1)
Predicted No Cross-Buy (0)	13,250	955
Predicted Yes Cross-Buy (1)	2,415	634

Table 17: Confusion Matrix for the SVM Model Using the SMOTE Sampling Technique on the Testing Set

In Figure 7, the variable importances for the most important predictor variables are visualized. In this figure, desktop logins could be considered as the most important variable.

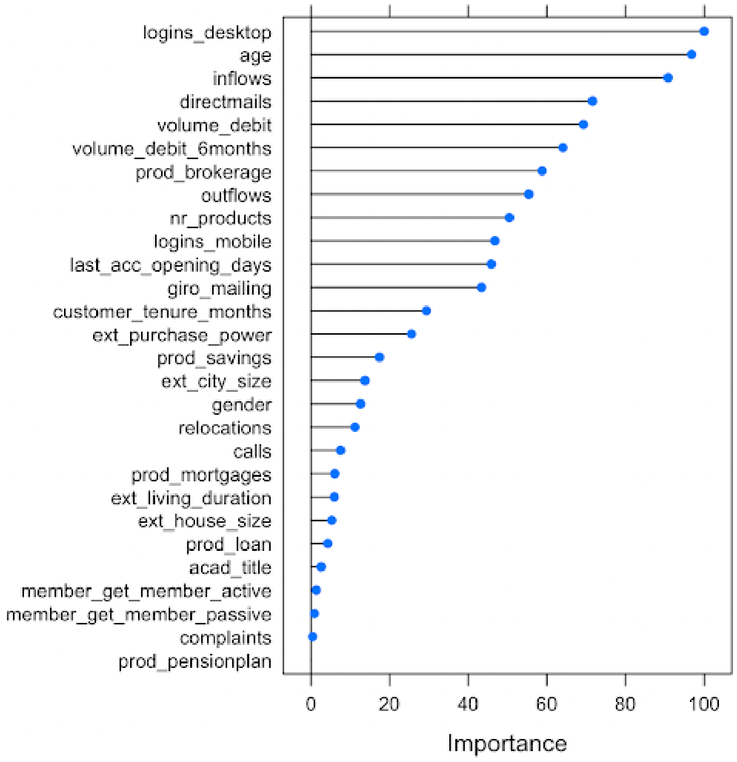

Figure 7: Variable Importance for SVM Using SMOTE Sampling

Lastly, in Table 18, the five most important predictor variables for the SVM model are listed. Similar to the kNN model, the variables desktop logins, age, and direct mailing are included as well. The variables age and direct mailing are related to the proposed research questions.

Variable	Importance
Desktop Logins	100.0%
Age	96.8%
Inflows	90.8%
Direct Mailing	71.6%
Total Debit	69.3%

Table 18: Top 5 Variable Importance for the SVM Model using SMOTE Sampling

Undersampling

For the undersampling technique the same procedure has been performed as for the SMOTE sampling model. The confusion matrix for the undersampling method can be found in Table 19.

	Actual No Cross-Buy (0)	Actual Yes Cross-Buy (1)
Predicted No Cross-Buy (0)	11,203	619
Predicted Yes Cross-Buy (1)	4,462	970

Table 19: Confusion Matrix for the SVM Model Using the Undersampling Technique on the Testing Set

In Figure 8, the variables are shown in order of importance for the undersampling technique.

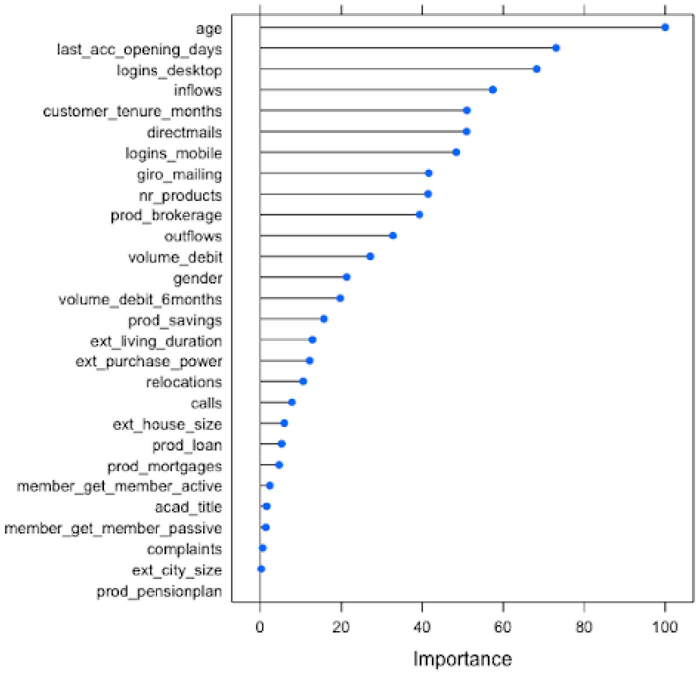

Figure 8: Variable Importance for SVM Using Undersampling

The final results for the undersampling technique are shown in Table 20. The variables age, desktop logins, and inflows show high importance for this sampling technique, as well as, for SMOTE sampling. In this case, age is the only variable that relates back to the research question.

Variable	Importance
Age	100.0%
Last Account	73.1%
Desktop Logins	68.3%
Inflows	57.5%
Customer Tenure	51.0%

Table 20: Top 5 Variable Importance for the SVM Model using Undersampling

Oversampling

Table 21 displays the confusion matrix for the SVM model using the oversampling method.

	Actual No Cross-Buy (0)	Actual Yes Cross-Buy (1)
Predicted No Cross-Buy (0)	10,995	620
Predicted Yes Cross-Buy (1)	4,670	969

Table 21: Confusion Matrix for the SVM Model Using the Oversampling Technique on the Testing Set

Secondly, in Figure 9 the ordered feature importance is shown for the independent variables, ordered from high to low.

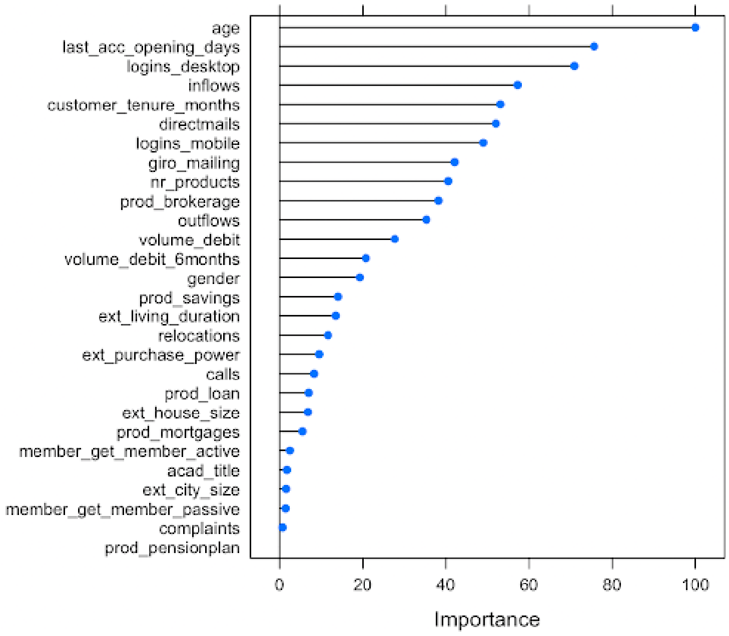

Figure 9: Variable Importance for SVM Using Oversampling

Ultimately, in Table 22 the five most important variables are shown with the importance percentage. Again, as in the two previous sampling methods, age, desktop logins, and inflows are present. Remarkably, the table exists of the same variables as for the undersampling technique, in the same order.

Variable	**Importance**
Age	100.0%
Last Account	75.6%
Desktop Logins	70.9%
Inflows	57.2%
Customer Tenure	53.1%

Table 22: Top 5 Variable Importance for the SVM Model using Oversampling

All in all, the results for the three sampling methods are reflected by the evaluation metrics show in Table 23, with the best-performing metrics marked in bold.

	SMOTE	**Undersampling**	**Oversampling**
Accuracy	**80.5%**	70.5%	69.3%
Specificity	39.9%	**61.0%**	**61.0%**
Sensitivity/Recall	**84.6%**	71.5%	70.2%
Precision	**20.8%**	17.9%	17.2%
F1 Score	27.3%	**27.6%**	26.8%
AUC-ROC	62.2%	65.5%	**65.6%**

Table 23: Results for the SVM Model on the Testing Set

From Table 23 it could be concluded that the SMOTE sampling technique was the best-performing one, as it showed the highest results for three out of six evaluation metrics. SMOTE sampling was the best-performing method looking at accuracy (80.5%), sensitivity or recall (84.6%), and precision (20.8%). As for the undersampling and oversampling methods, they both scored at 61.0% on specificity. Whereas, the undersampling method was the best-performing method in terms of F1 score (27.6%). The oversampling method performed best looking at AUC-ROC (65.6%). Therefore, it would be recommended when running a support vector machine on the dataset, to preprocess the data using the SMOTE sampling technique.

5.5 *Results Overview*

In Table 24 an overview is presented of all the results discussed in the sections above. The three models used in this research are presented next to each other, so they could be compared. The best-performing model is marked in bold for each evaluation metric.

For the SMOTE sampling technique, each model performs best in different metrics. The logistic regression model performs the best looking at specificity (77.9%) and AUC-ROC (68.4%). The kNN model performs best in accuracy (81.8%) and sensitivity or recall (86.8%). Lastly, SVM has the best results for precision (20.8%) and F1 score (27.3%).

Considering the undersampling technique, there is an equal distribution for best-performing models here as well. Again, each model here performs best for the exact same evaluation metrics. Logistic regression performs the best for specificity (63.6%) and AUC-ROC (71.5%). Equally, kNN shows the best performance in accuracy (72.1%) and sensitivity or recall (73.8%). Ultimately, SVM has the best performance looking at precision (17.9%) and F1 score (27.6%).

Looking at the oversampling technique, logistic regression outperforms the other two models. It shows the best evaluation metrics for five out of six metrics. It performs best looking at accuracy (72.5%), specificity (74.1%), precision (18.2%), F1 score (27.6%), and AUC-ROC (71.4%). Only for sensitivity or recall SVM outperforms the logistic regression model with a sensitivity or recall of 70.2%.

SMOTE Sampling	**LogReg**	**kNN**	**SVM**
Accuracy	75.1%	**81.8%**	80.5%
Specificity	**77.9%**	32.3%	39.9%
Sensitivity/Recall	47.5%	**86.8%**	84.6%
Precision	17.9%	19.7%	**20.8%**
F1 Score	26.0%	24.5%	**27.3%**
AUC-ROC	**68.4%**	59.5%	62.2%
Undersampling			
Accuracy	63.9%	**72.1%**	70.5%
Specificity	**63.6%**	55.1%	61.0%
Sensitivity/Recall	66.7%	**73.8%**	71.5%
Precision	15.7%	17.4%	**17.9%**
F1 Score	25.4%	26.5%	**27.6%**
AUC-ROC	**71.5%**	64.4%	65.5%
Oversampling			
Accuracy	**72.5%**	67.7%	69.3%
Specificity	**74.1%**	45.8%	61.0%
Sensitivity/Recall	56.8%	69.9%	**70.2%**
Precision	**18.2%**	13.3%	17.2%
F1 Score	**27.6%**	20.6%	26.8%
AUC-ROC	**71.4%**	57.8%	65.6%

Table 24: Model Comparison by Sampling Technique for Logistic Regression, k-Nearest Neighbors, and Support Vector Machine Fitted to the Testing Set

A distinction could not be made for the three models looking at SMOTE sampling and undersampling techniques. This would be based on the preference of certain evaluation metrics over the others. However, for the oversampling method logistic regression outperforms the kNN and SVM models for almost all of the evaluation metrics.

6 DISCUSSION

The goal of this research is to predict customer cross-buying behavior in the financial services industry. Moreover, this study aims to identify which customer behaviors and demographics influence a customer's cross-buying actions. The models logistic regression, kNN, and SVM are used to conduct this research. The problem statement for this thesis is as follows: *What customer behaviors and demographics affect the customer's cross-buying actions?*

In the research questions the specific effects of direct mailing campaigns, customer characteristics, city size and living duration, with regards to cross-buying are examined. Suggestions for these relationships could be found in previous research (Alves Werb & Schmidberger, 2021; Kumar et al., 2008; Morisada et al., 2018).

The baseline model used for this thesis reported an accuracy of 68.1%. Whereas the accuracy in the logistic regression model created for this research was 75.1% (SMOTE), 63.9% (undersampling), and 72.5% (oversampling).

The logistic regression model provided a way to establish a more interpretable model than the one created in previous research (Alves Werb & Schmidberger, 2021). All three undersampling methods that were used for the logistic regression model outperformed the baseline model. Compared to kNN and SVM, logistic regression was the best-performing model looking at the oversampling technique specifically. The variable importances derived from the logistic regression model were inconsistent for most variables. However, age was the most important predictor variable for all three sampling methods. These findings are in line with the research conducted by Dahana et al. (2020); Verhoef et al. (2001).

Secondly, the k-nearest neighbors model was adopted to predict cross-buying behavior. The kNN model outperformed the logistic regression model in terms of accuracy for the SMOTE sampling technique (81.8%), and the undersampling technique (72.1%). The kNN model generally performed best for accuracy and sensitivity or recall. For the variable importance analysis which has been conducted for this model, age, desktop logins, and inflows are considered as important predictor variables. Therefore, the findings for the kNN model were in line with the research of Dahana et al. (2020); Estrella-Ramón (2017); Estrella-Ramon et al. (2016); Verhoef et al. (2001).

As for the support vector machine, its performance was similar to the kNN model. For the SMOTE sampling and undersampling technique it presented a comparable performance in terms of the evaluation metrics. The SVM model outperformed the other two models for precision and F1 Score. For all three sampling methods the variables age, desktop logins, and inflows were presented as critical variables in order to predict our target variable. The variable importances are equal to the ones established for the kNN model, and the literature established in the paragraph above.

The potential influence of desktop logins was not captured in the proposed research questions. However, this finding was established in the research by Estrella-Ramon et al. (2016), where it was found that customers who carry out online banking behavior are more likely to cross-buy. Direct mailing was also found to be an important predictor, which was the fourth most important variable in the kNN and SVM model, both using SMOTE sampling. These findings relate to the research by Kumar et al. (2008), affirming that there exists a positive effect between cross-buying behavior and direct mailing campaigns. This results in an affirmative answer to *research question 1*, regarding the influence of direct mailing campaigns on cross-buying.

Besides that, a positive effect was found for age in relation to the target variable. Age showed high variable importances for kNN ranging from 99.4% to 100%. This positive effect could also be found for the support vector machines. Meaning that if the age of a customer increases, a customer would be more likely to cross-buy. Relating to the literature review conducted for this research this could be confirmed by Dahana et al. (2020); Verhoef et al. (2001), who found that age had a positive effect on cross-buying. Therefore, age would be the customer characteristic with the strongest predictive performance on cross-buy, relating back to *research question 2*, which questioned the effect of customer characteristics on cross-buying.

As for *research question 3*, research by Alves Werb and Schmidberger (2021) suggests that a strong substantial effect could be found for city size and living duration. However, the results from the models performed in this thesis would contradict this suggestive relationship. In the three models performed no effect was found for a positive relationship between either city size and living duration on cross-buying behavior. This could be due to the fact that the variables denoting city size and living duration were limited in the information they provided. It concerned bucketed variables, which did not capture the exact number as it was being translated into a range of numbers. This finding would be a useful addition to their research, as in the research by Alves Werb and Schmidberger (2021) it is suggested to further investigate the potential effects of city size and living duration.

Another variable relating to the financial behavior of customers, which had an unexpectedly positive effect on cross-buying is inflows. This was not something this research evolved around, however a substantial positive relationship between the number of inflows and customer cross-buying behavior had been found. It has been found by Dahana et al. (2020);

Estrella-Ramón (2017) that income would have a positive effect on cross-buying behavior. As the cross-buying variable in this research is defined by customers opening another checking account, a relationship could exist between opening a second checking account and a higher income. This could form a limitation to this research, as it unravels a substantial effect that could not be fitted into this research's framework.

Another limitation for this research could be the artificially changed target variable proportion in our training set, as it might not be representative of real-world data. As class imbalance in these types of customer data might be common, the dataset to which the models are applied to in this research could differ from that. Even though the testing set remains unchanged, it could be the case that the data on which the models are trained are not as representative of the real-world data as we would like. For future research it would therefore be suggested to assess data from various organizations to find out whether this imbalance is common for multiple organizations.

To conclude, this study contributes to the currently existing research within this framework as it has confirmed hypothetical relationships among certain customer behaviors and characteristics, and the probability of cross-buying. These relationships could now be confirmed for the financial services industry. The main findings of the positive effects of age, direct mailing, and desktop logins on customer cross-buying behavior is something that could be implemented in a financial services company's marketing strategy for targeting customers. If the findings of this thesis would be applied to a real-world financial services organization, marketing efforts for targeting potential cross-buying customers could be handled more rapidly and effectively.

7 CONCLUSION

The objective of this research was to create models in order to predict customer cross-buying behavior, which could be applied to customer data in financial services organizations. Therefore, certain customer characteristics and behaviors could be identified and customer profiles could be made which would be helpful in order to correctly target them.

Considering the research question which relates to customer characteristics, the general conclusion would be that age is one of the most important demographical predictors in this study. The older the customer, the more likely they are to cross-buy. Furthermore, online banking behavior and the number of direct mailing campaigns also have a positive effect on cross-buying behavior. Lastly, regarding the external environmental factors, the variables of city size and living duration have found to perform a minimal, positive effect on the target variable.

The results of this research would benefit marketeers and management teams of financial services organizations, as the customer characteristics and behaviors that were identified in this research allow them to effectively target potential customers.

For future research, it would be recommended to further investigate the relationship between inflows and cross-buying behavior. Moreover, as an addition to this research it would be interesting to develop a machine learning algorithm which directly identifies a potential customer for cross-buying as an outcome of the model.

8 CODE SOURCES

http://amunategui.github.io/smote/
https://datascience.stackexchange.com/questions/13971/standardization-normalization-test-data-in-r0
https://machinelearningmastery.com/feature-selection-with-the-caret-r-package/
https://www.machinelearningplus.com/machine-learning/caret-package/

REFERENCES

Alves Werb, G., & Schmidberger, M. (2021). Predictive modeling in marketing: Ensemble methods for response modeling. *Alves Werb, G., & Schmidberger, M.(2021). Predictive Modeling in Marketing: Ensemble Methods for Response Modeling. Die Unternehmung*, *75*(3), 376–396.

Blattberg, R. C., Malthouse, E. C., & Neslin, S. A. (2009). Customer lifetime value: Empirical generalizations and some conceptual questions. *Journal of Interactive Marketing*, *23*(2), 157–168.

Bradley, A. P. (1997). The use of the area under the roc curve in the evaluation of machine learning algorithms. *Pattern recognition*, *30*(7), 1145–1159.

Bzdok, D., Krzywinski, M., & Altman, N. (2018). Machine learning: supervised methods. *Nature methods*, *15*(1), 5.

Dahana, W. D., Miwa, Y., Baumann, C., & Morisada, M. (2020). Relative importance of motivation, store patronage, and marketing efforts in driving cross-buying behaviors. *Journal of Strategic Marketing*, 1–29.

Estrella-Ramón, A. (2017). Explaining customers' financial service choice with loyalty and cross-buying behaviour. *Journal of Services Marketing*.

Estrella-Ramon, A., Sánchez-Pérez, M., & Swinnen, G. (2016). How customers' offline experience affects the adoption of online banking. *Internet Research*.

Garland, R. (2004). Share of wallet's role in customer profitability. *Journal of Financial Services Marketing*, *8*(3), 259–268.

Gentsch, P. (2018). *Ai in marketing, sales and service: How marketers without a data science degree can use ai, big data and bots*. Springer.

Hallowell, R. (1996). The relationships of customer satisfaction, customer loyalty, and profitability: an empirical study. *International journal of service industry management*.

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263-1284. doi: 10.1109/TKDE.2008.239

Kamakura, W. A., Ramaswami, S. N., & Srivastava, R. K. (1991). Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *international Journal of Research in Marketing*, *8*(4), 329–349.

Knott, A., Hayes, A., & Neslin, S. A. (2002). Next-product-to-buy models for cross-selling applications. *Journal of interactive Marketing*, *16*(3), 59–75.

Kumar, V., George, M., & Pancras, J. (2008). Cross-buying in retailing: Drivers and consequences. *Journal of retailing*, *84*(1), 15–27.

Larivière, B., & Van den Poel, D. (2005). Predicting customer retention and profitability by using random forests and regression forests techniques. *Expert systems with applications*, *29*(2), 472–484.

Lemon, K. N., & Wangenheim, F. V. (2009). The reinforcing effects of loyalty program partnerships and core service usage: a longitudinal analysis. *Journal of Service Research*, *11*(4), 357–370.

Li, S., Sun, B., & Montgomery, A. L. (2011). Cross-selling the right product to the right customer at the right time. *Journal of Marketing Research*, *48*(4), 683–700.

Li, S., Sun, B., & Wilcox, R. T. (2005). Cross-selling sequentially ordered products: An application to consumer banking services. *Journal of Marketing Research*, *42*(2), 233–239.

Little, J. D. (2004). Models and managers: The concept of a decision calculus. *Management science*, *50*(12_supplement), 1841–1853.

Liu, T.-C., & Wu, L.-W. (2007). Customer retention and cross-buying in the banking industry: An integration of service attributes, satisfaction and trust. *Journal of financial services marketing*, *12*(2), 132–145.

Mansouri, S. (2021). Business cycles influences upon customer cross-buying behavior in the case of financial services. *Journal of Financial Services Marketing*, 1–21.

Molnar, C. (2020). *Interpretable machine learning*. https://christophm.github.io/interpretable-ml-book/interpretable-ml.pdf.

Morisada, M., Miwa, Y., & Dahana, W. D. (2018). Behavioral impacts of promotion-induced cross-buying: The moderating roles of age and gender. *Journal of Business Diversity*, *18*(2).

Patro, S., & Sahu, K. K. (2015). Normalization: A preprocessing stage. *arXiv preprint arXiv:1503.06462*.

Prinzie, A., & Van den Poel, D. (2008). Random forests for multiclass classification: Random multinomial logit. *Expert systems with Applications*, *34*(3), 1721–1732.

Ramyachitra, D., & Manikandan, P. (2014). Imbalanced dataset classification and solutions: a review. *International Journal of Computing and Business Research (IJCBR)*, *5*(4), 1–29.

Refaeilzadeh, P., Tang, L., & Liu, H. (2009). Cross-validation. *Encyclopedia of database systems*, *5*, 532–538.

Reinartz, W., Thomas, J. S., & Bascoul, G. (2008). Investigating cross-buying and customer loyalty. *Journal of Interactive Marketing*, *22*(1), 5–20.

Reinartz, W. J., & Kumar, V. (2003). The impact of customer relationship characteristics on profitable lifetime duration. *Journal of marketing*, *67*(1), 77–99.

Robitzsch, A. (2020). Why ordinal variables can (almost) always be treated as continuous variables: Clarifying assumptions of robust continuous and ordinal factor analysis estimation methods. In *Frontiers in education* (Vol. 5, p. 177).

Sanjar, K., Bekhzod, O., Kim, J., Paul, A., & Kim, J. (2020). Missing data imputation for geolocation-based price prediction using knn–mcf method. *ISPRS International Journal of Geo-Information*, *9*(4), 227.

Sun, J., Lang, J., Fujita, H., & Li, H. (2018). Imbalanced enterprise credit evaluation with dte-sbd: Decision tree ensemble based on smote and bagging with differentiated sampling rates. *Information Sciences*, *425*, 76–91.

Verhoef, P. C., Franses, P. H., & Hoekstra, J. C. (2001). The impact of satisfaction and payment equity on cross-buying: A dynamic model for a multi-service provider. *Journal of Retailing*, *77*(3), 359–378.

Zeithaml, V. A., Lemon, K. N., & Rust, R. T. (2001). *Driving customer equity: How customer lifetime value is reshaping corporate strategy*. Simon and Schuster.

9 APPENDIX A

Dependent Variable	Variable Definition
Cross-buy	Customer opened a checking account: 1 (yes), 0 (no)
Transaction Data	
Calls	Number of calls in last 180 days
Complaints	Number of complaints in last year
Customer Tenure	Number of months since customer onboarding
Inflows	Total volume of inflows on savings account from outside in the last 6 months (€)
Last Account	Number of days since last account opening
Desktop Logins	Number of logins in the last 180 days
Mobile Logins	Number of mobile sessions in the last 180 days
Number of Products	Total number of products (accounts)
Outflows	Total volume of outflows from savings account in the last 6 months
Loans	Number of consumer loan accounts
Mortgages	Number of mortgage accounts
Brokerage	Number of investment accounts
Pension Plan	Number of long term savings plans
Savings	Number of savings accounts
Relocations	Number of relocations/address changes in the last year
Total Debit	Total balances of all debit (savings) accounts (€)
Total Debit Six Months	Credit balance of all products from 6 months ago (€)
Marketing Efforts	
Direct Mailing	Total number of mailings in the last year
Giro Mailing	Received an email about opening a checking account: 1 (yes), 0 (no)

Table 25: Dataset Variable Definitions: Target Variable, Transaction Data, and Marketing Efforts (Reprinted from Alves Werb and Schmidberger (2021))

Customer Characteristics	
Academic Title	Does the customer have an academic title: 1 (yes), 0 (no)
Age	Customer's age in years
Gender	Customer's gender: 1 (male), 0 (female)
Get Member Active	Customer recommended a customer: 1 (yes), 0 (no)
Get Member Passive	Customer was recommended by a customer: 1 (yes), 0 (no)
City Size	City size: 1 (<5.000 inhabitants), 2 (5.000-10.000), 3 (10.001-20.000), 4 (20.001-50.000), 5 (50.001-100.000), 6 (100.001-200.000), 7 (200.001-500.000), 8 (>500.000)
House Size	Average number of households per building in the residential block: 1 (1-2 households), 2 (3-5), 3 (6-9), 4 (10-19), 5 (>19)
Purchase Power	Average purchase power in the residential block: 1 (extremely low), 2 (very low), 3 (low), 4 (average), 5 (high), 6 (very high), 7 (extremely high)
Car Power	Predominant vehicle category in the neighborhood: 1 (subcompact), 2 (compact), 3 (mid-size), 4 (full size), 5 (mixed)
Living Duration	Average duration of residence in the customer's building: 1 (0-1 year), 2 (1-2), 3 (2-3), 4 (3-4), 5 (4-5), 6 (5-6), 7 (6-8), 8 (8-10), 9 (more than 10 years)

Table 26: Dataset Variable Definitions: Customer Characteristics (Reprinted from Alves Werb and Schmidberger (2021))

10 APPENDIX B

Variable	Class	Score Range	Mean	Missing Data
Cross-Buy	Categorical			0.00%
Calls	Discrete	[0..58]	0.1046	0.00%
Complaints	Discrete	[0..8]	0.00353	0.00%
Customer Tenure	Discrete	[0..567]	140.2	0.00%
Inflows	Continuous	[1, 7113]	1180	0.47%
Last Account	Discrete	[1..15306]	3490	0.00%
Desktop Logins	Discrete	[1..3048]	6.422	0.00%
Mobile Logins	Discrete	[1..4259]	13.3	0.00%
Number of Products	Discrete	[1..17]	1.433	0.00%
Outflows	Continuous	[1, 7659]	6745	0.47%
Loans	Discrete	[0..4]	0.1218	0.00%
Mortgages	Discrete	[0..13]	0.1521	0.00%
Brokerage	Discrete	[0..10]	0.1938	0.00%
Pension Plan	Discrete	[0..11]	0.00408	0.00%
Savings	Discrete	[0..11]	0.9112	0.00%
Relocations	Discrete	[0..3]	0.0421	0.00%
Total Debit	Continuous	[0, 39990708]	22623	0.00%
Total Debit Six Months	Continuous	[1, 64881]	25762	2.98%
Direct Mailing	Discrete	[0..9]	0.4707	0.00%
Giro Mailing	Categorical			0.00%
Academic Title	Categorical			0.00%
Age	Continuous			0.00%
Gender	Categorical			0.00%
Get Member Active	Categorical			0.00%
Get Member Passive	Categorical			0.00%
City Size	Categorical			2.67%
House Size	Categorical			3.05%
Purchase Power	Categorical			4.57%
Car Power	Categorical			10.13%
Living Duration	Categorical			9.07%

Table 27: Variable Descriptive Statistics

11 APPENDIX C

Distribution of Categorical Variables

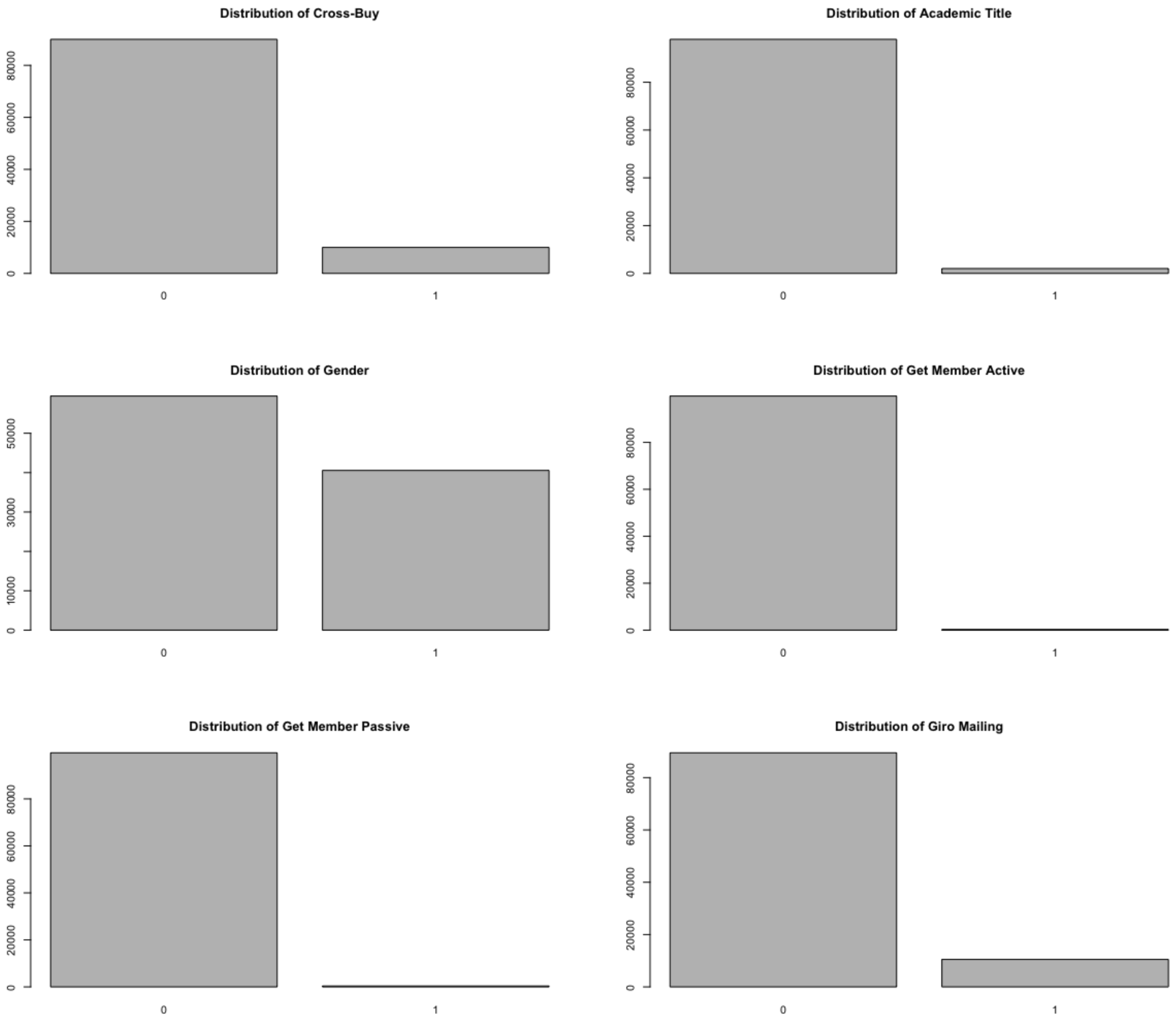

Figure 10: Variable Distributions for Categorical Variables

Distribution of Continuous Variables

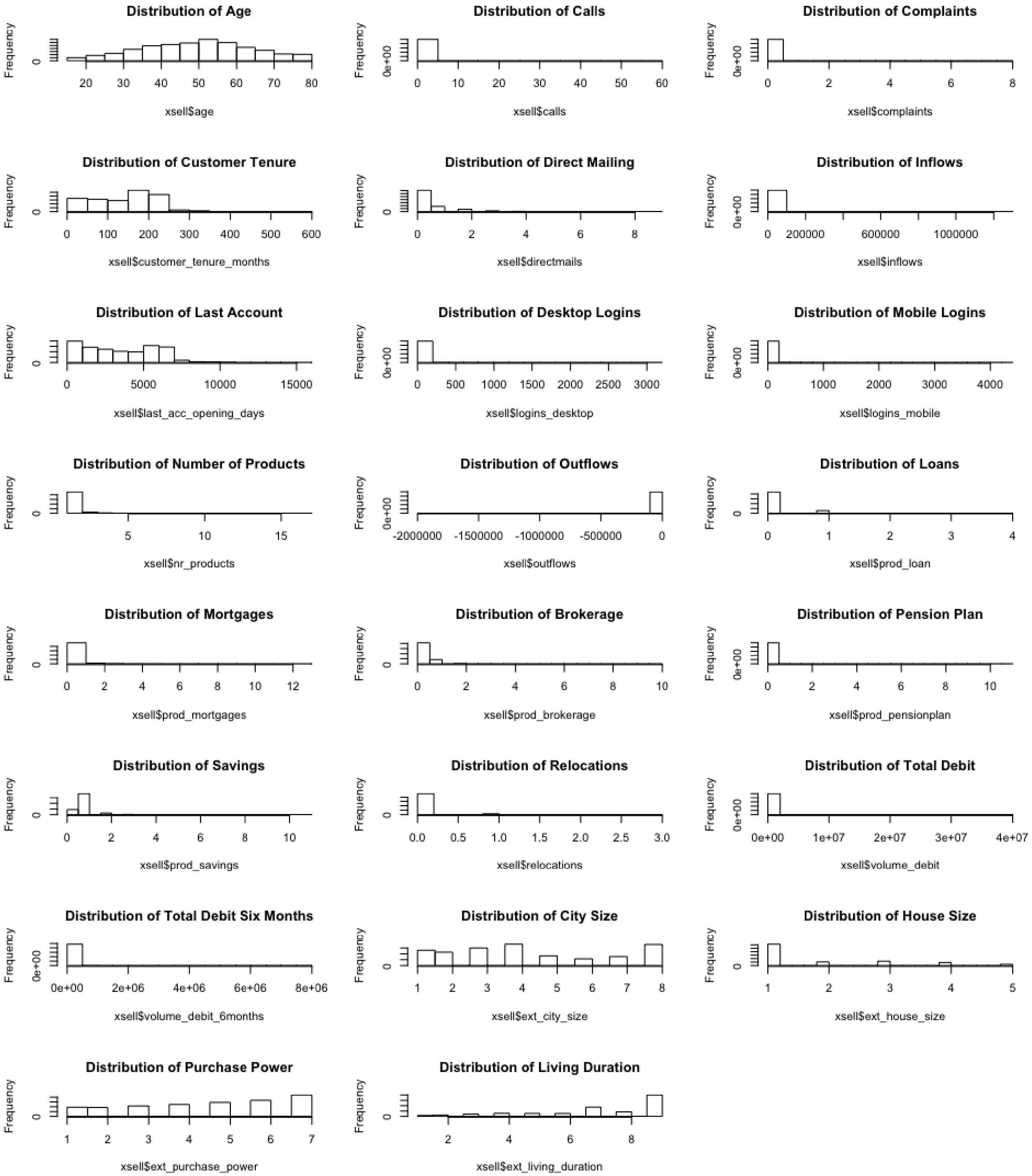

Figure 11: Variable Distributions for Continuous Variables