SMOTE meets Uplift Modeling:

Enhancing Uplift Analysis through Oversampling Techniques: A Comparative Study on Uplift Models in Cross-Selling

Seminar Paper / Term Paper

Submitted to Hon-Prof. Dr. Martin Schmidberger & Dr. Lennart Kraft

Faculty of Economics and Business, Goethe University Frankfurt, Frankfurt am Main

Manuel Scionti - scionti.manuel@hotmail.it - Erasmus Program

Abstract

This paper presents a comparative study investigating the impact of oversampling techniques, specifically SMOTE, on the performance of uplift models in cross-selling campaigns with high class imbalance. The study aims to improve the prediction of customers' likelihood to purchase a new credit account by exploring the application of oversampling techniques. The research covers uplift analysis, introduces commonly used uplift models, and discusses the chosen oversampling technique. Experimental results demonstrate the effectiveness of SMOTE in reducing class imbalance and enhancing uplift model performance. The study concludes by emphasizing the value of oversampling approaches in addressing high class imbalance and providing recommendations for future research and practical implementation in marketing campaigns.

Contents

List of Figures

List of Tables

1 Introduction

Uplift analysis, also known as the incremental response or true lift model, has emerged as a promising approach to marketing campaigns, particularly in the context of cross-selling.(Gutierrez, 2017) In contrast to traditional targeting methods, which focus on predicting the response of individual customers, uplift analysis aims to identify the individuals who are most likely to be influenced by a particular treatment or intervention, resulting in a desired outcome. By distinguishing between treatment effects and baseline effects, uplift analysis allows marketers to optimise their strategies by targeting only those customers who would not have taken the desired action without the intervention(Nyberg & Klami, 2023c). However, uplift analysis faces several challenges, one of which is the presence of high class imbalance in the target variable.(Nyberg & Klami, 2023c) In cross-selling scenarios, the majority of customers often do not respond to the campaign, resulting in a large imbalance between the positive and negative classes. This class imbalance can adversely affect the performance of uplift models, resulting in sub-optimal predictions and limited effectiveness of targeted marketing efforts. This paper focuses on the application of oversampling techniques, specifically the *Synthetic Minority Over-sampling Technique* (SMOTE), in uplift analysis for cross-selling campaigns to address the challenge of high class imbalance. The aim of this study is to investigate whether the incorporation of oversampling techniques can improve the performance of uplift models and mitigate the negative effects of class imbalance. We conduct a comparative study using a large banking dataset kindly provided by ING Bank containing information on customer demographics, transaction history and responses to marketing campaigns. The outcome variable of interest is whether customers purchase an additional product, specifically a new credit account. By comparing different uplift models with and without the application of oversampling techniques, we aim to understand the influence of oversampling on uplift model performance and determine whether oversampling can improve their predictive capabilities in the presence of class imbalance. The results of this study can help marketers improve the targeting and personalisation of their campaigns, ultimately leading to higher cross-selling success rates. The rest of this paper is organized as follows. In the next section, we provide a review of the existing literature on uplift analysis, oversampling techniques, and their application in addressing class imbalance. We then describe the dataset and experimental setup, including data preprocessing, feature engineering, and treatment assignment strategies. Subsequently, we present the uplift models and oversampling techniques employed in this study. The evaluation metrics used to assess the performance of the uplift models are defined, and the experimental results and analysis are discussed. Finally, we summarize the key findings, discuss their implications, and provide recommendations for future research and practical applications of oversampling techniques in marketing campaigns.

2 Related Work

In this section we lay the groundwork by building on two previous areas of research: uplift models and class imbalance. We give a brief introduction to both. This will provide the necessary background for understanding the rest of the paper.

2.1 Uplift Models

Commonly used uplift models in cross-selling analysis include S-Learner, T-Learner, and X-Learner. These models are also called metalearnes as they are based on the framework of machine learning algorithms and utilize different approaches to estimate individual treatment effects and predict uplift probabilities. To quote Künzel et al., (2019b):

> There is growing interest in estimating and analyzing heterogeneous treatment effects in experimental and observational studies. We describe a number of metaalgorithms that can take advantage of any supervised learning or regression method in machine learning and statistics to estimate the conditional average treatment effect (CATE) function. Metaalgorithms build on base algorithms—such as random forests (RFs), Bayesian additive regression trees (BARTs), or

neural networks—to estimate the CATE, a function that the base algorithms are not designed to estimate directly.

2.1.1 S-Learner

The S-Learner uplift model treats the uplift problem as a binary classification task. It trains a single model on the entire dataset, incorporating both treatment and control groups (Künzel et al.,2019b). The algorithm estimates the treatment effect by comparing the predicted outcomes of the treated and control groups. The pseudocode for the S-Learner algorithm is represented as follows:

```
Algorithm: S-Learner
Input: Features X, Treatments T, Outcomes Y
Output: Uplift predictions U

1. Combine the treated and control groups into a single dataset.
2. Train a binary classification model, such as logistic regression or random forest,
to predict the uplift.
3. Compute the predicted uplift probabilities for all instances.
4. Output the uplift predictions U based on the computed probabilities.
```

2.1.2 T-Learner

The T-Learner uplift model employs a two-step approach, where separate models are trained for the treated and control groups. In the first step, the algorithm estimates the treatment effect in each group independently. In the second step, it combines the estimated treatment effects to obtain the final uplift predictions. This approach allows for better capturing of heterogeneous treatment effects (Künzel et al.,2019b). It performs well if there are no common trends in the response under control and response under treatment and if the treatment effect is very complicated. Because data is not pooled across treatment groups, it is difficult for the T-learner to mimic a behavior (e.g. discontinuity) that appears in all the treatment groups. The pseudocode for the T-Learner algorithm is represented as follows:

```
Algorithm: T-Learner
Input: Features X, Treatments T, Outcomes Y
Output: Uplift predictions U

1. Train a model (e.g., regression or classification) on the treated group
to estimate the treatment effect.
2. Train a model on the control group to estimate the baseline effect.
3. Compute the difference between the two models' predictions as the uplift prediction.
4. Output the uplift predictions U based on the computed differences.
```

2.1.3 X-Learner

The X-Learner uplift model combines the strengths of both S-Learner and T-Learner by utilizing cross-validation. It divides the data into multiple folds, treating each fold as the validation set while training on the remaining folds. X-Learner estimates the treatment effects in each fold using both S-Learner and T-Learner approaches and averages the results to obtain the final uplift predictions (Künzel et al.,2019b). The X-learner can adapt to structural properties such as sparsity or smoothness of the CATE. (This is useful as CATE is often zero or approximately linear.) It is particularly effective when the number of units in one treatment group (often the control group) is much larger than in the other. The pseudocode for the X-Learner algorithm is represented as follows:

```
Algorithm: X-Learner
Input: Features X, Treatments T, Outcomes Y
Output: Uplift predictions U

1. Divide the data into K folds for cross-validation.
2. For each fold k:
   a. Train an S-Learner on the training set of fold k.
   b. Train a T-Learner on the training set of fold k.
   c. Compute the estimated uplift predictions for the validation set
   of fold k using both S-Learner and T-Learner.
3. Average the uplift predictions across all folds to obtain the final
uplift predictions U.
```

These uplift models provide different strategies for estimating individual treatment effects and predicting uplift probabilities. By utilizing these models, marketers can gain valuable insights into the impact of their campaigns and make informed decisions to optimize their cross-selling efforts.

2.2 Class imbalance

Class imbalance characterizes a scenario wherein the distribution of target classes within a dataset exhibits significant skewness, with one class representing the majority while the other class serves as the minority. Although class imbalance has been extensively studied within the domain of classification, it has received comparatively less attention in the context of uplift modeling. This imbalance poses formidable challenges in accurately modeling uplift effects, as conventional modeling techniques often exhibit a bias toward the majority class and struggle to capture the intricate patterns and predictive relationships inherent within the minority class. Consequently, the performance of uplift models may be compromised, resulting in biased estimations and diminished predictive capability. To tackle this issue, two primary techniques have been proposed: weighting and sampling; which comprises oversampling, undersampling, and synthetic sampling (Chawla et al. 2002). Weighting assigns higher weights to observations belonging to the minority class in the cost function to ensure their proper consideration by the algorithm. Oversampling involves generating multiple copies of minority class observations through resampling. Synthetic sampling generates novel and distinct observations based on the characteristics of existing observations. Conversely, undersampling entails the removal of certain observations from the majority class.

Among all these techniques S*ynthetic Minority Over-sampling Technique* (SMOTE) stands out (Nyberg & Klami, 2023d). This technique aim to mitigate the impact of class imbalance by generating synthetic minority samples, thereby harmonizing the dataset and enhancing the modeling of uplift effects. These approaches hold promise in improving the performance and efficacy of uplift models, allowing for more precise targeting of customers in cross-selling campaigns.

3 The oversampling process

In this approach, the class imbalance issue is addressed by incorporating instances from the minority class into the dataset. This is achieved through the replication of instances, either by random sampling or by employing intelligent algorithms. A notable oversampling technique that gained significant popularity in class imbalance classification is SMOTE, introduced by Chawla et al. (2002). Consequently, SMOTE has emerged as a prominent data processing and sampling algorithm within the fields of machine learning and data mining. Over the past 15 years, several extensions of SMOTE have been introduced, including:

- borderline-SMOTE
- safe-level SMOTE

- ADASYN
- cluster-SMOTE
- LVQ-SMOTE

3.0.1 SMOTE

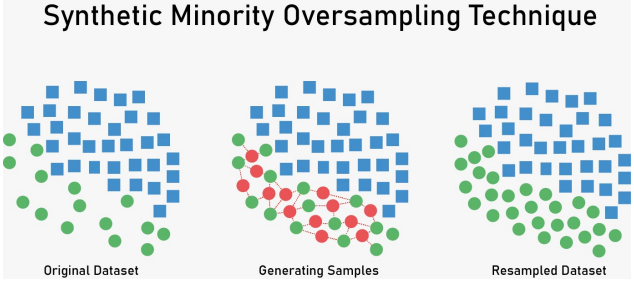


Fig. 1: Illustration of how SMOTE works

SMOTE, the *Synthetic Minority Oversampling Technique*, is a well-established data augmentation method widely used in machine learning, particularly in the context of imbalanced classification problems. Its main objective is to address the inherent challenge posed by imbalanced datasets, where the minority class has significantly fewer samples compared to the majority class. The functioning of SMOTE involves generating synthetic samples for the minority class to rectify the imbalance in class distribution. It achieves this by creating new synthetic instances along line segments that connect existing samples of the minority class. The algorithm selects a sample from the minority class and identifies its k nearest neighbors. Then, it randomly selects one of the neighbors and generates a synthetic example by interpolating between the chosen sample and the selected neighbor. This process continues until a desired balance between the minority and majority classes is achieved. The synthetic samples created by SMOTE enhance the representation of the minority class, providing more diverse training data for the classifier. This augmentation can significantly improve the classifier's ability to capture the intricate patterns and decision boundaries associated with the minority class, resulting in improved classification performance (Chawla et al. 2002). Due to its effectiveness, SMOTE has been widely adopted to address class imbalance challenges in various domains, including fraud detection, medical diagnosis, and anomaly detection. It offers a valuable approach to alleviate the obstacles inherent in imbalanced datasets and enhance the performance of classifiers when confronted with imbalanced classification tasks (Mohammed et al., 2020b).

Algorithm 1 SMOTE

Require: Minority class samples: *minority_samples*, Number of synthetic samples to generate: N
Ensure: Synthetic samples: *synthetic_samples*

synthetic_samples ← empty list
while $N > 0$ **do**
  random_minority_sample ← randomly select a sample from *minority_samples*
  k_nearest_neighbors ← find k nearest neighbors of *random_minority_sample*
  synthetic_sample ← create empty sample
  for each feature in *random_minority_sample* **do**
    difference ← randomly select a value between 0 and 1
    synthetic_feature_value ← *random_minority_sample[feature]* + *difference* × (*randomly selected neighbor[feature]* - *random_minority_sample[feature]*)
    synthetic_sample[feature] ← *synthetic_feature_value*
  end for
  synthetic_samples.append(*synthetic_sample*)
  $N \leftarrow N - 1$
end while
return *synthetic_samples* =0

4 Dataset

This study pertains to a comprehensive dataset comprising of 100,000 clients of a large retail bank, all of whom possess a Girokonto (payment account), but none have a lending product (Konsumentenkredit) at the start date t0. Our target variable for the analysis is denoted as 'xsell', indicating whether the clients proceeded to open a loan account at a later time t1 (xsell = 1) or not (xsell = 0). The dataset comprises a stratified sample of 10,000 buyers and 90,000 non-buyers.

The scope of this research is focused on cross-selling, a behavior indicating whether a given customer buys a second or third product, crucial for conducting efficient marketing campaigns and sustaining profitable customer relationships. The dataset includes various customer and account features such as the academic title, age, number of calls and complaints in the last year, customer tenure in months, gender, whether the client received a loan mailing, household income, number of desktop and mobile logins in the last 180 days, and marital status.

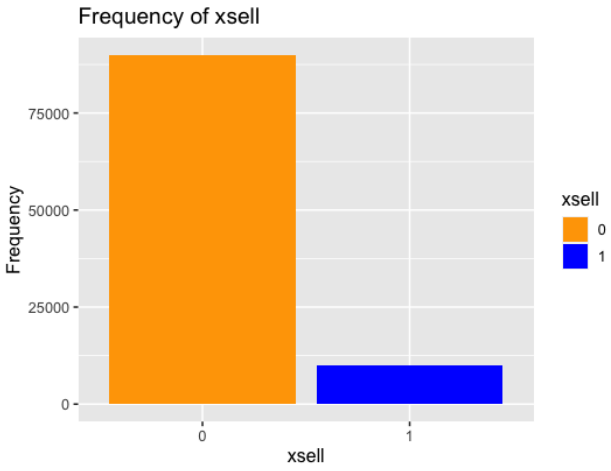


Fig. 2: Frequency of xsell

Additionally, the dataset includes features such as whether the customer recommended or was recommended by another client, the number of days the Girokonto balance was above or below zero in the last 90 days, the number of transactions made with a Giro card or a Visa card in the last 90 days, the total number of products/accounts the customer possesses, their occupation group, the number of relocations or address changes in the last year, the total volume of inflows and outflows from the savings account in the last six months, the number of different types of accounts like mortgage, investment and savings, the total balances of all debit and lending accounts, and whether a client opened a consumer loan within six months. Finally, the dataset integrates variables from external data sources, provided by external data providers such as Deutsche Post or Bertelsmann, merged mainly on the basis of addresses ensuring GDPR compliance.

Variables	Description
acad_title	Dummy: Does client have an academic title like Dr. or Prof. (0/1)
age	Customer's age in years
calls	Number of calls in the last 180 days
complaints	Number of complaints in the last year
customer_tenure_months	Number of months since customer onboarding
gender_sex	F = Single account holder female, M = Single account holder male, MF = Joint account
loan_mailing	Did the client receive a direct mail promoting consumer loans?
income	Household income if available, otherwise 0
logins_desktop	Number of logins in the last 180 days
logins_mobile	Number of mobile sessions in the last 180 days
marital_status	Marital status: VH = Married, GS = Divorced, LE = Single
member_get_member_recommender	Dummy if customer recommended a client
member_get_member_recommended	Dummy if customer was recommended by a client
nr_days_when_giro_above_0	Number of days when Griokonto was ¿0EUR (in the last 90 days)
nr_days_when_giro_below_0	Number of days when Griokonto was ¡0EUR (in the last 90 days)
nr_girocard_trx_90d	Number of transactions using Giro card (last 90 days)
nr_products	Total number of products (accounts)
occupation	Occupation groups
nr_relocations	Number of relocations in the last year/address changes
nr_visacard_trx_90d	Number of transactions using Visa card (last 90 days)
vol_eur_inflows	Total volume of inflows from savings account in the last 6 months
vol_eur_outflows	Total volume of outflows from savings account in the last 6 months
prod_mortgages	Number of mortgage accounts
prod_brokerage	Number of investment accounts
prod_savings	Number of savings accounts
vol_eur_debit	Total balances of all debit (savings) accounts
vol_eur_credit	Total balances of all debit (lending) accounts
xsell	Dummy if a client opened a consumer loan maximum six months later

Tab. 1: Variables descriptions

4.1 Preprocessing

The analysis proceeds in a series of interconnected phases: data loading, exploratory analysis and preprocessing, feature engineering, uplift machine learning model building, and interpretation of results.
Missing data within the dataset is handled by either replacing the missing values with zeroes or by using the mean of the respective variable. Depending on the variable and its distribution, either of these strategies can be a more suitable choice. This phase ensures the data is clean and ready to be used in subsequent steps.

Feature engineering follows, which involves the creation of new features from existing ones, to potentially enhance the predictive power of the model. Categorical variables are transformed into dummy variables, which are numerical and therefore suitable for machine learning algorithms. Furthermore, new variables are engineered based on existing ones, such as creating an overdraft variable which indicates if a client has used an overdraft within 90 days. Also, the creation of squared and cubic terms of age is performed, implying a nonlinear relationship between age and the response.

4.2 Variables selection

The process of variable selection involved utilizing a combination of *Stepwise selection* and *Lasso logistic regression*. As there is no definitive rule for selecting the ideal subset of variables (GuyonIsabelle & ElisseeffAndré, 2003), additional business considerations were taken into account. Ultimately, a final subset of optimal variables was determined by blending the results from both models.
The optimal subset of our predictors is the following:

Tab. 2: Selected predictors

Statistic	N	Mean	St. Dev.	Min	Max
age	100,000	40.427	12.126	18	65
logins_mobile	100,000	148.848	251.024	0	4,056
customer_tenure_months	100,000	93.548	75.310	0	510
overdraft	100,000	0.240	0.427	0	1
vol_eur_inflows	100,000	4,264.072	5,370.757	1	16,814
vol_eur_outflows	100,000	13,616.810	5,787.275	1	17,718
nr_girocard_trx_90d	100,000	19.555	29.398	0	345

5 Experiments and results

SMOTE is applied, resulting in the creation of two dataframes. The first dataframe represents the original unbalanced distribution of the target variable 'check-in' with a split of 90%/10%. The second dataframe, which includes oversampling, exhibits a more balanced distribution with a proportion of 60%/40% for 'xsell'. Once the data is prepared, the machine learning phase starts. The dataset is divided into a training set (80%) and a test set (20%). The same split is applied to both dataframes.

5.1 Unbalance dataframe

A logistic regression model is constructed using the training data, incorporating variables such as age, mobile logins, customer tenure, overdraft usage, euro inflow and outflow volumes, and giro card transactions in the last 90 days. To validate the model, predictions are made on the validation set using a response probability threshold of 0.1, in order to model the imbalance of the target variable. Deciles of the predicted cross-sell probability are computed to generate Lift and Gain charts. These charts, commonly utilized in marketing contexts, allow for quantifying the performance of predictive models.

5.1.1 Lift char & Gains table

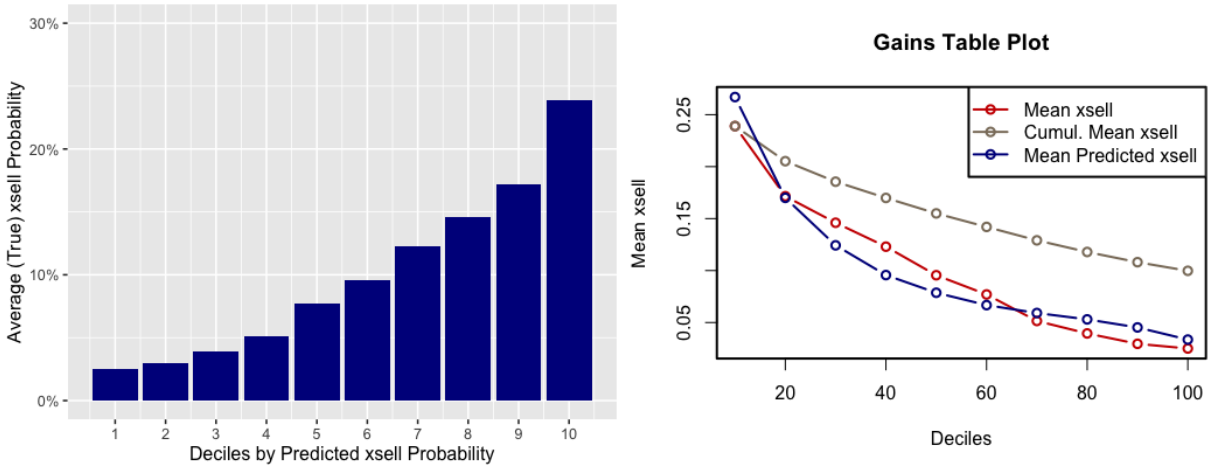
In the context of Uplift Modeling, the Lift Chart and Gains Table are two commonly used evaluation tools that help quantify the performance of a predictive model (Gubela et al., 2019b).

- **Lift Chart:** A Lift Chart (Figure 3a) illustrates the effectiveness of a model in terms of targeting individuals who are most likely to respond positively to a marketing intervention. It compares the cumulative response rate of a model against a random selection. The Lift Chart provides insights into the uplift achieved by the model, which is the ratio of the response rate with the model to the response rate without the model. The x-axis of the Lift Chart typically represents the percentage of the population targeted, ordered by the model's predicted uplift scores. The y-axis represents the uplift achieved, often measured as the response rate or the conversion rate. The higher the uplift value at a given percentage of the population targeted, the more effective the model is in identifying individuals who are likely to respond positively to the intervention. The Lift Chart helps determine the efficiency and effectiveness of targeting strategies and assists in making informed decisions regarding resource allocation (Gubela et al., 2019b).

- **Gains Table:** A Gains Table (Figure 3b) is a tabular representation of the cumulative gains achieved by a model in targeting individuals based on their predicted uplift scores. It provides a detailed breakdown of the response rate, cumulative response rate, and the percentage of the population targeted. The Gains Table is useful in understanding the incremental improvement achieved by a model compared to random targeting. The Gains Table typically consists of several columns. The first column represents the percentage of the population targeted (e.g., through a marketing campaign). The subsequent columns provide information such as the percentage of responders within that targeted population,

the cumulative percentage of responders, and the response rate. These columns help in assessing the effectiveness of the model at various stages of targeting and allow for comparison against random or baseline targeting strategies.

Both the Lift Chart and Gains Table provide visual and quantitative insights into the performance of an Uplift Model. They enable marketers to evaluate the model's ability to identify the most responsive individuals and make informed decisions about resource allocation and targeting strategies.



(a) Lift chart by xsell probabilities

(b) Gains table

Fig. 3: Comparison of Lift chart and Gains table

Additionally, in Table 3 the uplift resulting from "loan mailing" is conducted. This analysis involves evaluating the average probability of response based on age and calculating the uplift by measuring the difference in the average probability of response between the treatment and control groups.

Tab. 3: Exploration model variable Age

	N(Treat=0)	N(Treat=1)	Mean Resp.(Treat=0)	Mean Resp.(Treat=1)	Uplift
[18,31]	20,901	1,223	0.088	0.193	0.104
(31,39]	17,996	1,527	0.107	0.202	0.096
(39,51]	18,229	1,457	0.104	0.229	0.125
(51,65]	17,929	738	0.071	0.244	0.173

We can observe that the net uplift increases as the age group progresses. For the age group [18,31], the uplift is 0.1045, indicating a positive effect on the response probability when targeted with treatment. Similarly, for the age group (31,39], the uplift is 0.0958, suggesting a positive impact on the response probability compared to the control group. The uplift continues to increase for the age groups (39,51] and (51,65], with values of 0.1248 and 0.1727, respectively. These results indicate that targeting customers in higher age groups has a stronger positive effect on the response probability, suggesting that the variable "Age" plays a significant role in the uplift of the marketing campaign.

5.1.2 Causal Conditional Inference Forest (CCIF)

In Uplift Modeling, different types of learners are employed to estimate the individual treatment effects, which represent the causal impact of a treatment on an outcome. The three commonly used types of learners in Uplift Modeling are S-learners, T-learners, and X-learners. In the case of CCIF, it can be categorized as an X-learner (Guelman et al., 2015). CCIF builds an ensemble of decision trees, where each tree represents a causal model. It estimates the treatment effect by separately modeling the outcome for the treatment

and control groups, similar to T-learners. However, it also incorporates the advantages of S-learners by considering the entire population when constructing the ensemble.

Tab. 4: Causal Conditional Inference Forest (CCIF) - Performance in 10 segments

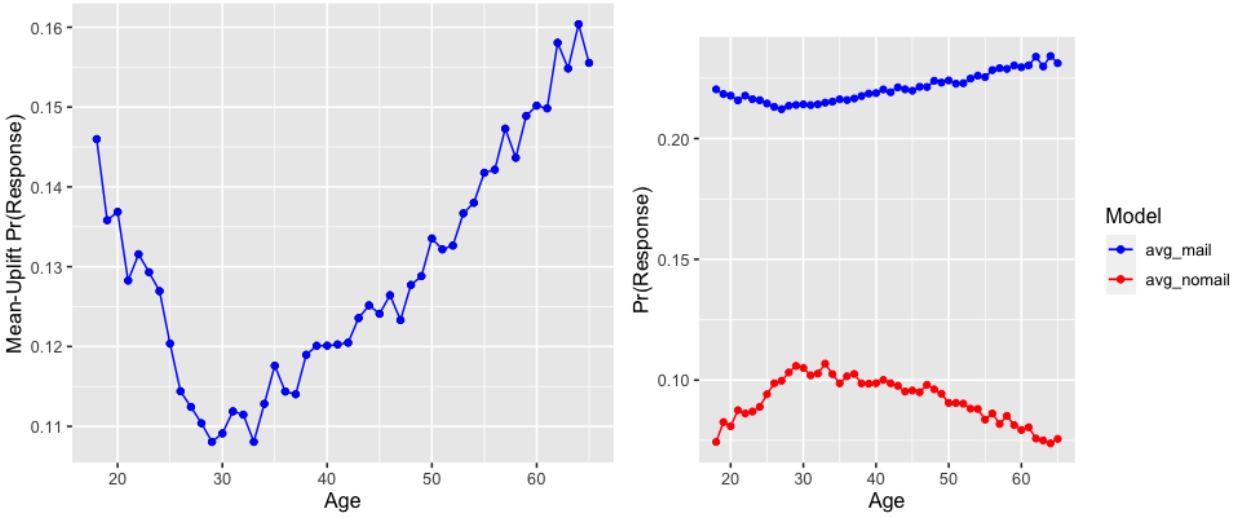
Group	N(Treat=0)	N(Treat=1)	N(Y=1, Treat=0)	N(Y=1, Treat=1)	Uplift(Treat=1 vs. Treat=0)
1	1990	10	0	36	-0.018090
2	1979	21	4	55	0.162684
3	1962	38	5	91	0.085198
4	1939	61	8	101	0.079059
5	1911	89	15	128	0.101559
6	1874	126	20	143	0.082423
7	1855	145	28	229	0.069653
8	1802	198	46	264	0.085819
9	1749	251	55	299	0.048169
10	1672	328	74	394	-0.010036

Table 4 presents the performance of the Causal Conditional Inference Forest (CCIF) in terms of uplift for different groups. In each row, we have information about a specific group identified by the number in the "group" column. The "N(Treat=1)" and "N(Treat=0)" columns represent the number of individuals in the treatment and control groups, respectively. The "N(Y=1, Treat=1)" and "N(Y=1, Treat=0)" columns show the number of positive responses in the treatment and control groups. Finally, the "Uplift(Treat=1 vs. Treat=0)" column displays the calculated uplift for each group. Looking at the uplift values, we can observe that for some groups, the uplift is positive, indicating a positive effect of the treatment on the response probability. For example, in group 2, the uplift is 0.162684, suggesting a significant improvement in the response rate for individuals in this group when exposed to the treatment. Similarly, group 5 shows an uplift of 0.101559, indicating a positive impact of the treatment. These results imply that targeting individuals in these specific groups has the potential to generate higher response rates and thus enhance marketing effectiveness. On the other hand, some groups exhibit negative uplift values, such as group 10 with an uplift of -0.010036. This suggests that the treatment may have a detrimental effect on the response probability for individuals in this group. Understanding such negative uplift values is crucial for marketing decision-making, as it helps identify segments where interventions may not be effective or could even have adverse effects. Overall, analyzing the uplift values provides valuable insights into the performance of the CCIF model and helps guide marketing strategies by identifying the groups that are most likely to respond positively to the treatment.

Variables importance

Tab. 5: Variable Importance - CCIF

Variable	Importance
Age	0.25
Logins Mobile	0.15
Customer Tenure (Months)	0.12
Overdraft	0.08
Volume of EUR Inflows	0.18
Volume of EUR Outflows	0.11
Number of Girocard Transactions (90 days)	0.11



(a) Illustration of CCIF with age mean response

(b) Illustration of CCIF with age response

Fig. 4: Illustration of CCIF performance with respect the variable 'age'

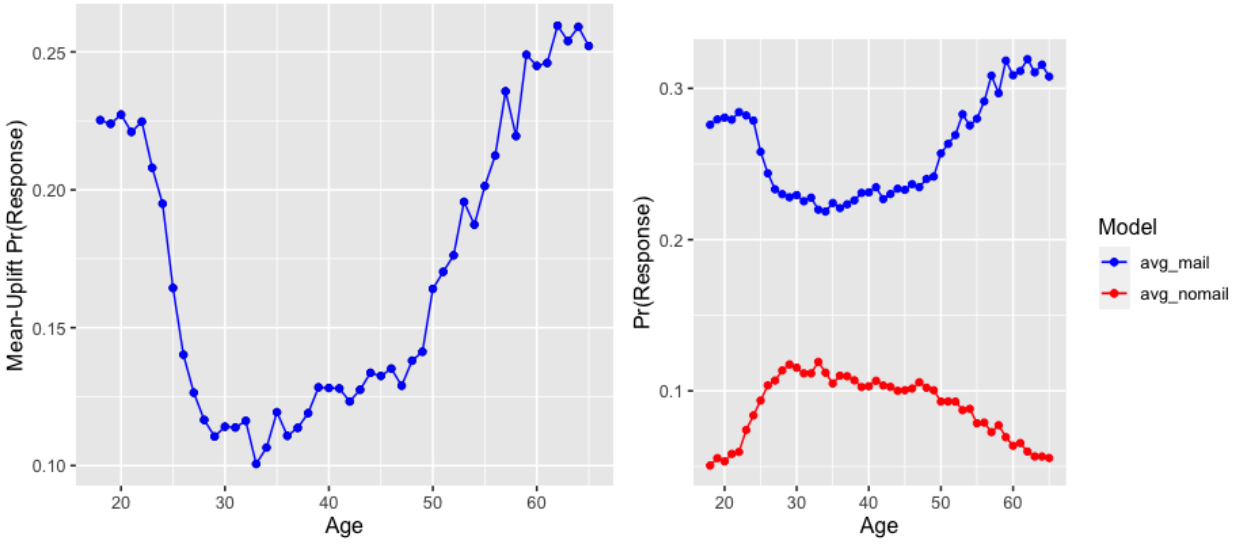
5.1.3 Uplift Random Forest

The Uplift Random Forest (URF) is an algorithm used in uplift modeling in cross-selling marketing. It combines the principles of random forests with the notion of uplift to predict the individual treatment effects and identify customers who are most likely to respond positively to a marketing intervention (Guelman et al., 2015). In the URF algorithm, an ensemble of decision trees is built by sampling bootstrap datasets and constructing trees that consider the treatment indicator variable, outcome variable, and covariates. Each tree in the ensemble represents a causal model, capturing the heterogeneity in treatment effects across different individuals (Guelman et al., 2015). During the prediction step, the URF calculates predicted outcomes separately for the treatment and control groups using the constructed trees. The uplift prediction is obtained by taking the difference between the predicted outcomes for the treatment and control groups. By aggregating the uplift predictions from all trees in the ensemble, an average uplift prediction is obtained. The Uplift Random Forest can be considered as an X-learner (Künzel et al., 2019c). It incorporates both the S-learner and T-learner frameworks in its approach. Similar to the T-learner, the URF models the outcome separately for the treatment and control groups to capture the heterogeneity in treatment effects. Additionally, like the S-learner, it considers the entire population and leverages the ensemble of trees to account for the overall treatment effect. The model is fitted using the training set and undergoes checks for balance between the Treatment and Control groups, as well as verifying the conditional independence of treatment and covariates. Net Information Value and Weight of Evidence calculations are performed. The model's predictions on the validation set are utilized to compute net uplift and assess the model's performance across ten segments in terms of uplift.

Tab. 6: Uplift Random Forest - Performance in 10 segments

Group	N(Treat=0)	N(Treat=1)	N(Y=1, Treat=0)	N(Y=1, Treat=1)	Uplift(Treat=1 vs. Treat=0)
1	1987	14	0	40	-0.020131
2	1977	22	5	67	0.193383
3	1972	28	4	79	0.102796
4	1931	69	8	121	0.053280
5	1917	83	11	136	0.061586
6	1864	136	30	175	0.126704
7	1854	146	27	212	0.070584
8	1804	196	40	252	0.064392
9	1734	266	52	286	0.030552
10	1693	307	78	372	0.034343

Table 6 presents the performance of the Uplift Random Forest (URF) model in 10 segments, providing valuable insights into the uplift achieved for different groups. Looking at the values, we can observe the following patterns: in group 1, the treatment had a negative uplift of -0.020131, indicating a decrease in the response probability compared to the control group. This suggests that the treatment might not be effective for individuals in this segment. Conversely, in group 2, the treatment resulted in a significant uplift of 0.193383. This indicates a substantial improvement in the response probability for individuals in this segment when exposed to the treatment. From a marketing perspective, targeting this group with the specific intervention employed by the URF model could yield highly positive outcomes. Similar positive uplifts are observed in groups 3 to 7, albeit with varying magnitudes. These groups exhibit uplifts ranging from 0.102796 to 0.126704, indicating a positive impact of the treatment on the response probability. In groups 8 to 10, the uplifts are relatively smaller, ranging from 0.030552 to 0.034343. While the uplifts may be less significant compared to other groups, they still suggest that the treatment has a positive effect on the response probability.



(a) RF model's performance (mean uplift) (b) RF model's performance (individual uplift)

Fig. 5: Illustration of RF performance with respect the variable 'age'

Variables importance

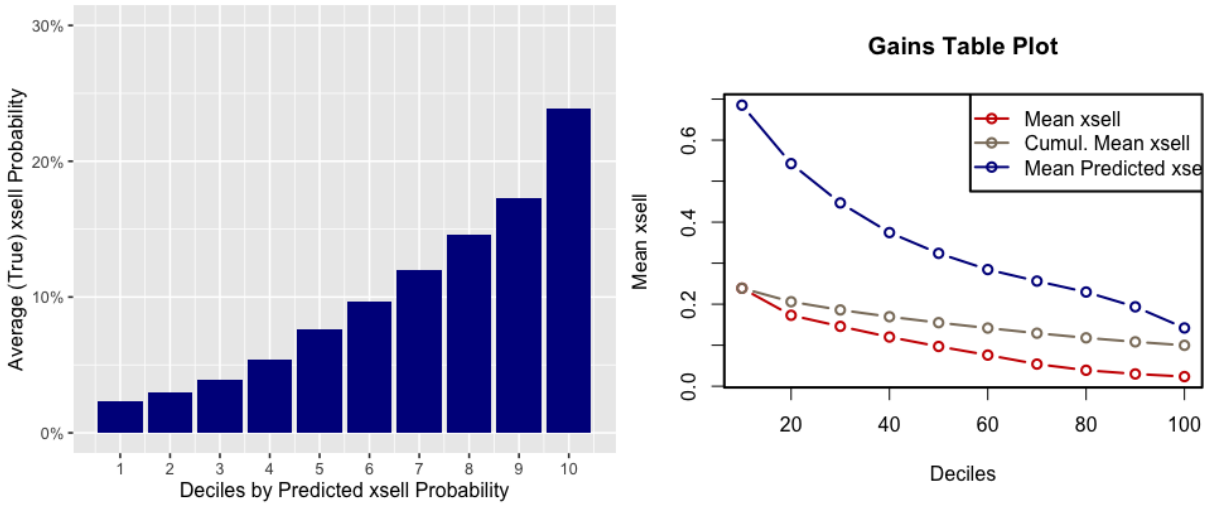
Tab. 7: Variable Importance - URF

Variable	Relative Importance
Customer Tenure (Months)	9.928067
Logins Mobile	8.703928
Age	7.772436
Volume of EUR Inflows	6.577957
Number of Girocard Transactions (90 days)	5.171046
Volume of EUR Outflows	4.818399
Overdraft	1.543085

5.2 Oversampled data frame

Tab. 8: Exploration model variable Age - SMOTE

Age Group	N(Treat=0)	N(Treat=1)	Mean Resp.(Treat=0)	Mean Resp.(Treat=1)	Uplift
[18,31]	25576	1935	0.3344	0.5214	0.1870
(31,38.4]	21986	2505	0.4351	0.5984	0.1633
(38.4,49.9]	23323	2678	0.4002	0.6434	0.2432
(49.9,65]	24387	1614	0.2982	0.6239	0.3257



(a) Lift chart by xsell probabilities (SMOTE applied)

(b) Gains table (SMOTE applied)

Fig. 6: Comparison of Lift chart and Gains table (SMOTE applied)

In Table 8 different age groups are analyzed in the context of an exploration model. The table provides information about the number of individuals in the control group (N(Treat=0)) and the treatment group (N(Treat=1)). Additionally, it presents the mean response probabilities in each group (Mean Resp.(Treat=0) and Mean Resp.(Treat=1)). The uplift column represents the uplift value, which indicates the difference in response probability between the treatment and control groups. In this case, the uplift values for each age group range from 0.1633 to 0.3257. A positive uplift value suggests that the treatment (in this case, using the SMOTE technique on the variable "Age") has a positive impact on the response probability compared to the control group. Overall, the table provides insights into the relationship between the age variable and the response probability, highlighting the potential effectiveness of using the SMOTE technique to enhance the uplift in marketing efforts.

5.2.1 Causal Conditional Inference Forest - SMOTE

Tab. 9: Causal Conditional Inference Forest (CCIF) SMOTE - Performance in 10 segments

Group	N(Treat=0)	N(Treat=1)	M.R.(Treat=0)	M.R.(Treat=1)	R.R.(Treat=0)	R.R.(Treat=1)	Uplift
1	12	1988	1	37	0.083333	0.018612	0.064722
2	36	1977	4	72	0.111111	0.036419	0.074692
3	42	1960	4	84	0.095238	0.042857	0.052381
4	86	1916	14	95	0.162791	0.049582	0.113208
5	91	1900	12	130	0.131868	0.068421	0.063447
6	131	1861	31	174	0.236641	0.093498	0.143143
7	161	2016	19	221	0.118012	0.109623	0.008389
8	176	1647	42	238	0.238636	0.144505	0.094131
9	256	1744	62	302	0.242188	0.173165	0.069022
10	276	1724	66	387	0.239130	0.224478	0.014652

Table 9 represents the performance of the Causal Conditional Inference Forest (CCIF) with SMOTE oversampling in 10 segments. Each row corresponds to a specific segment (group), and the columns provide relevant metrics. 'N(Treat=1)' and 'N(Treat=0)' represent the counts of individuals in the treatment (1) and control (0) groups, respectively. 'Mean Resp.(Treat=1)' and 'Mean Resp.(Treat=0)' represent the counts of individuals who responded positively (1) and did not respond (0) in the treatment and control groups, respectively. 'Resp. Rate(Treat=0)' and 'Resp. Rate(Treat=0)' represent the response rates (proportion of positive responses) in the treatment and control groups. 'uplift' indicates the uplift, which represents the difference in the response rates between the treatment and control groups. For example, in the first segment (group 1), there were 12 individuals in the treatment group and 1988 individuals in the control group. Only one individual in the treatment group responded positively, while 37 individuals in the control group responded positively. The response rate in the treatment group was 0.083333 (or 8.33%), while in the control group, it was 0.018612 (or 1.86%). The uplift for this segment is 0.064722 (or 6.47%), indicating a positive impact of the treatment.

Variable Importance

Tab. 10: Variable Importance - CCIF (SMOTE)

Variable	Relative Importance
logins_mobile	1.4112682
customer_tenure_months	0.9880306
vol_eur_inflows	0.7469857
vol_eur_outflows	0.6617714
age	0.4970875
nr_girocard_trx_90d	0.3891323
overdraft	0.1070380

In Table 10 the relative importance of different variables is presented. The "Relative Importance" column showcases the significance of each variable in predicting the outcome using the CCIF model with SMOTE. The values in the "Relative Importance" column range from 0.1070380 to 1.4112682. A higher value indicates a higher level of importance for that particular variable in predicting the target variable or outcome. In this case, variables such as logins_mobile, customer_tenure_months, and vol_eur_inflows have relatively higher importance compared to others like overdraft. The relative importance values provide insights into which variables have a stronger influence on the outcome within the context of the CCIF model with SMOTE. This information can be used to prioritize variables for further analysis or feature selection in the uplift modeling process.

Tab. 11: Variable Importance - Uplift Random Forest - SMOTE

Variable	Relative Importance
logins_mobile	9.6444142
customer_tenure_months	7.7086847
nr_girocard_trx_90d	7.4074500
age	7.3811943
vol_eur_inflows	6.7906385
vol_eur_outflows	6.3088539
overdraft	0.8327058

5.2.2 Uplift Random Forest - SMOTE

Table 11 displays the variable importance in the Uplift Random Forest model with SMOTE oversampling.

The "Relative Importance" values indicate the contribution of each variable in predicting uplift. A higher value indicates a greater impact on the model's performance. In this table, we observe that "logins_mobile" has the highest relative importance, followed by "customer_tenure_months," "nr_girocard_trx_90d," and "age." These variables play a significant role in predicting the uplift, as they have higher relative importance compared to other variables such as "vol_eur_inflows," "vol_eur_outflows," and "overdraft". Moreover, this table provides insights into the importance of each variable in the Uplift Random Forest model with SMOTE oversampling, helping researchers and practitioners understand which features have the most influence on predicting uplift in their specific context.

Tab. 12: Uplift Random Forest SMOTE - Performance in 10 segments

Group	N(Treat=0)	N(Treat=1)	M.R.(Treat=0)	M.R.(Treat=1)	R.R.(Treat=0)	R.R.(Treat=1)	Uplift
1	15	1985	1	34	0.066667	0.017128	0.049538
2	15	1985	2	71	0.133333	0.035768	0.097565
3	31	1969	3	72	0.096774	0.036567	0.060207
4	59	1941	13	95	0.220339	0.048944	0.171395
5	107	1893	11	115	0.102804	0.060750	0.042054
6	124	1876	16	172	0.129032	0.091684	0.037348
7	168	1832	33	217	0.196429	0.118450	0.077979
8	213	1787	46	289	0.215962	0.161724	0.054239
9	255	1745	67	323	0.262745	0.185100	0.077645
10	280	1720	63	352	0.225000	0.204651	0.020349

In Table 12 are displayed the uplift performance of the Uplift Random Forest model in 10 different segments. Looking at the values in the "uplift" column, we can observe that the uplift values vary across the different segments. Positive uplift values indicate that the treatment group has a higher response rate compared to the control group, suggesting that the treatment has a positive effect on the target outcome. In this table, we can see that segments 2, 4, 7, and 9 have relatively higher uplift values compared to the other segments. This implies that the treatment has a more significant impact on the target outcome in these segments. On the other hand, segments 5 and 6 have relatively lower uplift values, suggesting a lower impact of the treatment in these segments.

6 Results & discussion

In this study on the influence of SMOTE on uplift models for cross-selling in a digital bank, we conducted experiments comparing the performance of two models: the Causal Conditional Inference Forest (CCIF) and the Uplift Random Forest (URF). We evaluated both models without applying SMOTE and with the application of SMOTE to address the issue of high class imbalance in the target variable. The global net uplift effect will be used as metrics to compare these different settings. The global net uplift effect is a measure that quantifies the overall impact of the treatment on the target outcome (Nyberg Klami, 2023e).

In our experiments, we obtained the following net uplift effects:

- CCIF model without SMOTE: 0.123649
- Uplift Random Forest model without SMOTE: 0.1539789
- CCIF model with SMOTE: 0.2512693
- Uplift Random Forest model with SMOTE: 0.2678449

Comparing these net uplift effects, we observe that both models exhibit a positive uplift effect. However, the models with SMOTE applied demonstrate higher net uplift effects compared to the models without. This finding suggests that SMOTE plays a crucial role in enhancing the performance of uplift models, particularly in the context of cross-selling where the target variable suffers from high class imbalance. Our experimental results highlight the positive effect of oversampling on uplift modeling. The higher net uplift effects obtained from the models with SMOTE indicate a more substantial impact of the treatment on the target outcome. In conclusion, our empirical experiments demonstrate that SMOTE is a promising solution for improving the performance of uplift models in scenarios with high imbalance. By addressing the class imbalance issue, SMOTE enhances the models' ability to identify the target population that benefits the most from the treatment, leading to more effective cross-selling strategies in the digital retail banking industry. However, further studies and empirical evidence will be needed to confirm these results, which are currently based more on empirical evidence than on solid theoretical foundations.

7 Conclusion

In conclusion, our study focused on the influence of SMOTE on uplift models for cross-selling in a digital bank. We compared the performance of two models, the Causal Conditional Inference Forest (CCIF) and the Uplift Random Forest (URF), both with and without the application of SMOTE to address the issue of high class imbalance in the target variable. Based on our empirical experiments and analysis, we can draw the following conclusions:

1. Class imbalance is a common challenge in uplift modeling for cross-selling, where the response rates between the treated and control groups may vary significantly. This class imbalance can lead to biased estimations and hinder accurate model training.
2. The application of SMOTE, a technique that oversamples the minority class by creating synthetic samples, proves to be an effective approach for mitigating the class imbalance problem. By balancing the class distribution, SMOTE improves the performance of uplift models and enables them to capture the underlying patterns and relationships in the data more accurately.
3. Our results demonstrate that SMOTE has a positive impact on the net uplift effect, which quantifies the overall influence of the treatment on the target outcome. The uplift models with SMOTE exhibit higher net uplift effects compared to those without SMOTE, indicating that SMOTE enhances the models' ability to identify the target population that benefits the most from the treatment.
4. The positive effect of SMOTE on uplift modeling has practical implications for digital banks and other industries where targeted marketing and personalized recommendations are crucial. By leveraging SMOTE and uplift models, businesses can optimize their cross-selling efforts and tailor their marketing campaigns to the individuals most likely to respond positively to the treatment.

References

[1] Apt, C. (2010). The role of Machine learning in business optimization. In International Conference on Machine Learning (pp. 1–2). https://icml.cc/Conferences/2010/papers/903.pdf

[2] Ascarza, E. (2018). Retention Futility: Targeting High-Risk Customers Might be Ineffective. Journal of Marketing Research, 55(1), 80–98. https://doi.org/10.1509/jmr.16.0163

[3] Chawla, N. V., Bowyer, K. W., Hall, L. J., Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling technique. Journal of Artificial Intelligence Research, 16, 321–357. https://doi.org/10.1613/jair.953

[4] Choi, Y., & Choi, J. S. (2022). How does Machine Learning Predict the Success of Bank Telemarketing? Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-1695659/v1

[5] Devriendt, F., Berrevoets, J., & Verbeke, W. (2021). Why you should stop predicting customer churn and start using uplift models. Information Sciences, 548, 497–515. https://doi.org/10.1016/j.ins.2019.12.075

[6] Gubela, R. M., Bequé, A., Lessmann, S., & Gebert, F. (2019). Conversion uplift in e-commerce: A Systematic benchmark of modeling strategies. International Journal of Information Technology and Decision Making, 18(03), 747–791. https://doi.org/10.1142/s0219622019500172

[7] Guelman, L., Guillén, M., Pérez-Marín, A. M. (2015). Uplift Random forests. Cybernetics and Systems, 46(3–4), 230–248. https://doi.org/10.1080/01969722.2015.1012892

[8] Gutierrez, P. (2017, July 4). Causal Inference and Uplift Modelling: A Review of the literature. PMLR. https://proceedings.mlr.press/v67/gutierrez17a.html

[9] GuyonIsabelle, ElisseeffAndré. (2003). An introduction to variable and feature selection. Journal of Machine Learning Research. https://doi.org/10.5555/944919.944968

[10] Hu, J. (2022). Customer feature selection from high-dimensional bank direct marketing data for uplift modeling. Journal of Marketing Analytics, 11(2), 160–171. https://doi.org/10.1057/s41270-022-00160-z

[11] Jacob, D. J. (2021). CATE Meets ML - Conditional Average Treatment Effect and Machine Learning. Social Science Research Network. https://doi.org/10.2139/ssrn.3816558

[12] Künzel, S. R., Sekhon, J. S., Bickel, P. J., & Yu, B. (2019). Metalearners for estimating heterogeneous treatment effects using machine learning. Proceedings of the National Academy of Sciences, 116(10), 4156–4165. https://doi.org/10.1073/pnas.1804597116

[13] Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020). Machine Learning with Oversampling and Undersampling Techniques: Overview Study and Experimental Results. https://doi.org/10.1109/icics49469.2020.239556

[14] Nyberg, O., & Klami, A. (2023). Exploring uplift modeling with high class imbalance. Data Mining and Knowledge Discovery, 37(2), 736–766. https://doi.org/10.1007/s10618-023-00917-9

[15] Rombaut, E., & Guerry, M. (2020). The effectiveness of employee retention through an uplift modeling approach. International Journal of Manpower, 41(8), 1199–1220. https://doi.org/10.1108/ijm-04-2019-0184

[16] Rzepakowski, P., & Jaroszewicz, S. (2012). Uplift modeling in direct marketing. Journal of Telecommunications and Information Technology, 43–50. http://dlibra.itl.waw.pl/dlibra-webapp/Content/1229/ISSN_1509-4553_2_2012_43.pdf

[17] Sołtys, M., Jaroszewicz, S., & Rzepakowski, P. (2014). Ensemble methods for uplift modeling. Data Mining and Knowledge Discovery, 29(6), 1531–1559. https://doi.org/10.1007/s10618-014-0383-9

[18] Olaya D, Coussement K, Verbeke W (2020) A survey and benchmarking study of multitreatment uplift modeling. Data Min Knowl Disc 34(2):273–308 Olaya et al. (2020)

[19] Zhao, Z., & Harinen, T. (2019). Uplift Modeling for Multiple Treatments with Cost Optimization. https://doi.org/10.1109/dsaa.2019.00057

[20] Ngai, E. W., & Wu, Y. (2022). Machine learning in marketing: A literature review, conceptual framework, and research agenda. Journal of Business Research, 145, 35–48. https://doi.org/10.1016/j.jbusres.2022.02.049

A Appendix

The analysis code is available here:
https://github.com/mnlscn/predicting-modelling-INGBank/blob/main/final_report_code.r

B Statutory Declaration

An das Ende der Bachelorarbeit/ Masterarbeit ist eine ehrenwörtliche Erklärung über die Ver- wendung benutzter Hilfsmittel und das Zitieren zu setzen. Sie hat folgenden Wortlaut:

"*Ich versichere hiermit, dass ich die vorliegende Arbeit selbständig und ohne Benutzung anderer als der angegebenen Quellen und Hilfsmittel verfasst habe. Wörtlich übernommene Sätze oder Satzteile sind als Zitat belegt, andere Anlehnungen, hinsichtlich Aussage und Umfang, unter Quellenangabe kenntlich gemacht. Die Arbeit hat in gleicher oder ähnlicher Form noch keiner Prüfungsbehörde vorgelegen und ist nicht veröffentlicht. Sie wurde nicht, auch nicht auszugs- weise, für eine andere Prüfungs- oder Studienleistung verwendet.*"

Ort, Datum

Frankfurt Am Main 14/07/2023

Unterschrift