Predicting Customer Cross-buying Behavior using Ensemble Machine Learning Methods

Seminar Thesis

Submitted to
Hon.Prof. Dr. Martin Schmidberger
Faculty of Economics and Business Administration
Johann Wolfgang Goethe University
Frankfurt am Main

by
Ngoc Khanh Uyen Tran
Program: M.Sc. Management Science
Matriculation Number: 8518013
E-Mail: s0657641@stud.uni-frankfurt.de

Seminar: Data Mining in Marketing:
Data Driven Customer Analytics with Machine Learning

Semester: Summer 2025

Abstract

This study investigates the effectiveness of ensemble machine learning methods for predicting customer cross-buying behavior and identifies which factors influence the targeted cross-buying decision. Throughout different preprocessing and feature engineering techniques, Random Forest and Gradient Boosting models were implemented to predict the cross-buying likelihood from a dataset of 100,000 customers. Performance evaluation on validation data demonstrates that XGBoost algorithm achieves better discriminative power with AUC of 0.795 and GINI coefficient of 0.589, while Random Forest provides a more balanced precision-recall rates. Variable importance measures analysis recognized income as the most influential predictor across both models, followed by internet banking login frequency and customer tenure. These findings suggest customer segmentation strategies based on income classification, along with real-time digital engagement monitoring systems and customer lifecycle-specific marketing approaches for enhancing cross-selling campaign effectiveness.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

RF	Random Forest
GB	Gradient Boosting
XGB	Extreme Gradient Boosting
TDL	Top Decile Lift
ROC	Receiver-operating Characteristic Curve
AUC	Area Under the ROC Curve
VIM	Variable Importance Measures

1 Introduction

Cross-buying behavior, defined as propensity of customers to purchase additional products or services from the same provider, represents a significant opportunity for increasing firms' profitability and competitive advantage. Traditional approaches to cross-selling prediction have relied primarily on demographic and rule-based segmentation, which often fail to capture the complex, non-linear relationships in customer behavior data (Li et al., 2005). On the other hand, ensemble methods have been recognized as a more optimal approach compared with single classifiers to predict customer behavior with robust performance and enhanced accuracy (Bogaert & Delaere, 2023).

This study implements ensemble machine learning methods, specifically Random Forest and Gradient Boosting algorithms, to address the research question: *"What are the key factors that drive customer's decision to cross-buy?"*. The primary objectives include developing predictive models for cross-buying behavior prediction, identifying the most influential factors through variable importance measures, and translating model insights into actionable marketing implications for cross-selling campaigns.

2 Literature Review

2.1 Predicting Customer Cross-buying Behavior

Cross-buying, the purchase of multiple product categories from a single provider, has been a focus of research due to its potential impact on customer loyalty and firm profitability. Predicting customer cross-buying is essential for an effective marketing campaign with minimal wasted efforts, especially in financial services sector where the likelihood of cross-buying is considered as relatively rare. With the application of machine learning, predictive modeling has become particularly valuable for predicting customer churn and loan outcomes in finance (Zhang, 2025). Beyond traditional machine learning algorithms, advanced analytics like ensemble methods, deep learning and neural networks can be applied to customer data to predict cross-selling responses for various financial products (Dutta & Bhattacharya, 2019). Understanding the factors affecting cross-buying intention is essential for strategy development, as financial institutions can better allocate resources and improve their cross-selling strategies, leading to an enhanced competitiveness in the financial sector (Fan et al., 2011).

The table below presents some of the key existing studies on predicting customer cross-buying behavior and which factors are the most influential to the target outcome.

Table 1: Literature review on predicting customer cross-buying behavior

Studies	Methodology	Key Findings
Estrella-Ramón (2017)	Applying logistic regression on a dataset of 2,187 randomly selected customers from a large bank, including features related to individual loyalty, cross-buying behaviors, and demographic indicators	The study examines the impact of customer loyalty and cross-buying behavior on their choice of financial services from the same provider, with level of services usage and previous service usage were recognized as the main predictor of additional purchase intention.
Verhoef and Donkers (2005)	Using a dataset of 3,317 customers from a financial-services provider, including explanatory variables such as customer social demographics and transactional behavior. The study analyzes the effect of acquisition channels on retention and cross-selling, predicting the retention rates and cross-buying probabilities.	The direct-mail acquisition channel performs poorly on both retention and cross-selling, while radio and TV perform poorly only on retention. The firm's website generally performs well for retention, except for automobile insurance. Acquisition channels have a significant impact on customer loyalty but a weaker effect on cross-buying.
Tung and Carlson (2015)	Developing a research model based on literature with structural equation modeling to test relationships among 269 customers of retail banks in Hong Kong	Customer cross-buying intentions are primarily associated with image conflicts about the provider's abilities to deliver high quality financial services from different activities. High levels of customer loyalty are also a key factor in cross-buying intentions.
Fan et al. (2011)	Developing a research model to assess factors affecting cross-buying intention of bancassurance based on Importance Performance Analysis	Payment equity and experience are perceived as the most important factor. Other factors with moderate influences are image, service convenience, interpersonal relationships, trust, product variety and pricing.

2.2 Predictive Modeling using Ensemble Methods

Ensemble methods have gained significant reputation for its strength in predictive modeling for marketing applications, particularly in forecasting customer behavior such as cross-buying, churn, and purchasing patterns. These techniques have demonstrated strong performance compared to traditional approaches. Various ensemble strategies, including bagging, boosting, stacking, and voting, have been explored to enhance predictive accuracy (Abbasimehr et al., 2014; Nguyen et al., 2023). For instance, Random Forest, a popular ensemble method, has shown exceptional results in identifying cross-buyers while maintaining high accuracy (Werb & Schmidberger, 2021). Specifically for churn prediction, boosting techniques have been recognized as resulting to the best outcomes (Abbasimehr et al., 2014). Recent studies have consistently ranked ensemble methods higher than single classifiers across multiple performance metrics, including AUC, expected maximum profits, accuracy, F1 measure, and top-decile lift (Bogaert & Delaere, 2023).

The below table shows some of the empirical studies on applications of ensemble methods, specifically in predicting behavioral outcomes for marketing and customer analytics.

Table 2: Literature review on predictive modeling using ensemble methods

Studies	Methodology	Key Findings
Werb and Schmidberger (2021)	Applying ensemble machine learning and interpretability methods on predictive modeling to predict customer cross-buying likelihood in finance services sector and identify influential variables on targeted behavior	Random forest is recognized as the most optimal model, as identified 73.3% of cross-buyers with an accuracy of 72.5%. The study also demonstrated the use of interpretability methods, such as VIM, PDP, ICE and LIME, to enhance understanding of outputs from the complex ensemble methods.
Nguyen et al. (2023)	Proposing a tailored model for customer behavior prediction (churn and purchasing behavior) using ensemble learning strategies (stacking and voting), which combining Random Forest, CNN, and Boosting algorithms with hard voting, as well	The proposed model achieves high performance on four datasets: Campaign (AUC: 94.82%, F1: 68.29%), Cell2Cell (AUC: 86.82%, F1: 79.99%), Bank (AUC: 87.81%, F1: 89.11%), and Online Shoppers (AUC: 94.55%, F1: 70.79%)

	as optimizing classifier weights using an evolutionary algorithm. The study evaluates model performance on four datasets: Campaign, Cell2Cell, Bank, and Online Shoppers	
Abbasimehr et al. (2014)	Comparative assessment of four ensemble methods: Bagging, Boosting, Stacking, and Voting. Base learners: C4.5 Decision Tree, Artificial Neural Network, Support Vector Machine, RIPPER. Sampling techniques: Oversampling, Synthetic Minority Over-sampling Technique (SMOTE)	The study shows that SMOTE did not increase predictive performance. Ensemble learning is recognized as significantly improved individual base learners' performance in terms of AUC, sensitivity, and specificity. Boosting was the best ensemble method, with Boosting RIPPER and Boosting C4.5 being the top performers.
Kumar et al. (2023)	Building a model to predict customer churn likelihood from a dataset acquired from Kaggle model using different ensemble algorithms.	Random Forest model was identified as the best fit due to its high accuracy, specificity, AUC, and F1 score.

3 Methodology

3.1 Ensemble Machine Learning Methods

Ensemble learning represents a powerful paradigm in machine learning that combines multiple base learners to achieve more powerful predictive performance compared to individual models. The fundamental principle of ensemble methods involves combining multiple models to reduce prediction variance, mitigate overfitting, and enhance generalization capability across unseen data (Breiman, 1996; Dietterich, 2000). Based on results from existing literature on predicting customer behavior with machine learning algorithms, this study implements two ensemble methods: (1) Random Forest and (2) Gradient Boosting to predict customer cross-buying behavior.

3.1.1 Random Forest

According to Breiman (2001), Random Forest (RF) is a bagging-based ensemble method that constructs multiple decision trees using bootstrap sampling and random feature selection at each split node. The algorithm builds independent decision trees, where each tree t is trained on a bootstrap sample drawn with replacement from the original training dataset. At each internal node, a random subset of features is selected from the total features for classification tasks. The final prediction for classification is determined through majority voting across all trees. This aggregation mechanism reduces prediction variance through the law of large numbers while maintaining low bias.

RF offers several advantages for predictive modeling, including implicit feature selection through random subspace sampling, robust handling of missing values, and computational efficiency through parallel tree construction. Compared with other methods, RF can handle mixed data types efficiently, especially with categorical variables with high cardinality, and provides built-in measures of prediction confidence through out-of-bag error estimation (Liaw & Wiener, 2002).

3.1.2 Gradient Boosting

According to Friedman (2001), Gradient Boosting (GB) implements a sequential ensemble approach that builds models iteratively, with each subsequent model correcting errors made by the ensemble of previous models. The boosting procedure follows a gradient descent approach in function space, where at each iteration, a new "weak", meaning base learner, is fitted to the negative gradient of the loss function that is computed for each training instance. This gradient-based optimization enables the algorithm to focus learning capacity on difficult-to-predict instances, progressively reducing prediction errors through iterative refinement.

This study specifically implements XGBoost (Extreme Gradient Boosting), a more optimized version of GB that applies several algorithmic enhancements including regularized objective functions, second-order gradient information, and advanced tree pruning strategies. The XGBoost (XGB) framework provides better performance for prediction tasks, with column block structure for parallel tree learning, cache-aware access patterns for improved computational efficiency, and handling of sparse features and missing values (Chen & Guestrin, 2016).

3.2 Performance metrics

For marketing analytics, the evaluation of predictive models requires metrics that translate statistical performance into actionable insights for effective campaign and resource allocation. Therefore, this study considers three following metrics as primary: TDL, GINI coefficient, AUC. High values in these metrics mean that marketing resources can be focused on the most promising customer segments, maximizing return-on-investment.

Top Decile Lift (TDL) measures how effectively a model identifies high-probability check-in customers within the top 10% of predicted probabilities compared to random selection (Berry & Linoff, 2004). This metric directly addresses the practical marketing question: *"How much better is our targeting compared to random customer selection?"*. For instance, a TDL of 2.0 indicates the model identifies twice as cross-buying customers in the top decile as random selection would achieve.

Meanwhile, GINI Coefficient provides a comprehensive measure of model discrimination across all probability thresholds, related to the Area Under the ROC Curve (AUC) through the relationship $GINI = 2*AUC - 1$ (Fawcett, 2006). This metric quantifies the model's ability to distinguish between cross-buying and no cross-buying observations, with values range from 0 (random performance) to 1 (perfect discrimination).

On the other hand, F1-Score balances precision and recall considerations, particularly relevant given the class imbalance where cross-buying represent 10% of observations. The F1-score computes the harmonic mean of precision and recall: $F1 = 2*(Precision*Recall)/(Precision + Recall)$, providing a single metric that accounts for both false positive and false negative costs (Sokolova & Lapalme, 2009). However, for imbalanced datasets, such as the one utilized in this study, one class significantly outnumbers the other and imposes challenges for binary classifiers, which often resulting in high accuracy but low precision and recall for minority classes (Shahzad Ashraf et al., 2020; Sumaira Ashraf & Toqeer Ahmed, 2020). Accuracy and Specificity are considered less important because they can be misleading in an imbalanced dataset, in which high accuracy can be achieved by simply predicting the majority class, whereas Specificity focuses on correctly identifying the negative class while the primary goal of this study is to correctly identify the minority class, such as customers who cross-buy.

3.3 Interpretation Method

In order to understand the underlying decision-making mechanisms, this study utilizes Variable Importance Measures (VIM) to quantify the contribution of individual features to the predictive performance of ensemble models, providing actionable insights into the factors driving customer cross-buying behavior. This method computes numerical measures to indicate how much influence each variable has on the predictive performance of a model. For instance, RF calculating importance based on the total reduction in node impurity achieved by splits on each variable across all trees in the ensemble, providing confidence intervals for importance estimates and assessment of ranking stability. Whereas XGB using gain-based importance measures that quantify the average improvement in loss function achieved by splits on each feature across all boosting iterations (Breiman, 2001; Genuer et al., 2010; Chen & Guestrin, 2016).

Variable importance scores are normalized to a 0-100 scale for interpretability, with 0 (not important at all) and 100 (the most important), suggesting customer segmentation strategy based on likelihood to cross-buy while maintaining mathematical rigor in the underlying importance calculations.

4 Data

In this study, the dataset was acquired from a major banking and financial service institution in Germany, including 100,000 observations with 34 variables ranging from customer demographics and transaction behavior, with target variable *'xsell'* showing whether the client purchased loan after marketing campaign or not. The details of variables and their definitions are provided in the Appendix: Variable Description.

4.1 Preprocessing and Feature Engineering

Initial data preprocessing involves imputation of missing values in the dataset, which was being done accordingly to the nature and context of each variable. Firstly, for transaction volume variables, *'vol_eur_inflows'* and *'vol_eur_outflows'*, were inspected with 489 missing values for each variable. According to financial perspective, this study assumes that missing entries in these variables likely indicated the absence of transactions. Therefore, all missing values were replaced with zero, preserving the interpretability of the features while avoiding potential bias from imputation. Secondly, for continuous variables, the variable *'age_at_onboarding'* contained 3 missing values, whereas 'ext_house_size' and 'ext_purchase_power' had 419 and 464 missing values, respectively. This study assumes that these data are missing at random, and subsequently, missing values were imputed using the mean of the observed data for each variable. This method helps to maintain the overall distribution. All imputed values were rounded to the nearest integer to align with the variable's scale. After these procedures, the dataset was checked again for missing values, confirming that all entries were completely imputed.

Features were also engineered to enhance the dataset's predictive power and analytical value. For instance, the original categorical variable *'marital_status'* was transformed into a binary indicator variable, 'married'. Clients identified as married *('marital_status == "VH")* were assigned a value of 1, and all others marital status were classified as 0. This transformation enables straightforward inclusion in models that require numerical input and facilitates interpretability. Furthermore, to fully exploit the information encoded in the *'marital_status'* categorical variable, dummy variables for each status level were generated using the '*dummy.code'* function from the *'psych'* package. The resulting dummy variables were converted to factors and appended to the main dataset.

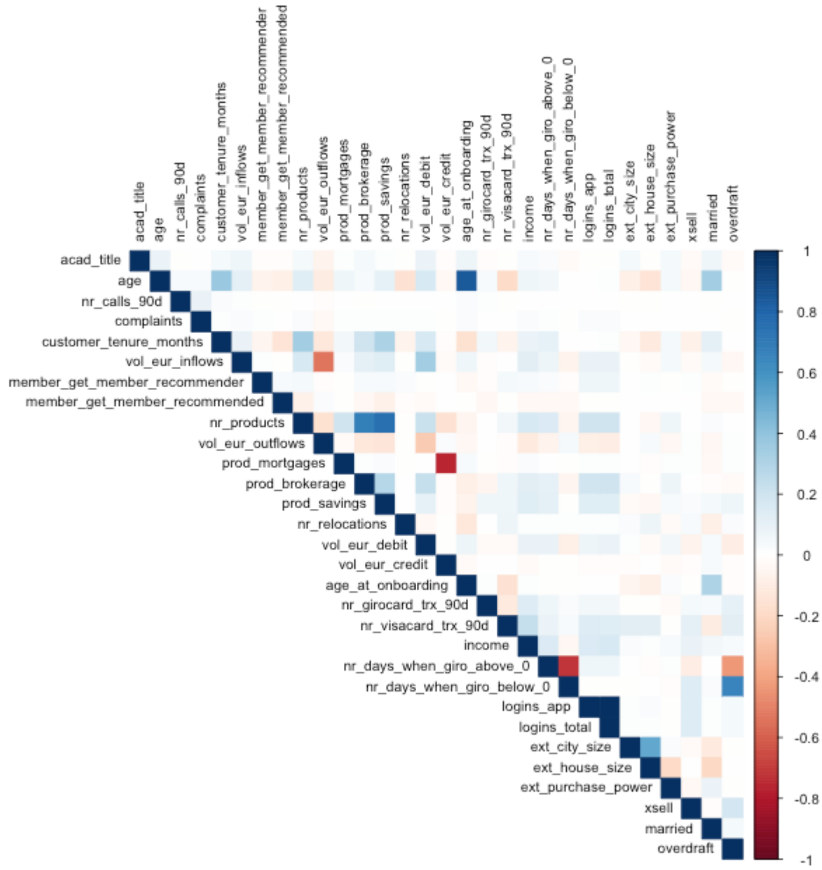
The variable *'overdraft'* was also created to capture whether a client had utilized their overdraft facility within the last 90 days. Specifically, if *'nr_days_when_giro_below_0'* was greater than zero, the client was flagged as having used overdraft (value 1); otherwise, the value was indicated as 0.

4.2 Correlations

After preprocessed, the clean dataset was reloaded for descriptive statistics and correlation analysis. This examination is needed to check for the relationships among numerical variables, in which to identify potential multicollinearity and support feature selection for predictive modeling later.

Pairwise Pearson correlations were calculated among numerical variables in the dataset, including demographic, behavioral, transactional, and external attributes, as well as the target variable *('xsell')*. The matrix displays pairwise Pearson correlation coefficients, ranging from -1 (perfect negative linear relationship) to +1 (perfect positive linear relationship), with color gradients from red (negative) to blue (positive).

Figure 1: Correlations of numeric variables



Most variable pairs exhibit weak correlations, as indicated by the predominance of light colors, suggesting that the dataset is not heavily affected by multicollinearity. However, some variables stand out with stronger correlations. For instance, the number of products ('nr_products') shows moderate positive correlations with financial variables such as 'vol_eur_inflows' and 'vol_eur_outflows', suggesting that customers holding more products tend to have higher transaction volumes. Similarly, 'nr_girocard_trx_90d' and 'nr_visacard_trx_90d' are positively correlated, indicating that customers who are active with one card type are likely to be active with another. There is a notable positive relationship between 'logins_app' and 'logins_total', since app logins may as well contribute to total logins.

Two 'nr_days_when_giro_above_0' and 'nr_days_when_giro_below_0' variables show a moderate negative correlation, reflecting the mutually exclusive nature of positive and negative account balances. External variables such as 'ext_city_size', 'ext_house_size', and 'ext_purchase_power' show weak to moderate positive correlations with each other, suggesting a general trend that customers in larger cities may have larger houses and greater purchase power.

The target variable ('xsell') displays only modest correlations with most predictors, suggesting that the predictors capture complementary information and that the cross-buying likelihood is not determined by any single feature. However, slightly stronger associations may be observed with variables such as 'nr_products', 'income', and 'overdraft', indicating their potential relevance for predictive modeling.

4.3 Descriptive Statistics

Descriptive statistics were calculated for all numerical variables, including measures such as mean, median, standard deviation, minimum, and maximum. The summary table below provides a comprehensive overview of the central tendency and dispersion within the cross-buying dataset.

Table 3: Descriptive statistics of numeric variables

Statistic	N	Mean	St. Dev.	Min	Max
acad_title	100,000	0.021	0.143	0	1
age	100,000	41.411	13.303	16	72
nr_calls_90d	100,000	0.014	0.235	0	20
complaints	100,000	0.013	0.145	0	11
customer_tenure_months	100,000	110.041	84.909	0	440
vol_eur_inflows	100,000	7,371.719	26,074.400	0.000	2,000,000.000
member_get_member_recommender	100,000	0.026	0.160	0	1
member_get_member_recommended	100,000	0.020	0.140	0	1
nr_products	100,000	2.608	1.223	1	17
vol_eur_outflows	100,000	-6,023.904	22,812.690	-1,858,000.000	0.000
prod_mortgages	100,000	0.026	0.224	0	13
prod_brokerage	100,000	0.394	0.579	0	12
prod_savings	100,000	1.050	0.705	0	9
nr_relocations	100,000	0.099	0.315	0	4
vol_eur_debit	100,000	28,472.050	99,170.160	0.000	13,705,492.000
vol_eur_credit	100,000	-2,472.402	25,695.830	-3,104,629.000	0.000
age_at_onboarding	100,000	32.221	12.401	-4	70
nr_girocard_trx_90d	100,000	15.116	29.097	0	459
nr_visacard_trx_90d	100,000	37.694	50.191	0	1,034
income	100,000	1,942.814	2,398.713	0	168,451
nr_days_when_giro_above_0	100,000	57.098	18.134	0	65
nr_days_when_giro_below_0	100,000	4.916	13.938	0	65
logins_app	100,000	106.518	150.465	0	7,672
logins_total	100,000	108.886	151.393	0	7,674
ext_city_size	100,000	4.759	2.441	1	8
ext_house_size	100,000	2.119	1.313	1	5
ext_purchase_power	100,000	4.432	1.964	1	7
xsell	100,000	0.100	0.300	0	1
married	100,000	0.385	0.487	0	1
overdraft	100,000	0.227	0.419	0	1

Customer demographics variables such as *'age'* and *'age_at_onboarding'* have means of 41.4 and 32.2 years, respectively, with substantial standard deviations (13.3 and 12.4), indicating a wide age range in the customer population. The minimum for age_at_onboarding is -4, suggesting either data entry error or imputation artifact, which required further data validation. Binary variables such as *'acad_title'*, *'member_get_member_recommender'*, *'married'*, and *'overdraft'* have means far below 0.5, indicating that these characteristics are relatively rare among the sampled population. For instance, only 2.1% of customers hold an academic title, and 38.5% are married.

For customer engagement metrics, *'nr_calls_90d'* (mean = 0.01) and *'complaints'* (mean = 0.013) are relatively low, with most customers having little to no engagement such as calls or complaints. However, the maximum number of calls is 20 and complaints is 11 shows that there is still a small subset of customers are highly engaged with the firm. The *'customer_tenure_months'* variable has a mean of 110 months (over 9 years) with a large standard deviation (84.9), reflecting considerable heterogeneity in customer longevity, from new customers (minimum = 0) to long-term clients (maximum = 440 months).

Financial activity measures such as *'vol_eur_inflows'* and *'vol_eur_outflows'* exhibit considerable skewness with standard deviations over 20,000, with inflows are up to €2,000,000 and outflows are as low as -€1,858,000. The high variance and extreme values indicate the presence of outliers or a right-skewed distribution, likely driven by a small number of high-value customers or transactions. Transaction frequency, represented by *'nr_girocard_trx_90d'* and *'nr_visacard_trx_90d'*, also shows significant variability (standard deviations of 29.1 and 50.2), with some customers having over 1,000 transactions with Visa card in 90 days, highlighting diverse usage patterns within the population. Income levels are also widely dispersed (mean = €1,942, standard deviation = €2,399), with the maximum reaching €168,451, again indicating extreme outliers.

Product-related variables such as *'nr_products'*, *'prod_mortgages'*, *'prod_brokerage'*, *'prod_savings'* display moderate means and low standard deviations, with the majority of customers holding between one and three products. *'prod_brokerage'* and *'prod_savings'* show a mean of 0.39 and 1.05, respectively, with maximum values of 12 and 9, suggesting a subset of customers with diverse financial portfolios.

External variables such as *'ext_city_size'*, *'ext_house_size'*, and *'ext_purchase_power'* have means of 4.76, 2.12, and 4.43, respectively, with relatively low standard deviations, suggesting limited variation and likely discrete categorical encoding.

The target variable *'xsell'* shows a mean of 0.10 and standard deviation of 0.3, indicating that 10% of customers attempted cross-buying, showing class imbalance issue which will be addressed in the following section.

4.4 Data Sampling

To ensure robust model development and evaluation, the following steps were undertaken: random splitting of the dataset, sampling to address class imbalance, setting cross-validation for model training and evaluation.

The cleaned dataset after preprocessing was first randomly shuffled to eliminate any potential ordering effects. A fixed random seed *(set.seed(12345))* was used to guarantee reproducibility of results. The shuffled data *(xsell_random)* was then splitted into two subsets with the ratio of 80:20. The first 80,000 observations were allocated to the training set *(xsell_train)*, representing 80% of the dataset. The remaining 20,000 observations constituted the validation set *(xsell_valid)*, accounting for 20% of the dataset.

Upon inspection, results show that there is a significant class imbalance issue within target variable, with only 10% of customers purchased loan, which may lead to bias in model prediction as the algorithms tend to predict the majority class (customer did not purchase loan). To mitigate this, the distribution of the target variable was applied data sampling method using the *'ovun.sample'* function with both oversampling and undersampling strategies *(method = "both")*, targeting an equal class distribution *(p = 0.5)* to achieve 50% observations with loan purchased and 50% of observation with loan were not purchased. The total number of samples was preserved at the original training set size.

Importantly, this technique was attempted on training set only, and no resampling was performed on the validation set, thereby ensuring unbiased assessment and avoid overly optimistic model performance.

For model training, this study also utilized a 10-fold cross-validation to provide reliable estimates of predictive power and ensure performance against overfitting. The training control object was configured as follows: 10-fold cross-validation *(method = "cv", number = "10")*, random search over hyperparameters for the best combination *(search = "random")*, final predictions saved for analysis *(savePredictions = "final"*), enabled classification tasks *(classProbs = "TRUE")*, two-class summary statistics *(summaryFunction = "twoClassSummary")*, and allowed parallel processes for speed up computation *(allowParallel = "TRUE"*).

5 Model Training and Hyperparameter Tuning

The predictive modeling strategy in this study incorporated two state-of-the-art ensemble machine learning methods: (1) Random Forest and (2) Gradient Boosting (XGBoost). Both models were trained on the balanced training dataset, employing 10-fold cross-validation and systematic hyperparameter tuning to optimize predictive performance. Prior to model training, the target variable *('xsell')* was encoded as a binary factor with two levels: *"No" (0)* and *"Yes" (1)*. This encoding ensures compatibility with binary classifiers and appropriate handling of class probabilities during prediction.

The following sections detail the modeling process, including features used and hyperparameter settings.

5.1 Random Forest

The RF implementation was conducted using the *'train'* function from the *caret* package, with the following key parameters: method as *"rf"* for RF algorithm, metric for hyperparameter optimization: *"ROC"* which is suitable for binary classification with imbalanced classes, number of trees *('ntree')* as 50 for each configuration for balancing computational efficiency and predictive stability, cross-validation control as the implemented 10-fold cross-validation in the *'ctrl'* object, ensuring robust estimation of generalization performance. Hyperparameter optimization focused on the number of predictor variables randomly sampled at each split *('mtry')*. A grid search was performed across four values: 5, 10, 20, and 30. This process identifies the optimal *'mtry'* value for maximizing the ROC metric, thereby enhancing model discrimination. To accelerate training and reduce computational time, parallel processing was applied using a cluster occupying all but one available core, which is particularly efficient for computing on a large dataset and an extensive cross-validation loop.

5.2 Gradient Boosting

The model was fitted using the *'train'* function from the *caret* package, with the *"xgbTree"* method specifying XGBoost as the algorithm, along with metrics for hyperparameter optimization as *"ROC"* and 10-cross validation as defined in the '*ctrl'* object, to estimate generalization error and minimize overfitting risk. A set of hyperparameter configuration was selected as the best combination after trial-and-error as follows: number of boosting rounds *('nrounds')* as 150, learning rate *('eta')* as 0.07, maximum tree depth *('max_depth')* as 3, minimum child weight *('min_child_weight')* as 3 , column subsampling *('colsample_bytree')* as 1, row subsampling *('subsample')* as 1, and regularization term *('gamma')* as 0. This hyperparameter setting balances model complexity and generalization, with moderate depth and learning rate to avoid overfitting while retaining sensitivity to complex patterns. Parallel processing was applied to reduce computational time for model training, with a cluster utilized across available CPU cores to comprehend a large dataset and cross-validation folds.

6 Performance Evaluation

After training, both RF and XGB models was used to generate predicted probabilities for the validation dataset *('xsell_valid')*. Predictions exceeding a threshold of 0.5 were classified as positive cross-sell outcomes *("1")*, while those below this threshold were classified as negative *("0")*. These predictions were formatted as factors for subsequent evaluation and comparative analysis.

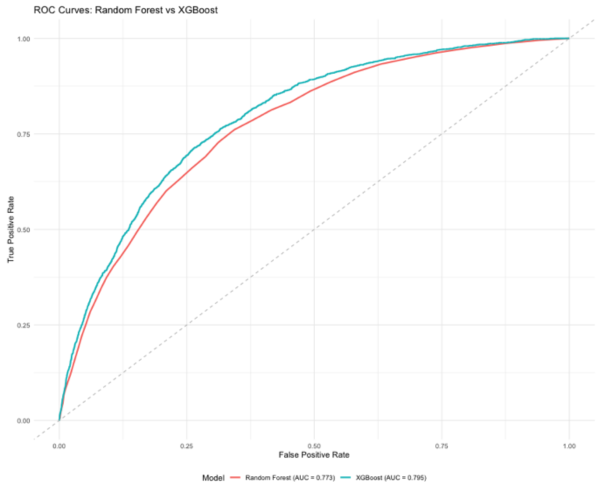
The predictive performance of the RF and XGB models was assessed using key metrics relevant to marketing analytics as indicated in the methodology section. The table below summarizes the performance metrics for both the Random Forest (RF) and XGBoost (XGB) models on the validation dataset.

Table 4: Performance metrics of Random Forest and XGBoost models

Metric	Random Forest Model	XGBoost Model
TDL	3.248	3.494
GINI	0.547	0.589
AUC	0.773	0.795
F1	33.100	33.564
Precision	31.950	21.693
Recall	34.336	74.135
Accuracy	86.155	70.725
Specificity	91.897	70.347

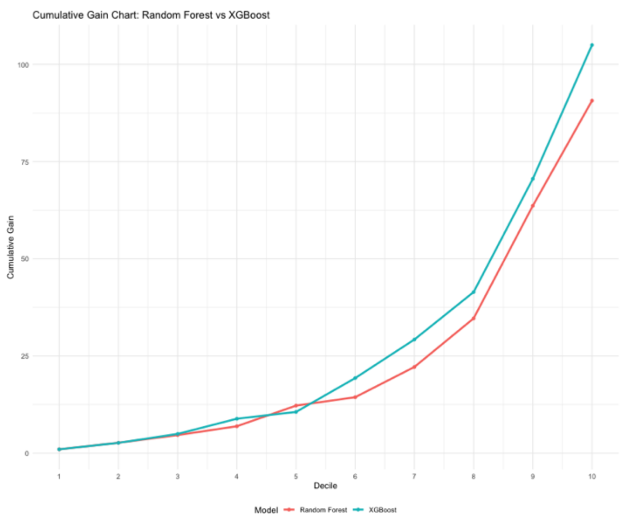
The performance metrics show that XGB consistently outperformed RF across the prioritized metrics. The XGB model achieved higher TDL (3.494 vs. 3.248), GINI (0.589 vs. 0.547), and AUC (0.795 vs. 0.773), indicating better discriminative power and capability to identify customers that are likely to respond to cross-buying marketing campaigns. The separation from the diagonal reference line (random classifier) in the ROC curve also show XGBoost demonstrating better ability to distinguish responders from non-responders.

Figure 2: ROC curves of Random Forest and XGBoost models



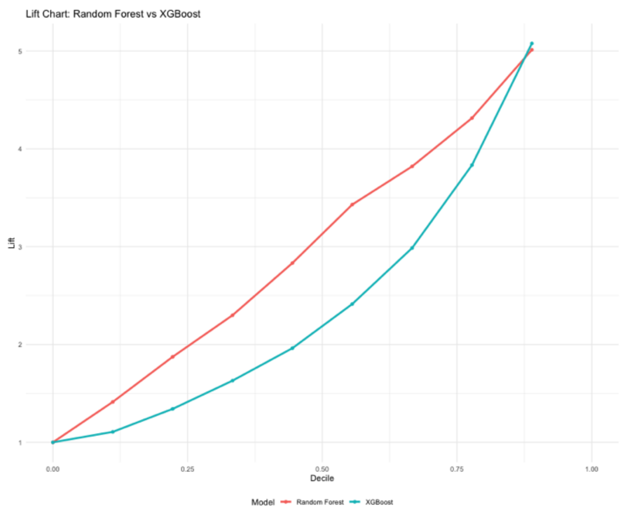
The F1 scores for both models are comparable (33.564 for XGB and 33.100 for RF), reflecting a relatively equal effectiveness in identifying positive cases. However, XGB has a notable trade-off between precision and recall metrics. While its recall rate (74.135) far exceeds that of RF (34.336), indicating a better ability to identify true positives, the precision is much lower (21.693), suggesting that a relatively large proportion of predicted positives are false. This could be explained by the class imbalance within target variable of the validation dataset, in which majority of customers did not attempt cross-buying, leading to the predictive modeling built on balanced training dataset was not able to predict perfectly to the actual context of cross-buying behavior. Nevertheless, this still can be advantageous in contexts where maximizing campaign reach is prioritized over minimizing false positives, especially for initiatives targeting a broad customer base with very little likelihood of positive cases such as cross-buying or fraud detection.

Figure 3: Cumulative Gain Chart of Random Forest and XGBoost models



The cumulative gain chart plots the proportion of actual positives captured as the proportion of targeted customers increases. XGB indicates a higher cumulative gain curve, especially in the upper deciles, meaning it is more effective at capturing a larger share of responders when a larger portion of the customer base is targeted. This suggests that XGB is more efficient for broader marketing campaigns where the objective is to maximize overall reach and response volume.

Figure 4: Lift chart of Random Forest and XGBoost models



On the other hand, RF maintains a more balanced and higher rate in precision and specificity (91.897 vs. 70.347), but at the expense of recall. RF has higher lift across most deciles, particularly in the top deciles, indicating that it is more effective at concentrating positives in its highest-scoring groups. In marketing perspective, this means that if a marketing campaign targets the top deciles of predicted customers, the RF model would deliver a higher proportion of true positives than XGB in those segments. This suggests that RF may be more suitable for scenarios where resource allocation is constrained, and the cost of false positives is significant.

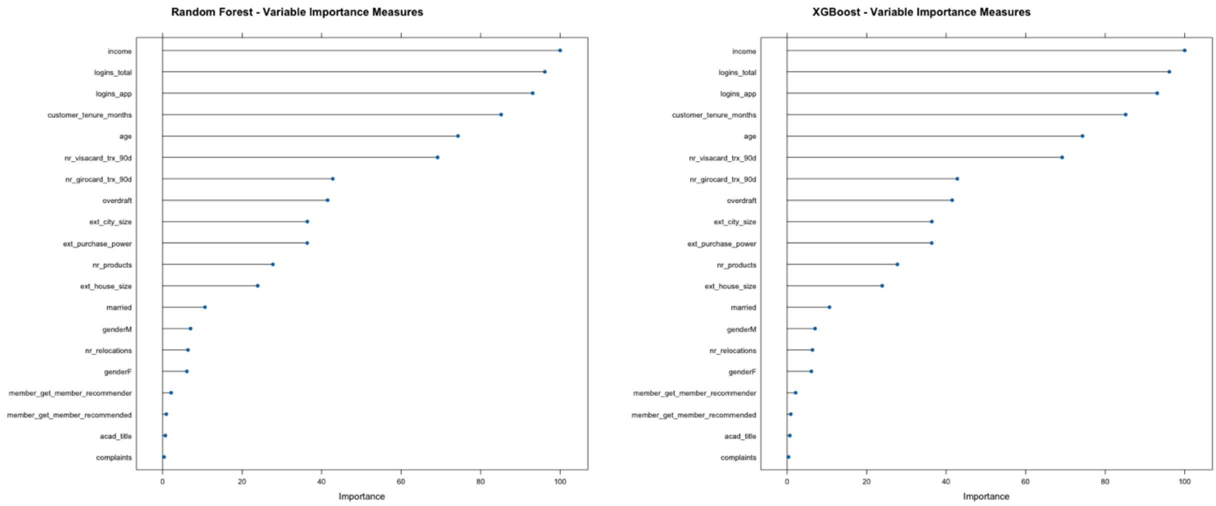
Therefore, with each model has its own strength and limitation, the choice between models depends on the strategic objectives and tolerance for false positives within the marketing program. XGBoost may be the preferred model for initiatives that are focused on maximizing customer engagement and campaign responsiveness, whereas RF may be more beneficial for marketing campaigns with limited budgets and require stricter resource allocation.

7 Interpretation and Discussion

With both RF and XGB models showing consistency in their feature rankings for variable importance measures (VIM), this section discusses most influential predictors and suggests its implications in supporting strategic business decision-making.

7.1 Variable Importance

Figure 5: Variable importance measures



The VIM analysis identifies income as the most influential predictor of cross-buying behavior, achieving the highest importance scores in both models. The prominence of income suggests that customer segmentation strategies should prioritize income-based categorization for cross-selling marketing initiatives. Login frequency emerge as the second and third most important features, with '*logins_total'* and '*logins_app'* demonstrating nearly equivalent importance scores, approximately 93 and 95. This digital engagement reflects the pattern in which platform interaction, specifically internet banking and mobile app usage, serves as a proxy for customer purchase intention.

Customer tenure *('customer_tenure_months')* ranks fourth in importance, indicating the relationship between customer lifecycle stage and purchase diversification. This finding may be related to customer lifetime value framework, in which relationship duration correlates with trust development and willingness to explore additional products.

Regarding demographics and transaction behavior, the *'age'* variable demonstrates moderate importance, around 75, suggesting generational differences in cross-buying behavior, which may be an insight for age-segmented marketing approaches, as different demographic background could differ cross-buying behavior. Visa card usage frequency *('nr_visacard_trx_90d')* and Giro card transactions *('nr_girocard_trx_90d')* show moderate predictive power, with around 65 and 70 respectively, indicating that transactional behavior also correlate with customer intention for cross-buying.

7.2 Marketing Implications

As income is recognized as the most influential predictor, for cross-selling approaches around income-based customer segments, high-income and upper-middle-income customers (top quartile) should be prioritized for frequent and consistent efforts. These customers demonstrate the highest cross-buying likelihood and, thus, generate the most return-on-invest for marketing campaigns.

Furthermore, the high importance of login metrics suggests the need for efficient digital engagement strategies. Real-time monitoring systems should track login frequency patterns to identify optimal timing for delivering cross-selling messages. Customers with increased digital engagement represent better likelihood for cross-buying, whereas declining engagement patterns may signal potential churn risk requiring retention interventions before cross-selling attempts. Additionally, mobile banking strategies become more important given the relatively equivalent importance compared with total number of logins for internet banking. Thus, firms should pay more attention in mobile banking functionality, as well as user interface and experience optimization, along with customer segment scoring based on login frequency and session duration. For instance, push notification and reminder systems can be leveraged to introduce relevant product offerings during identified high-engagement timing for better likelihood of customer response.

VIM analysis also shows the significant role of customer tenure in marketing approaches. Specifically, for new customers with tenure of 0-12 months or moderately established customers with tenure of 1-3 years, firms should focus on building relationship before introducing cross-selling initiatives, as early cross-selling attempts on new customers are likely not optimal and may create overwhelming experiences. Long-tenure customers with tenure over 3 years are the most beneficial segment when it comes to cross-buying behavior, therefore, firms should invest on marketing initiatives to drive customer awareness and consideration of additional purchase intention.

From the above-mentioned analysis, firms should take into consideration of variable importance identified from VIM to optimize their marketing campaign efforts to prioritize top-decile segments that are likely to response to cross-selling promotions. It also suggests reallocating marketing resources toward income-based segmentation, digital engagement improvement initiatives, and tenure-based relationship management systems.

8 Conclusion and Outlook

This study demonstrates the effectiveness of ensemble machine learning methods in predicting customer cross-buying behavior. The comparative analysis shows that XGB model achieves strong discriminative power (AUC = 0.795, GINI = 0.589), whereas RF provides a more balanced precision-recall trade-off. Thus, model selection relies on marketing objectives, in which XGB model is more optimal for maximizing reach and engagement, while RF is better for campaign with resource-limited allocation. Variable importance analysis identifies income as the most influential predictor, followed by digital engagement metrics and customer tenure, providing foundations for segmentation strategies in cross-selling campaigns.

For future research, advanced ensemble methods such as stacking and voting can be applied, further combining the predictive power of RF and XGB to leverage enhanced performance. Moreover, there is also potential direction for temporal modeling enhancements using time-series ensemble methods to capture dynamic patterns and behavioral changes over time. To improve understanding of key predictors, interpretability advancement through methods such as SHAP and LIME could provide instance-level explanations supporting individual customer decisions beyond aggregate variable importance measures (Lundberg & Lee, 2017). Cross-industry validation studies could examine the ensemble method effectiveness in predicting cross-buying behavior in other sectors such as telecommunications and insurance. Comparative analyses across different regulatory environments and cultural contexts would strengthen the external validity of findings and identify sector-specific marketing optimization strategies.

The continued advancement of ensemble methods for customer behavior prediction promises opportunities for enhancing marketing intelligence, as well as providing decision-making support for an enhanced campaign effectiveness.

References

Abbasimehr, H., Setak, M., & Tarokh, M.J. (2014). A comparative assessment of the performance of ensemble learning in customer churn prediction. *Int. Arab J. Inf. Technol.*, 11, 599-606.

Ashraf, S., & Ahmed, T. (2020). Machine Learning Shrewd Approach For An Imbalanced Dataset Conversion Samples. *Journal of Engineering and Technology*, 11.

Ashraf, S., Saleem, S., Ahmed, T., Aslam, Z., & Muhammad, D. (2020). Conversion of adverse data corpus to shrewd output using sampling metrics. *Visual Computing for Industry Biomedicine and Art*, *3*(1).

Berry, M. J., & Linoff, G. S. (2004). *Data mining techniques: For marketing, sales, and customer relationship management* (2nd ed.). Wiley.

Bhattacharya, S., & Dutta, S. (2019). Cross selling of investment products and services: a case study of leading financial services organisation. *International Journal of Business Forecasting and Marketing Intelligence*, *5*(2), 241.

Bogaert, M., & Delaere, L. (2023). Ensemble Methods in Customer Churn Prediction: A Comparative Analysis of the State-of-the-Art. *Mathematics*, *11*(5), 1137.

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

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5-32.

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-794.

Chiang, K. F., Tze, L. L., Chieh, T. Y., & Yu, H. L. (2011). Factors of cross-buying intention - bancassurance evidence. *African Journal of Business Management*, *5*(17), 7511–7515.

Dietterich, T. G. (2000). Ensemble methods in machine learning. *Multiple Classifier Systems*, 1857, 1-15.

Estrella-Ramón, A. (2017). Explaining customers' financial service choice with loyalty and cross-buying behaviour. *Journal of Services Marketing*, *31*(6), 539–555.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, 29(5), 1189-1232.

Genuer, R., Poggi, J. M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, 31(14), 2225-2236.

Kumar, S., Soundarapandiyan, K., & Meera, S. (2023). Comparative Study of Customer Churn Prediction Based on Data Ensemble Approach. *2023 Intelligent Computing and Control for Engineering and Business Systems (ICCEBS)*, 1–10.

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.

Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3), 18-22.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, 30, 4765-4774.

Nguyen, T. M., Le, T. A., & Nguyen, T. H. (2023). A flexible framework for customer behavior prediction based on ensemble learning. *Proceedings of the 12th International Symposium on Information and Communication Technology*, 126–134.

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427-437.

Tung, B., & Carlson, J. (2015). Examining determinants of cross buying behaviour in retail banking. *International Journal of Quality & Reliability Management*, *32*(8), 863–880.

Verhoef, P. C., & Donkers, B. (2005). The effect of acquisition channels on customer loyalty and cross-buying. *Journal of Interactive Marketing*, *19*(2), 31–43.

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

Zhang, X. (2025). Financial Customer Behavior Prediction Based on Machine Learning: A Comprehensive investigation. *ITM Web of Conferences*, *73*, 02004.

Appendix: Variable Description

Variable Name	Description
acad_title	Client has a name prefix like Dr. or Prof.
age	Customers age in years
nr_calls_90d	Number of telephone contacts initiated by client
complaints	Number of complaints within one year
customer_tenure_months	Customer tenure since onboarding in months
gender	Gender
vol_eur_inflows	Sum of inflows in EUR (gross) on savings accounts
marital_status	Marital status, with „VH“ means „verheiratet“ (married)
member_get_member_recommender	Dummy variable, whether the client has recommended another client via member get member program
member_get_member_recommended	Dummy variable, whether the client was recommended by another client via member get member program
nr_products	Number of products (accounts) the clients holds
vol_eur_outflows	Sum of outflows in EUR (gross) on savings accounts
prod_mortgages	Number of mortgages the clients holds if any
prod_brokerage	Number of IP products the clients holds if any
prod_savings	Number of variable savings products the clients holds if any
nr_relocations	Number of distinct addresses the client had over one year
vol_eur_debit	sum of savings volumes the clients have on savings products
vol_eur_credit	sum of loan volumes the clients have on lending products
age_at_onboarding	Customers age in years when onboarded
loan	Number of loan products if any
nr_girocard_trx_90d	Number of transactions the client made via Girocard (local debit card, known as “EC Karte”) within 90 days if any
nr_visacard_trx_90d	Number of transactions the client made via VISA card within 90 days if any
income	Salary if known
nr_days_when_giro_above_0	Number of days the client's payment account was above 0€ within 65 days
nr_days_when_giro_below_0	Number of days the client's payment account was below 0€ (overdraft usage) within 65 days
logins_app	Number of app logins within 90 days
logins_total	Number of total logins including internet banking within 90 days
ext_city_size	External variable: density of inhabitants, with 1 = very low (rural area) and 7 = very high (city)
ext_house_size	External variable: house size
ext_purchase_power	External variable: purchase power, with 1=very low and 7 = very high
xsell	Target variable, whether the client purchase a loan after marketing campaigns

Goethe-Universität | 60629 Frankfurt am Main
Fachbereich Wirtschaftswissenschaften | Prüfungsamt
(Stand: 04.06.2018)

GOETHE UNIVERSITÄT FRANKFURT AM MAIN

Ehrenwörtliche Erklärung/ Statutory Declaration

An das Ende von schriftlichen Arbeiten (bspw. Seminararbeiten, Bachelor- oder Masterarbeiten) ist eine ehrenwörtliche Erklärung zu setzen. Sie hat folgenden Wortlaut:

Ehrenwörtliche Erklärung

"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 auszugsweise, für eine andere Prüfungs- oder Studienleistung verwendet."

Ort, Datum: Frankfurt, 21.07.2025 Unterschrift: ……………………………………

Bei schriftlichen Arbeiten, die zusätzlich in elektronischer Form einzureichen sind, ist jedoch folgender Wortlaut zu verwenden:

Ehrenwörtliche Erklärung

"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 auszugsweise, für eine andere Prüfungs- oder Studienleistung verwendet. Zudem versichere ich, dass die von mir abgegebenen schriftlichen (gebundenen) Versionen der vorliegenden Arbeit mit der abgegebenen elektronischen Version auf einem Datenträger inhaltlich übereinstimmen."

Ort, Datum: …………………………………… Unterschrift: ……………………………………

A statutory declaration is to be included at the end of every written work (e.g. Seminar paper, Bachelor's or Master's thesis).
The translation is as follows:

Statutory Declaration

"I herewith declare that I have composed the present thesis myself and without use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The thesis in the same or similar form has not been submitted to any examination body and has not been published. This thesis was not yet, even in part, used in another examination or as a course performance."

Place, Date: Frankfurt, 21.07.2025 Signature: ……………………………………

In case of written work, which is also to submit on a data carrier, the translation is as follows:

Statutory Declaration

"I herewith declare that I have composed the present thesis myself and without use of any other than the cited sources and aids. Sentences or parts of sentences quoted literally are marked as such; other references with regard to the statement and scope are indicated by full details of the publications concerned. The thesis in the same or similar form has not been submitted to any examination body and has not been published. This thesis was not yet, even in part, used in another examination or as a course performance. Furthermore I declare that the submitted written (bound) copies of the present thesis and the version submitted on a data carrier are consistent with each other in contents."

Place, Date: …………………………………… Signature: ……………………………………

AACSB ACCREDITED