DBA5106 FOUNDATION OF BUSINESS ANALYTICS

CLASSIFICATION: CUSTOMER CHURN PREDICTION

GROUP PROJECT 2

SEMESTER 1 AY 2024/2025

PREPARED BY GROUP 26

A0297564L	PAN MINWEN
A0297417R	KUAN SHENG CHEN
A0296776A	SHOTA SANADA
A0296770N	KHWANCHAT PHUTPHITHAK

AY2024/2025 SEMESTER 1

DBA5106: FOUNDATIONS of BUSINESS ANALYTICS

Group Project 2

Classification: Prediction on Customer Churn

Group 26

A0297564L	PAN Minwen
A0297417R	Kuan Sheng Chen
A0296776A	Shota Sanada
A0296770N	Khwanchat Phutphithak

I. Executive Summary

Customer retention is critical for businesses to maintain brand penetration and reduce the cost of acquiring new customers. This project analyzes customer churn using a dataset from the banking industry, applying four machine learning models: Logistic Regression, Explainable Boosting Machine (EBM), Random Forest, and CatBoost. Each model's performance and interpretability were evaluated to balance prediction accuracy with actionable insights.

Among the models, EBM emerged as the recommended model due to its ability to achieve high predictive accuracy (86%) while maintaining transparency. It identified critical parameters such as tenure, age, and activity level as the main determinants of churn. Meanwhile, CatBoost and Random Forest exhibited strong performance. Their dependence on SHAP values for interpretation rendered them less accessible for prompt business applications, while Logistic Regression failed to accommodate non-linear correlations. This analysis underscores the necessity of balancing interpretability with performance to guide strategic interventions effectively.

The findings illustrate the significance of interpretable models such as EBM in informing client retention tactics. By concentrating on major features, companies can execute targeted strategies such as retaining existing customers with loyalty programs, re-engaging inactive members with marketing campaigns or promotions, and various offerings by demographic profile, like personalized marketing and engagement initiatives, to diminish churn and enhance client loyalty, hence promoting long-term profitability.

II. Introduction

Brand penetration, defined as the number of consumers interacting with a brand, is a crucial metric for all company sizes. While acquiring new customers drives growth, it comes at a high cost due to extensive marketing investments, increasing the cost per acquisition. However, many customers fail to maintain long-term engagement, forcing businesses to continually invest in new customer acquisition, creating a cycle of substantial spending.

To prevent this cycle, businesses recognize the importance of customer retention alongside acquisition. Retaining customers sustains brand penetration and reduces costs by minimizing the need for constant recruitment efforts. Predictive models have become essential for understanding customer behavior and detecting churn, enabling timely interventions to enhance retention. However, while these models often achieve high accuracy, their complexity can hinder interpretability, challenging decision-makers who require actionable insights.

This report evaluates customer churn using data from Kaggle (Badole, 2023), comparing glassbox models, which prioritize interpretability, with blackbox models, known for higher predictive power but limited transparency. SHAP (SHapley Additive exPlanations) bridges the interpretability gap in blackbox models. The analysis explores trade-offs between accuracy and clarity, providing practical recommendations that balance precise predictions with actionable insights.

III. Problem Statement

The dataset used for analysis is derived from a banking context, where churn prediction is crucial for customer retention and brand loyalty. The data comprises multiple customer-related features, including account balance, age, geographic location, number of products used, activity status, and estimated salary. These features collectively capture the behavioral, demographic, and financial characteristics that may affect a customer's decision to stay with or leave the bank. In this scenario, factors like high account balances combined with low engagement, such as minimal product usage or inactivity, may signify a churn risk. Subsequently, we will apply various machine learning approaches and utilize SHAP for each to assess the interpretability of results.

IV. Methodology

We commence with an initial data exploration, concluding that missing value imputation is ignorable and that label imbalance needs to be addressed. To handle this issue, we employ a zero-R model as a benchmark and four machine learning algorithms—Logistic Regression, EBM, Random Forest, and CatBoost—to evaluate the performance matrix of their original forms and adjusted forms through grid search. Ultimately, we introduce explainability tools to assess the contributions of the features on both a global and local scale.

A. Exploratory data analysis

In this phase, we examined the target variable and detected an imbalance, with class 1 (departure) being underrepresented. Furthermore, we analyzed feature correlations and identified that specific features exhibited a stronger linear relationship with the target variable, guiding our feature selection and engineering strategies to improve model performance.

B. Data pre-processing

During the data processing phase, we standardize numerical features to ensure they have a mean of zero and a standard deviation of one, hence enhancing the efficacy of numerous machine learning methods. We convert categorical features into a binary representation by one-hot encoding for model training. Furthermore, we employ the Synthetic Minority Over-sampling Technique (SMOTE) to rectify class imbalance by creating synthetic samples for the minority class, thereby improving the model's capacity to learn from underrepresented data.

C. Machine Learning Models Selection

1. Zero-R Baseline model

The baseline model is a zero-R model, which forecasts the churn rate to be the same as the majority value – zero, in our case. The performance matrix of the baseline model will then be compared with those of various machine learning algorithms.

2. Logistic Regression

Logistic Regression is a robust machine learning model suitable for binary classification tasks that assess the probability of specific outcomes using predictor variables influencing the likelihood of a customer departing from the service. Coefficients in Logistic Regression improve interpretability by quantifying the influence of each feature, allowing decision-makers to determine the significance of factors and the strength of their impact on outcomes. The method's straightforward nature and adaptability in managing large data sets make Logistic Regression essential for data-informed decision-making. It enables companies to identify high-risk clients susceptible to attrition and allocate resources for their retention via targeted incentives or proactive support.

3. Explainable Boosting Machine (EBM)

The Explainable Boosting Machine (EBM) is a sophisticated machine learning model engineered to achieve a compromise between superior predicted accuracy and interpretability. It leverages generalized additive models (GAMs) to generate a series of transparent, additive functions easily understandable to others. Essential characteristics of EBM encompass its capacity to manage both numerical and categorical data, its resilience to overfitting through techniques like bagging, and its intrinsic interpretability, enabling users to visualize and comprehend the impact of each feature on the model's predictions. This renders EBM especially advantageous in applications where model transparency and trust are essential.

4. Random Forest

Random Forest is a resilient machine learning technique for categorization problems, rendering it particularly effective for predicting customer churn. Creating numerous decision trees and consolidating their outputs integrates intricate feature interactions and mitigates overfitting, resulting in increased predictive accuracy. Its capacity to manage varied information, covering numerical qualities such as balance and categorical ones like geography, augments its versatility.

Nonetheless, its complexity as a blackbox model constrains interpretability. The model's predictions stem from the combination of multiple decision trees, complicating the comprehension of how particular variables influence specific outcomes, hence obscuring the factors driving churn. SHAP is utilized to evaluate feature significance and derive insights to determine affecting customer choices.

5. CatBoost

CatBoost is a robust gradient-boosting algorithm that handles both categorical and numerical data, making it ideal for customer churn prediction. Techniques like ordered boosting minimize overfitting while maintaining high accuracy and effectively capturing complex relationships in diverse datasets, such as customer demographics and transactional behavior. With built-in handling of missing values and categorical variables, CatBoost simplifies data preparation and is computationally efficient, making it well-suited for large-scale datasets.

However, as a blackbox model like Random Forest, its complexity limits interpretability, making it challenging to identify how features influence predictions. To address this, tools like SHAP are applied to enhance interpretability and provide insights into the importance of features.

D. Performance evaluation

Concerning the evaluation, we use the following two measures:

- **F1-Score:** The harmonic means of precision and recall balance these metrics, offering a thorough measure of the model's accuracy, especially in the presence of class imbalance.
- **Accuracy:** The proportion of accurately predicted observations to the total observations. Nonetheless, imbalanced datasets may be misleading.

To enhance comprehension of the model's predictions, we employ explanatory tools such as SHAP and the integrated functions of the EBM. These techniques allow us to obtain insights into both global and local feature importance, facilitating our understanding of how individual features affect the model's decisions throughout the entire dataset and for a specific prediction.

V. Model Results and Interpretation

A. Data exploration.

The descriptive statistics, which are presented in Appendix 1, consist of the numerical data, the histograms and boxplots of numerical distributions, and the plots of the discrete data. The dataset does not contain any missing value, therefore obviating the necessity for imputation. The positive labels constitute 80% of the data, signifying a severe imbalance. The standard deviations of the numerical features vary in scale, indicating the necessity for standardization.

B. Model Result

1. Logistic Regression

We implemented an initial model to understand the baseline, focusing on overall accuracy and recall for the minority class. The model attained 80% overall accuracy and 59% accuracy for class 0, but exhibited difficulties with the minority class 1, achieving only 22% recall and an F1-Score of 0.32, indicating a considerable bias towards the majority class. To rectify this, we employed SMOTE and Random Undersampling for data equilibrium. According to Appendix 2.1, both approaches substantially enhanced class 1 recall to 69%; however, SMOTE exhibited greater precision (41%), F1-Score (0.51), test accuracy (72%), and weighted F1-Score (0.74). SMOTE was chosen as the final method due to its balanced performance and robustness.

2. Explainable Boosting Machine (EBM)

The Explainable Boosting Machine (EBM) demonstrated robust performance, with an overall accuracy of 86% on both the initial and test set evaluations. The initial model showed high precision (0.88) and recall (0.94) for class 0, signifying its efficacy in identifying true negatives. For class 1, the precision was lower at 0.72, and recall was 0.54, suggesting the model struggles to identify true positives accurately. The F1-Scores reflect this disparity, with 0.91 for class 0 and 0.62 for class 1. The macro and weighted averages suggest a balanced performance; nevertheless, the lower recall for class 1 highlights a potential area of improvement. Hyperparameter adjustment using grid search did not substantially modify these metrics, demonstrating the model's durability and constraints in managing class 1 predictions.

3. Random Forest

We initially implemented a Random Forest classifier and evaluated its performance. The initial Random Forest model achieved 86% validation accuracy but struggled with minority class recall (44%), indicating ineffective identification of churned consumers. Hyperparameter adjustment via grid search marginally enhanced the model's accuracy to 87% without increasing minority recall. To tackle class imbalance, we applied SMOTE, Random Undersampling, and a balanced Random Forest. The SMOTE enhanced Random Forest, which provided the best trade-off, achieving 83% validation accuracy and improving minority class recall to 59%, while maintaining a reasonable precision at 54%, outperforming other methods that compromised majority class performance. The test set achieved an accuracy of 84%, with a precision of 62% and a recall of 63% for the minority class. Appendix 4.1 contains thorough performance comparisons.

4. CatBoost

According to the results in Appendix 5.1, the initial model exhibited great accuracy with manual feature transformations but struggled with class imbalance, leading to a poor recall rate for churners (0.51). To address this, three approaches were applied: Random Oversampling, which replicates minority class samples, generated the best performance with the greatest F1-Score (0.63) and enhanced recall (0.61), effectively balancing the dataset while preserving essential information. In contrast, Random Undersampling and Balanced Random Forest compromised the majority class precision and overall F1-Score to boost minority recall. Therefore, Random Oversampling is the most effective method, offering a superior balance between recall and overall performance.

Metrics	**Baseline**	**Glassbox Model**		**Blackbox Model**	
	Zero-R	*Logistic Regression*	*Explainable Boosting Machine (EBM)*	*Random Forest*	*CatBoost*
Precision (0/1)	0.79/0.00	0.90/0.41	0.88/0.72	0.90/0.54	0.90/0.65
Recall (0/1)	1.00/0.00	0.72/0.69	0.94/0.54	0.88/0.59	0.91/0.61
F1-Score (0/1)	0.88/0.00	0.80/0.51	0.91/0.62	0.89/0.56	0.90/0.63
Micro-weighted accuracy	0.79	0.72	0.86	0.83	0.85

Table 1: Performance evaluations of each model

C. Model Interpretability

Each model offers a unique perspective on identifying the most influential features driving customer churn, leveraging different methodologies to rank these features:

Ranking (Most important)	**Glassbox Model**		**Blackbox Model**	
	Logistic Regression (Co-efficient)	***Explainable Boosting Machine (Global Mean Absolute Score)***	***Random Forest (Global SHAP Value)***	***CatBoost (Global SHAP Value)***
Rank 1	Age (0.514)	Tenure (1.845)	Age (0.144)	Age (1.415)
Rank 2	IsActiveMember (-0.296)	Age (1.640)	NumOfProducts (0.113)	NumOfProducts (1.209)
Rank 3	Geography_Germany (0.252)	NumOfProducts (0.794)	IsActiveMember (0.062)	IsActiveMember (0.766)
Rank 4	Balance (0.202)	Geography_Germany (0.321)	Balance (0.042)	Balance (0.582)
Rank 5	Gender_Male (-0.176)	Balance & NumOfProduct (0.284)	Geography_Germany (0.041)	Gender (0.407)

Table 2: Top five influence features of each model

The Logistic Regression model identified key drivers by quantifying the linear connection between each factor and the target variable using coefficients. The coefficients' magnitude and sign reflect the influence's importance and direction. Age, IsActiveMember, and Geography_Germany are the significant contributors, where Age significantly increases churn risk, and active membership reduces it. Logistic Regression makes it interpretable and ideal for understanding straightforward feature impacts.

Using global mean absolute SHAP values, the Explainable Boosting Machine (EBM) ranks features by their average contribution across predictions. Unlike Logistic Regression, EBM models non-linear relationships and interactions. Compared to the other models, tenure is the most critical factor. This highlights EBM's ability to capture patterns that simpler models might disregard.

Random Forest and CatBoost use global SHAP values to determine how each feature affects model predictions. CatBoost, a gradient-boosting model, excels at categorical features and complex feature interactions, while Random Forest prioritizes traits using decision trees. Despite their methodological differences, both models identify similar top features, such as Age, NumOfProducts, IsActiveMember, and Balance, highlighting their consistent relevance to churn predictions. However, CatBoost assigns greater SHAP values to Age and NumOfProducts than Random Forest, reflecting its ability to model subtle interactions and non-linear effects more effectively than Random Forest. These findings highlight both models' ability to detect churn factors and CatBoost's ability to handle complex data structures.

Besides global scale interpretation, each model offers local-scale interpretation, facilitating insights into specific predictions. Logistic Regression offers straightforward local explanations through coefficients that illustrate feature contributions, whereas EBM identifies non-linear patterns and interactions. Random Forest and CatBoost utilize SHAP values to decompose predictions into individual features' contributions, explaining each feature's importance on the result. These interpretations enhance understanding and actionable strategy.

VI. Final Recommendation

Based on the analysis, we recommend the Explainable Boosting Machine (EBM) as the most suitable model for bank churn prediction based on our dataset, effectively balancing predictive accuracy and interpretability. Achieving a validation accuracy of 86%, EBM matches the efficacy of sophisticated blackbox models such as CatBoost, while preserving a transparent framework that offers actionable insights. Its enhanced precision for the minority class (72%) and F1-Score (62%) further illustrate its effectiveness in detecting high-risk clients compared to Logistic Regression and Random Forest.

In contrast to blackbox models that depend on external tools like SHAP for interpretability, EBM intrinsically offers transparent explanations of feature significance, allowing decision-makers to comprehend the primary factors influencing churn. The integration of predictive capability and transparency renders EBM an optimal selection for business contexts, necessitating both precision and clarity, ensuring that retention strategies are data-driven and actionable.

The model finding highlights the importance of customer engagement, as Tenure and IsActiveMember are the primary determinants influencing customer churn rate. Companies can utilize the data to develop the right strategies, such as loyalty programs or campaigns, to prevent consumers’ departure. Furthermore, a cross-selling plan helps reduce customer attrition, as the number of items offered is critical; fewer products may result in diminished engagement or dependence on the company's services. Moreover, the model indicates that demographic factors such as age and residence in Germany are the primary determinants of customer attrition. This encompasses customer behaviour influenced by diverse lifestyles, necessitating distinct strategies such as targeted promotions for specific age segments or localized campaigns specifically in Germany.

VII. Conclusion and Research Limitations

In conclusion, the study investigates customer churn prediction within the banking sector through a comprehensive analysis of various machine learning models. While blackbox models like Random Forest and CatBoost provide high accuracy, their reliance on interpretability tools like SHAP limits their applicability in business decision-making. Conversely, the Explainable Boosting Machine (EBM) balances robust predictive performance and inherent interpretability, positioning it as the most effective model. Primary factors influencing churn are tenure, age, and activity level. Among the models, EBM is the most appropriate for balancing predictive power and actionable insights since interpretable models can facilitate targeted strategies for enhancing customer retention while stressing the necessity of balancing accuracy and transparency in churn prediction.

Regarding the limitations of this research, the dataset sourced from a single bank may restrict the generalizability of findings to other industries or regions. Utilizing oversampling techniques such as SMOTE to address class imbalance enhanced recall for the minority class; however, it also raised concerns regarding potential biases stemming from synthetic data. Furthermore, the use of SHAP for interpreting black-box models, although practical, requires substantial computational resources and domain expertise, which restricts its usability. The study also needs to account for temporal changes in customer behavior and missing insights from time series data. Future research should investigate varied datasets, create more accessible interpretability tools, and assess the models' performance over time to enhance applicability and enrich understanding of customer churn.

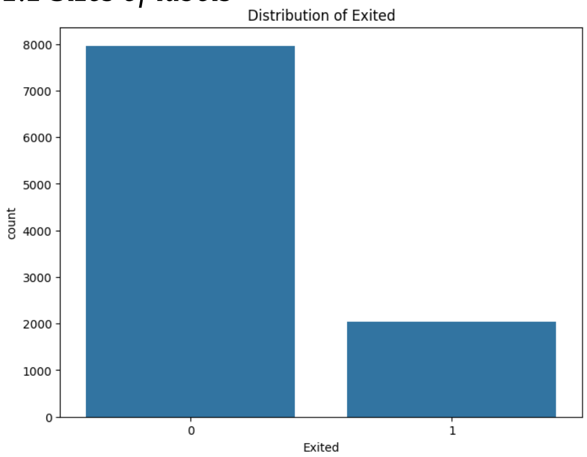
VIII. Reference

Badole, S. (2023). *Bank customer churn prediction dataset*. Kaggle. Retrieved November 24, 2024, from https://www.kaggle.com/datasets/saurabhbadole/bank-customer-churn-prediction-dataset.

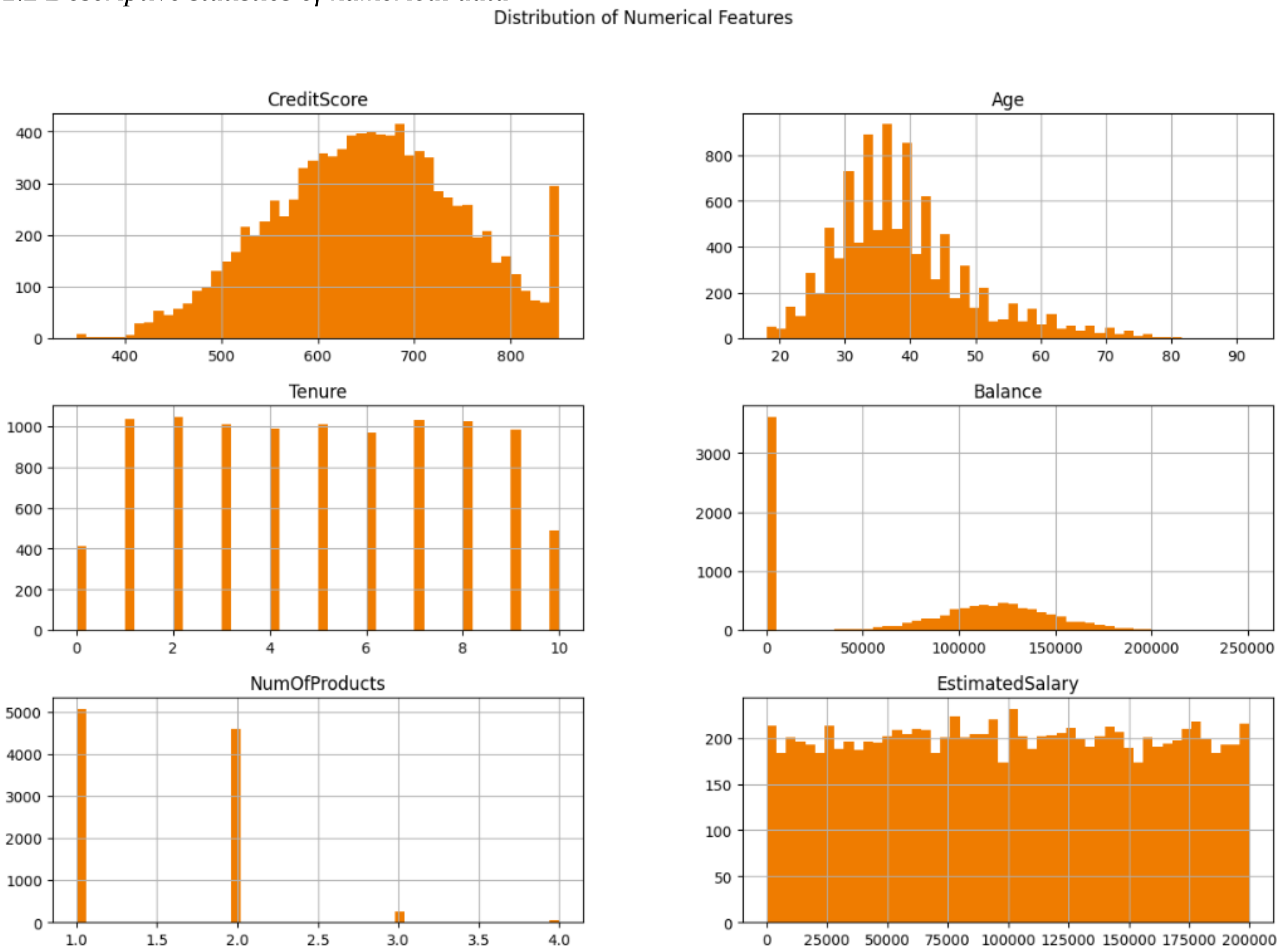
IX. Appendix

1. Exploratory data analysis

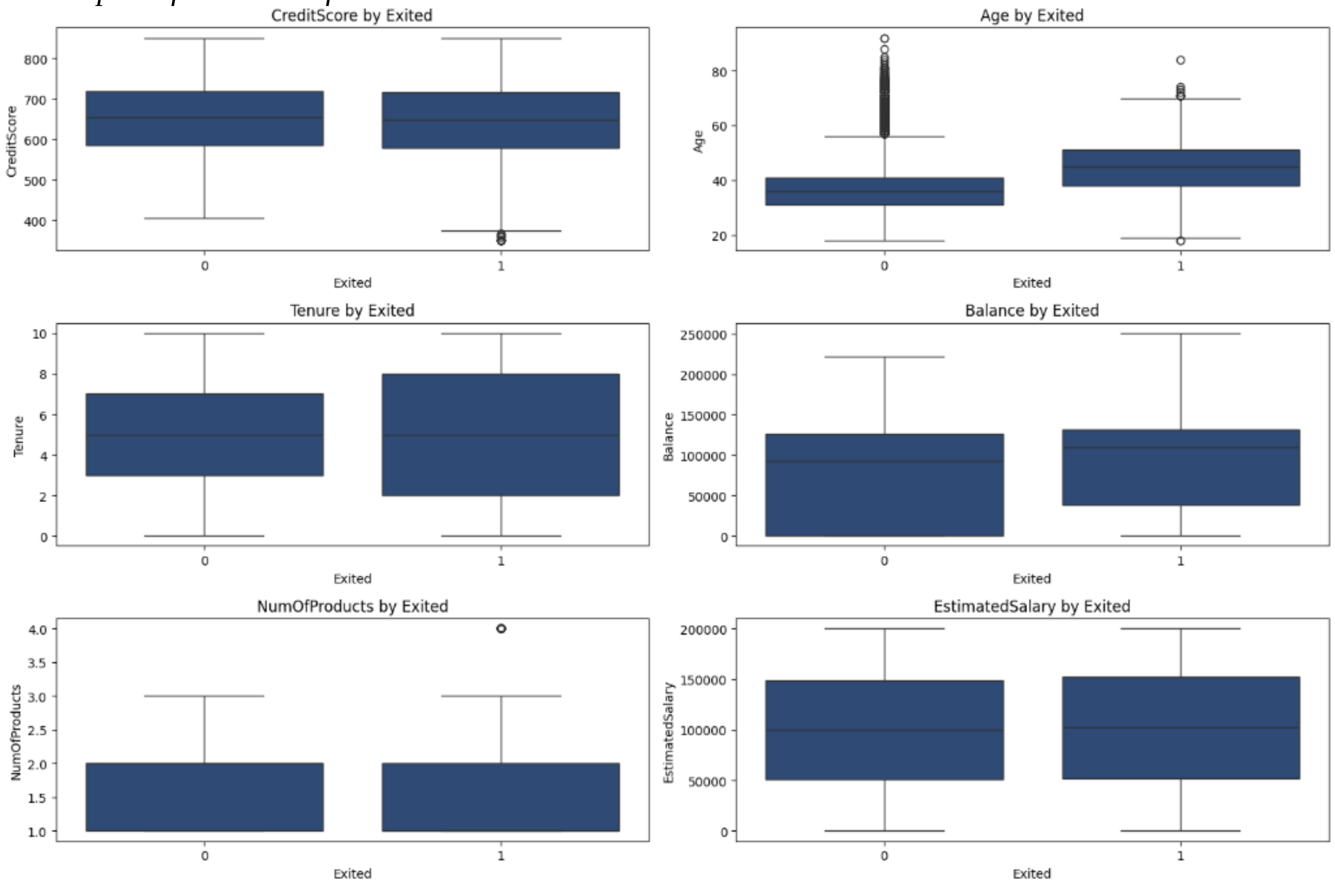
1.1 Sizes of labels



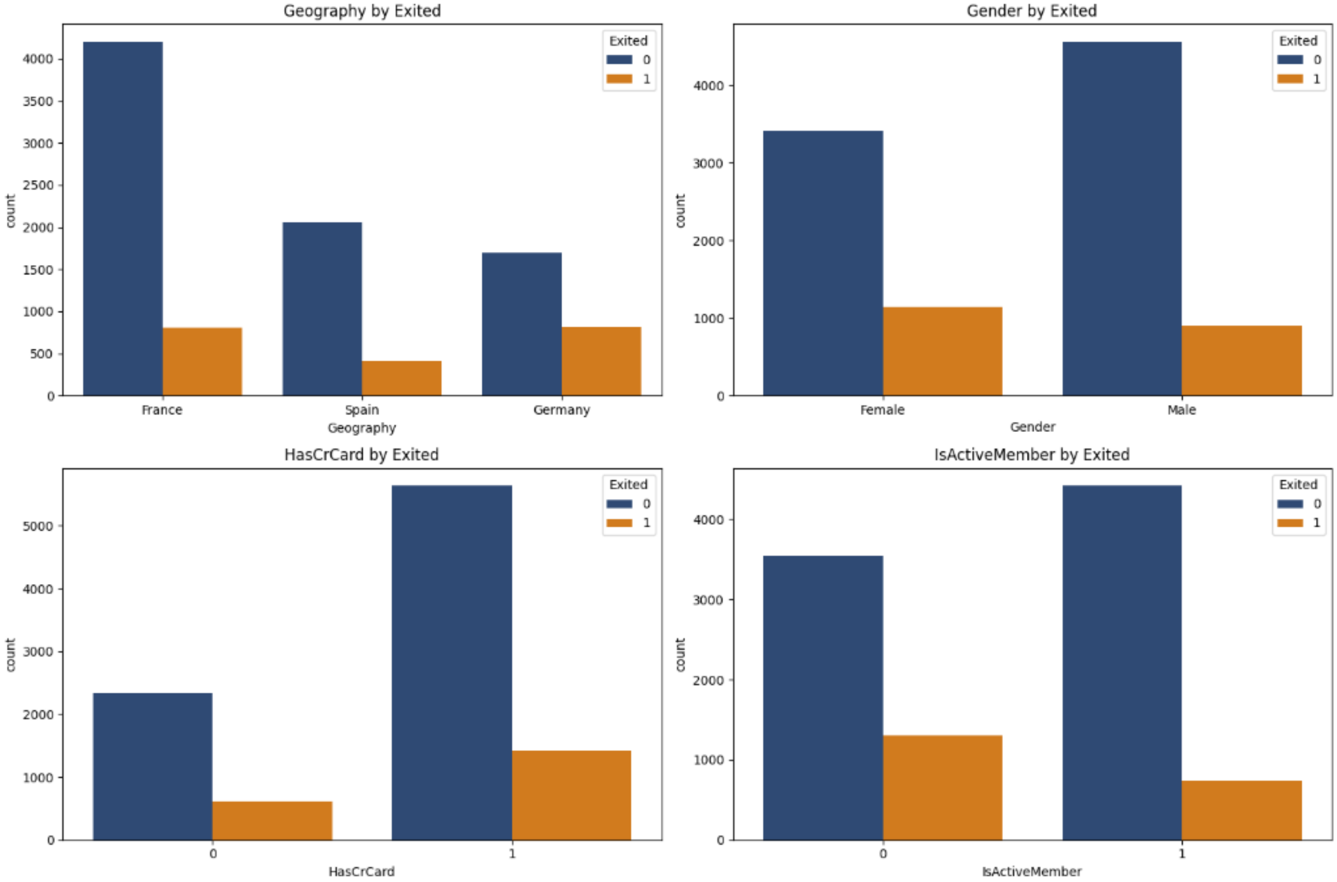
1.2 Descriptive statistics of numerical data



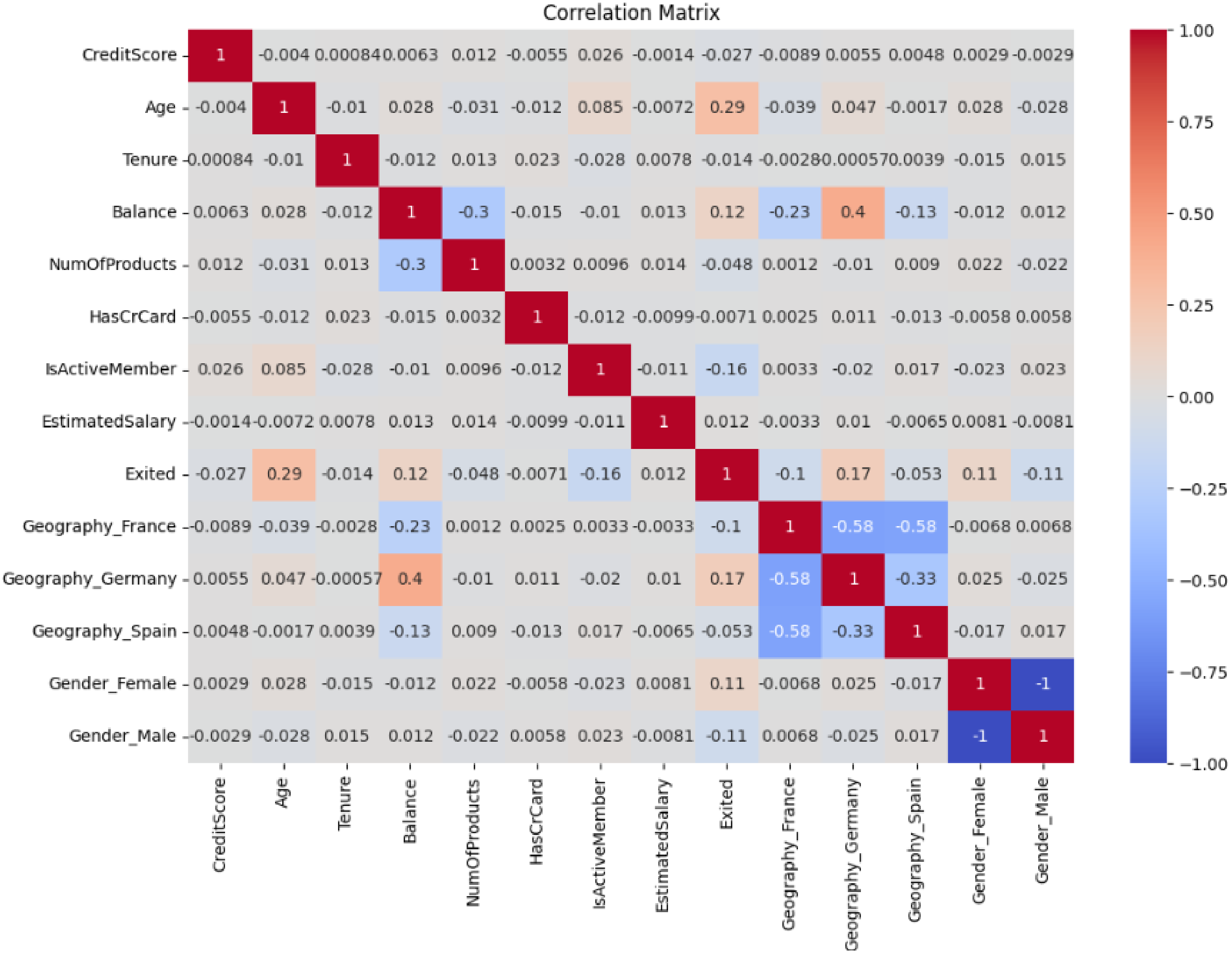
1.3 Boxplot of numerical features



1.4 Distributions of categorical features



1.5 Correlation matrix (note that the categorical features are one-hot encoded and that the target is exit)

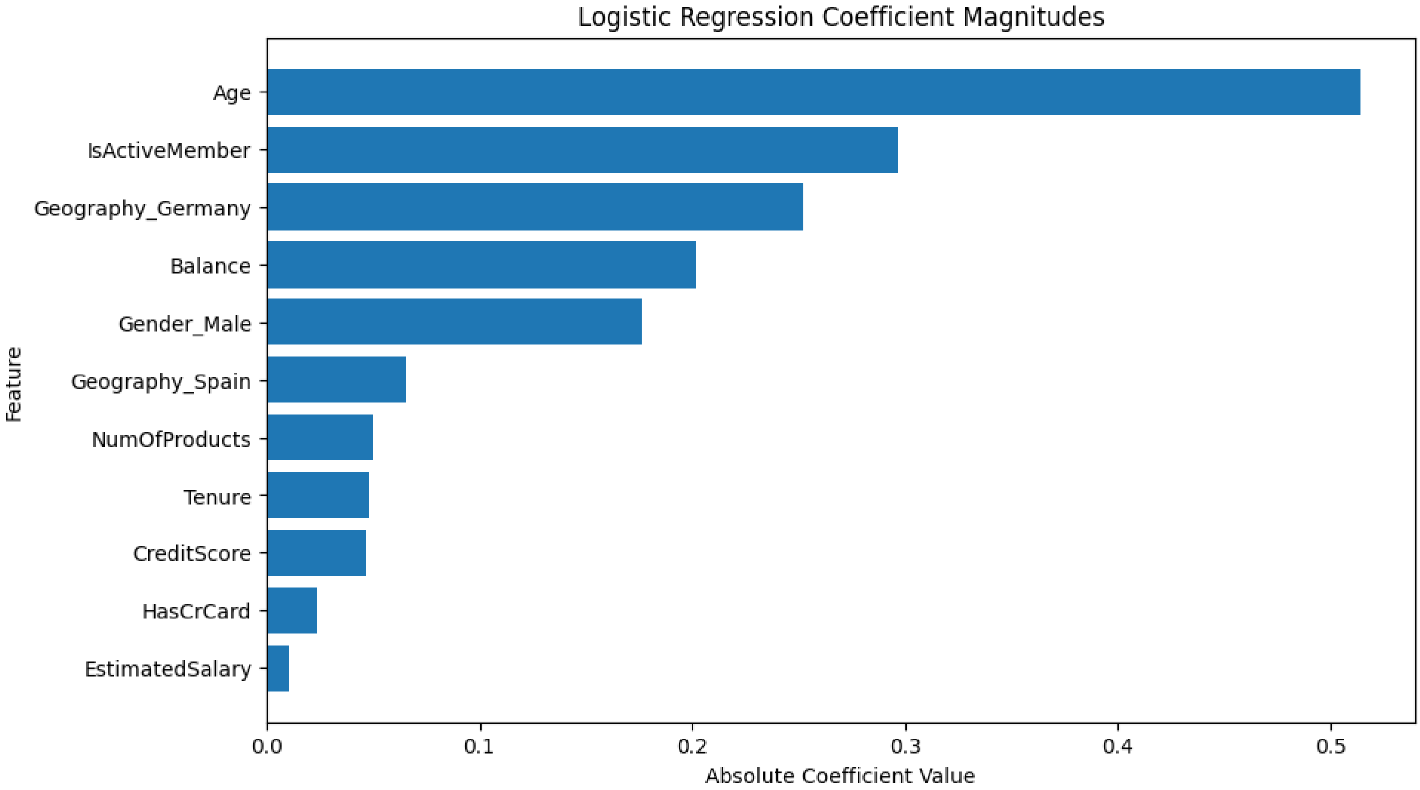


2. Logistic Regression

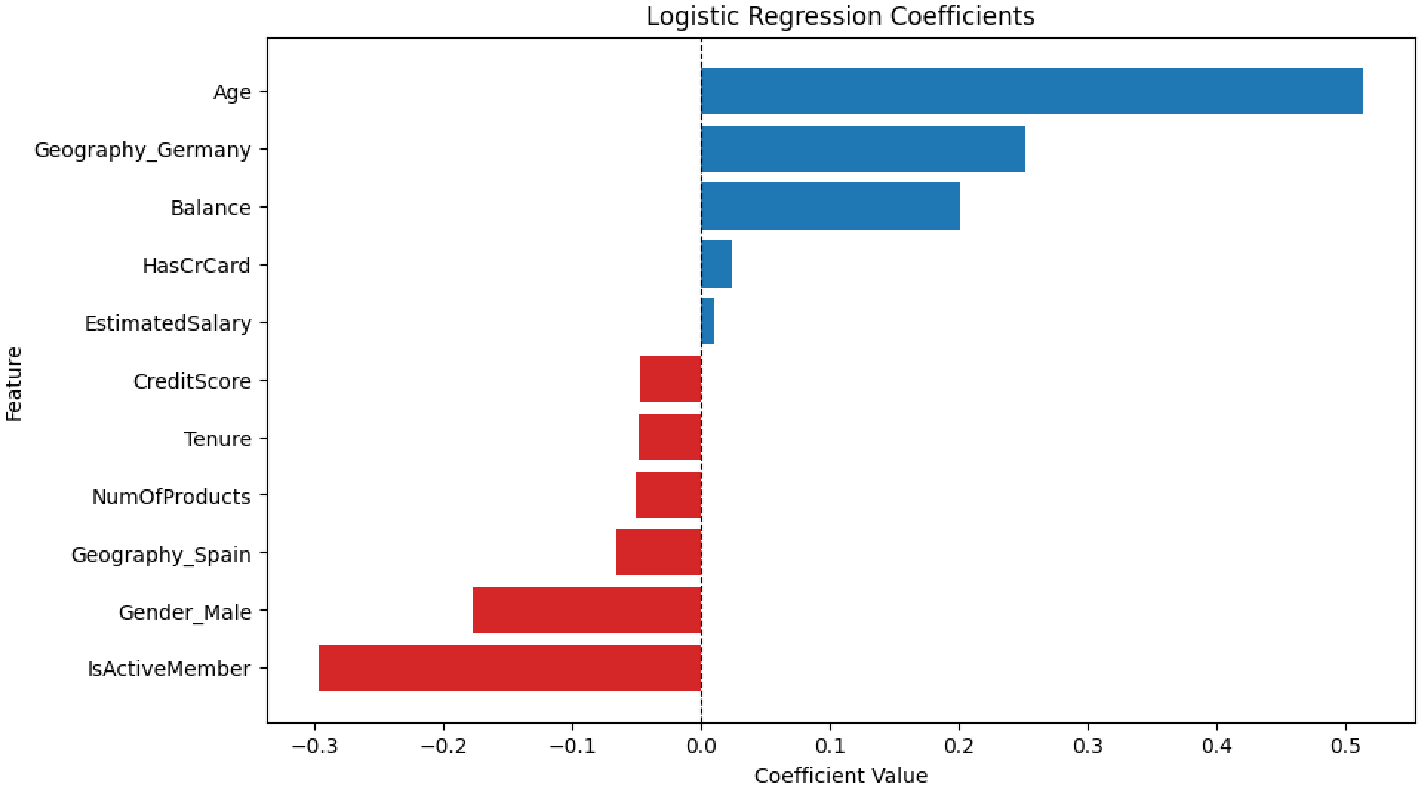
2.1 Logistic Regression – performances of different models

MODEL	VALIDATION ACCURACY	MINORITY PRECISION	MINORITY RECALL	F1-SCORE (MINORITY)
Initial Logistic Regression Model	80%	59%	22%	32%
Grid-Search Optimization	80%	61%	21%	31%
Random Undersampling	71%	40%	69%	51%
SMOTE-Enhanced RF (Final)	72%	41%	69%	51%

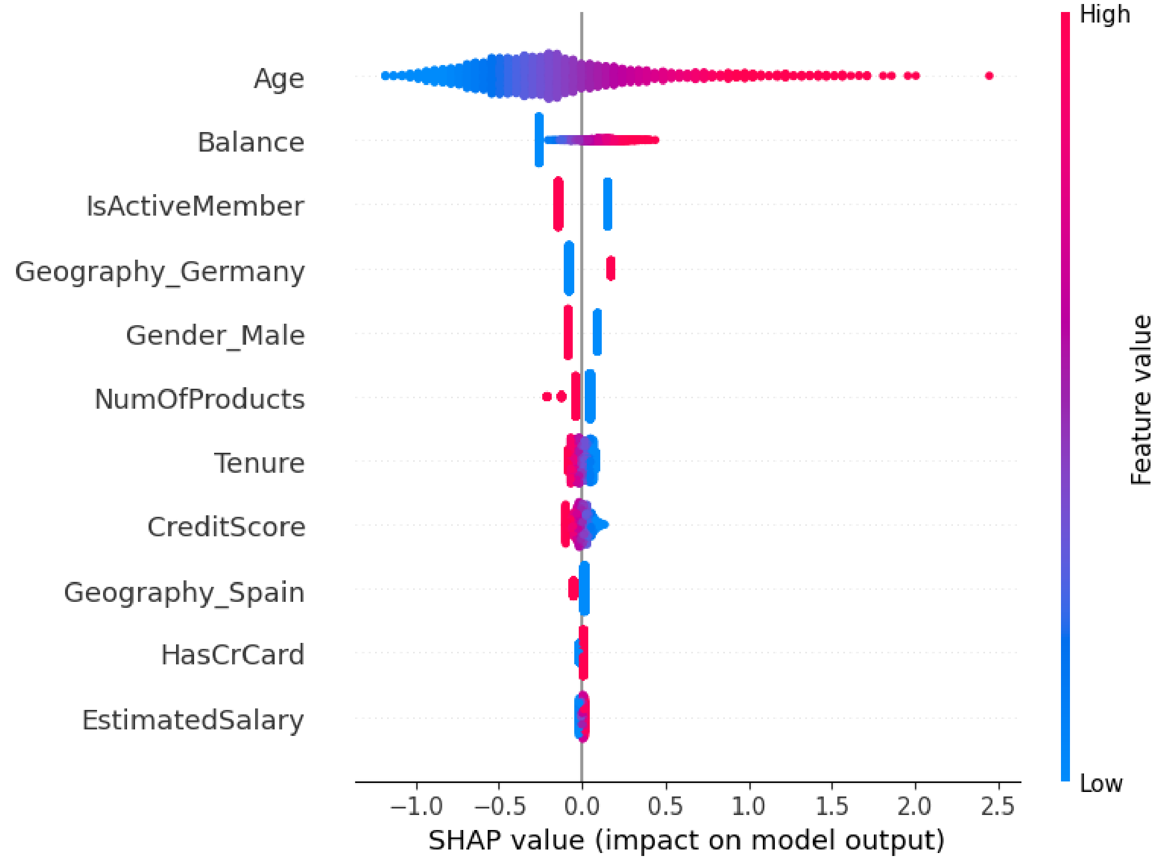
2.2 Logistic Regression – coefficients magnitude



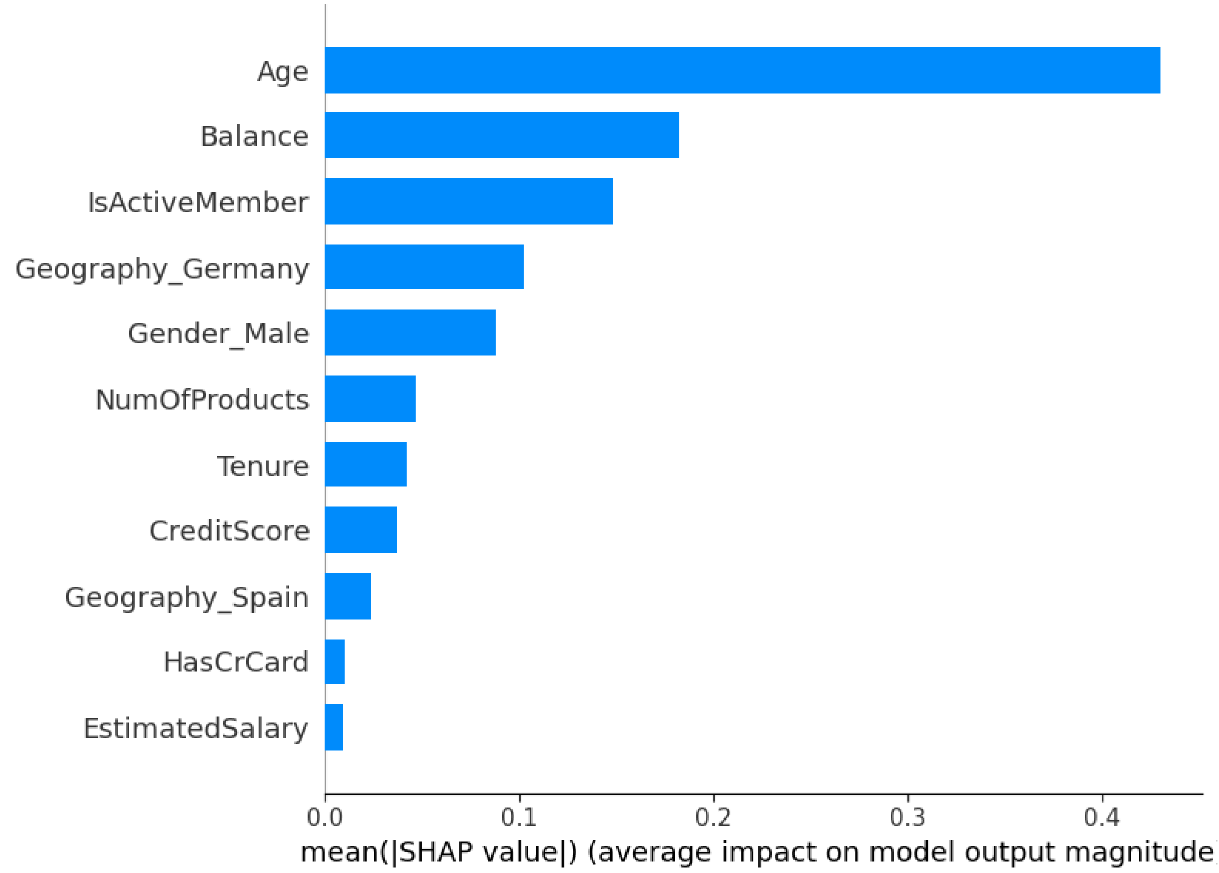
2.3 Logistic Regression – Coefficient with sign



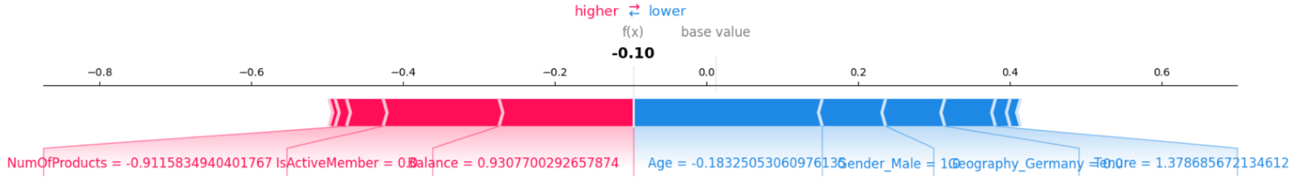
2.4 Logistic Regression – global SHAP analysis for class 1



2.5 Logistic Regression - global SHAP analysis (bar chart) for class 1



2.6 Logistic Regression – an example of local SHAP analysis

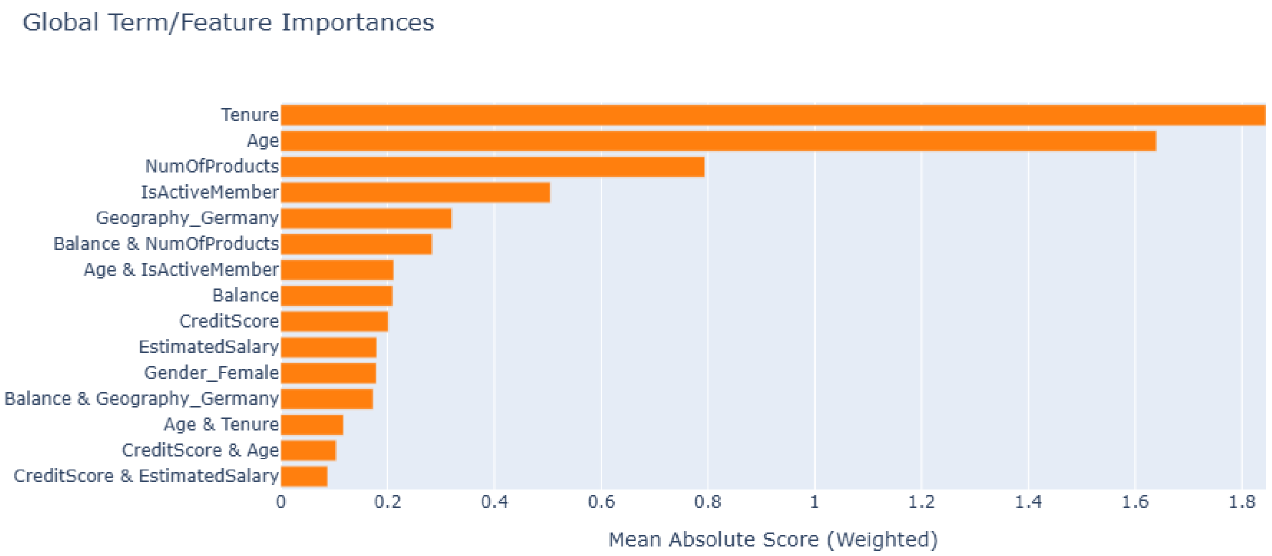


3. Explainable Boosting Machine (EBM)

3.1 Explainable Boosting Machine – performances of different models

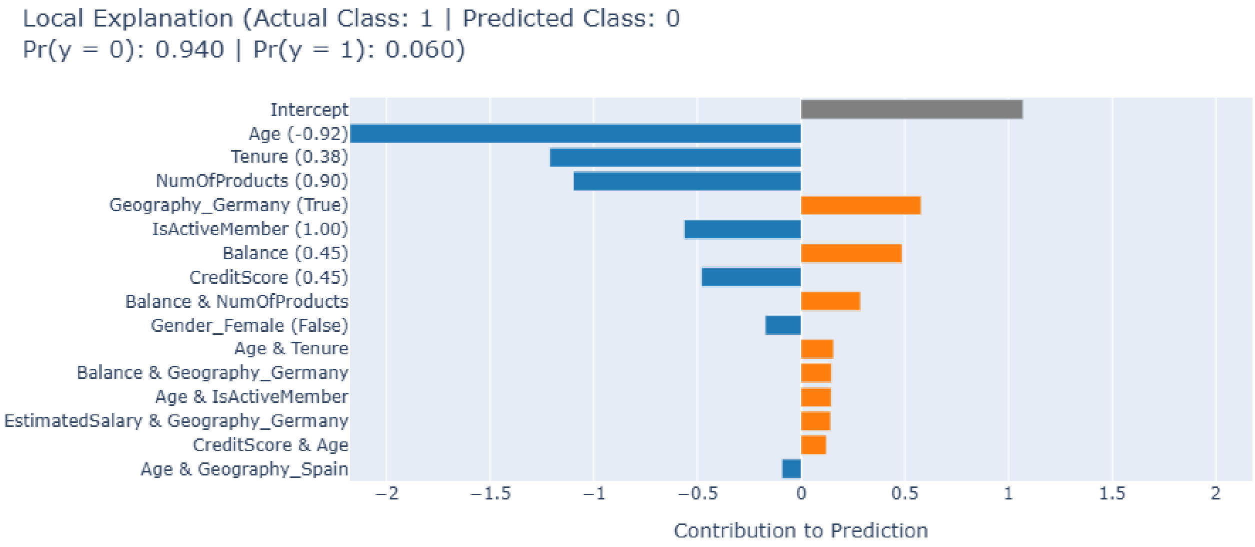
MODEL	VALIDATION ACCURACY	MINORITY PRECISION	MINORITY RECALL	F1-SCORE (MINORITY)
Initial Balanced Random Forest Classifier Sampling	87%	81%	51%	63%
SMOTE-Enhanced RF	86%	72%	54%	62%
Grid-Search Optimization (Final)	86%	72%	54%	62%

3.2 Explainable Boosting Machine – Global Mean Absolute Score



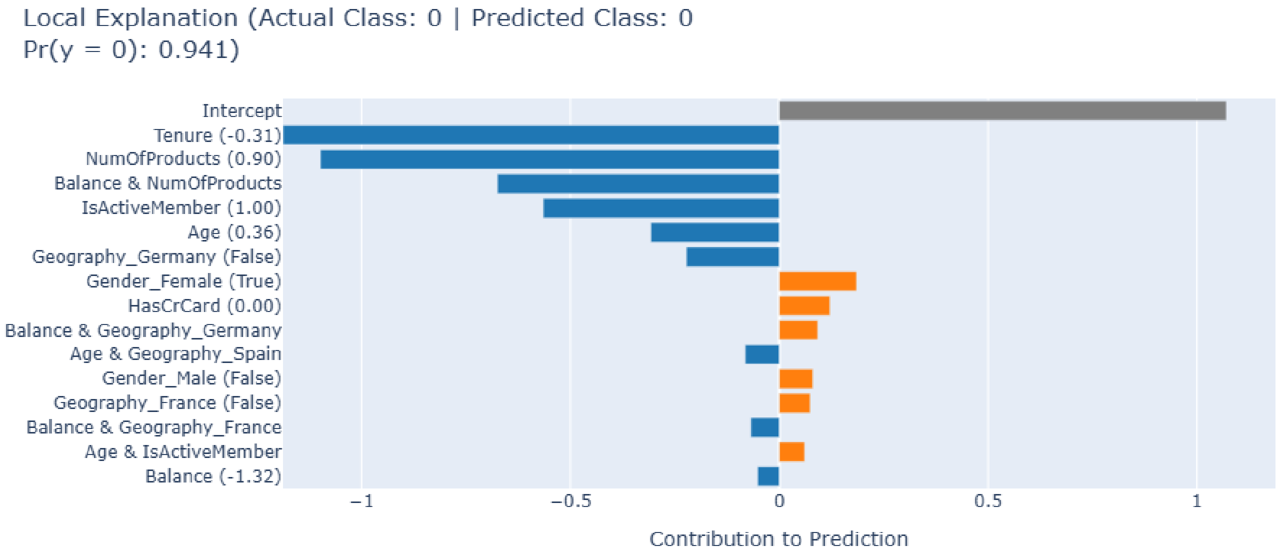
3.3 Explainable Boosting Machine – local importance example 1

As shown in this graph, this individual's age, tenure, and number of products negatively influence the prediction, thus making the predicted target zero while the true target value is one.


Local Explanation (Actual Class: 1 | Predicted Class: 0
Pr(y = 0): 0.940 | Pr(y = 1): 0.060)

3.4 Explainable Boosting Machine – local importance example 2

In this case, the machine learner makes a correct prediction.

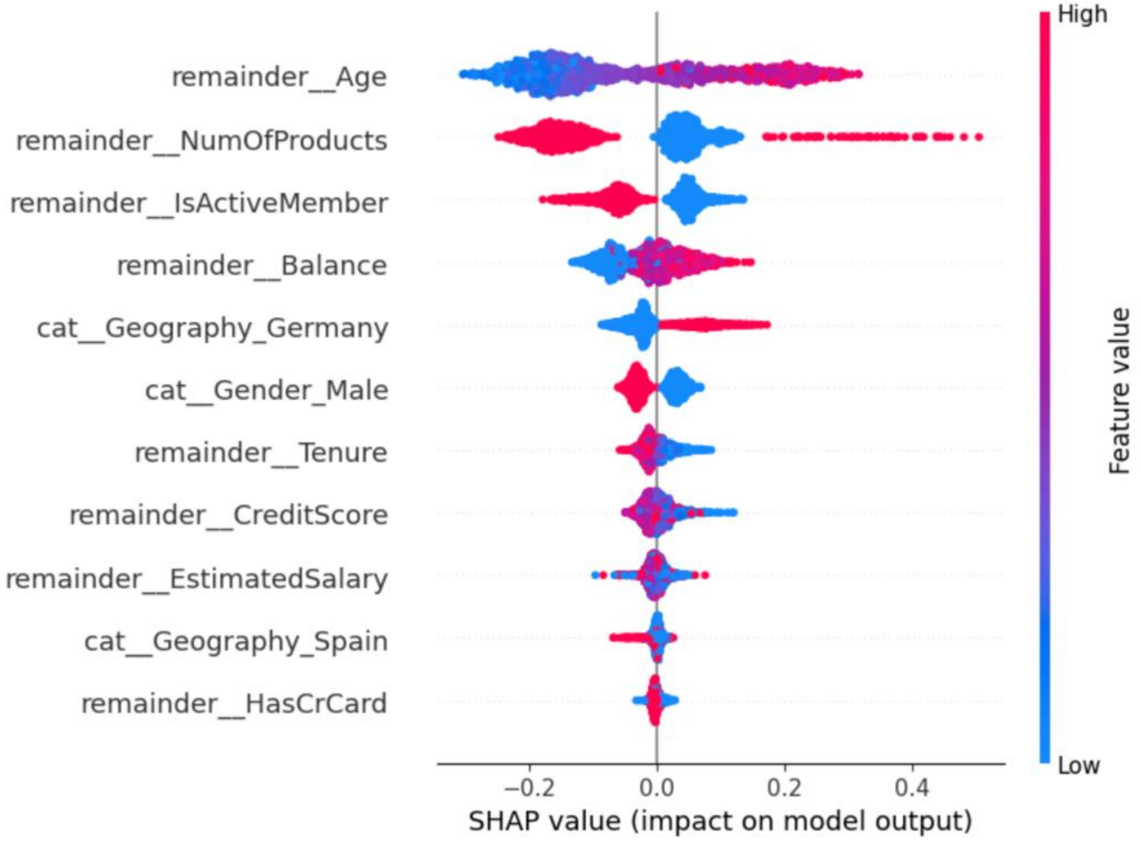

Local Explanation (Actual Class: 0 | Predicted Class: 0
Pr(y = 0): 0.941)

4. Random Forest

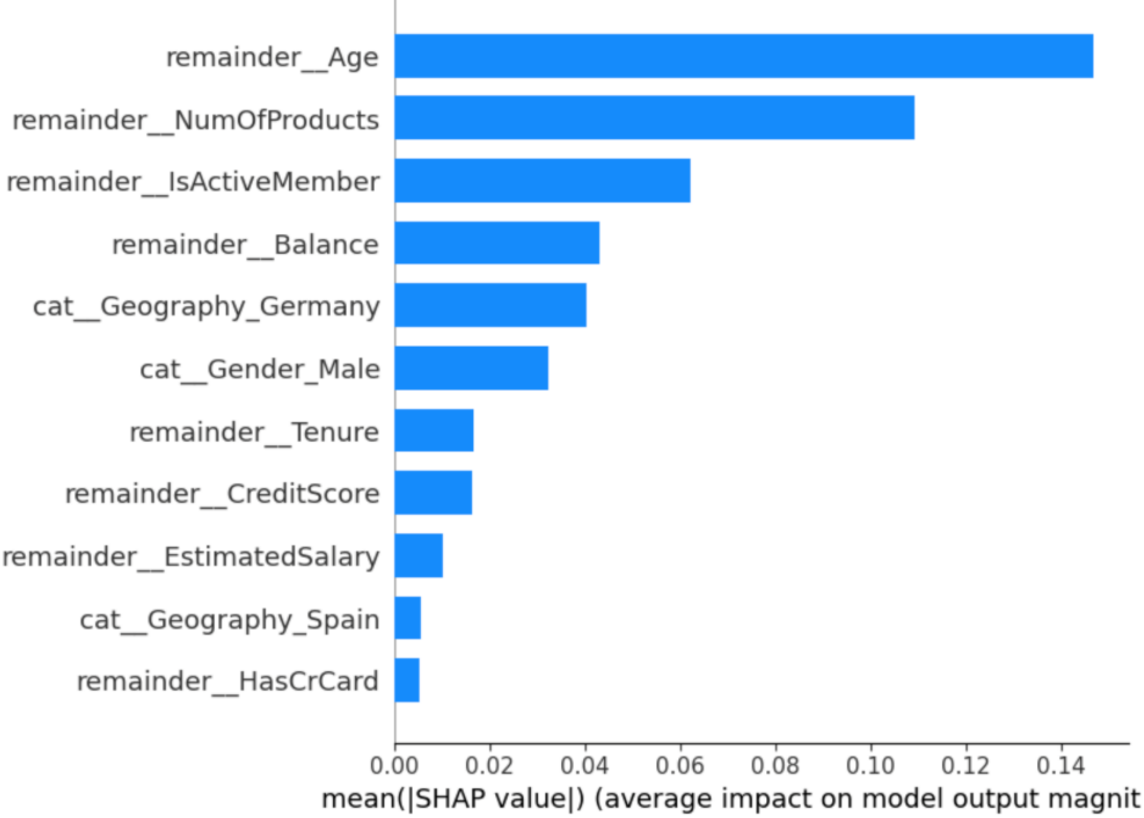
4.1 Random Forest- model performance comparison

MODEL	VALIDATION ACCURACY	MINORITY PRECISION	MINORITY RECALL	F1-SCORE (MINORITY)
Initial Random Forest	86%	74%	44%	55%
Grid-Search Optimized RF	87%	77%	42%	55%
SMOTE-Enhanced RF (Final)	83%	54%	59%	56%
Random Undersampling	78%	45%	72%	55%
Balanced Random Forest	78%	45%	71%	55%

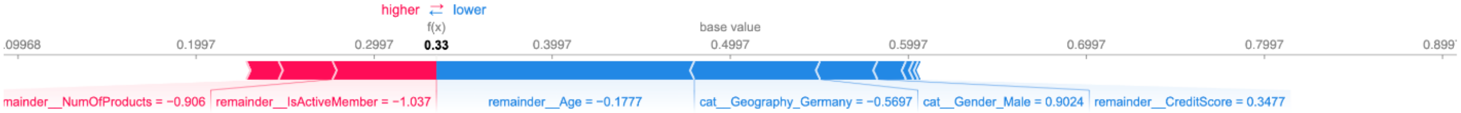
4.2 Random Forest – global SHAP analysis for class 1



4.3 Random Forest – SHAP Analysis (Bar Chart) for Class 1



4.4 Random Forest: SHAP Analysis (local) for Class 1 – One Case

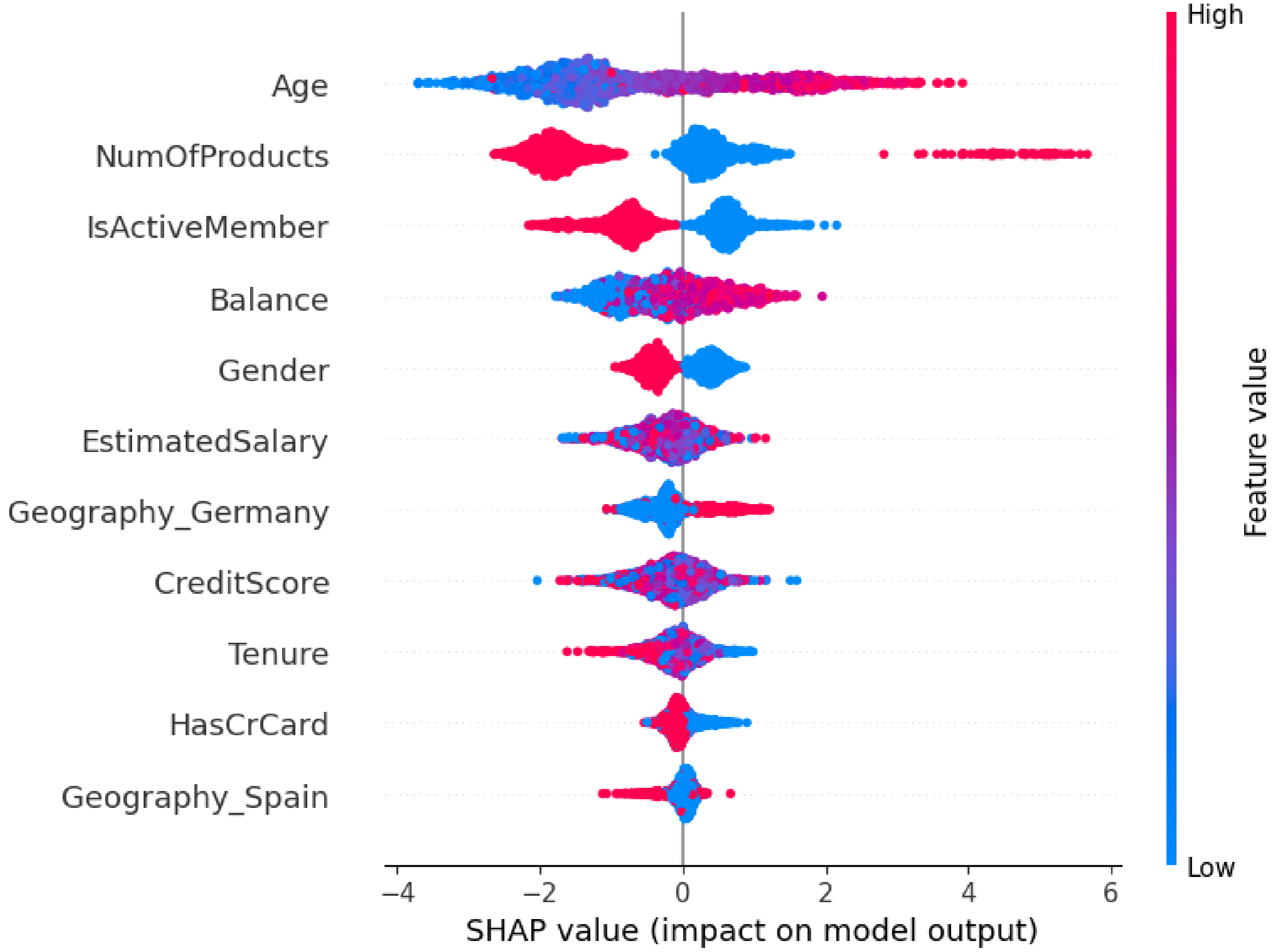


5. CatBoost

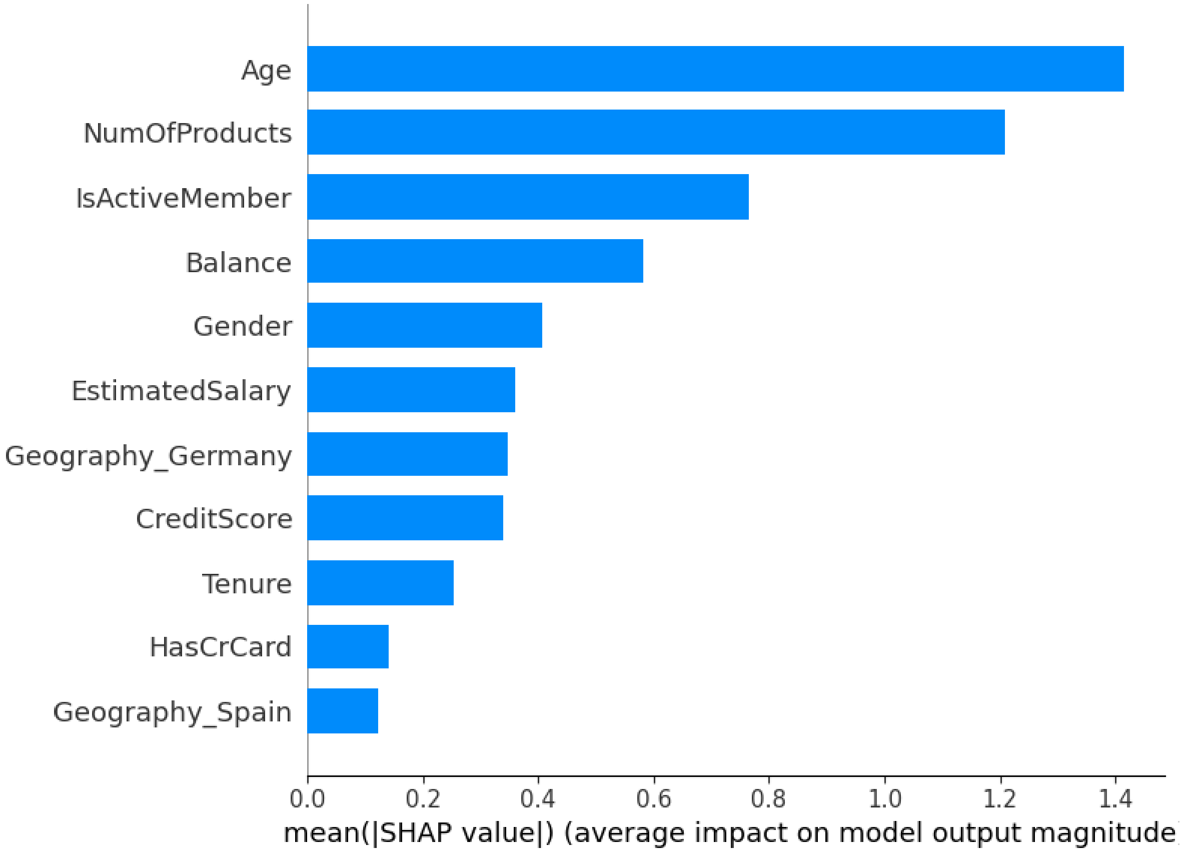
5.1 CatBoost – Model Performance Comparison

MODEL	VALIDATION ACCURACY	MINORITY PRECISION	MINORITY RECALL	F1-SCORE (MINORITY)
Initial CatBoost Model	86%	81%	49%	61%
Balanced Random Forest	79%	51%	77%	61%
Random Undersampling	80%	52%	77%	62%
Random Oversampling	83%	57%	77%	65%
Grid-Search Optimization (Final)	85%	65%	61%	63%

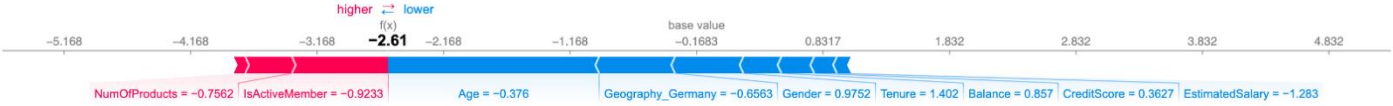
5.2 CatBoost– global SHAP analysis for class 1



5.3 CatBoost– global SHAP analysis for class 1



5.4 CatBoost – local SHAP Analysis for One Case


higher ⇄ lower
f(x)
base value
NumOfProducts = −0.7562
IsActiveMember = −0.9233
Age = −0.376
Geography_Germany = −0.6563
Gender = 0.9752
Tenure = 1.402
Balance = 0.857
CreditScore = 0.3627
EstimatedSalary = −1.283