Group Project

Customer Churn Analysis Report

BM04BIM - Big Data Management and Analytics

2024 - 2025

Honor code

By submitting this assignment, I affirm the following:

1. All work presented in this assignment is my own. I have not collaborated with others or copied work from any unauthorized source.
2. If I used AI tools like ChatGPT, Co-Pilot, etc., I only sought guidance or clarification. Any generated content has been fully understood and appropriately modified to align with the assignment.
3. I understand the submitted code and can explain my work if asked.

I declare that I have read, understood, and agree to abide this honor code.

Student Name	Student Number
Felix Masselter	735890fm
Thy Cao	673415dc
Marius van den Assem	495750ma
Long Le	735090al
Yécine Mezghani	745573ym

Date: 29.11.2024

1. Data Pre-processing

1.1. Outliers

IQR and box plots are employed to identify outliers of continuous variables in the dataset. Looking at *Appendix A*, there are outliers in *credit_score* and *age*. To determine an appropriate method, we examined the churn rate among the outliers of these 2 variables.

churn	number of observations (*credit_score*)	number of observations (*age*)
0	0	255
1	16	65

Table 1. Number of Observations in Churn by Credit Scores and Age

For *credit_score*, all 16 outlier observations indicate positive churn. We retain this data as it might be significant in predicting churn rates.

For *age*, the churn rate aligns more consistently with the dataset. Given the large number of outliers (N = 320), removing them could result in significant information loss and affect predictive models. Hence, we applied the log transformation to minimize the impact of outliers while retaining them. This approach is validated by examining the distribution of age before and after transformation (*Appendix B*), with the results showing a normal distribution after the transformation.

1.2. Irrelevant Variables

customer_id was removed due to its irrelevance to the quantitative analysis.

1.3. Categorical Variables

Two categorical variables, *gender* and *country*, were one-hot encoded for our analysis. During our one-hot encoding process, we left out the reference category, in this case, France and Female to avoid perfect collinearity.

2. Exploratory Data Analysis

2.1. Overall Churn Rate

The overall churn rate is 20.36%, showing an imbalance in the data (*Appendix Q*).

2.2. Churn Rate among Demographic Variables

2.2.1. Gender

The churn rate among females is 24.98%, which is higher than the rate among males, 16.49%, see *Appendix C*.

2.2.2. Age

We classified *age* by groups for interpretation, including the total number of observations in each group for better comparison. *Appendix D* reveals high churn rates among customers aged 40 to 70. Most of them are higher than 30%, especially the 50 - 60 group with a churn rate of 55.57%. The churn rates among the 40 - 70 age group significantly exceed the average churn rate of **20.36%** and the rest of the dataset. This age group comprises approximately 39% of the dataset with a number of observations of 3,486, suggesting *age* could be a significant predictor of churn rate.

2.2.3. Credit Scores

Appendix E shows an interesting insight that all customers with credit scores below 400 churned. However, this group only accounts for a small number (19 observations) of the customers so the overall impact is low. In general, there is no identifiable relationship between churn rate and credit scores.

2.2.4. Country

Appendix F reveals that customers from Germany have the highest churn rate of 33.32%, compared to the churn rates of Spain and France, with 16.7% and 16.15%, respectively.

2.2.5. Estimated Salary

Appendix G shows that customers with high salaries (> €140,000) are slightly more likely to churn. However, overall, there are no significant differences in churn rates between customers with different estimated salaries.

2.3. Churn Rate and Tenure

Appendix H shows that there is a weak negative relationship between tenure and churn rate. This is validated by the **Pearson correlation** of **-0.37.** This shows that longer-term customers are slightly less likely to churn.

2.4. Interesting Patterns

We yield some interesting findings from the analysis of the data:

- **Credit Score:** Customers with a low credit score (< 400) have a 100% churn rate.
- **Balance**: Customers with extremely high balances are more likely to churn. Customers with 0 balances are the least likely to churn, and they account for a significant number (N = 3,241). (*Appendix J*)
- **Products Number:** Customers who own more products (3 and 4) are much more likely to churn. More than 80% of customers owning 3 products churned, while 100% of customers owning 4 products churned. (*Appendix K*)

- **Active Status**: Customers who are inactive are more likely to churn (26.6% compared to 14.4%). (*Appendix L*)

3. Model Building

A frequent problem with customer churn datasets is that the data is often imbalanced which can be seen also during the exploratory data analysis with 80% of observations of the dataset being non-churners. This section will show how we approached the model-building process.

3.1. Models Selection for Comparison

Firstly, the problem at hand is a supervised problem with a structured dataset. This is why we decided to use CatBoost, XGBoost, Random Forest, Logistic Regression and K-NN to predict churns. We went with 5 models in total because every model exhibits different learning approaches. For instance, logistic regression looks for linear relationships, tree-based models (Catboost, XGBoost, Random Forest) capture nonlinear interactions and K-NN looks more at the similarity between customers. This can also mitigate the weaknesses of each of the models. Furthermore, having 5 models allows us to better understand which features drive customer churn by comparing the feature importance between the different models (*Appendix O*).

3.2. Data Preparation for Model Training

We used the pre-processed dataset to train the model, *churn* is defined as the target variable (y) while the others are the features (X). Furthermore, a train-test split ratio of 80-20 with stratification was used to keep the class distribution across splits.

3.3. Addressing Class Imbalance

We applied the Synthetic Minority Over-sampling Technique with Edited Nearest Neighbours (SMOTE-ENN), which is a combination of SMOTE and Edited Nearest Neighbors (ENN) (Chandaliya, 2023). If we used only SMOTE, it could create noisy and irrelevant synthetic samples, leading to overfitting. Hence, SMOTE-ENN is a better option, tackling the weakness by removing noisy samples from the dataset. This method combines the oversampling of the minority class (in this case churners) with the undersampling of the majority class, creating a more balanced dataset while removing noise.

3.4. Hyperparameters Tuning

Each model was tuned with the help of the grid search. In addition, 5-fold cross-validation was used to make the parameter selection more robust. During the hyperparameter tuning, we aimed to optimize the F1 score as it balances precision and recall.

3.5. Model Training and Evaluation

We trained each model using the best parameters found through grid search. These parameters were then used for predicting test data. Subsequently, the model was evaluated based on the following metrics: Accuracy, Precision, Recall, F1 Score and ROC AUC.

3.5.1. K-NN (All Features) vs K-NN (Reduced Features)

Through feature importance analysis using permutation importance, we discovered that *credit_score* and *estimated_salary* had negative mean importance scores, indicating they provide minimal predictive value and potentially introduce noise to the model. This led us to create K-NN including all features and K-NN with reduced features, which showed that the K-NN with reduced features outperformed the complete Model on every metric. (see appendix). After conducting a paired T-Test, we confirmed that the reduced model demonstrates a significantly higher Mean F1 Value than the complete model (p-value = 0.0077), validating our feature selection approach. Consequently, we went forward with the K-NN model with reduced features.

3.6. Results Interpretation

3.6.1. Strongest Predictors

The heat map in *Appendix N* shows that *log_age*, *active_member*, *products_number*, *gender_Male* and potentially also *balance* were the strongest predictors of customer churns as they were ranked top 5 for the majority of models.

3.6.2. Evaluation Metrics among Five Models

Model	Accuracy	Precision	Recall	F1 Score	ROC AUC
CatBoost	0.7906	0.4889	0.6639	0.5632	0.8298
XGBoost	0.7989	0.5043	0.6448	0.5659	0.8245
Random Forest	0.7944	0.4956	0.6120	0.5477	0.8072
Logistic Regression	0.7139	0.3739	0.6038	0.4619	0.7377
KNN (reduced)	0.7672	0.4533	0.7022	0.5509	0.7946

Table 2. Comparison between Evaluation Metrics of CatBoost, XGBoost, Random Forest, Logistic Regression and KNN (Reduced)

The reduced KNN model has a high recall (0.7022), indicating that it is the best model for identifying customers who will actually churn. Depending on the business context, missing a potentially churning customer (FN) can be more costly than incorrectly flagging a loyal customer (FP).

On the other hand, XGBoost has the highest accuracy (0.7989), which is less crucial in a churn problem but it also has the best precision (0.5043) and a strong F1 score (0.5659) when comparing it to the other models. As the F1 metric includes both recall and precision, it

allows us to balance catching potential churners and avoiding false alarms. The higher precision is also crucial because false positives can lead to unnecessary intervention costs and even irritate satisfied customers. Lastly, XGBoost also demonstrates the highest ROC AUC score (0.8245), which means that it has the best overall ability to distinguish between churners and non-churners.

3.6.3. Model Selection for Predicting Churns

The choice between these models boils down to the business context. For instance, if the cost of intervention is high, XGBoost might be a better option but if the cost of losing a customer is high then KNN might be preferable since it catches more potential churners. Due to limited business context, we would opt for XGBoost as it is overall the better model with a strong F1 score, highest precision and highest ROC AUC score.

4. Recommendations

4.1. Three Recommendations to Manage Customer Churn

Based on the three most important features from the heat map in *Appendix N* and domain knowledge, we developed 3 recommendations to help EBI better manage customer churn: an age-based retention strategy, a competitive bundle pricing strategy and an active member engagement program.

4.1.1. Age-based Retention Strategy

The age-based retention strategy will target the age group most vulnerable to churn, which lies between the ages of 40 to 70 years old (< 39% churn rate) with the 50 - 60 being most affected. (55.57% churn). The strategy would consist of developing tailored financial solutions for pre-retirement and retirement phases. This could include creating a specialized estate planning department to better accommodate this customer segment. Furthermore, the bank should analyze if any of their physical banking branches were closed as these could have affected this age segment to churn. If this is the case, then the age-based retention strategy should include potentially opening up these stores because the age tends to have lower digital literacy than the younger customer segment.

4.1.2. Competitive Bundle Pricing Strategy

The higher churn among customers with 3-4 products could be explained by pricing issues, more specifically competitors may offer better rates for similar product bundles. To start with, EBI should conduct a competitive price analysis by benchmarking each product's pricing against competitors and comparing bundle pricing with competitor offerings. In addition, they can analyze the price sensitivity at different product combination levels. According to the findings, the bank can optimize their prices by for example introducing discounts for owning multiple products and thus creating cost-effective bundle packages. This strategy will

lead to reduced churn in multi-product customers and better retain the price-sensitive customers.

4.1.3. Active Member Engagement Program

Our analysis reveals a concerning trend of increasing churn propensity in inactive members. To address this, we recommend implementing a comprehensive Active Member Engagement Program focusing on three key pillars: (1) an early warning system that is able to identify declining activity patterns before complete disengagement, (2) targeted re-engagement campaigns for inactive users with personalized incentives to restore banking activity, and (3) offering a streamlined digital banking experience to reduce barriers to engagement. This program addresses the root cause of inactive member churn by providing proactive intervention before customers fully disengage.

4.2. Proactive Measures

4.2.1. Churn Likelihood in Current Customers

As there are 1,000 observations, the churn likelihood in current customers is shown in a separate CSV file named "*Group2_PartB_Churnlikelihood*". We visualized the results in *Figure 1*.

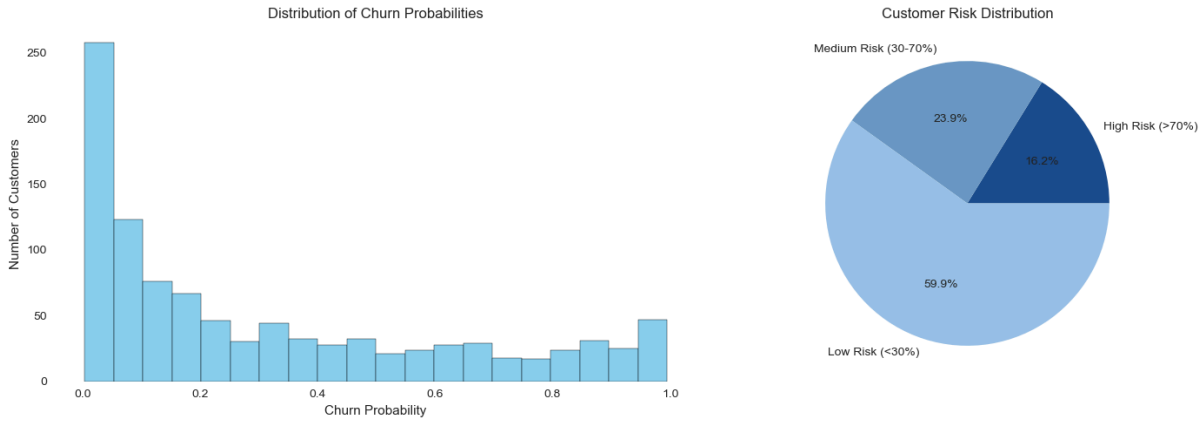


Figure 1. Bar Graph of Churn Probabilities Distribution and Pie Chart of Customer Risk Distribution

Due to limited context, we classified the churn likelihood based on our assumption:

- High Risk Customers (>70%): 162 (16.2%)
- Medium Risk Customers (30-70%): 239 (23.9%)
- Low Risk Customers (<30%): 599 (59.9%)

The bank plans to target customers with a high likelihood of churn (in this case we assumed >70%), which means that it would target 162 customers via telemarketing.

4.2.2. Profit Maximization Calculations

We used "*ebi_base_customers.csv*" to calculate the optimal thresholds for two scenarios retaining a customer valued at €5 and €10. These thresholds were optimized to maximize

expected profit by balancing the cost of intervention (€1) against the benefit of retaining customers.

First, we created 2 cost-benefit matrices. Each matrix has 4 quadrants representing the benefit/loss of predicting correctly/incorrectly a churned customer, see *Table 3* and *Table 4*.

Customer Valued at €5		**Predicted**	
		No Churn	**Churn**
Actual	**No Churn**	€0	-1 €
	Churn	- €5	€4

Table 3. Cost-Benefit Matrix of Predicting Churn (€5 Retention Value)

Customer Valued at €10		**Predicted**	
		No Churn	**Churn**
Actual	**No Churn**	€0	- €1
	Churn	- €10	€9

Table 4. Cost-Benefit Matrix of Predicting Churn (€10 Retention Value)

We do not yet know if our predictions are correct, which is why we could not create a confusion matrix for the "*ebi_exp_customers.csv*" dataset. However, we could optimize the threshold for classifying a customer as a churner to maximize expected profit for "*ebi_base_customers.csv*". The findings are an optimal threshold of 0.152 for €5 retention value and 0.071 threshold for €10 retention value. As shown in *Appendix M*, our optimal threshold halves when doubling the value of the customer retention because we have to classify more aggressively potential churners as missing churners are more costly at the €10 retention value and since our false positive loss remains the same at -€1. Based on these thresholds we classified 54.3% of customers to churn at the 0.152 threshold level and 69.2% of customers to churn at the 0.071 threshold level. These thresholds conflict with a much higher threshold set in 4.2.1. because the bank most likely assumed that targeting the customers with a high likelihood to churn would lead to the highest profit which it does not. Then, we examined the confusion matrices based on the optimized thresholds, see *Figure 2*.

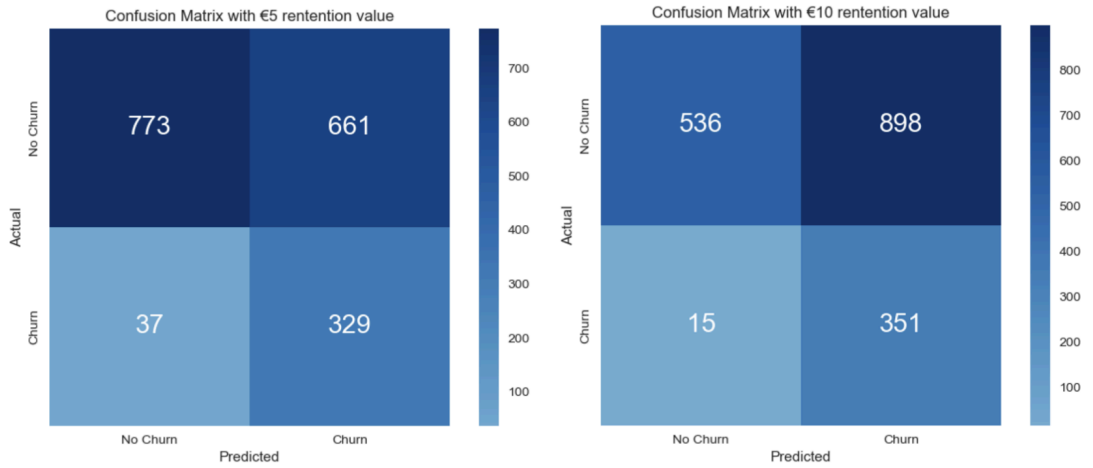


Figure 2. Confusion Matrices of Predicting Churn (€5 Retention Value vs. €10 Retention Value)

These predictions stem from our test data set (N=1,800) but now we have a new data set (N=1,000) so we aim to estimate a new confusion matrix based on the same proportions as the original confusion matrix except that we now have 1,000 total observations (Assuming that the "*ebi_exp_customers.csv*" and "*ebi_base_customers.csv*" do not have inherent differences). The calculations for the proportions at €5 retention value are as follows:

- True Negatives (TN): $TN\ Proportion = \frac{TN}{Total} = \frac{773}{1800} \approx 0.4294$
- False Positives (FP): $FP\ Proportion = \frac{FP}{Total} = \frac{661}{1800} \approx 0.3672$
- False Negatives (FN): $FN\ Proportion = \frac{FN}{Total} = \frac{37}{1800} \approx 0.0206$
- True Positives (TP): $TP\ Proportion = \frac{TP}{Total} = \frac{329}{1800} \approx 0.1828$

Then, as the dataset has 1,000 observations, we multiplied each proportion by 1,000:

- True Negatives (TN):
  $Estimated\ TN = TN\ Proportion \times 1,000 = 0.4294 \times 1,000 \approx 429$
- False Positives (FP):
  $Estimated\ FP = FP\ Proportion \times 1,000 = 0.3672 \times 1,000 \approx 367$
- False Negatives (FN):
  $Estimated\ FN = FN\ Proportion \times 1,000 = 0.0206 \times 1,000 \approx 21$
- True Positives (TP):
  $Estimated\ TP = TP\ Proportion \times 1,000 = 0.1828 \times 1,000 \approx 183$

The computation methods are the same at €10 retention value and shown in *Appendix P*.

From these results, the confusion matrices of the data set "*ebi_exp_customers.csv*" are as follows:

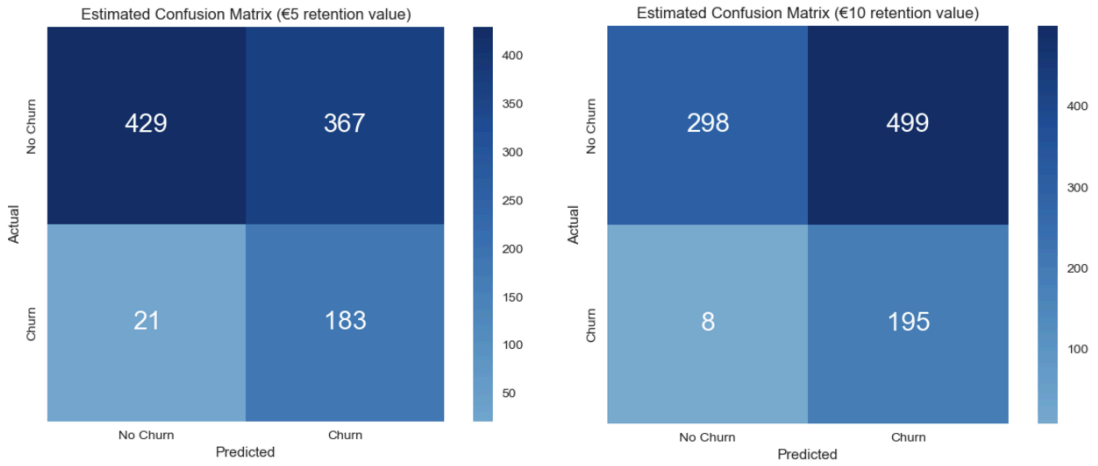


Figure 3.Estimated Confusion Matrices of Predicting Churn in Current Customer (€5 Retention Value vs. €10 Retention Value)

Therefore, we can calculate the estimated expected profit for each scenario, see *Table 5* and *Table 6*.

Customer Valued at €5	Calculation	Loss/Profit
True Positives	$Estimated\ TP \times Benefit/Loss = 183 \times €4$	€732
False Positives	$Estimated\ FP \times Benefit/Loss = 367 \times €\text{-}1$	- €367
False Negatives	$Estimated\ FN \times Benefit/Loss = 21 \times €\text{-}5$	- €105
Total Profit/Loss		**€260**

Table 5. Loss/Profit Calculation (€5 Retention Value)

Customer Valued at €10	Calculation	Loss/Profit
True Positives	$Estimated\ TP \times Benefit/Loss = 195 \times €9$	€1,755
False Positives	$Estimated\ FP \times Benefit/Loss = 499 \times €\text{-}1$	- €499
False Negatives	$Estimated\ FN \times Benefit/Loss = 8 \times €\text{-}10$	- €80
Total Profit/Loss		**€1,176**

Table 6. Loss/Profit Calculation (€10 Retention Value)

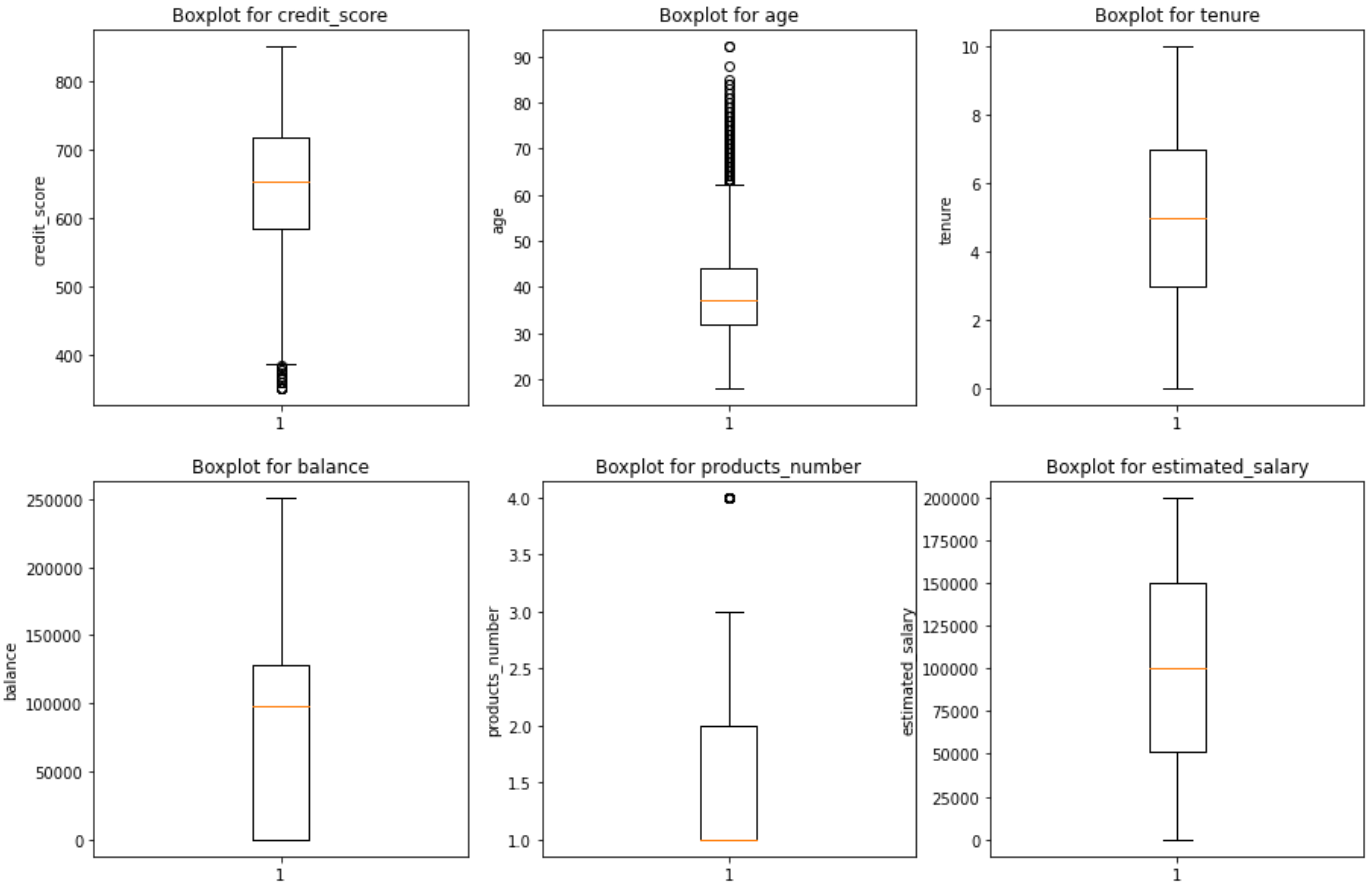
Based on the profit-maximizing threshold, we recommend EBI to target 54.3% (n=543 where $p > 0.152$) of customers in the *"ebi_exp_customers.csv"* dataset at a €5 retention value which would yield an estimated total profit of €260. At the €10 retention value, we recommend to target 69.2% (n=692 where $p > 0.071$) of customers, which would yield an estimated total profit of €1,176.

References

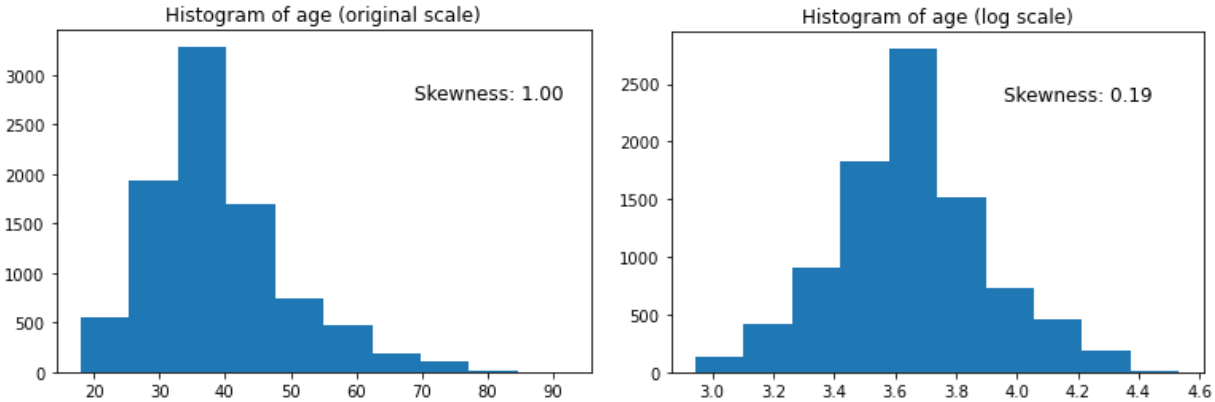
Chandaliya, P. (2023, February 28). SMOTE vs SMOTE-ENN: Which is more effective for Churn Prediction in Imbalanced Banking Data. *Medium*. https://pranav-c.medium.com/smote-vs-smote-enn-which-is-more-effective-for-churn-prediction-in-imbalanced-banking-data-b289414366a0

Appendices

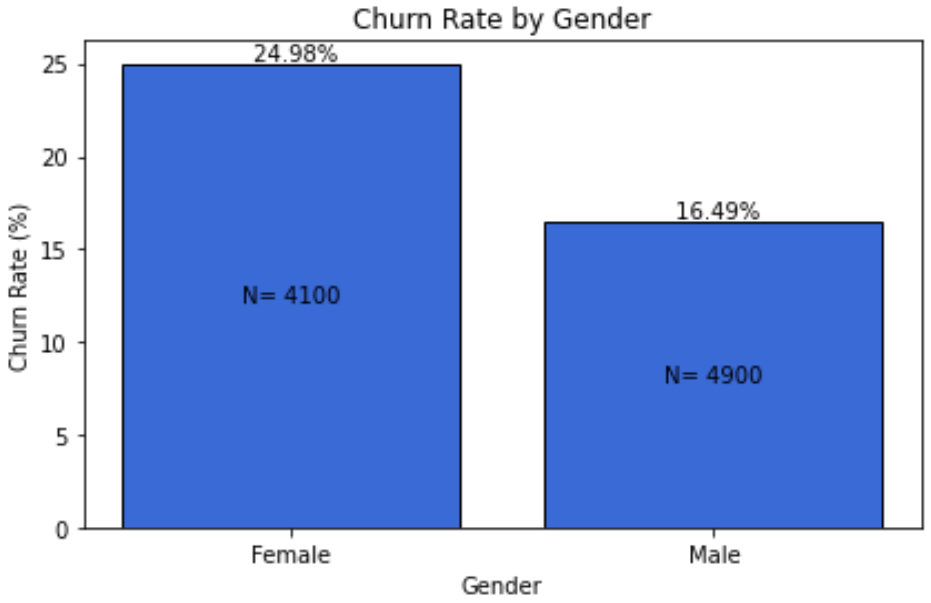
Appendix A. Box Plots of Credit Score & Churn, Age & Churn, Tenure & Churn, Balance & Churn, Number of Products & Churn, Estimated Salary & Churn



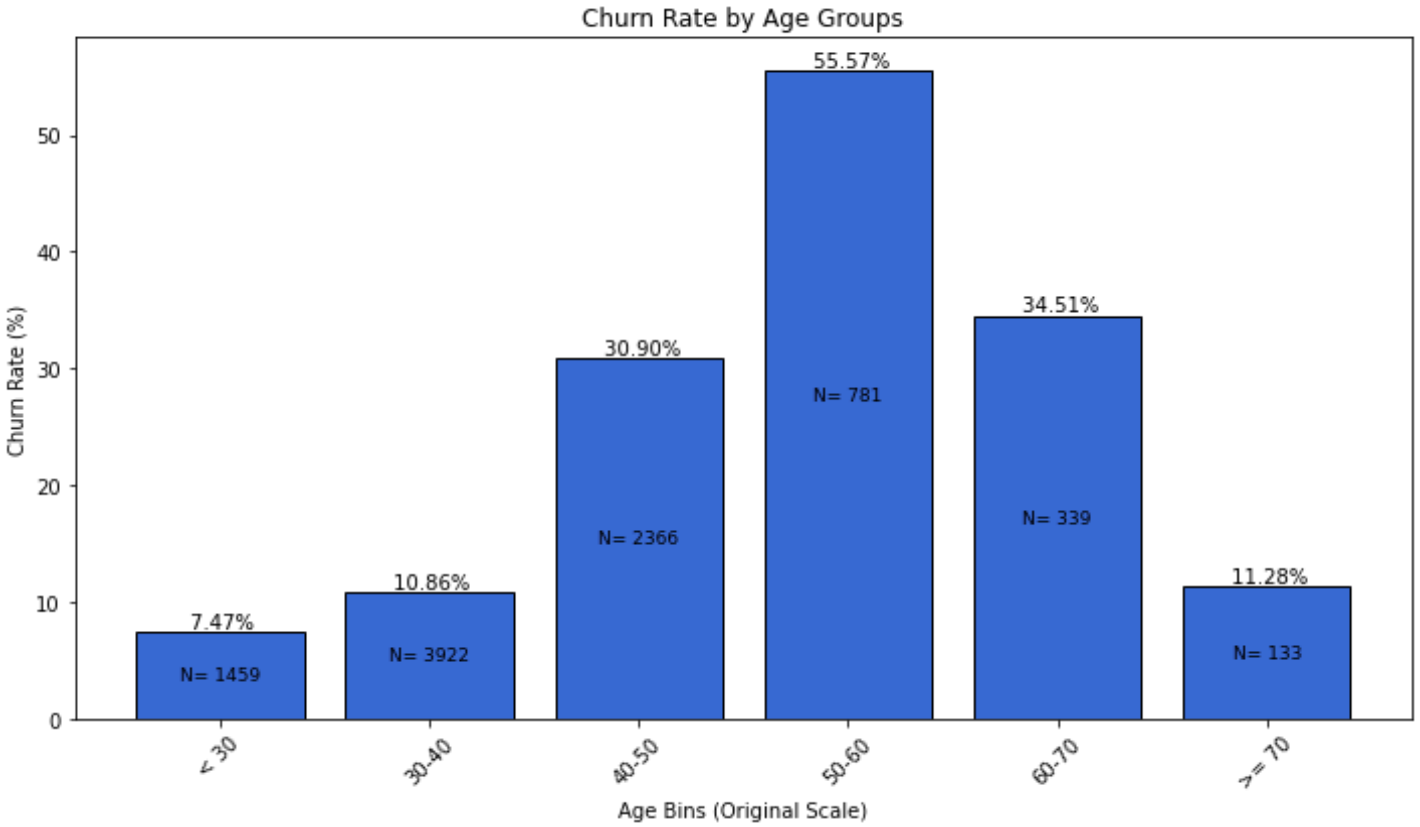
Appendix B. Histogram of Age & Churn Before and After Log Transformation



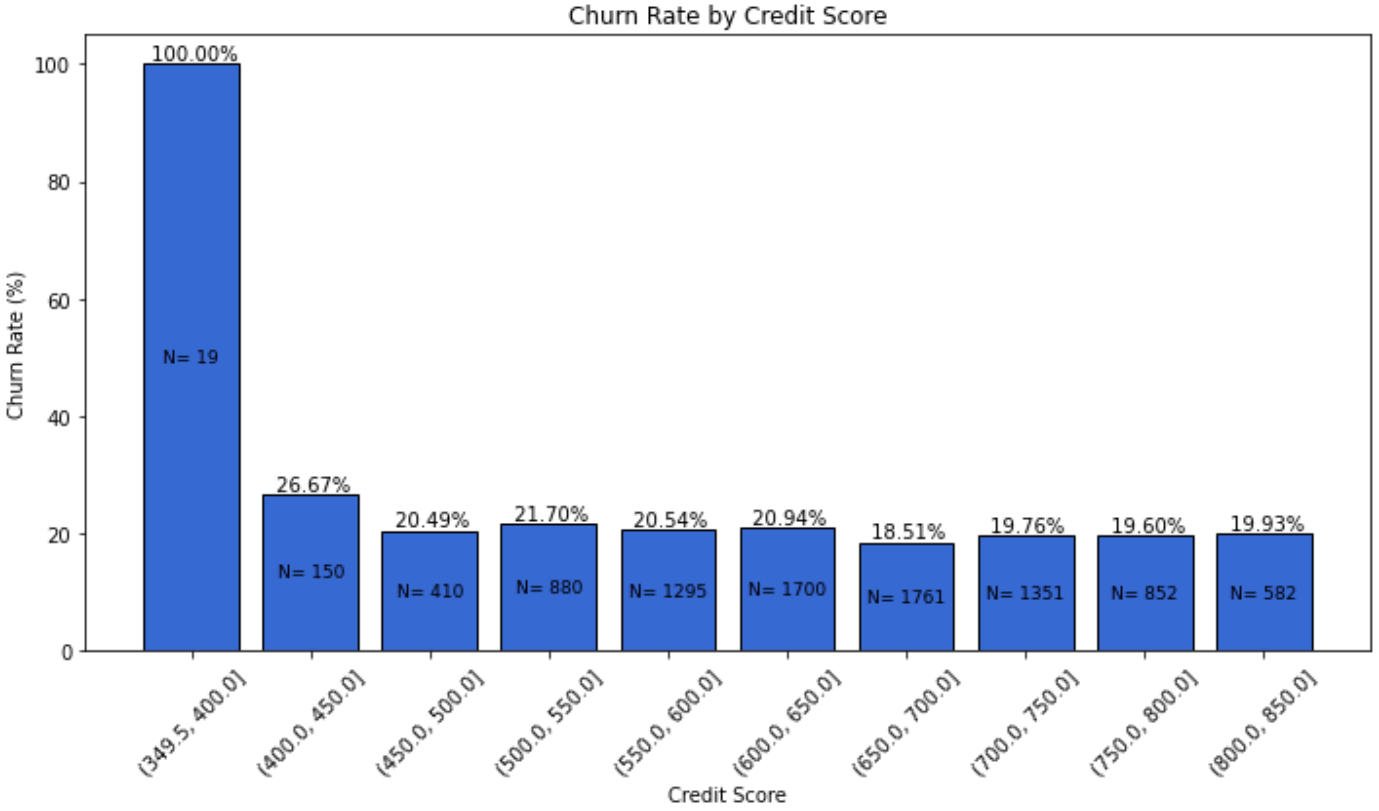
Appendix C. Bar Graph of Churn Rate by Gender



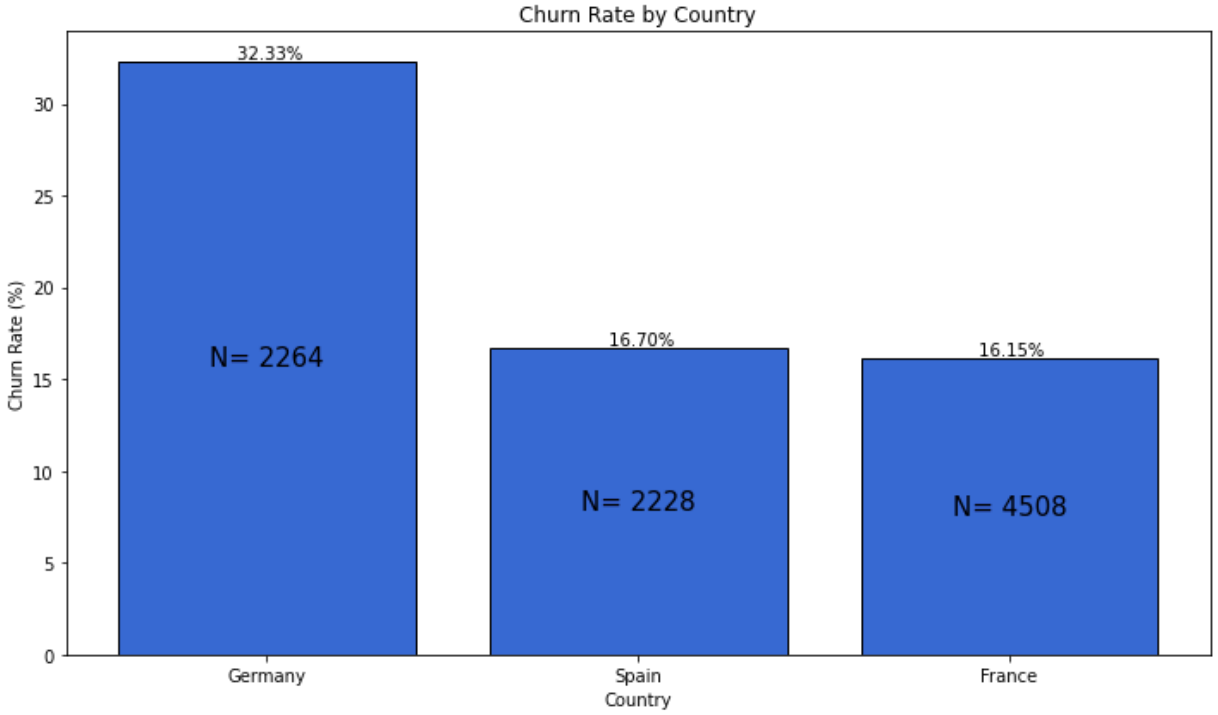
Appendix D. Bar Graph of Churn Rate by Age Groups



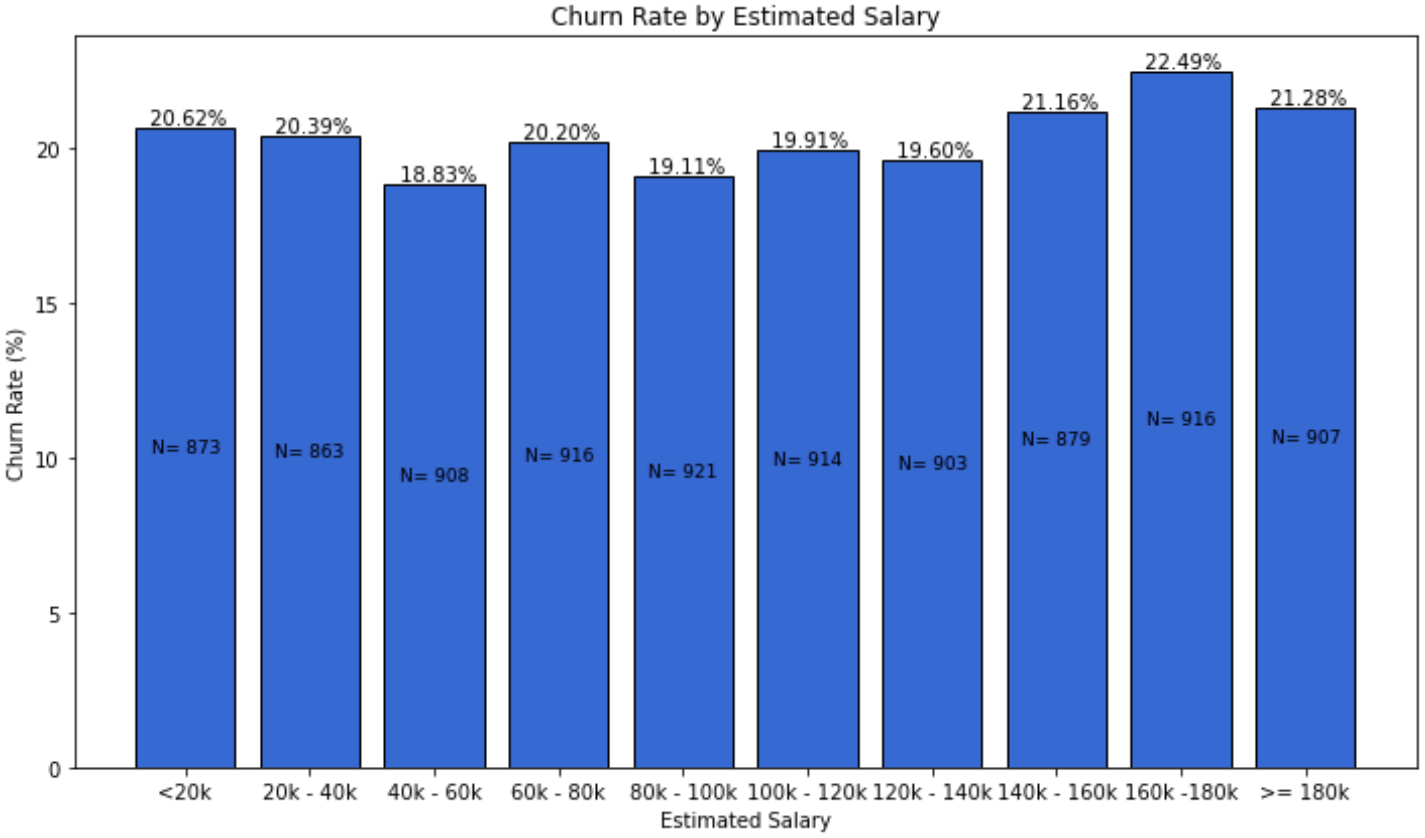
Appendix E. Bar Graph of Churn Rate by Credit Score



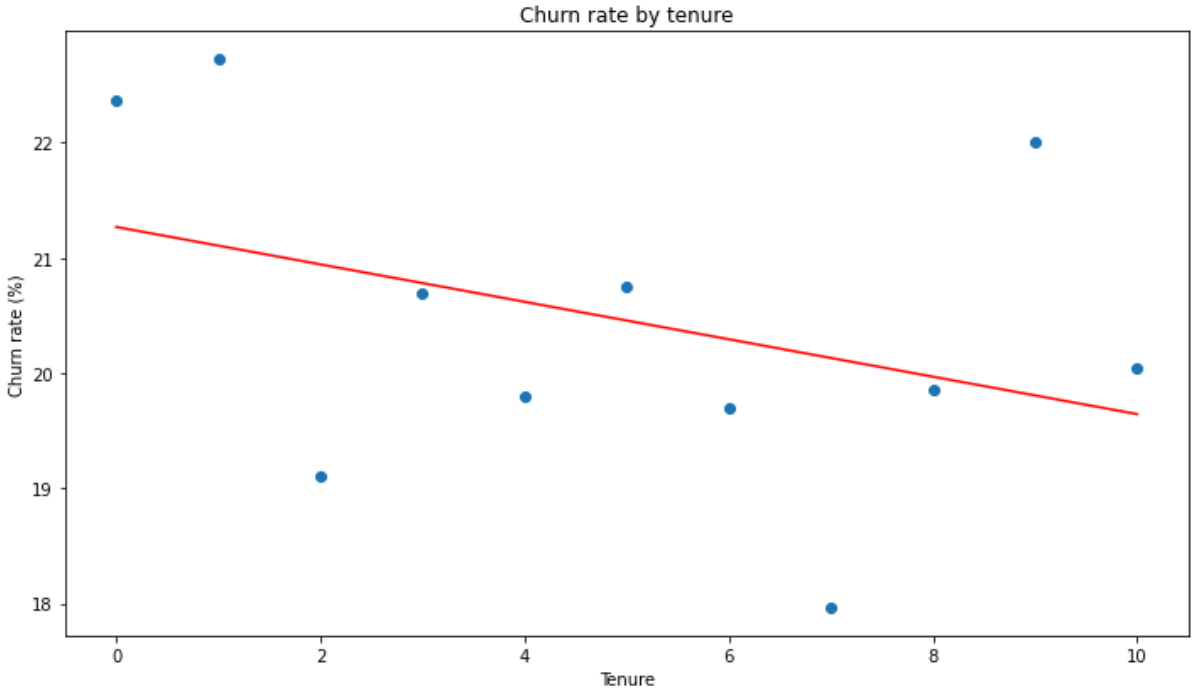
Appendix F. Bar Graph of Churn Rate by Country



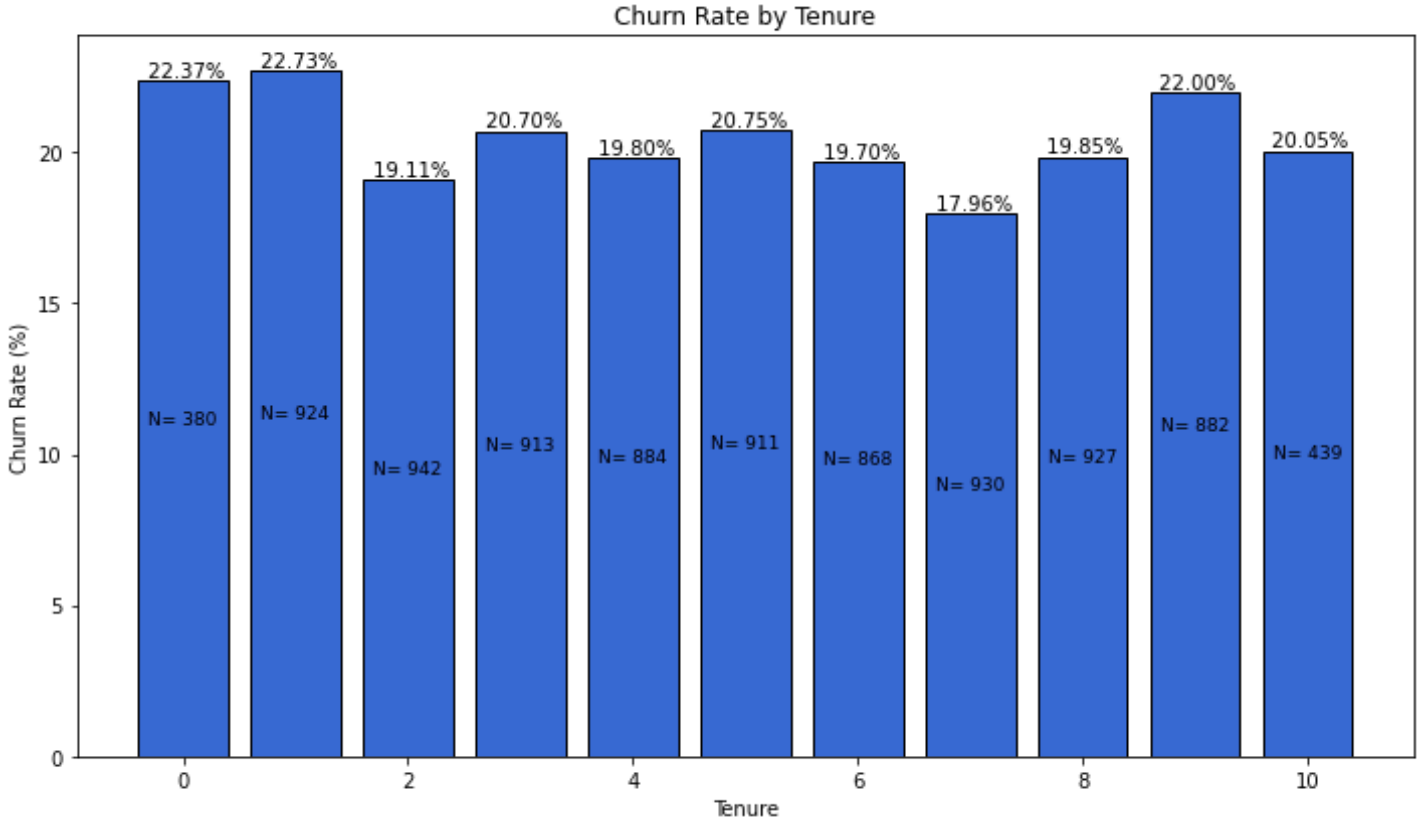
Appendix G. Bar Graph of Churn Rate by Estimated Salary



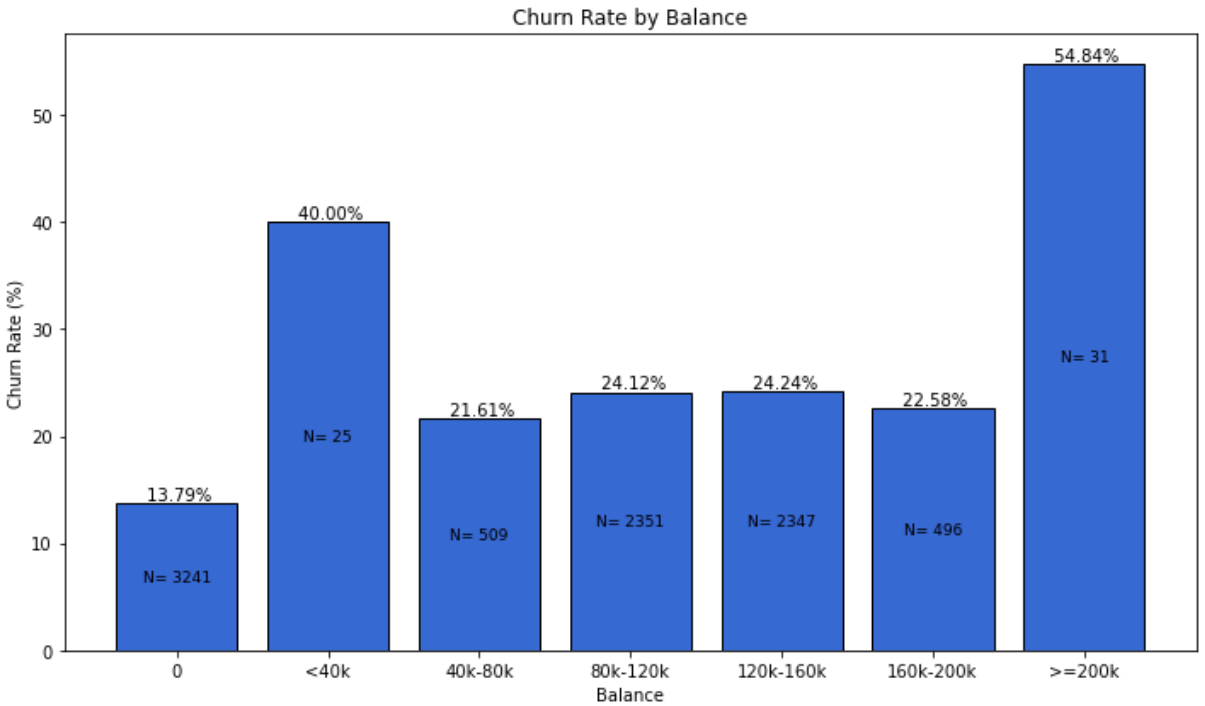
Appendix H. Scatter Plot of Churn Rate by Tenure



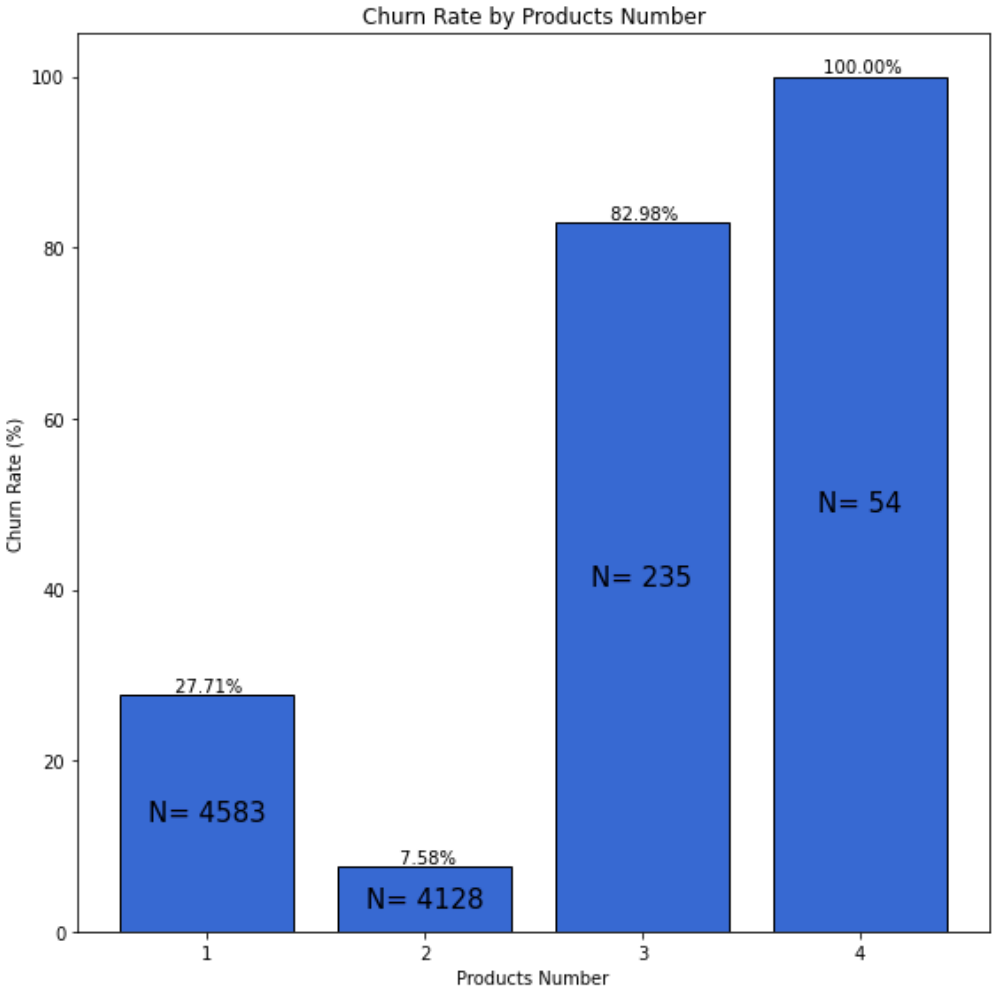
Appendix I. Bar Graph of Churn Rate by Tenure



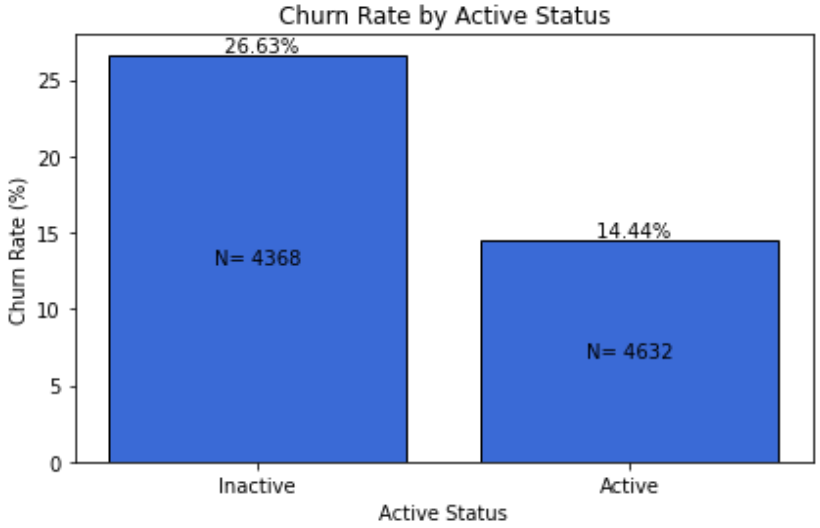
Appendix J. Bar Graph of Churn Rate by Balance



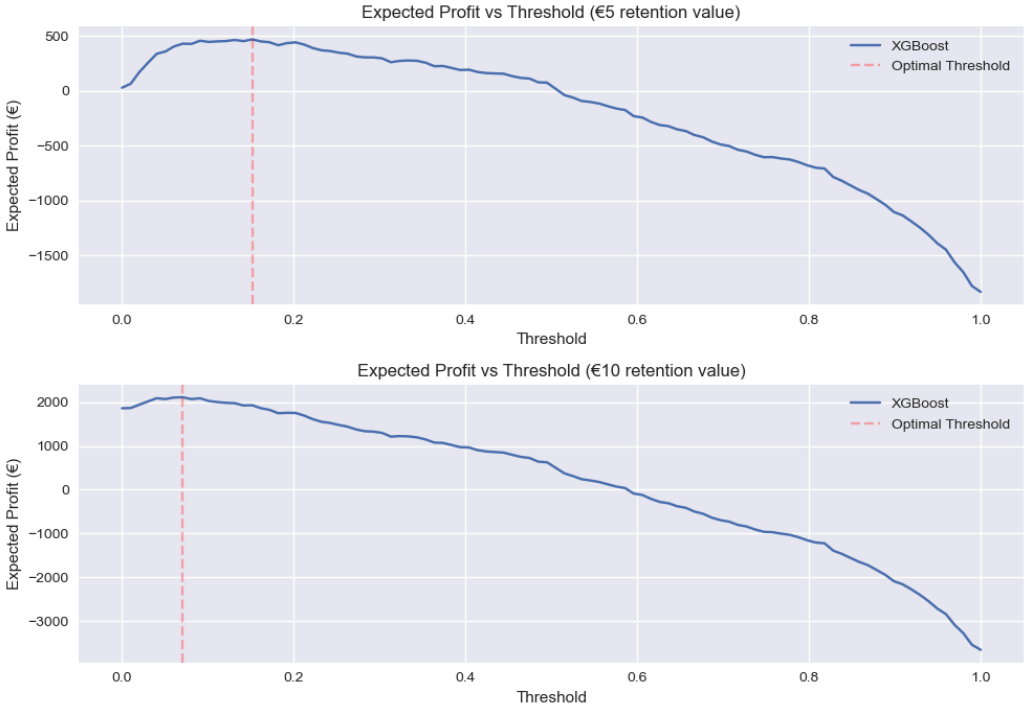
Appendix K. Bar Graph of Churn Rate by Numbers of Product



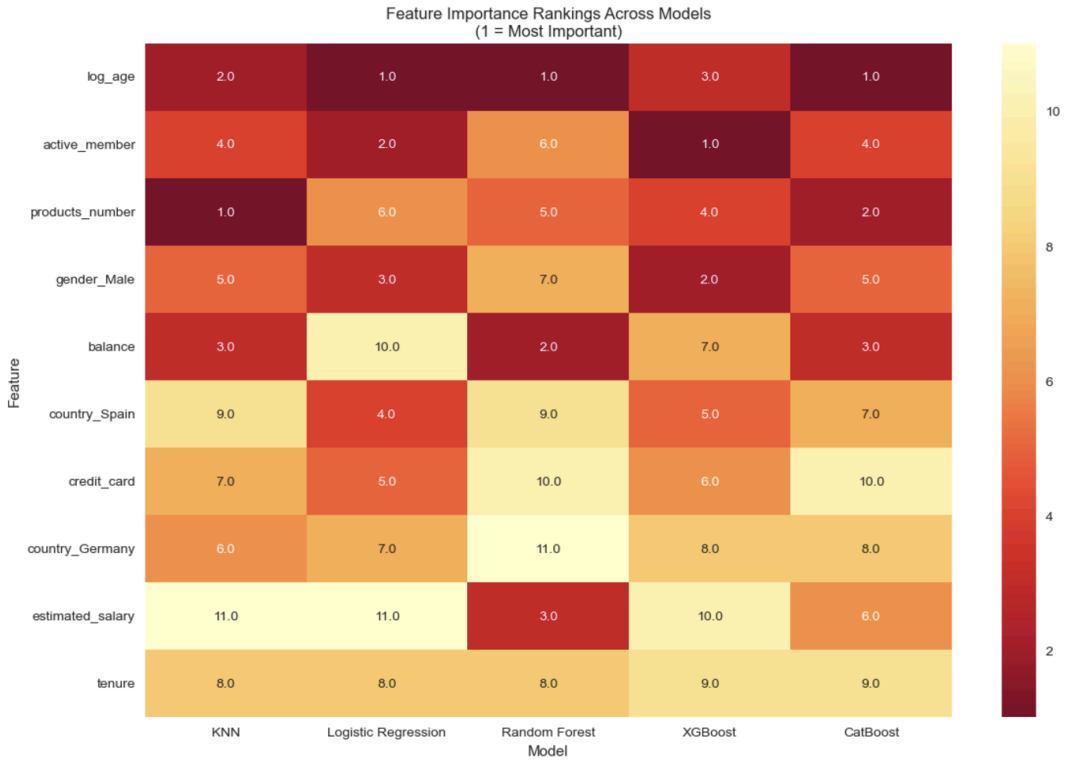
Appendix L. Bar Graph of Churn Rate by Active Status



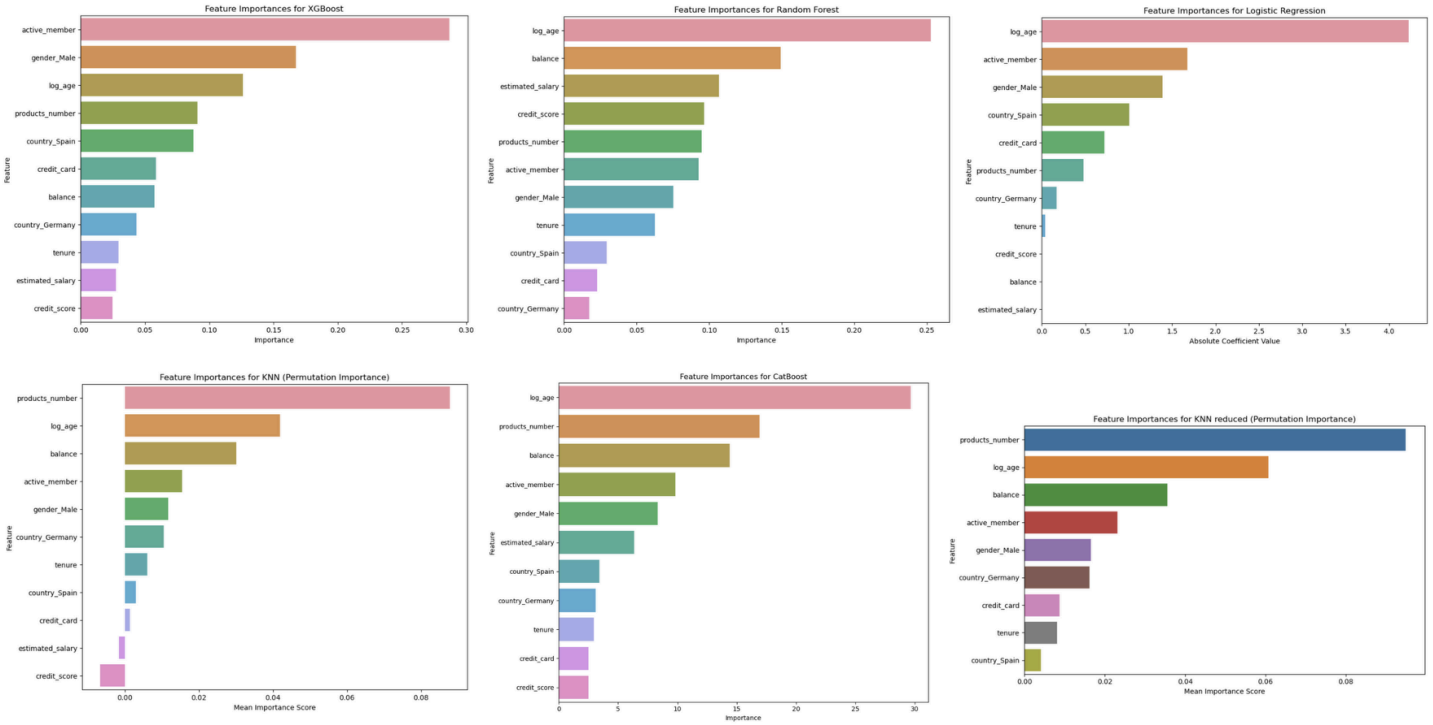
Appendix M. Line Graphs of Expected Profit and Threshold in Two Scenarios (€5 Retention Value and €10 Retention Value)



Appendix N. Heat Map of Ranked Feature Importance among Five Models



Appendix O. Comparison of Ranked Feature Importances among Five Models



Appendix P. Proportion Calculations for Churn Prediction at €10 Retention Value

The calculations are the same as the confusion matrix with €10 retention value:

- True Negatives (TN): $TN\ Proportion = \frac{TN}{Total} = \frac{536}{1800} \approx 0.2978$
- False Positives (FP): $FP\ Proportion = \frac{FP}{Total} = \frac{898}{1800} \approx 0.4989$
- False Negatives (FN): $FN\ Proportion = \frac{FN}{Total} = \frac{15}{1800} \approx 0.0083$
- True Positives (TP): $TP\ Proportion = \frac{TP}{Total} = \frac{351}{1800} \approx 0.1950$

Then, as the dataset has 1,000 observations, we multiplied each proportion by 1,000:

- True Negatives (TN):
  $Estimated\ TN = TN\ Proportion \times 1,000 = 0.2978 \times 1,000 \approx 298$
- False Positives (FP):
  $Estimated\ FP = FP\ Proportion \times 1,000 = 0.4989 \times 1,000 \approx 499$
- False Negatives (FN):
  $Estimated\ FN = FN\ Proportion \times 1,000 = 0.0083 \times 1,000 \approx 8$
- True Positives (TP):
  $Estimated\ TP = TP\ Proportion \times 1,000 = 0.1950 \times 1,000 \approx 195$

Appendix Q. Overall Churn Rate