Credit Card churn prediction

Quoc Frejter 430201, Katarzyna Dybcio 430189

Problem description

- **Background**
  The business manager, that is taking care of consumer credit card portfolio, is concerned about the increasing number of customers leaving their credit card services.
- **Problem**
  With only 16.07% of customers exhibiting churn, predicting churn is challenging due to class imbalance.
- **Aim of the project**
  Develop a predictive model to identify potential churners, empowering the bank to proactively engage and retain customers through personalized services.

Dataset

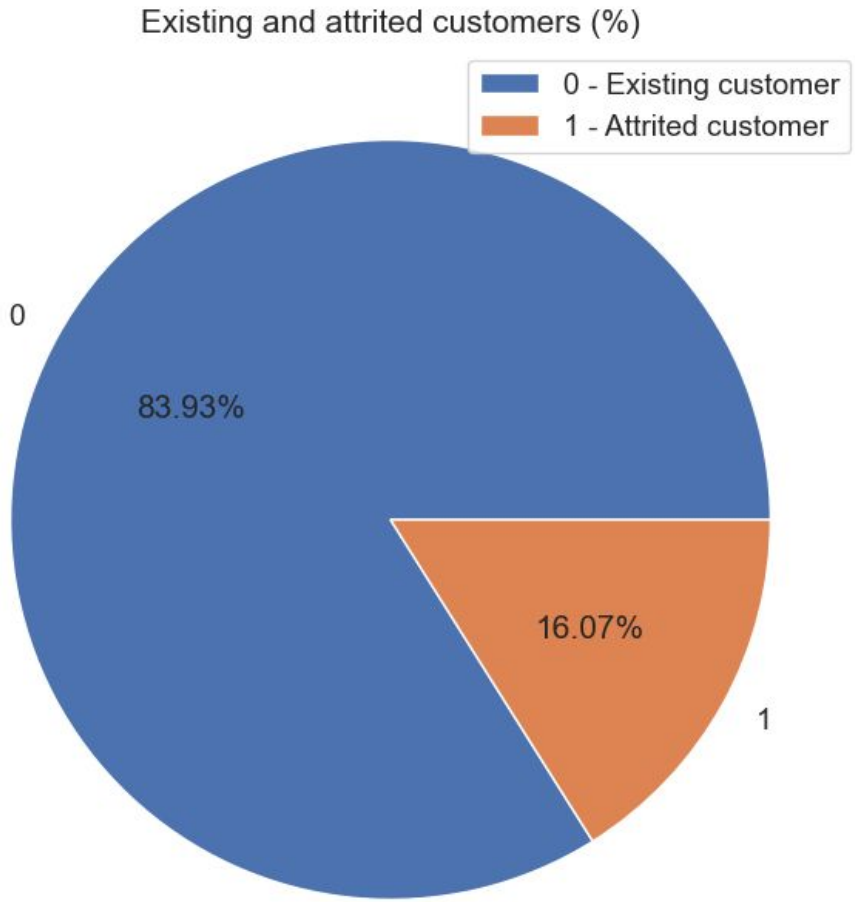
- **Target Variable: Attrition_customer**

  Internal event (customer activity) variable - if the account is closed then 1 else 0

- **Independent Variables: 20 features**

  'CLIENTNUM', 'Attrition_Flag', 'Customer_Age', 'Gender', 'Dependent_count', 'Education_Level', 'Marital_Status', 'Income_Category', 'Card_Category', 'Months_on_book', 'Total_Relationship_Count', 'Months_Inactive_12_mon', 'Contacts_Count_12_mon', 'Credit_Limit', 'Total_Revolving_Bal', 'Avg_Open_To_Buy', 'Total_Amt_Chng_Q4_Q1', 'Total_Trans_Amt', 'Total_Trans_Ct', 'Total_Ct_Chng_Q4_Q1', 'Avg_Utilization_Ratio'

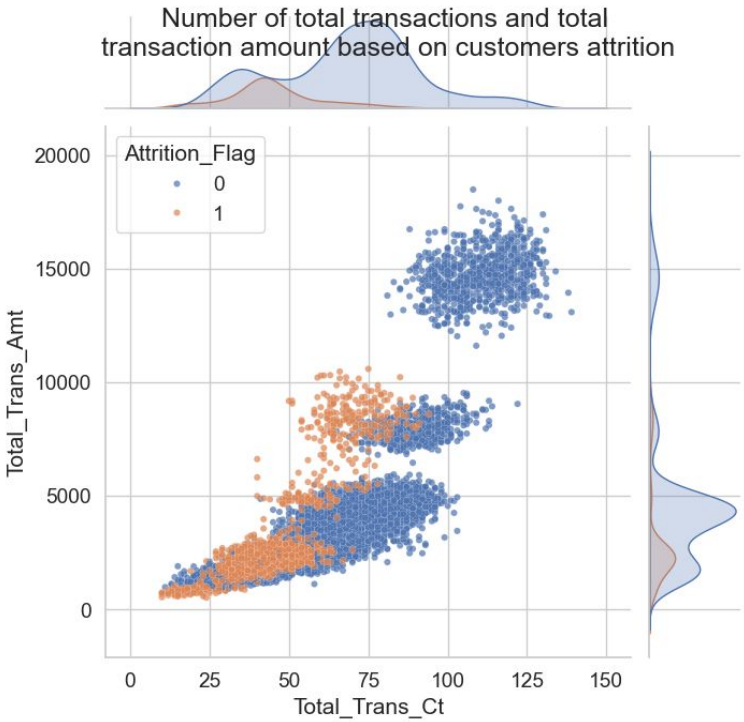
- **10127 observations**

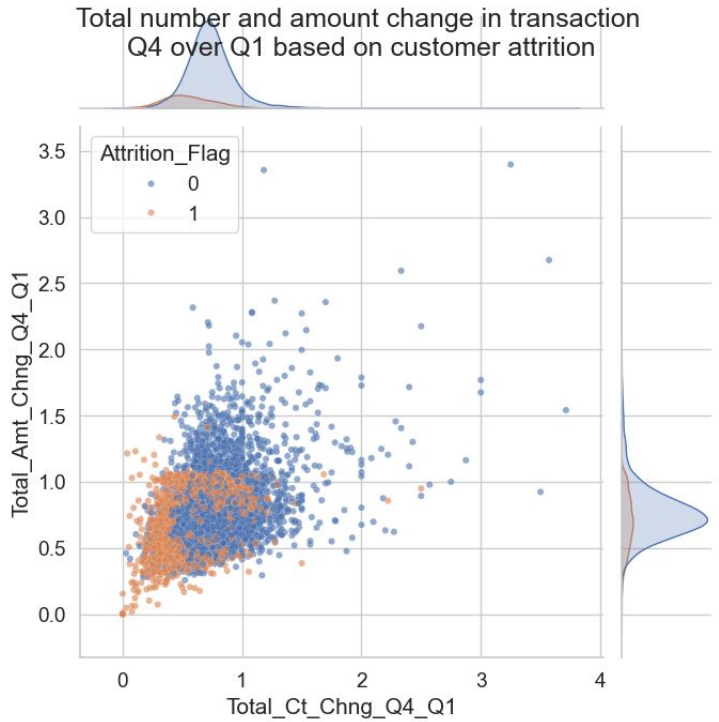


Target Variable

- Highly imbalanced dataset

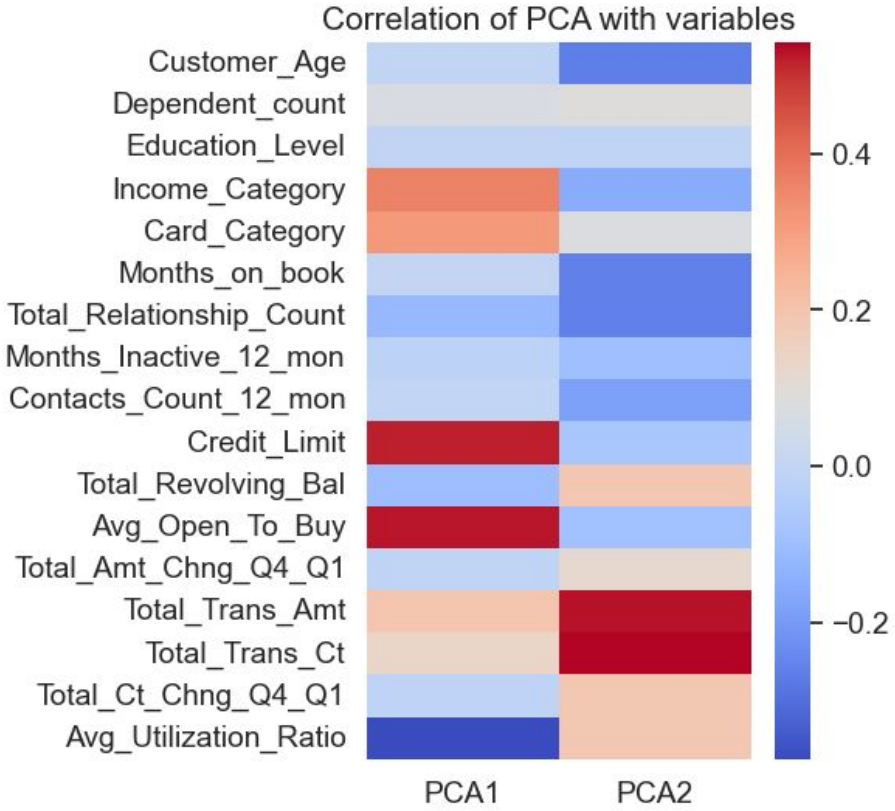
Attrition_Flag	Count	Percent of all
0	8500	84%
1	1627	16%





Customers who have opted out of the bank's services have significantly fewer completed transactions, and these transactions involve a lower average amount of funds.

K-means, PCA - clustering of customers



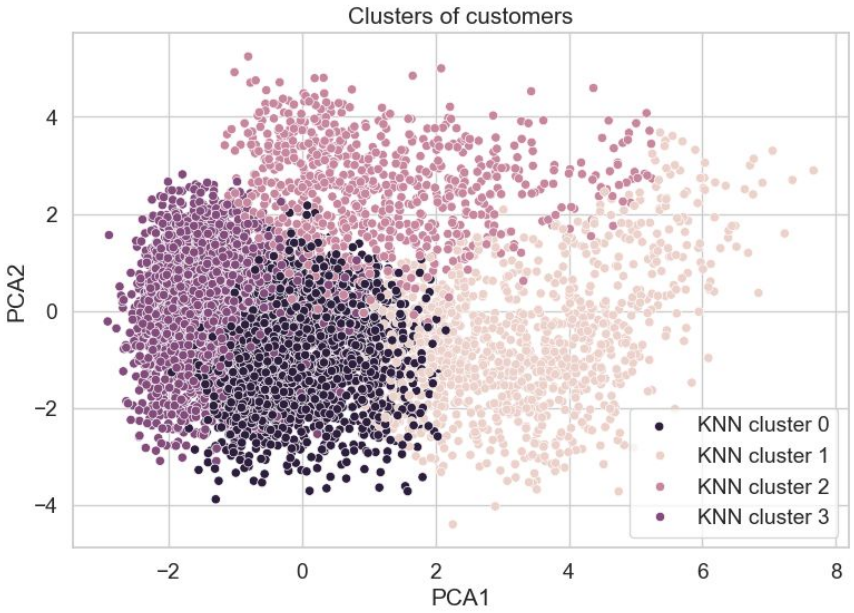
High PCA1

- High income, "premium" card, high credit limit but low utilization
- Wealthy customers who exhibit a low utilization ratio.

High PCA2

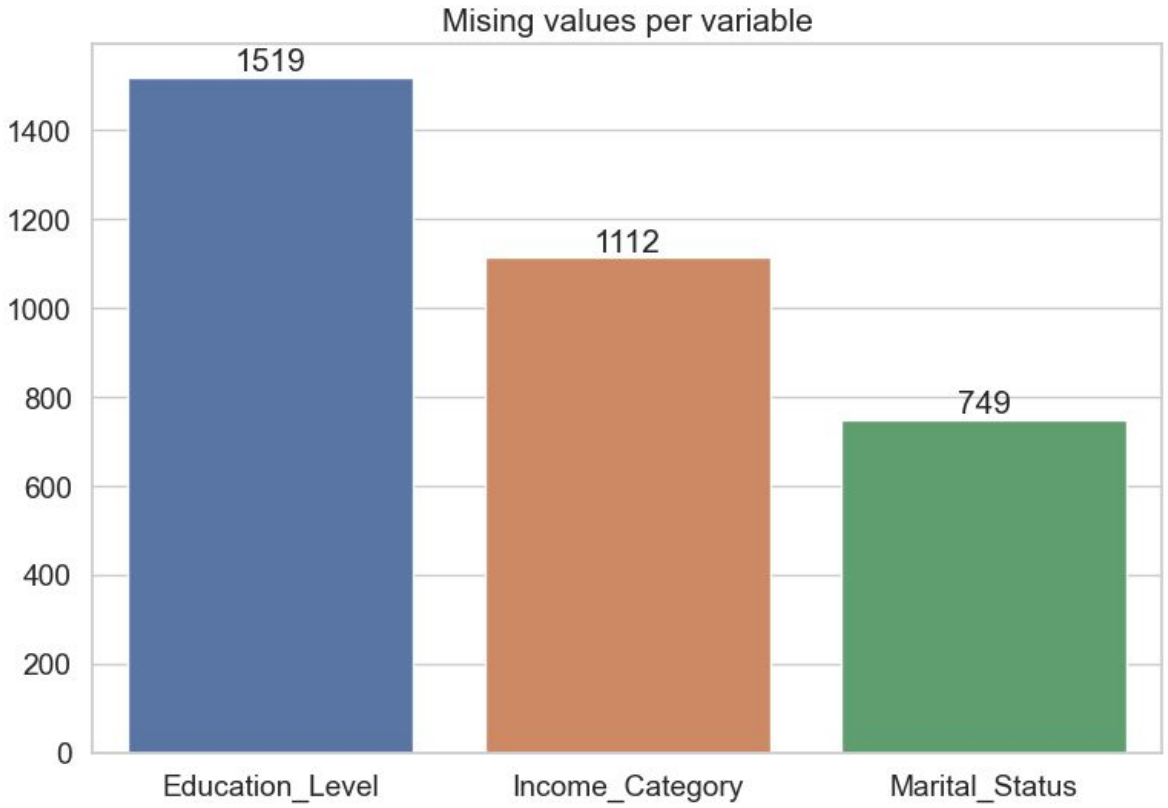
- Young, low income, low number of month on book, low number of products, high transaction count and amount
- Young/new customers who actively use their cards

Plot of clusters



- On this plot we can see, that employing K-Nearest Neighbors (KNN) and Principal Component Analysis (PCA) for modeling the attrition flag could pose challenges. We can see that customers are quite mixed up and our algorithm would have problems dividing the group.

Missing values



- We have missing values only for Education level, Marital status and income category
- Imputation methods
  Education_level - imputed using KNN
  Income_category - imputed using KNN
  Marital_Status - Filled missing values with most frequent

Steps for success

Train test split (Stratified 80/20 split)

imputation (KNN, mean)

Feature engineering and selection

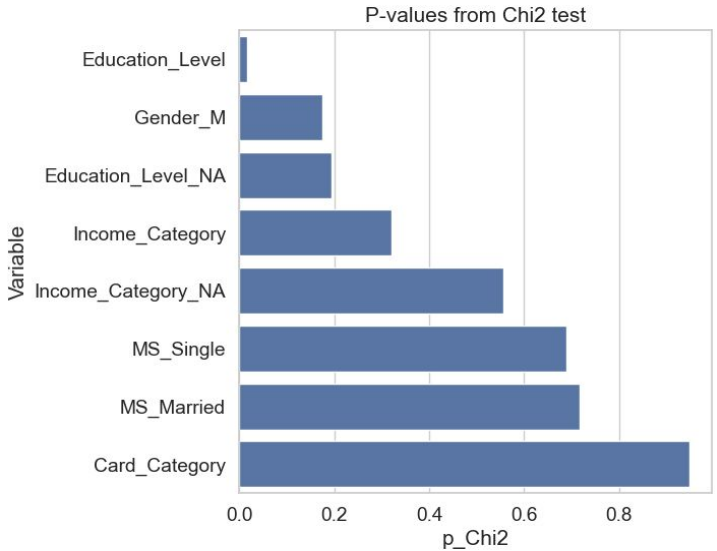
Oversampling (SMOTE)

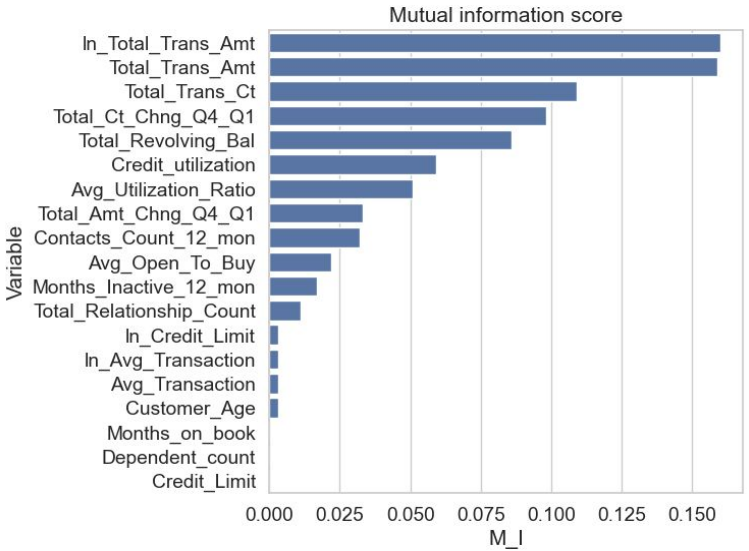
Logit
SVM
Random Forest
KNN

Feature engineering and selection

Transformations:

- Categorical variables: one-hot encoding, ordinal encoder
- Numerical variables: log transformation, standard scaler
- Interactions: average transaction, account utilization





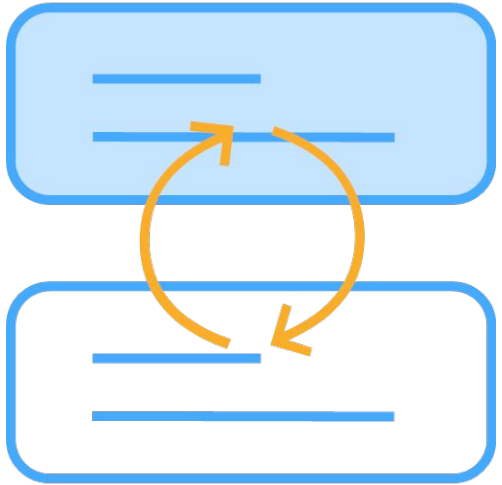
Hyperparameter tuning

Method: **Grid search**

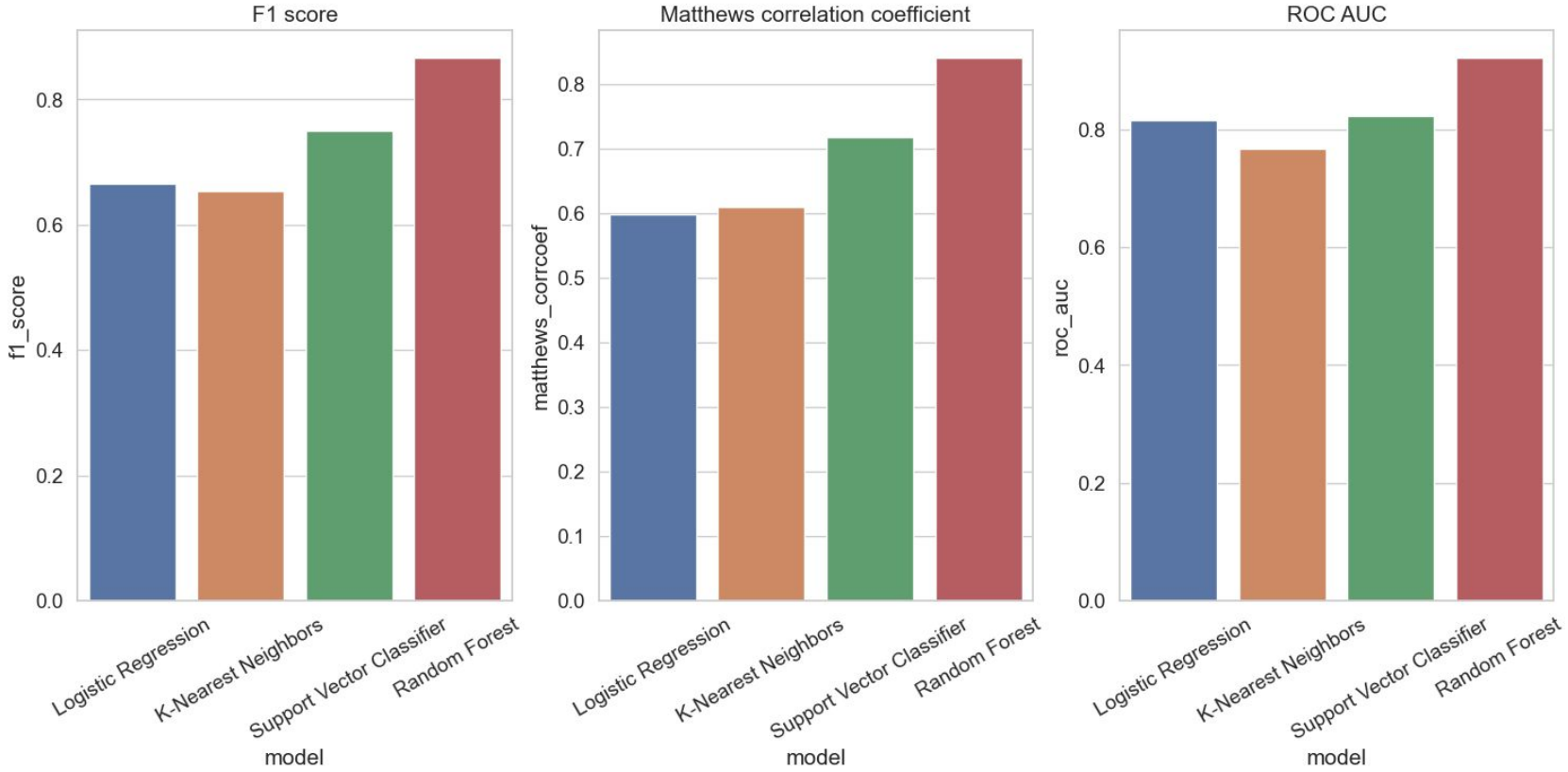
Cross validation: **Stratified K-fold** (8 splits)

Scoring measures:

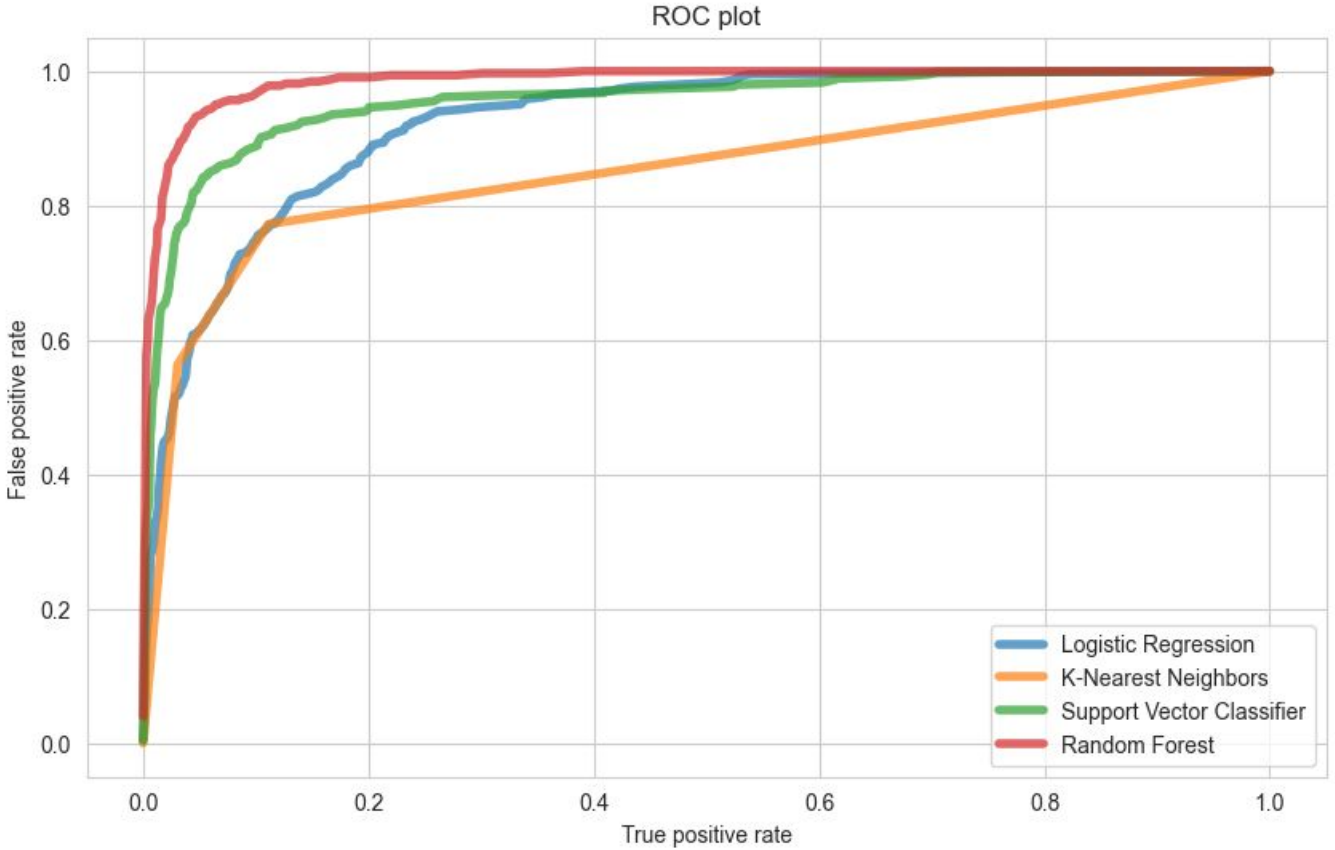
- **F1 score**
- **Matthews correlation coefficient**
- ROC AUC (additional)



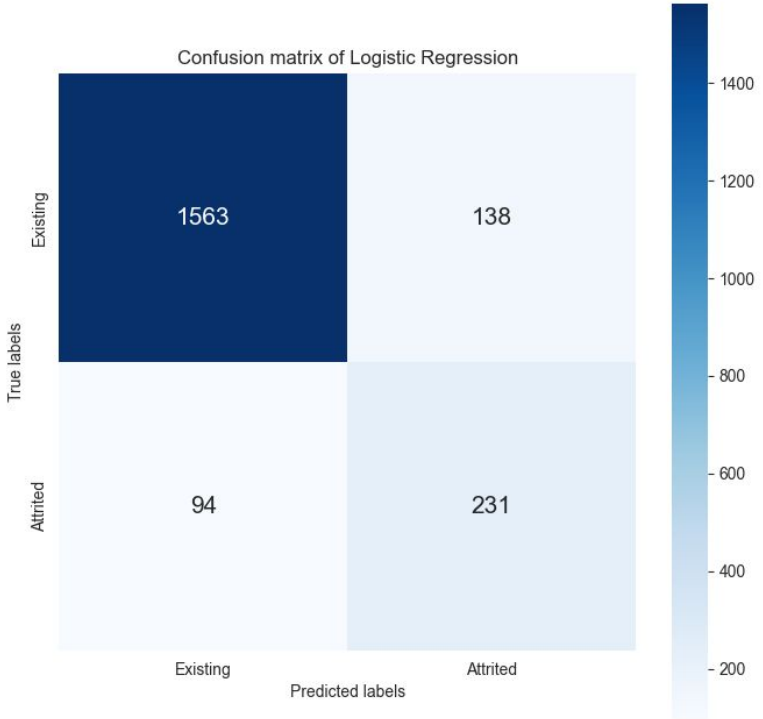
Models performance

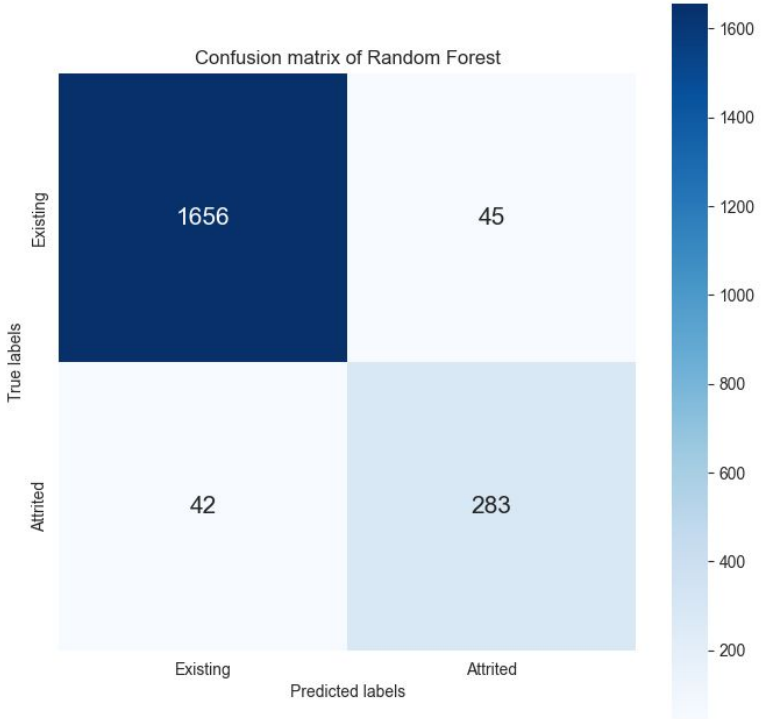


Models performance



Logistic Regression vs Random Forest





Summary

- Additionally, Random Forest correctly identified the highest number of customers(238) who actually churned from the bank services
- In almost every model, the best results were achieved with data augmented using the SMOTE method. That is because our data was highly imbalanced.
- For the Random Forest, it was beneficial to reduce the number of variables when training the model what was confirmed by the test statistics.
- Based on this, we can conclude that the Random Forest model is the most effective method for our database.

Thank you!