Churn Modelling for Bank

Aslihan Celik

MS Computational Science & Engineering

GT ID: 903356676

Objectives

- Predicting the churn for bank as a binary classification task
  - Churn: If a customer will leave the bank or stay
- Understand the relationship between the customers attributes and their potential to leave the bank
- Help banks to better understand their customers and maximize the number of customers
- Help banks to take early action to prevent their customers from leaving

Dataset

- There are 12 features including
  - Age
  - credit score
  - tenure
  - balance
  - number of products
  - estimated salary
  - customer ID
  - Surname
  - Geography
  - Gender
  - has credit card
  - is active member.

- “Exited “ is the binary variable as 1 indicating that the customer is left and 0 as customer stayed.

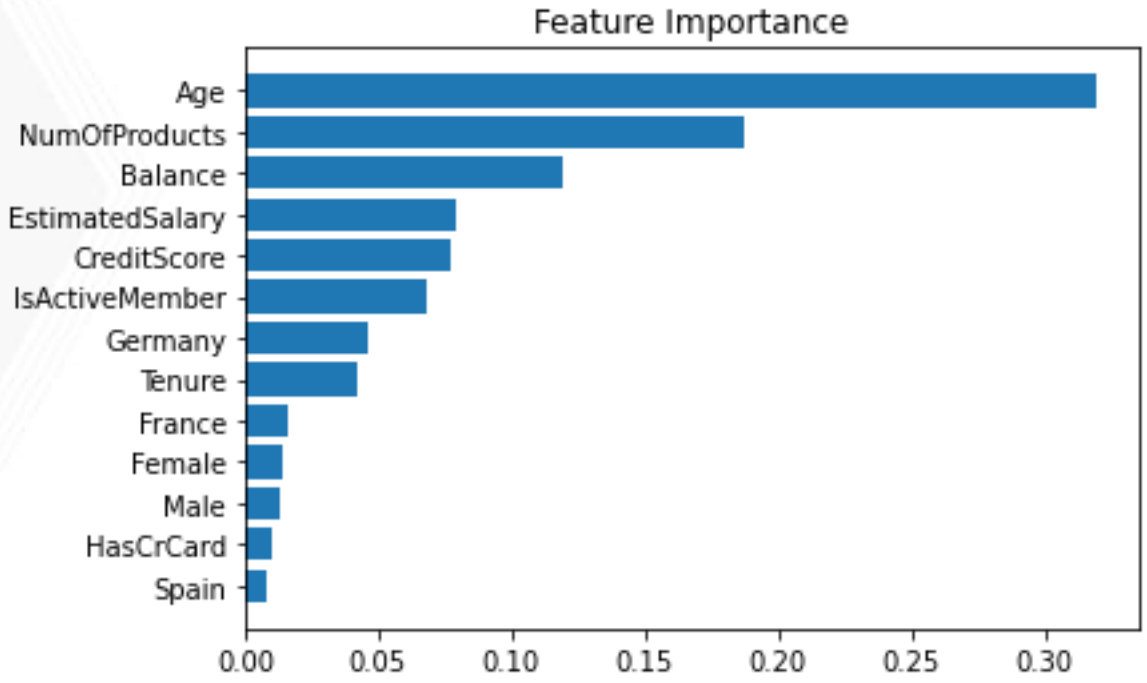
- 10000 data points in the dataset.

Preprocessing

- Dataset balanced as both classes having 2037 instances each.
- Standardization is applied.
- One hot encoding for categorical variables

Exploratory Data Analysis

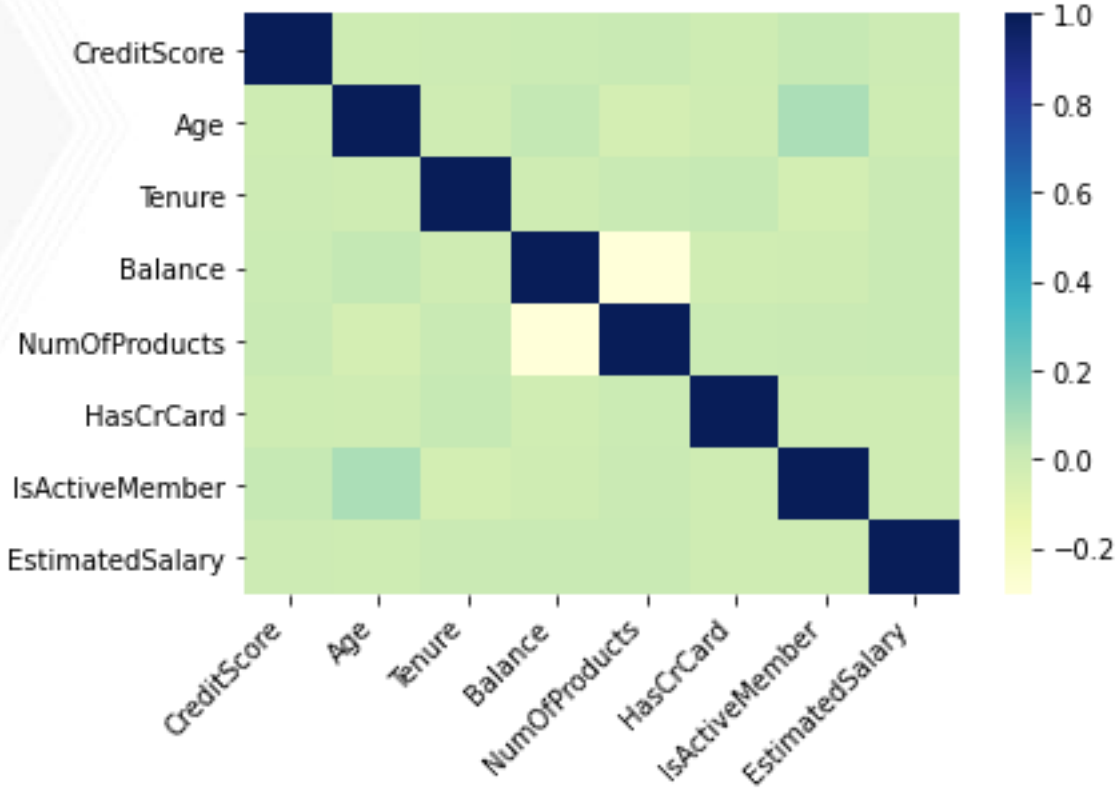
- Age is the most important feature in predicting the churn for a bank followed by the number of products the customer uses as the second most important



Georgia Tech
CREATING THE NEXT

Exploratory Data Analysis

- There is no multicollinearity among independent variables.
- Little linear correlation between variables Age and IsActiveMember.
- There is also low correlation between variables NumOfProducts and Balance.



Georgia Tech
CREATING THE NEXT

Modeling

- 4 different models are implemented that are:
  - SVM
  - Logistic Regression
  - Decision Tree
  - Random Forest

- Hyperparameter Tuning is performed for each model with 5-fold cross validation in order to optimize the estimator parameters.

Model Selection

- Since the dataset is balanced after the preprocessing accuracy can be compared.

- As it matches our objective we choose to look at F1 score in order to select the model.

    - For the objective of this project we are interested in both precision and recall. F1 score as it is the weighted average of precision and recall.

        - Precision: of all the customers that labeled as exited the bank, how many actually exited.
        - Recall: of all customers truly exited the bank how many did we actually labeled

- According to the previous slide Decision Tree is selected as the best performing model for churn prediction for a bank.

Table 7: Scores for Balanced Dataset

Model	Accuracy	F1 Score
SVM	0.78	0.77
LR	0.70	0.70
DT	0.80	0.80
RF	0.80	0.79

Conclusions

- Age is the most important feature to predict the churn (if the customer will leave the bank or stay)
- Models that have nonlinear decision boundaries can outperform the models with linear decision boundary in the case that dataset might not be linearly seperable
- For the binary classification task of predicting the churn, Decision Tree resulted in best performance