**Final Project Report**

**09.06.2021**

**Eren Özbek**

I am studying Business Administration and Economics and I am doing my internship in one of the biggest banks in Turkey. Therefore, for my final project I have picked one of the most common real-world problem that is challenging for many businesses which is churn prediction. For banks it is critical to identify users who are at risk to churn before they leave for their revenue. Even slight variations in the churn rates can affect their profits. I have used a the [dataset](https://www.kaggle.com/adammaus/predicting-churn-for-bank-customers) specific for the prediction of the bank customer churn.

In this project I aimed to identify and create features which have an effect on customer churn and build a model for predicting a customer is going to churn or not. So that the output can be used to create user specific incentives or discounts to circumvent their leave.

The dataset has 13 columns:

* Customer ID : personal identification
* Surname: personal identification
* Credit Score: Numerical Value
* Geography: Categorical Value
* Gender: Categorical Value
* Age: Numerical Value
* Tenure: Numerical Value
* Balance: Numerical Value
* Number of Products: Numerical Value
* HasCrCard: 0,1
* IsActiveMember: 0,1
* Estimated Salary: Numerical Value
* Exited: 0,1

I have applied Exploratory Data Analysis to understand the relationship between churn and the provided information. From the Exited column I have counted churn ratio as 20%.

Chart, pie chart

Description automatically generated

Then I have dropped the irrelevant columns and preprocess the data for the modeling. In the pre-processing I have used one hot encoding to deal with categorical data. I have created features which I think likely be an indicator for the churn:

* SalaryAgeRatio : indicator of a customer's socio-economic status
* CreditScoreSalaryRatio: spending behavior and willingness to save
* BalanceSalaryRatio: amount of customer’s willingness to save
* TenureBalanceRatio: shows whether a customer leave after the credit payment or stay within the bank system

A picture containing graphical user interface

Description automatically generated

From the Exploratory Data analysis, it can be seen that customers who churn are older. This can be an actionable idea for a business strategy like making services for older people. For the tenure, people who churn has a wide distribution showing that they are either spending less time or very long time with bank compared to retained people.

I have tested several machine learning models: Primal Logistic Regression, Logistic Regression with Polynomial Degree 2, Random Forest Classifier, and SVM. Since the churn set is small, to correctly identify this class F1 score used as the model performance criteria. When there is a population with 20% churn rate (true population) then predicting all population as non-churn will give still high accuracy with 80%. This is penalized in F1 score which is defined as the harmonic mean of precision and recall where precision is computing the proportion of positive class identifications being the correct ones and Recall is computing the proportion of actual positive class samples being identified correctly.

The best F1 score for predicting the churn of the models I have tried is Logistic Regression with Polynomial Degree 2, which has F1 score 0.58. The highest accuracy score is Random Forest Classifier with 0.87. As it can be seen from the accuracy metrics, they are high but the F1 score that is important for predicting the churn population should be improved.