Capstone2 Bank Churners Analysis

In this project, supervised models will be created to predict which users are more likely to churn and stop using bank. The most important factors that prevent users from churning and keep them retained in the system will be identified.

The dataset (Churn-Dataset) contains information about service usage, customer plans, and the target variable which indicates whether a user churned or not. Dataset composed of 10127 observations and 23 columns.

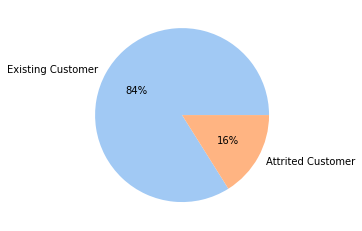


Figure 1: Attrition flag distribution

The target variable in the dataset is “attrition\_flag” column. For the target variable while the rate of existing customer is 83.9%, the rate of attrited customer is 16.1%.

In the dataset there are categorical columns 'gender', 'education level', 'marital status', 'income category', and 'card category'. The rate of target column almost is not change in the subcategories of gender, education level, marital status, income level and card category. This means that gender, education level, marital status, income level and card category have less effect on the target variable. It is concluded the same results after the model evaluations. It is interesting that, demographic data’s of customer has no effect on churn. Especially, income has no effect on churn.

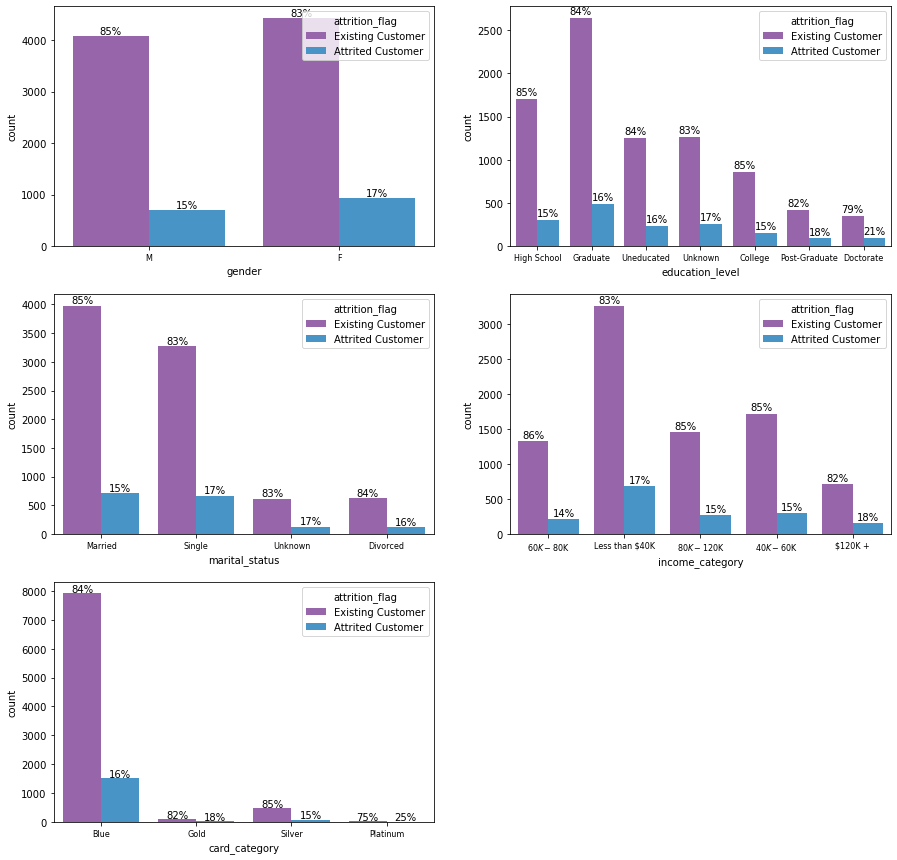


Figure 2: Categorical Column Counts

The features total\_trans\_ct, total\_trans\_amt, and total\_revolving\_bal are more influential on target column. While the average total transaction amount of attrited customers is 3095, existing customers is 4654. While the average total transaction count of attrited customers is 44.93, existing customers is 68.67. Also, total\_revolving\_bal is an effect on target column. While the average total revolving balance of attrited customers is 672.8, existing customers is 1256.6. The effect of these columns is also seen in the feature importance of models.

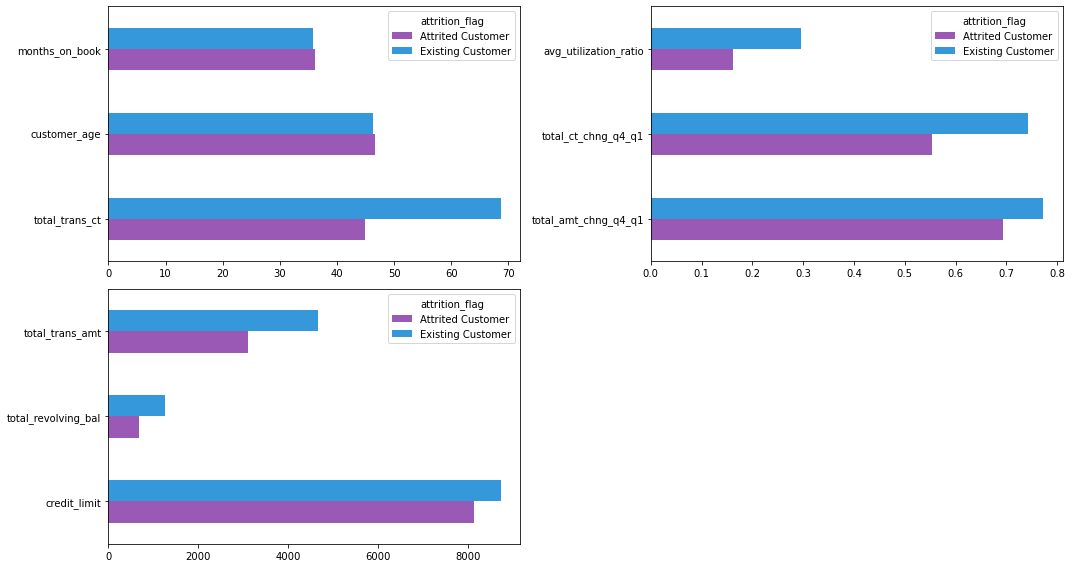
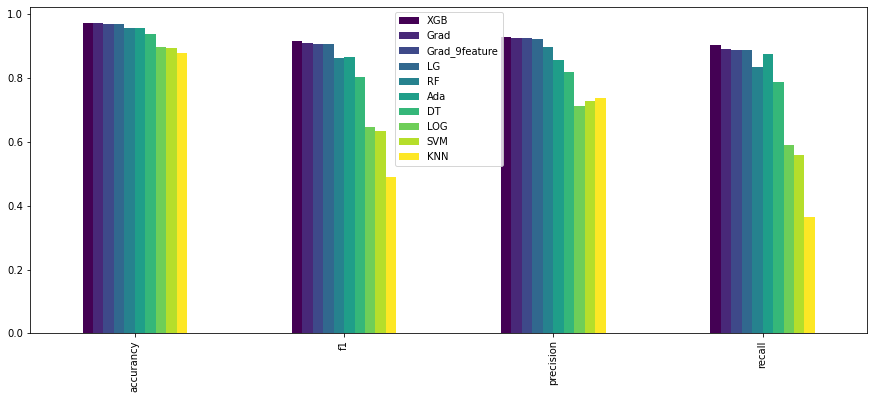
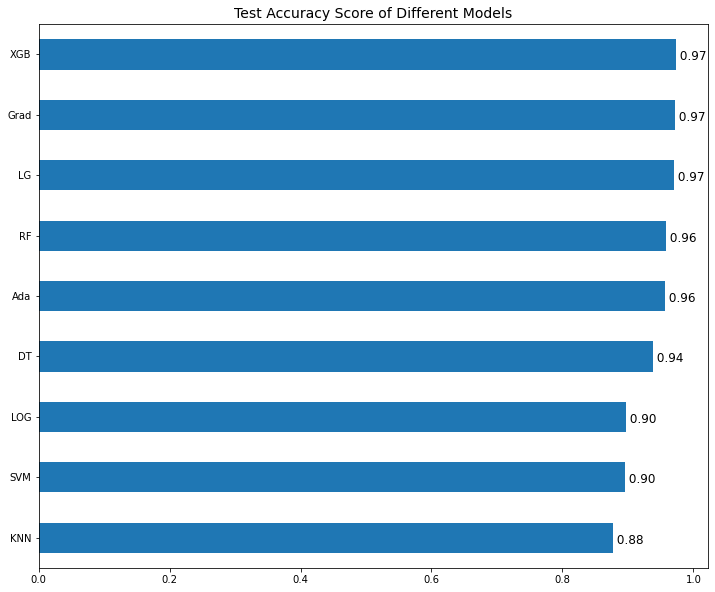
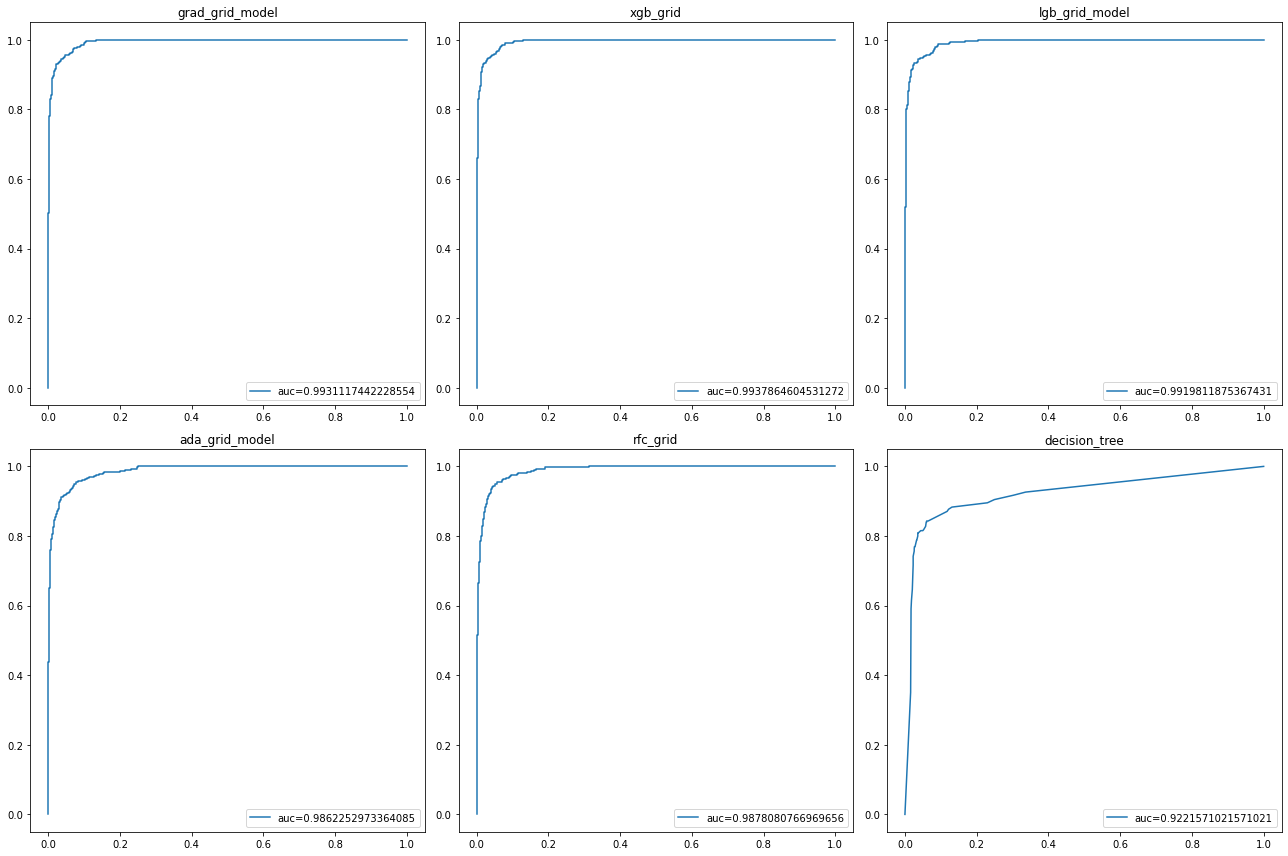


Figure 3: Means of Numeric Columns

9 model applied to the data. Three methods stand out among the methods. These methods are XGBoost, GradBoost, LightGBM. Although the scores of XGBoost is a bit higher than the other two, XGBoost is more close to overfitting. AUC score of GradBoost is also high and very close to 1. GradBoost can be chosen the best model. In the analysis of feature importance also, Gradboost uses less features in the model. It is also important.

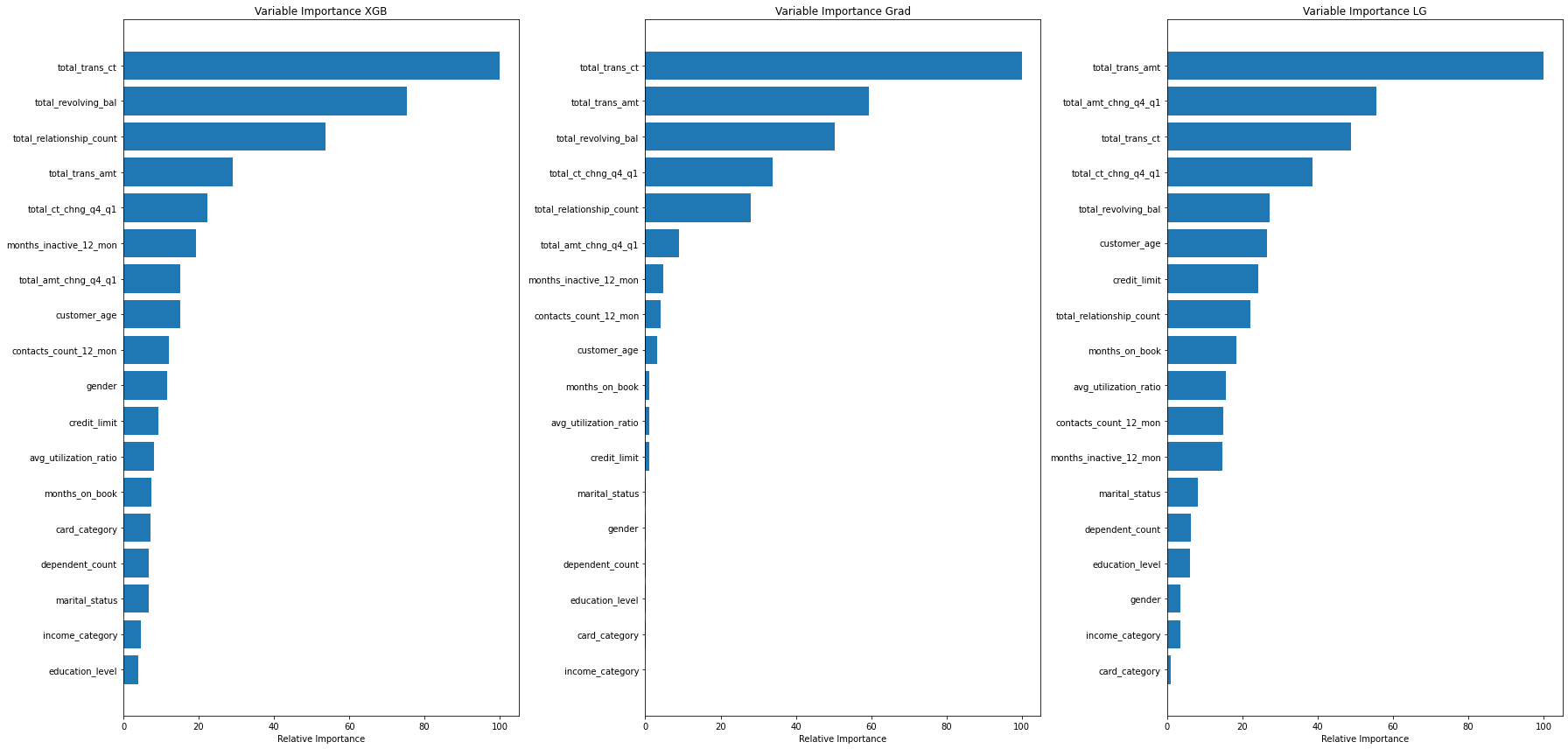






The most important features in the model are total\_trans\_ct, total\_trans\_amt, total\_revolving\_bal. Also the features total\_ct\_chng\_q4\_q1, total\_amt\_chng\_q4\_q1, total\_relationship\_count, months\_inactive\_12\_mon are important for the models. Customer age has a small effect on target. Categorical variables such as income, gender, educational level, marital status, card category has a very small effect on target.

* total\_trans\_ct, total\_trans\_amt, total\_revolving\_bal features of alarmed customers should be constantly checked. It can be determined alarm tresholds for these features.
* Special campaigns should be made for customers are more likely to churn.



In the gradboost model 9 features are most important. A gradientboost model with 9 features was fitted the data and had an accurancy score of 0.971 is very close to XGB and Gradboost (all features) scores. So this model can be chosen the best model.

A function was created to filter the clients are more likely to churn.