Credit Card Default Prediction

Subhash

July 2023

Problem Statement: This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the K-S chart to evaluate which customers will default on their credit card payments

Business Objective: The goal is to predict the likelihood of customers defaulting on their credit card payments in Taiwan. This prediction is essential for risk management purposes.

Significance: By accurately predicting payment default, businesses can proactively identify customers at risk and implement appropriate measures to minimize financial losses and improve customer retention.

Who Should Care?

Credit Card Companies

Commercial Banks

* Image source: Google image

Approach Overview

Data Cleaning

Understand and Clean

- Find information on undocumented columns values
- Clean data to get it ready for analysis

Data Exploration

Graphical and Statistical

- Exam data with visualization
- Verify findings with statistical tests

Predictive Modeling

Machine Learning

- Logistic Regression
- Random Forest
- SVM
- XGBoost

Data Overview

•Dataset Description: The project utilized a comprehensive dataset containing information on credit card customers. The dataset includes various attributes such as demographic details, payment history, credit limit, and bill amounts for multiple months.

The dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in Taiwan from April 2005 to September 2005. It consists of 30,000 rows and 25 columns. I observed that there are no duplicate values and misssing values/null values in the dataset. There are 24 Independent features and Default payment next month is the target variable. All independent features are of Int data type. There are categorical variables like Sex, Education and Marriage. Remaining all independent variables are Numerical.

•Imbalanced Data: One important characteristic of the dataset is its class imbalance, with a majority of non-defaulters and a minority of defaulters. This required to address the imbalance during the modeling process.

There are 25 variables:
ID: ID of each client
LIMIT_BAL: Amount of given credit
SEX: Gender 1=male, 2=female
EDUCATION: 1=graduate school, 2=university, 3=high school, 0, 4, 5, 6=others)
MARRIAGE: Marital status 1=married, 2=single, 3=divorce, 0=others
AGE: Age in years
History of Past Payments
PAY_0: Repayment status in September, 2005
-2: No consumption -1: Paid in full 0: The use of revolving credit 1 = payment delay for one month 2 = payment delay for two months 8 = payment delay for eight months 9 = payment delay for nine months and above.
PAY_2: Repayment status in August, 2005 (scale same as above)
PAY_3: Repayment status in July, 2005 (scale same as above)
PAY_4: Repayment status in June, 2005 (scale same as above)
PAY_5: Repayment status in May, 2005 (scale same as above)
PAY_6: Repayment status in April, 2005 (scale same as above)

Amount of bill Statement
BILL_AMT1: Amount of bill statement in September, 2005
BILL_AMT2: Amount of bill statement in August, 2005
BILL_AMT3: Amount of bill statement in July, 2005
BILL_AMT4: Amount of bill statement in June, 2005
BILL_AMT5: Amount of bill statement in May, 2005
BILL_AMT6: Amount of bill statement in April, 2005

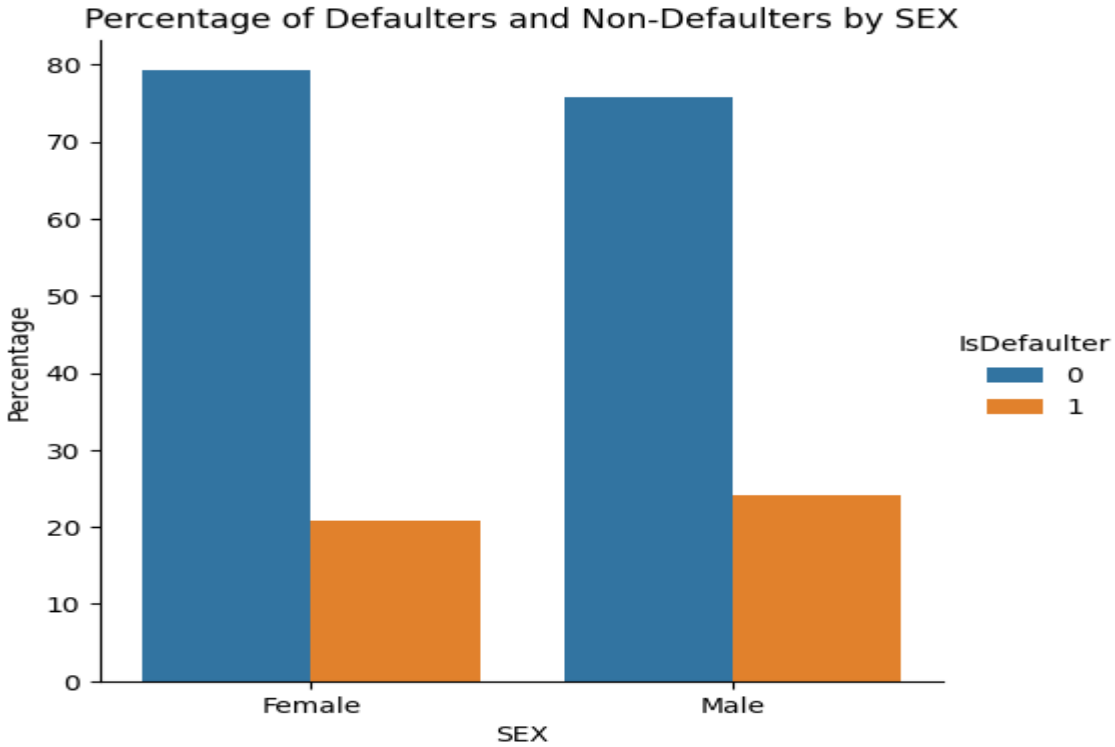
Amount of Previous Payments -Previous amount Paid
PAY_AMT1: Amount of previous payment in September, 2005
PAY_AMT2: Amount of previous payment in August, 2005
PAY_AMT3: Amount of previous payment in July, 2005
PAY_AMT4: Amount of previous payment in June, 2005
PAY_AMT5: Amount of previous payment in May, 2005
PAY_AMT6: Amount of previous payment in April, 2005

default payment next month: Default payment
1=yes, 0=no

Exploratory Data Analysis

What demographic factors impact payment default risk?

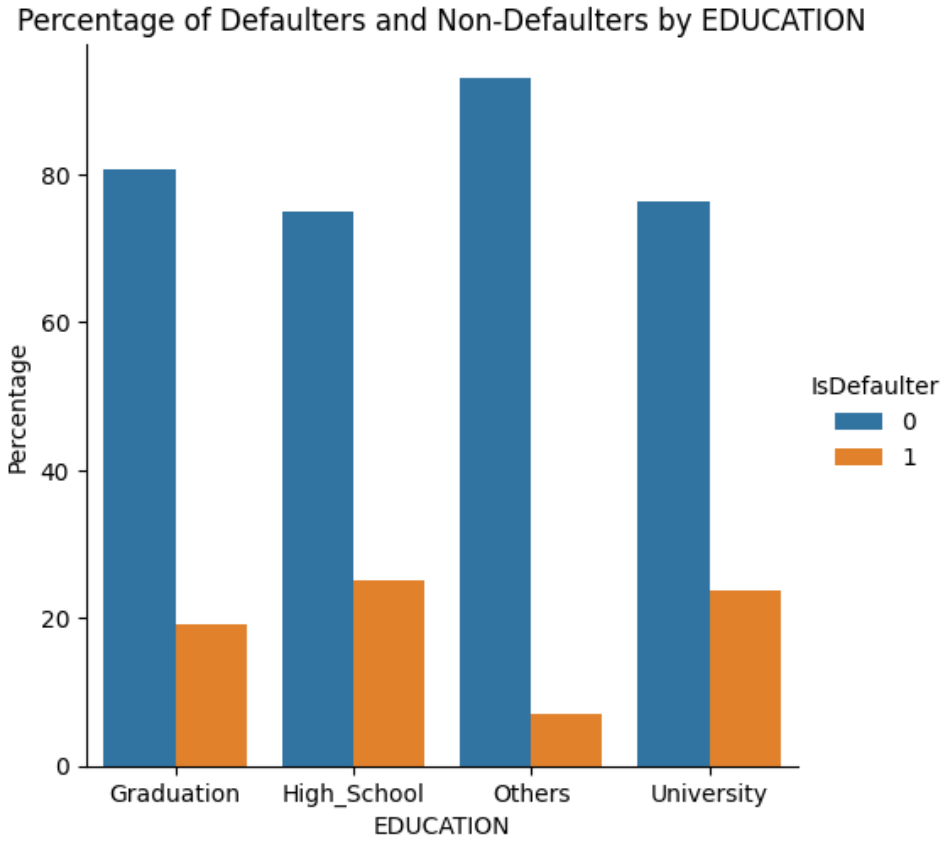
Gender Variable



Males are tending to be defaulting more than female. Only a slight difference.

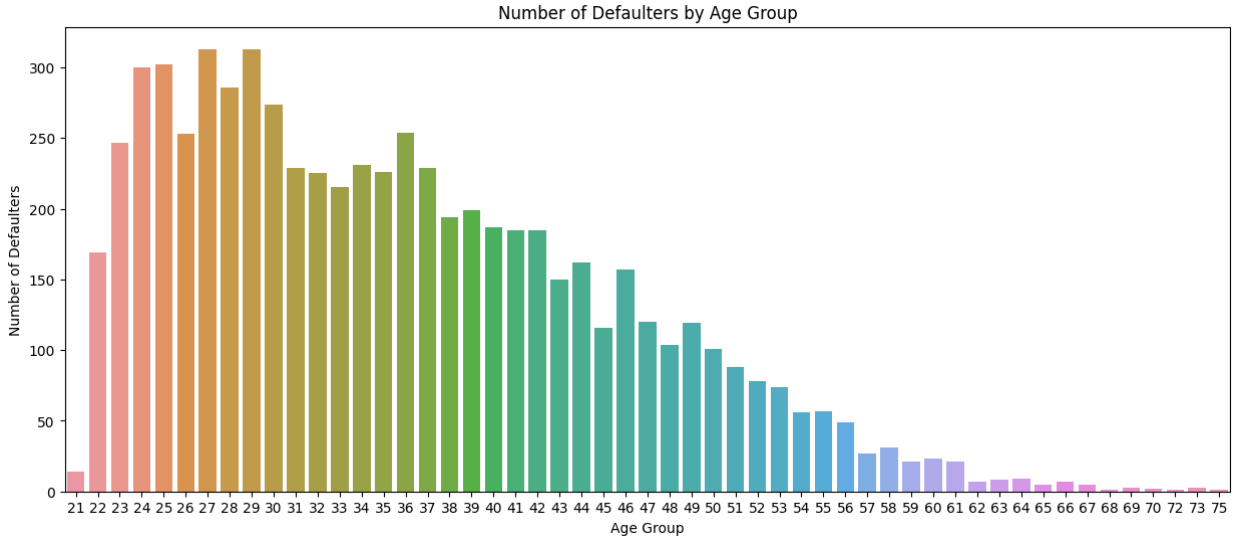
Education Variable

- Higher education level, chance of defaulters decreasing.
- Others are less in number of total customers

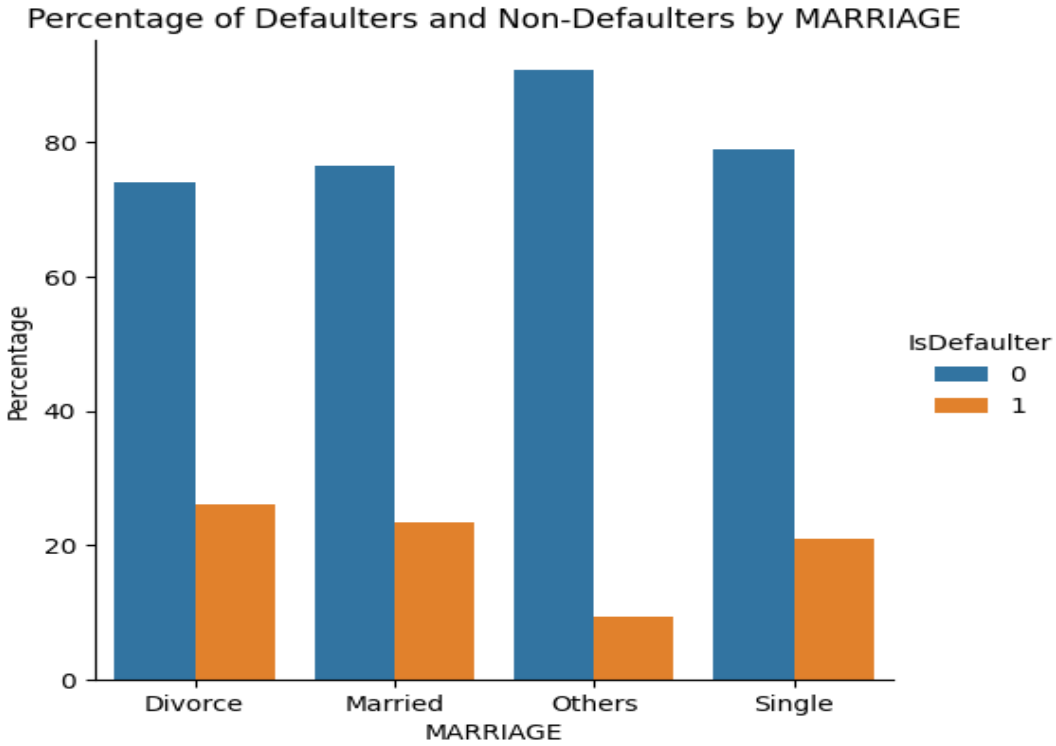


Age Variable

- More number of defaulters lies in age 20-40



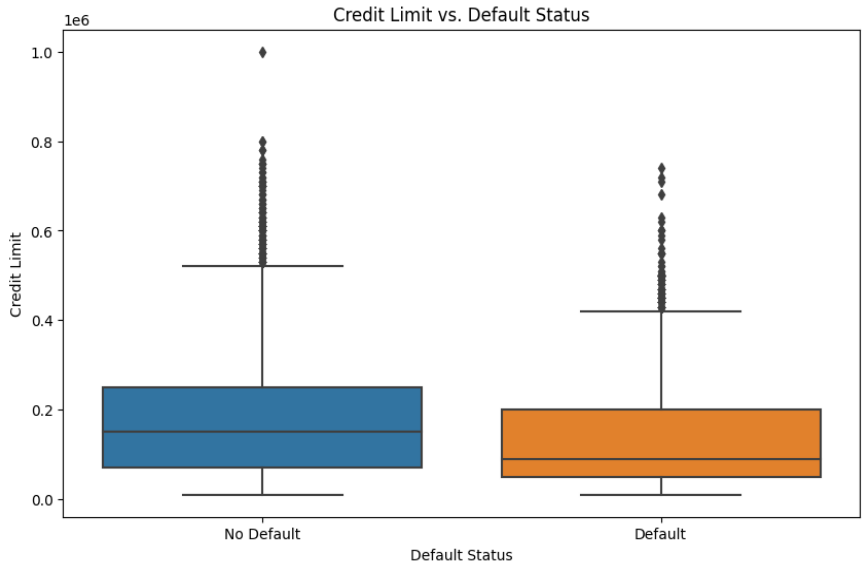
Marital Status Variable



Divorced people are more defaulters than married and single

Credit Limit Variable

- **Higher** credit limits,
- **lower** default risk.



EDA Summary

- Demographic factors that impact default risk are:
  - Education: Higher education is associated with lower default risk.
  - Age: Customers aged 20-40 have high default risk.
  - Sex: Females have lower default risk than males in this dataset.
  - Credit limit: Higher credit limit is associated with lower default risk.

Predictive Modeling

What recall, f1 score and ks scores can the models achieve?

Modeling Overview

Define Problem:	Supervised learning / binary classification
Imbalanced Classes:	78% non-default vs. 22% default
Models Applied:	Logistic Regression / Random Forest /SVM/ XGBoost

Modeling Steps

Data Preprocessing

- Feature selection
- Feature engineering
- SMOTE oversampling
- Train-test data splitting (75%/20%)
- Training data rescaling

Fitting and Tuning

- Start with default model parameters
- Hyperparameters tuning
- Measure recall, f1score, precision, roc_auc_score and ks statistic on training data and test data

Model Evaluation

- Models testing
- Recall, f1 score, ks statistic score comparison.
- Compare within the 4 models.(along with tuned models)

Outlier Treatment: The outlier treatment technique used is capping, which involves setting the extreme values (outliers) to specific percentiles of the data. Specifically, the 10th percentile and 90th percentile are used as lower bound and upper bound to cap the values.

Categorical Encoding: I did Binning for AGE variable and Label Encoding. Each age group is now represented by an integer from 0 to 5. Performed one-hot encoding on the SEX, EDUCATION, and MARRIAGE columns using the pd.get_dummies() function.

Feature Generation: BILL_AMT_SEPT, BILL_AMT_AUG, BILL_AMT_JUL, BILL_AMT_JUN, BILL_AMT_MAY, BILL_AMT_APR are highly correlated to each other. So, I am creating a new feature which average of bill amounts.

Feature Selection: After dealing with correlation, multi-collinearity and seeing the feature importance using Random Forest and decision tree classiffiers, These are my important features 'LIMIT_BAL', 'AGE', 'PAY_SEPT', 'PAY_AUG', 'PAY_JUL', 'PAY_JUN', 'PAY_MAY', 'PAY_APR', 'PAY_AMT_SEPT', 'PAY_AMT_AUG', 'PAY_AMT_JUL', 'PAY_AMT_JUN', 'PAY_AMT_MAY', 'PAY_AMT_APR', 'SEX_Female', 'SEX_Male', 'EDUCATION_Graduation', 'EDUCATION_High_School', 'EDUCATION_University', 'MARRIAGE_Divorce', 'MARRIAGE_Married', 'MARRIAGE_Single', 'AVG_BILL_AMT'

Before SMOTE : 0 :23364, 1:6636
After SMOTE: 0: 23364, 1:23364

Now shape of the dataset is (46728, 23)

Hyperparameters Tuning

- **Randomized Search** on Random Forest ,XGB and SVM
- **Grid Search** on Logistic Regression on limited parameters combinations.

Model Comparisons

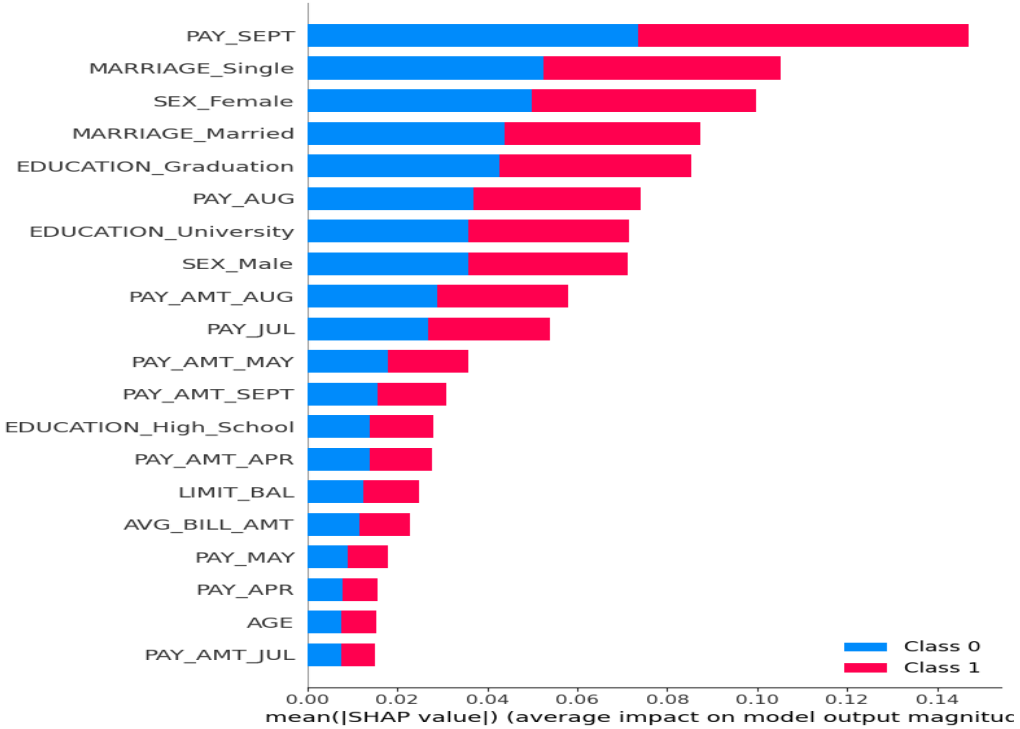
	Metric	Logistic_Regession	LR_Tuned	Random_forest	RF_Tuned	SVM	SVM_Tuned	XGB	XGB_Tuned
0	Recall_train	0.739843	0.753785	0.997372	0.854408	0.779955	0.796354	0.868293	0.994972
1	Recall	0.736312	0.748934	0.828415	0.812894	0.769401	0.776394	0.799591	0.827051
2	Precision_train	0.740000	0.754000	0.997000	0.854000	0.780000	0.796000	0.961000	0.995000
3	Precision	0.736000	0.749000	0.828000	0.813000	0.769000	0.776000	0.901000	0.827000
4	Accuracy_train	0.834731	0.832877	0.996918	0.891971	0.850197	0.861496	0.916424	0.994065
5	Accuracy	0.832734	0.829738	0.867745	0.852594	0.842493	0.844633	0.855247	0.861924
6	F1_score	0.815451	0.815337	0.862776	0.846988	0.830602	0.833776	0.847203	0.857395
7	ROC_AUC_score	0.890300	0.890300	0.928998	0.922627	0.901129	0.905590	0.920268	0.928007
8	KS_statistic	0.673062	0.667101	0.736743	0.708607	0.689967	0.695667	0.714044	0.728330

I finalized Tuned Random forest Model to be my best model based on the evaluation metrics recall, F1 score, KS statistic.

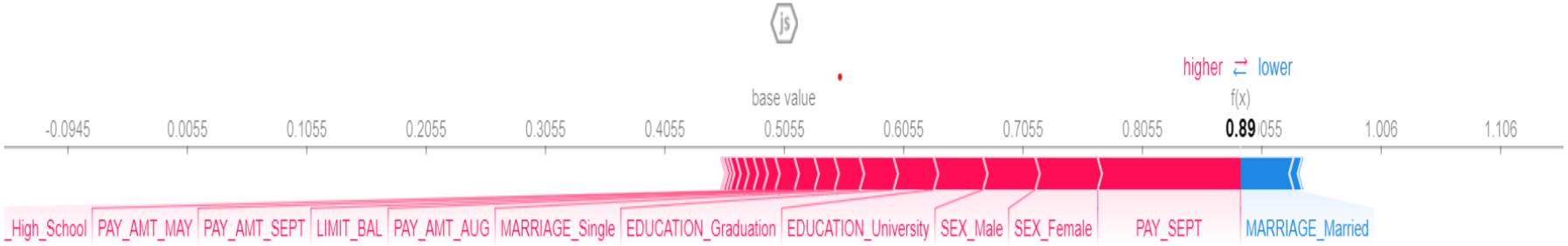
Even though XGB tuned model is giving high recall, f1 score and Ks, by observing the metrics on train data and test data, the model is likely overfitting. There is too much difference in train and test, model is almost learning everything from train data. For this reason, i disregard this model.

Tuned Random forest is performing well on the chosen metrics than remaining models Logistic Regression and SVM.

Feature Importance - SHAP



Model Explainability for a record




```
print(x_sample.columns)
print()
print(x_sample_v[0])

print(f'class of first record is {y_sample_v[0]}')
```

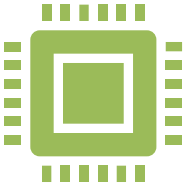
```
Index(['LIMIT_BAL', 'AGE', 'PAY_SEPT', 'PAY_AUG', 'PAY_JUL', 'PAY_JUN',
       'PAY_MAY', 'PAY_APR', 'PAY_AMT_SEPT', 'PAY_AMT_AUG', 'PAY_AMT_JUL',
       'PAY_AMT_JUN', 'PAY_AMT_MAY', 'PAY_AMT_APR', 'SEX_Female', 'SEX_Male',
       'EDUCATION_Graduation', 'EDUCATION_High_School', 'EDUCATION_University',
       'MARRIAGE_Divorce', 'MARRIAGE_Married', 'MARRIAGE_Single',
       'AVG_BILL_AMT'],
      dtype='object')

[-0.90840017  1.          1.6580188  -0.01387084  0.03099906  0.08216394
  1.87768545  0.15720544 -0.34922121 -0.40507853  0.08180482 -0.66898474
 -0.51785608 -0.42203399  0.          0.          0.          0.
  0.          0.          1.          0.         -0.05661587]
class of first record is 1
```

Conclusions

Random Forest has the best recall, f1_score and Ks statistic.

XGBoost can be used if precision is utmost important.



Model can be improved with more data and computational resources like income of a person feature.

Thank you!