Contents

Customer Default Identification Report

Introduction:

Acting as a provider of credit cards is one of the main activities of financial institutions such as Credit One. Practically, the funds of a Credit one are mainly used for lending activities. On the other hand, a bank will face a huge loss when a loan turns default. Therefore, banks always pay much attention to detect and predict the default behaviors of their customers.

Project objective:

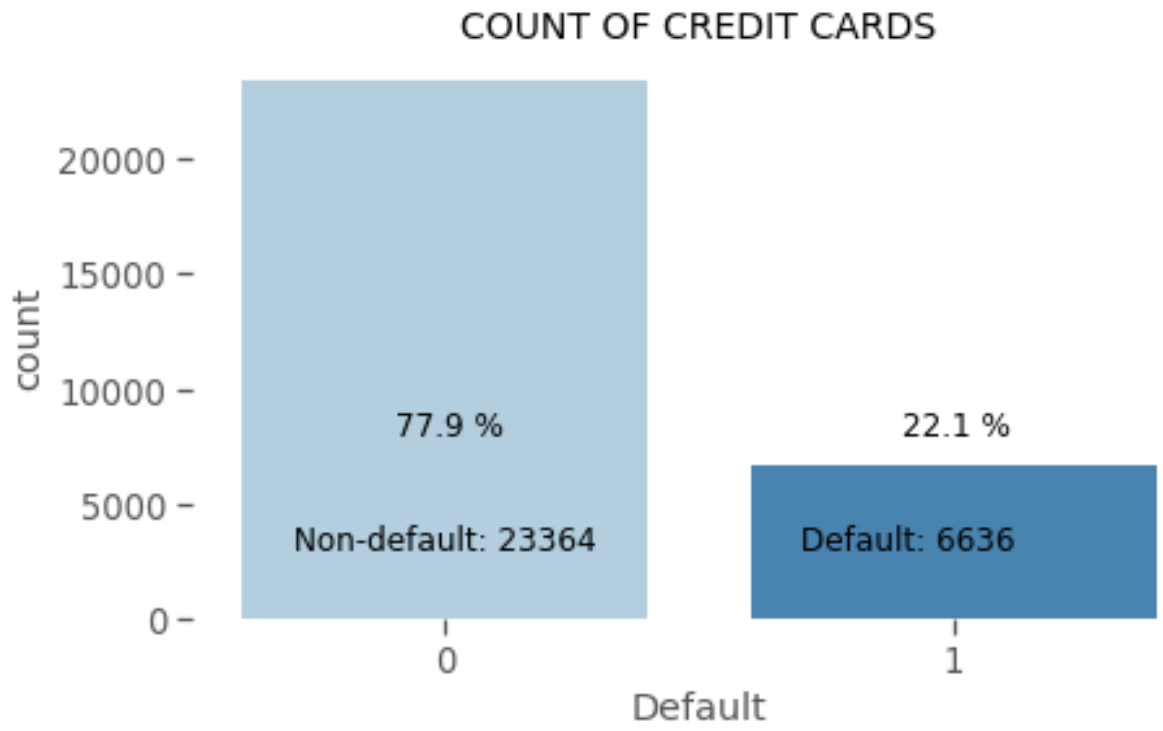
In order to prevent a credit cards from turning default, banks need to figure out how to make predictions based on customers' behaviors. Machine learning models appear to be one of the most effective solutions for predicting loans default. Therefore, the objective of this project is to build predictive models for credit card default predictions and to explore the impact of customer behavioral factors on making predictions further.

Data Analysis Framework and workflow

Find the real and actionable Business question	Formulate a hypothesis-driven Analysis plan	Collect relevant Data based on the Analysis	Derive Insights using Machine Learning & statistcis	Derive KPI's with actionable Recommendation
1. How do you ensure that customers can/will pay their loans? 2. Can we approve customers with high certainty?	Default rate is dependent on following 1. Customer Payment History 2. Customer transaction history 3. Customer demographics 4.Micro economic conditions 5.Macro economic conditions	Collect demographic and historical financial transactional information	1. Exploratory Data Analysis 2. Build and evaluate Models	Select model based on performance 1.Accuracy 2.Precision 3.Recall & 4.F1 score Impact of Credit default • Financial Loss • Customer Attiration

Exploratory Data Analysis

From this sample of 30,000 credit card holders, there were 6,636 default credit cards; that is, the proportion of default in the data is 22,1%.



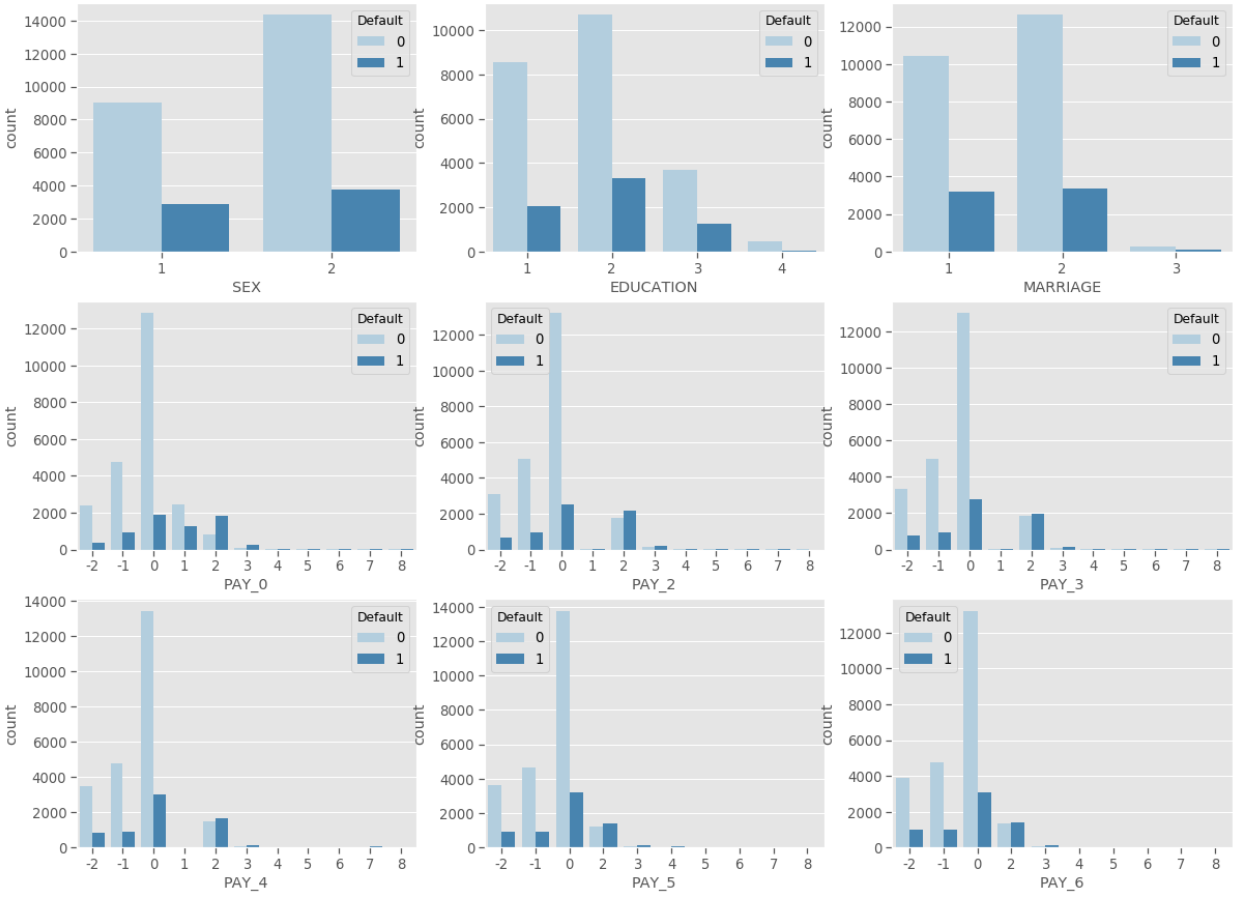
Variables exploration

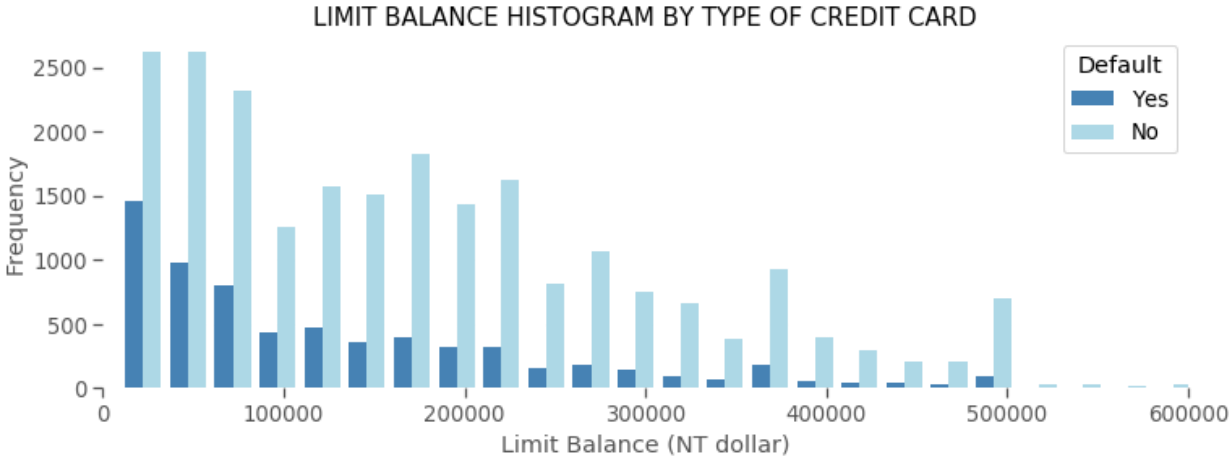
Gender - Sometimes gender can be useful when making predictions, so we plot the gender distribution of non default and defaulted credit card as well as the default proportion of males and females.

Demographic - Demographic data also matters in analysis and making predictions. Some of them can be strong predictors when making predictions on default.

Frequency of explanatory variables by defaulted and non-defaulted cards

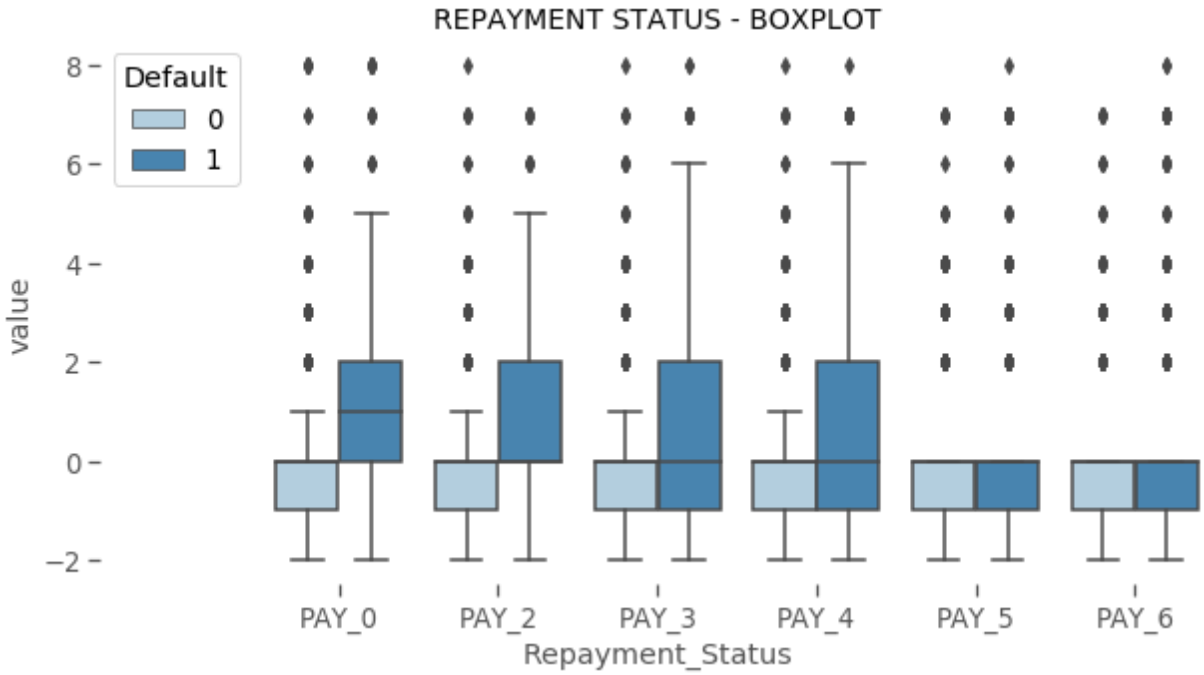
FREQUENCY OF CATEGORICAL VARIABLES (BY TARGET)





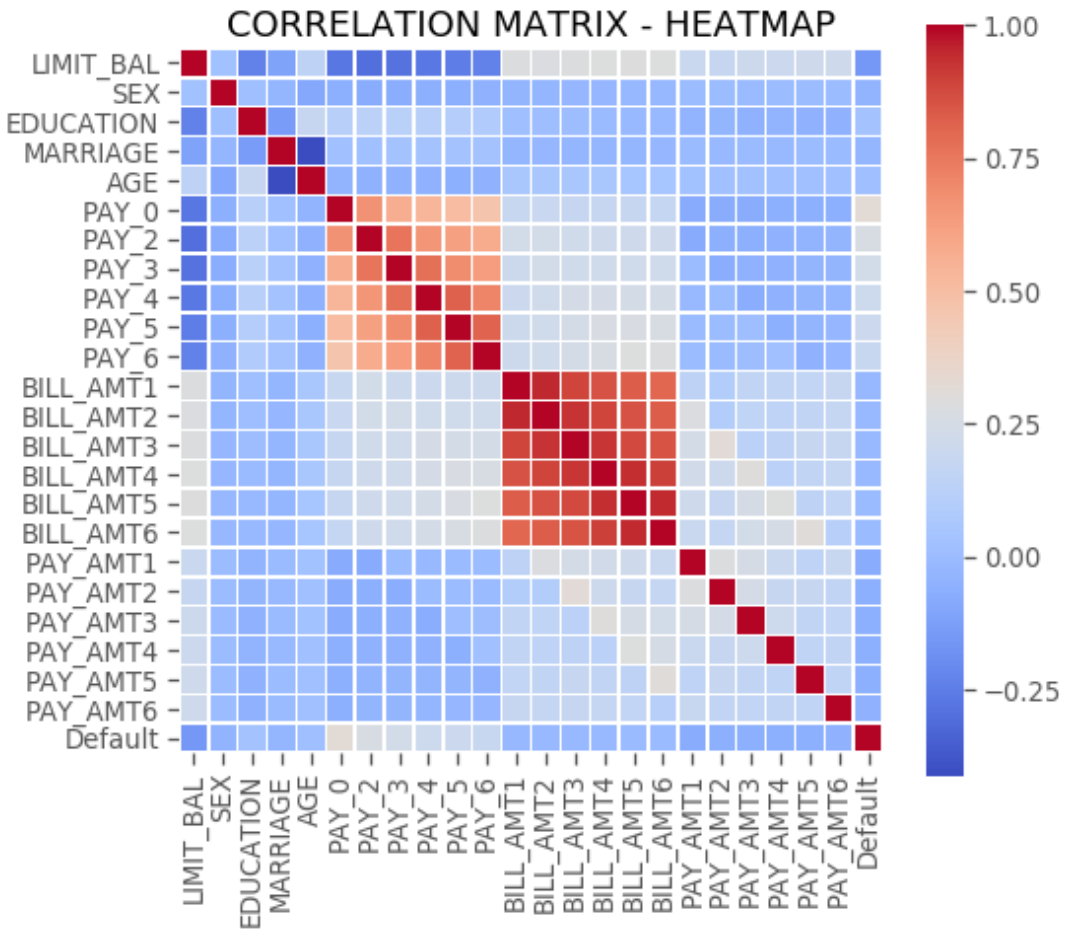
Summary of findings

- There are 30,000 credit card clients.
- The average value for the amount of credit card limit is 167,484 NT dollars. The standard deviation is 129,747 NT dollars, ranging from 10,000 to 1M NT dollars.
- Education level is mostly graduate school and university.
- Most of the clients are either married or single (less frequent the other status).
- Average age is 35.5 years, with a standard deviation of 9.2.
- As the value 0 for default payment means 'not default' and value 1 means 'default', the mean of 0.221 means that there are 22.1% of credit card contracts that will default next month.
- It seems that PAY_0 (Repayment status in September) and PAY_2 (Repayment status in August) have more discriminatory power the repayment status in other months.



Correlation Analysis

A correlation matrix of all variables is shown in the heatmap below. The only feature with a notable positive correlation with the dependent variable 'Default' is re-payment status during the last month (September). The highest negative correlation with default occurs with Limit_Balance, indicating that customers with lower limit balance are more likely to default. It can also be observed that some variables are highly correlated to each other, that is the case of the amount of bill statement and the repayment status in different months.



The heatmat shows that features are correlated with each other (collinearity), such us like PAY_0,2,3,4,5,6 and BILL_AMT1,2,3,4,5,6. In those cases, the correlation is positive.

Uncorrelated data are potentially more useful: discriminatory!

Feature Selection

Recursive Feature Elimination (RFE) is based on the idea to repeatedly construct a model and choose either the best or worst performing feature, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. The goal of RFE is to select features by recursively considering smaller and smaller sets of features.

Most important features (RFE):

- Repayment status in September (PAY_0)
- Amount of bill statement in September (BILL_AMT1)
- Amount of previous payments in August (PAY_AMT2)

Build and Evaluate Predictive Models

Based on the understanding and the exploration of the dataset, we will build a supervised machine learning model using Python and Scikit-learn.

1. Categorical values to binary variables

Notice that after joining the tables together, there are some columns have categorical values which need to be converted to binary variables.

- Education - The categories 4:others, 5:unknown, and 6:unknown can be grouped into a single class '4'.
- Similarly, the column 'marriage' should have three categories: 1 = married, 2 = single, 3 = others but it contains a category '0' which will be joined to the category '3'.

2. Model selection with cross-validation

In this section, we will try a few supervised models with 10-fold cross validation using all the feature columns and the default hyper-parameter setting. The purpose is to select a model that fits the dataset better.

The classification models used for this analysis are:

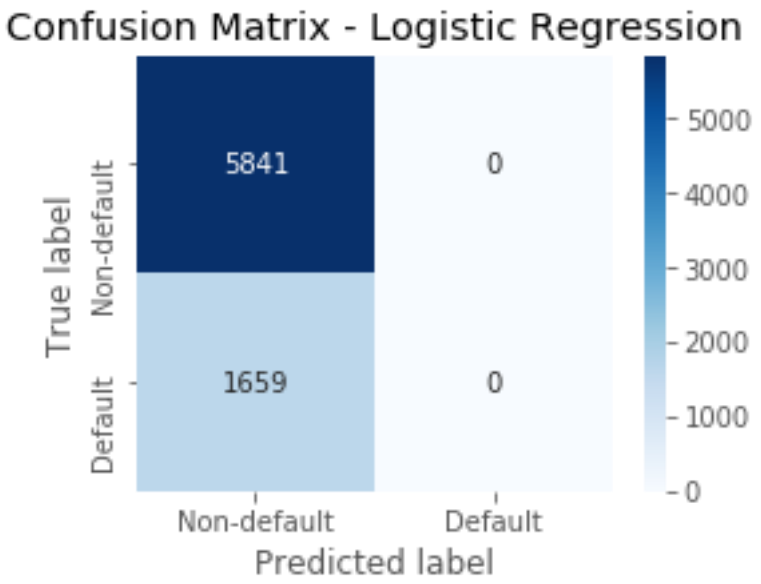
1. Logistic Regression,
2. Decision Tree and
3. Random Forest Classifier.

To build machine learning models the original data was divided into features (X) and dependent variable (y) and then split into train (75%) and test (25%) sets. Thus, the algorithms would be trained on one set of data and tested out on a completely different set of data (not seen before by the algorithm).

2.1 Logistic regression

Logistic Regression is one of the simplest algorithms which estimates the relationship between one dependent binary variable and independent variables, computing the probability of occurrence of an event

```
Accuracy: 0.7788

              precision    recall  f1-score   support

           0       0.78      1.00      0.88      5841
           1       0.00      0.00      0.00      1659

   micro avg       0.78      0.78      0.78      7500
   macro avg       0.39      0.50      0.44      7500
weighted avg       0.61      0.78      0.68      7500


Average 5-Fold CV Score: 0.7787 , Standard deviation: 0.0001
```



The model has not power predicting default credit cards. However, it can be observed that the average accuracy of the model is about 78%, which demonstrates that this metrics is not appropriate for the evaluation of this problem.

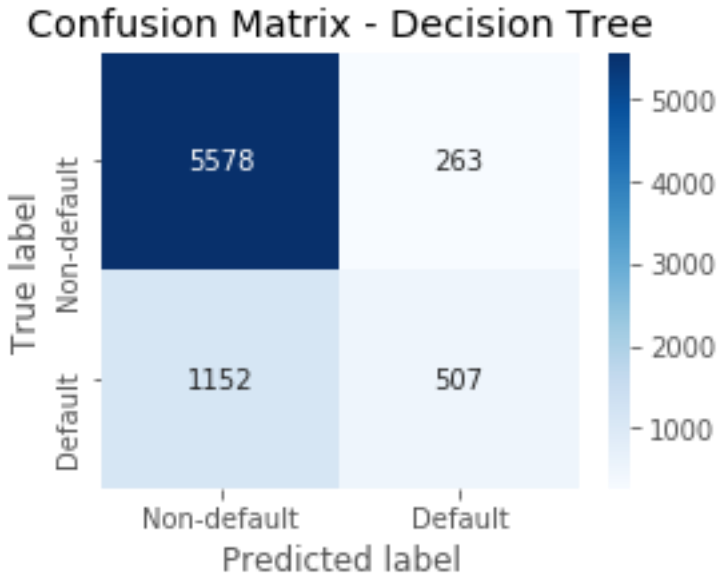
Due to the poor performance of logistic regression, other linear models may perform similarly. Therefore, we try a tree-based model next.

2.2 Decision Tree classifier

Decision Tree is another very popular algorithm for classification problems because it is easy to interpret and understand. An internal node represents a feature, the branch represents a decision rule, and each leaf node represents the outcome. Some advantages of decision trees are that they require less data preprocessing, i.e., no need to normalize features. However, noisy data can be easily overfitted and results in biased results when the data set is imbalanced.

```
Accuracy: 0.8113333333333334
```

```
              precision    recall  f1-score   support
           0       0.83      0.95      0.89      5841
           1       0.66      0.31      0.42      1659

   micro avg       0.81      0.81      0.81      7500
   macro avg       0.74      0.63      0.65      7500
weighted avg       0.79      0.81      0.78      7500

Average 5-Fold CV Score: 0.8136 , Standard deviation: 0.0058
```



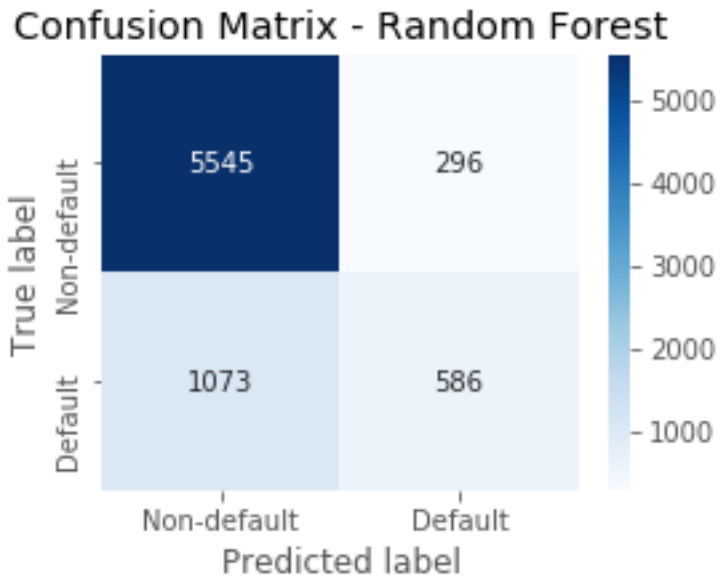
The performance of the decision tree model improved compared to the logistic regression model showed previously. However, the recall is still low (0.31).

2.3 Random forest classifier

Random forest classifier is comprised of multiple decision trees. It creates different random subset of decision trees from the training set as its predictors and selects the best solution by means of voting. As a result, the Random Forest model avoids overfitting problems.

```
Accuracy: 0.8174666666666667

              precision    recall  f1-score   support

           0       0.84      0.95      0.89      5841
           1       0.66      0.35      0.46      1659

   micro avg       0.82      0.82      0.82      7500
   macro avg       0.75      0.65      0.68      7500
```

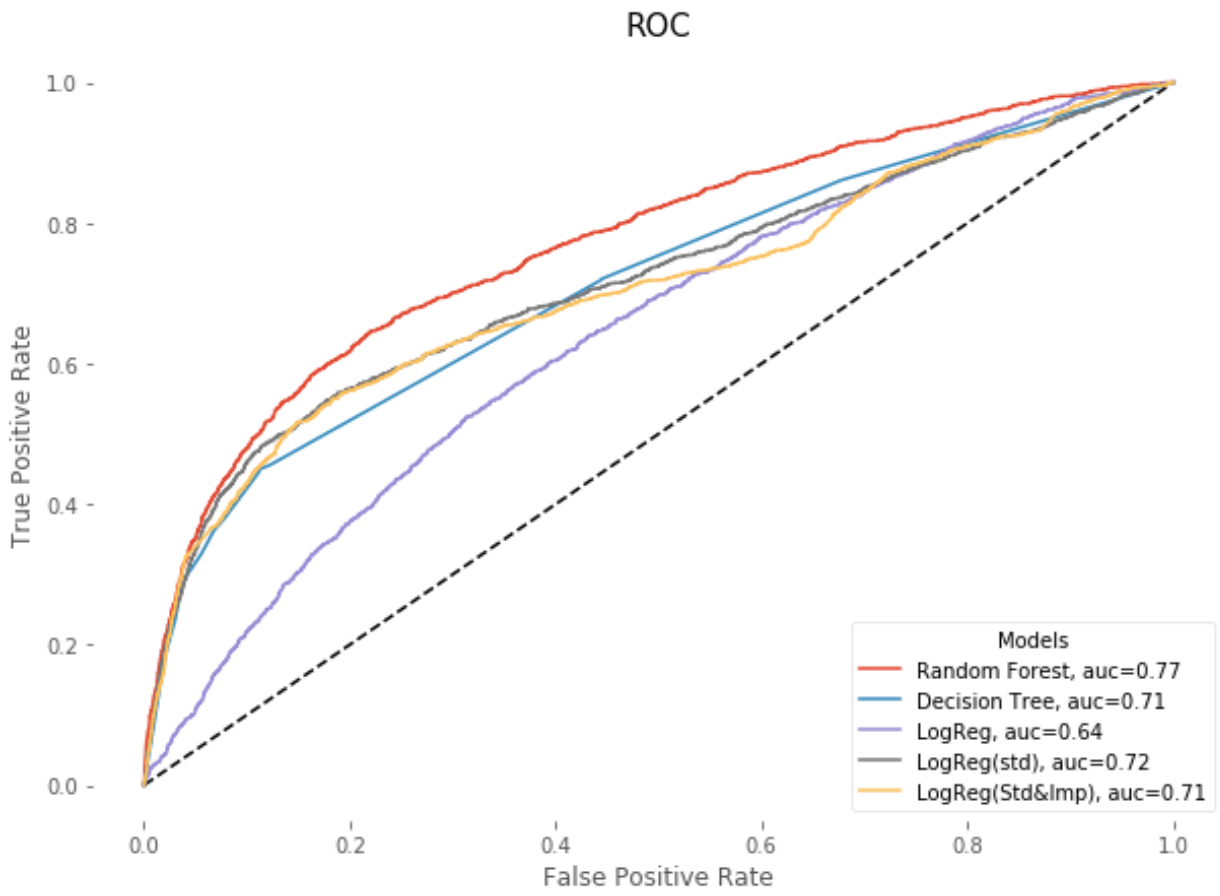
```
weighted avg        0.80      0.82      0.80      7500

Average 5-Fold CV Score: 0.8204 , Standard deviation: 0.0096
```



Random forest classifier performs a lot better than logistic regression. It seems we are on the right track. However, the model is still overfitted, and we need to tune the hyper-parameters later. Next, we will try a gradient boosting classifier as a comparison to this random forest classifier.

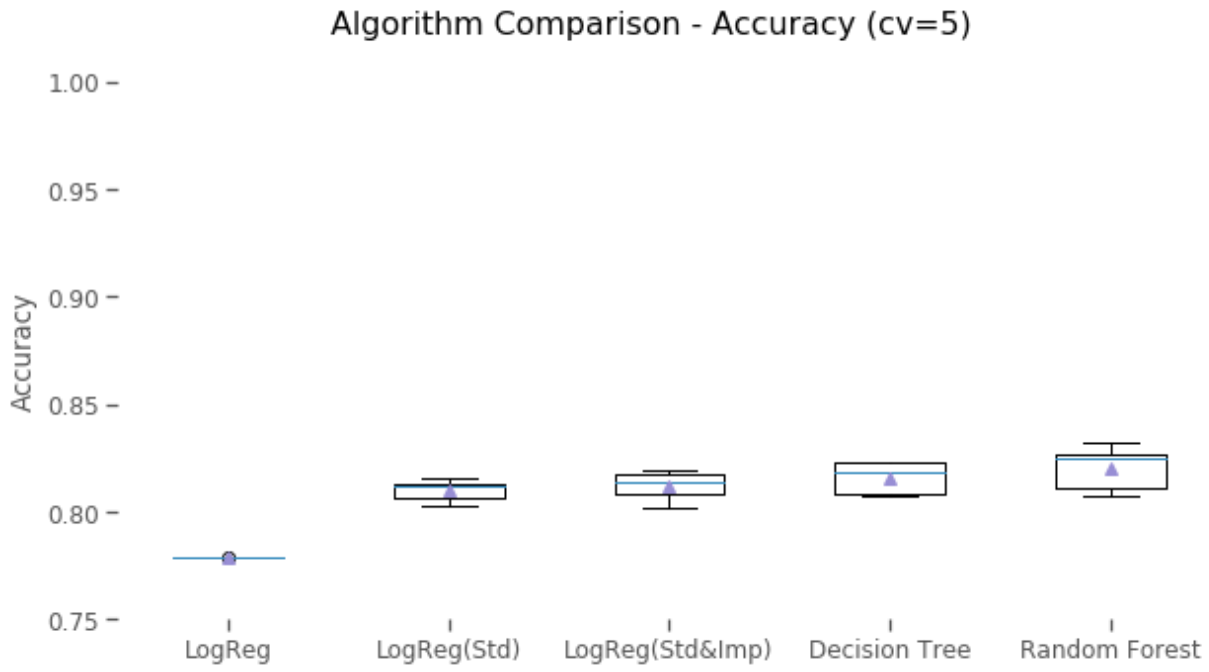
ROC curve comparison

The highest accuracy is obtained for the **Random Forest Classifier model**, with a value of 0.77. This means there is 77% chance that the model will be able to distinguish between default class and non-default class.



3. Model Accuracy comparison

The best accuracy is obtained for the Random Forest Classifier with a mean accuracy of 0.82, yet it is the model with higher variation (0.0096). In general, all models have comparable mean accuracy. Nevertheless, because the classes are imbalanced (the proportion of non-default credit cards is higher than default) this metric is misleading. Furthermore, accuracy does not consider the rate of false positives (non-default credits cards that were predicted as default) and false negatives (default credit cards that were incorrectly predicted as non-default). Both cases have negative impact on the bank, since false positives leads to unsatisfied customers and false negatives leads to financial loss.



Model performance comparison

Precision, Recall, F1-score

Model	Data	Precision	Recall	F1
Logistic Regression	Standardized	0.79	0.81	0.77
Logistic Regression	Important features	0.79	0.81	0.78
Decision Tree	original	0.80	0.82	0.79
Random Forest	original	0.80	0.82	0.80

Recommendations

In this project, we built a supervised machine learning model from scratch for predicting credit card default. We preprocessed the data for exploration and modeling, and trained a Random forest classifier for default prediction using Scikit-learn. The model, with the best hyper-parameter, has an good performance with a 0.89 F1 score for default and 0.39 F1 score for not default.

Linear models such as logistic regression did not fit the dataset well whereas tree-based models like random forest classifier and gradient boosting classifier can provide decent performance .