CreditOne

Customer Default Analysis

Carlos Barth

7/28/20

Introduction

Businesses rely on Credit One to assess the credit worthiness of their customers seeking loans. A high rate of loan defaults costs our customers thousands of dollars, weakens our business relationships with them and ultimately damages our reputation in the industry.

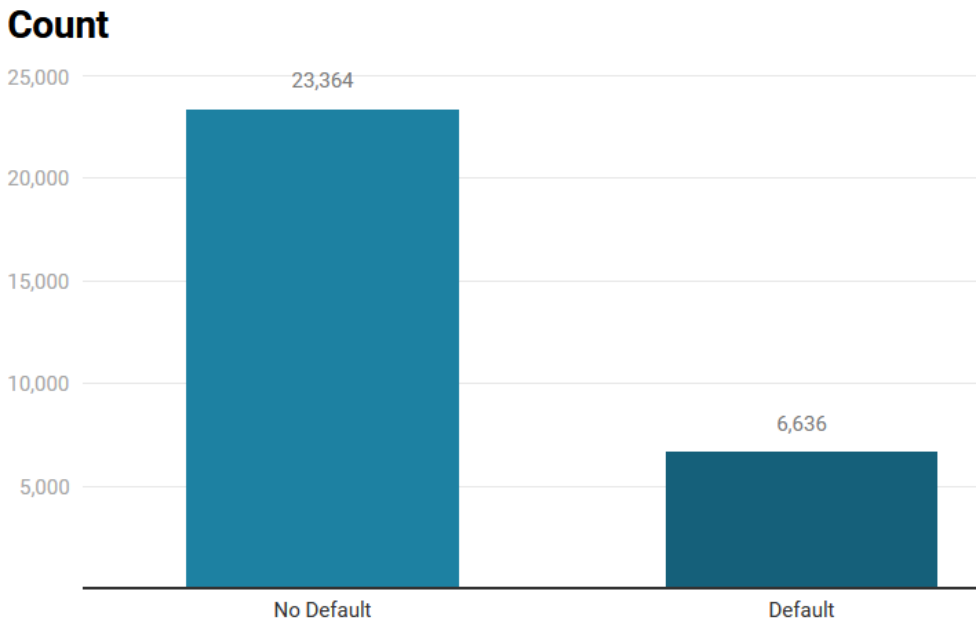
Credit One's Data Science team has been tasked with evaluating a sample of our customer's data set containing known account defaults. Better detection and earlier identification of conditions that lead to default would enable us to make better recommendations to our customers.

Objective

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.

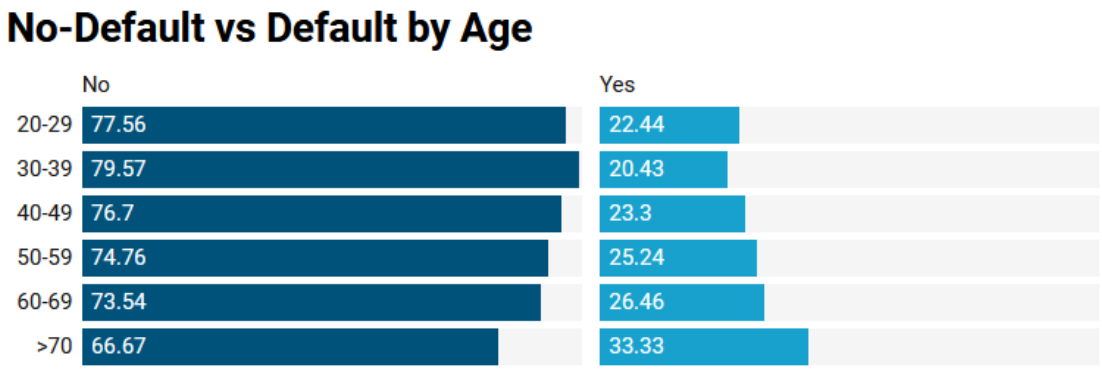
Exploratory Data Analysis

We have a sample of 30,000 records, where almost 78% are No-Default loans and 22% defaulted.

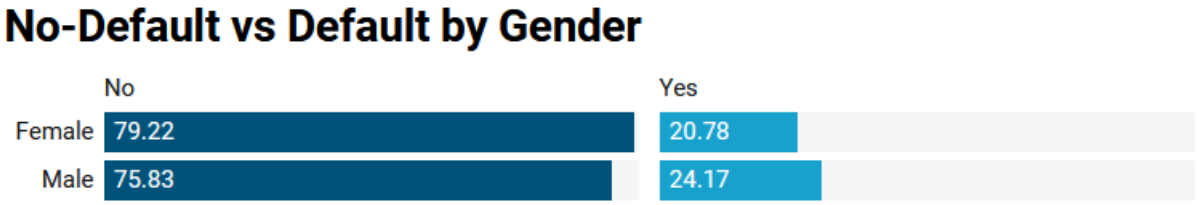


Attributes:

Age: We found that people between 20 and 49 tend to NO defaults their loans and older people tend to do it, this could be for different reasons, for example the range of people who tend to default (>70) could be because they lost their jobs and it's complicated get a new one and pay their debts.



Gender: In this case the gender can be useful when making predictions, so we plot the gender distribution of no-default and default loans.

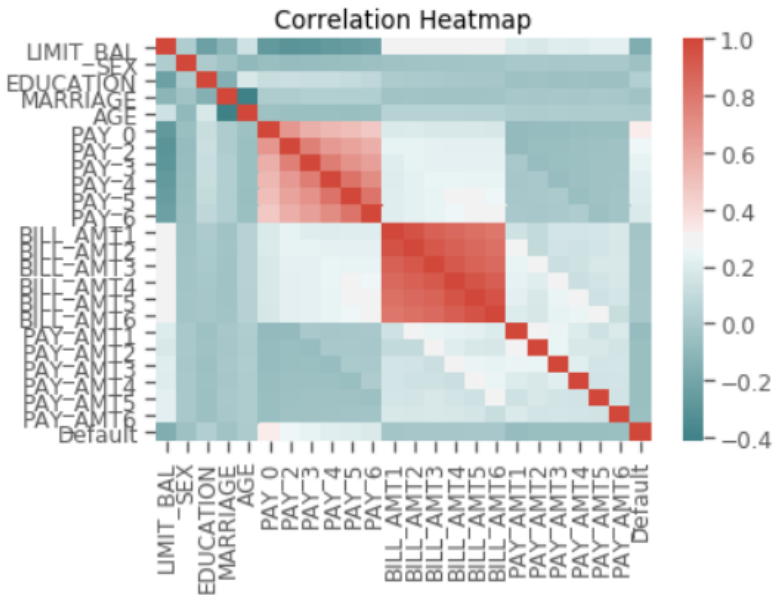


Summary of findings

- There are 30,000 credit loans for clients.
- The average value for credit card limit is $167,484. The standard deviation is $129,747, ranging from $10,000 to $1M.
- 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 loan contracts that will default next month.

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.



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

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. Categorical values to binary variables

In this section, we will try a few supervised models with 5-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:

- Logistic Regression
- Decision Tree and
- 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).

- **Logistic regression**

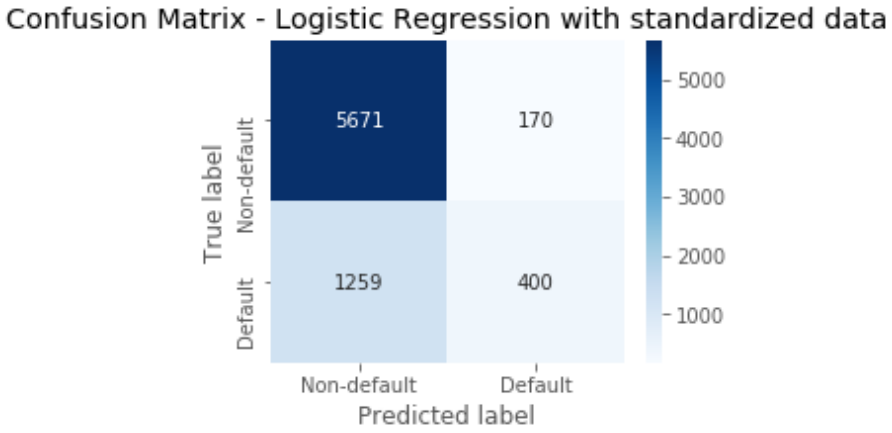
```
Accuracy: 0.7788

              precision    recall  f1-score   support

           0       0.78      1.00      0.88      5841
           1       0.00      0.00      0.00      1659

    accuracy                           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: 0.7788 , Standard deviation: 0.0001
```



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.

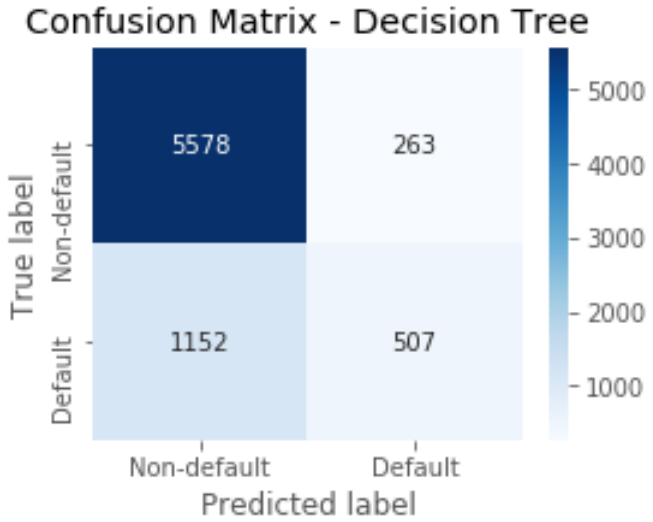
We try a tree-based model next.

- **Decision Tree classifier**

```
Accuracy: 0.8113333333333334
```

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



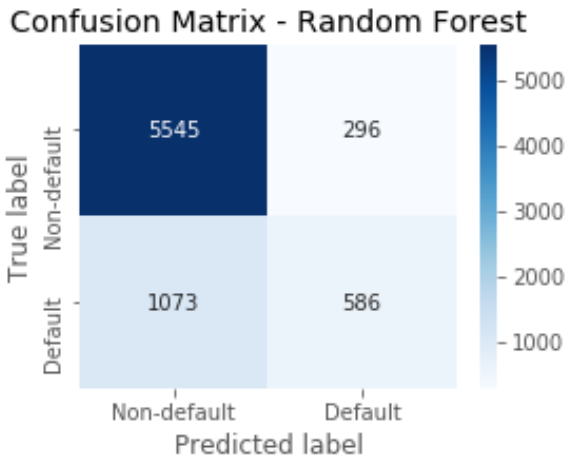
The performance of the decision tree model improved compared to the logistic regression model showed previously. However, the recall is still low (0.31).

- **Random forest classifier**

```
Accuracy: 0.8174666666666667
```

	precision	recall	f1-score	support
0	0.84	0.95	0.89	5841
1	0.66	0.35	0.46	1659
accuracy			0.82	7500
macro avg	0.75	0.65	0.68	7500
weighted avg	0.80	0.82	0.80	7500

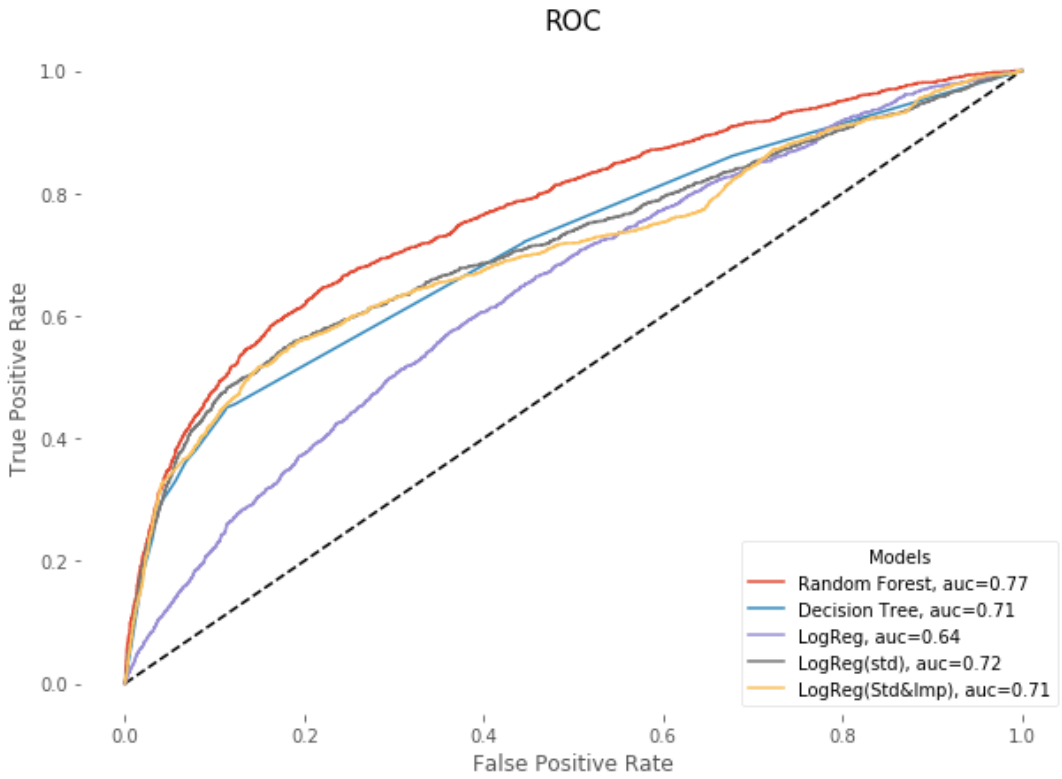
```
Average 5-Fold CV Score: 0.8203 , Standard deviation: 0.0093
```



Random forest classifier performs a lot better than logistic regression, However, the model is still overfitted, and we need to tune the hyper-parameters later.

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.



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 (0.0093). In general, all models have comparable mean accuracy. Nevertheless, because the classes are imbalanced (the proportion of non-default 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 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

Model Comparison

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

Recommendations

- Using predictive models seem to improve the ability to detect potential defaults for our customers. We recommend that we use our model on a larger data set and, after confirming similar results to those found here, implement this model for future client evaluation.
- Besides the information analyze before, is strongly recommended to start collecting next data to create a more robust credit score calculation and be more accurate to predict the default.
- In this project, we built a supervised machine learning model from scratch for predicting credit loans default. We pre-processed the data for exploration, modeling and trained different classifiers for default prediction using Scikit-learn.
- The model, with the best hyper-parameter, has a good performance with a 0.89 F1 score for default and 0.39 F1 score for not default (Random Forest).
- Linear models such as logistic regression did not fit the dataset well whereas tree-based models like random forest classifier and decision tree classifier can provide decent performance.