Marcelo Jimenez

Final Report

Introduction

The data analytics team was responsible of finding a much better way to understand how much credit to allow someone to use or, at the very least, if someone should be approved or not. We were given a data set with 25 features and close to 30,000 labels. This got reduced to only 24 features and close to 29,800 labels accounting for duplicates and inconsequential features. Initial EDA showed very low correlation between most features, but the significant ones are shown in the correlation map (Figure 1). Furthermore, it is worth noting there are more women and men in the data (Figure 1) and that there is more 'not default' customers than 'default', possibly making the subsequential modelling biased towards 'not default'.

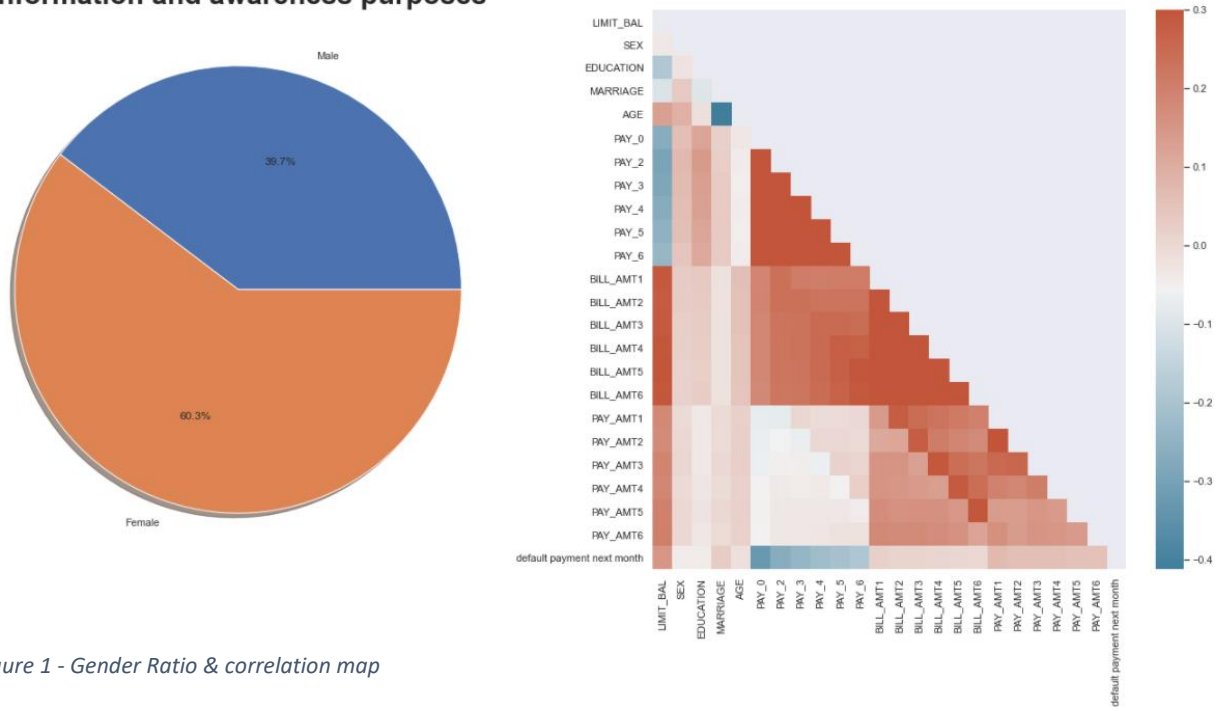

Figure 1 - Gender Ratio & correlation map

Modelling

Both regression and classification models including cross validation algorithms were used for the creation of the models. In summary, regression models failed to accurately represent the data and a classification approach was used. Random Forest classifier proved essential for this purpose and achieved a 96% accuracy for predicting the default status of a customer, this is depicted in Figure 2.

```
Random Forest Classifier - Accuracy: 0.95961955
```

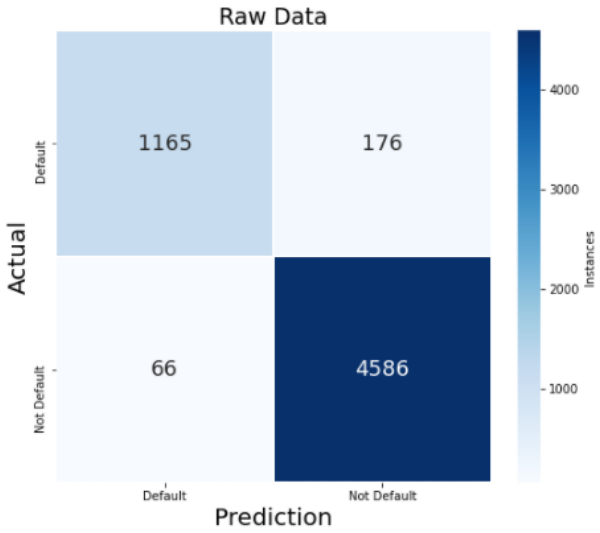

Figure 2 - Confusion Matrix

Furthermore, tried experimenting using a second set of data (Drops) which ignores the values of $500,000 credit limit since it appears to be an anomaly in the data, an artificial limit set by the company which the team thought would not help the model at all. However, there is a 15% difference between the accuracy of the raw data and the drops data, making the raw data much more accurate. Because of this, we will not exclude any data from the model.

Also, the team attempted to predict the best limit balance to give a new customer using the same classifier algorithm. However, the accuracy of the predictions was so low that it could be compared to a coin toss (Figure 3). This shows the limits of the machine learning classifier and the team concluded the most we can predict with certainty is whether a customer will default or not in the next month.

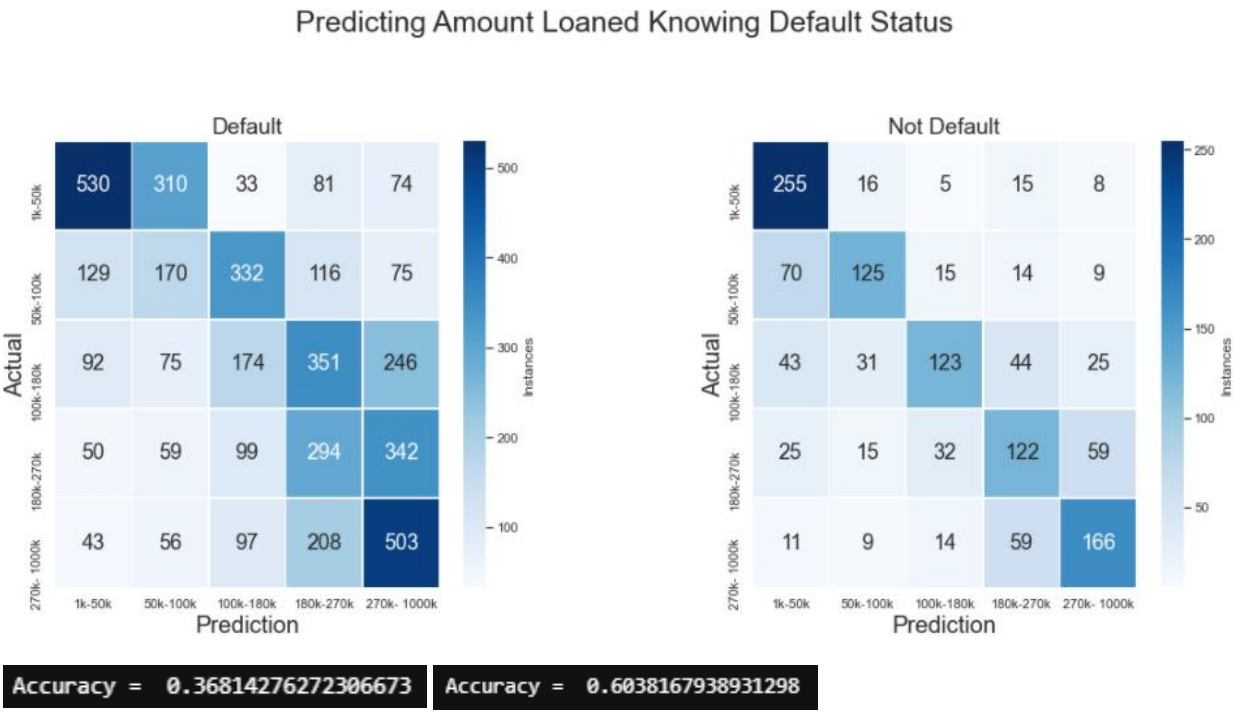

Figure 3- Accuracy scores of failed Limit Balance Predictions

Additional questions:

How do you ensure that customers can/will pay their loans?

The company cannot change the spending habits of customers, but it can limit the credit limit each customer gets. Based on this, it is vital to create a model that accurately predicts whether a customer will default or not the next month.

Can we approve customers with high certainty?

Yes, we can predict whether a customer will default next month and subsequently assess whether they should be approved or not. The current model can do this with a 96% accuracy.