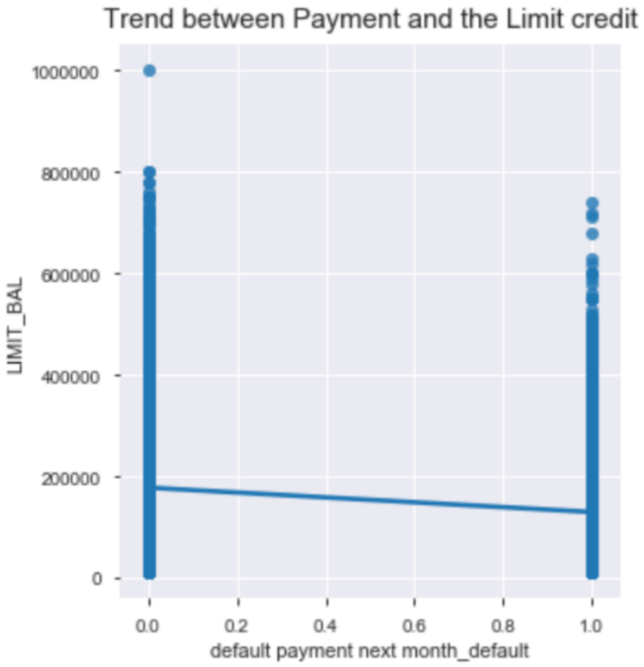
Credit One Project - Report

Credit one is having issues, increasing in customer default rates which could end losing revenue and customers for its clients and of course for Credit One.

One important fact, coming from the data provided, is that 70% of the customer's credit limit goes from $10,000 to $200,000, and from the 6,636 customers (22% of the total) who will default 78.6% will fall in this range. Therefore, a lot of attention need to be done in this range.

Furthermore, to emphasize the point mentioned above, it seems to be a general linear trend between Payment and the Limit credit. The visualization below shows that the higher the limit is, the less chance he/she would default.



In order to ensure customer will pay the loans, a model need to be created with a good performance around 70% that will help Credit One to provide a better service improving what it is havening today.

There are different algorithms that can be used to help Credit One to improve with high accuracy the way they approve customers. For instance, Random Forest Regressor, Linear Regression, Support Vector Regression, etc. Furthermore, in case those algorithms won't work in the way it is expected, the data can be discretized and use classification algorithms, such as Random Forest Classifier, Gradient Boosting Classifier, Support Vector Classifier, etc

Before selecting the model, feature selection needs to be done in order to know what variables are the best to use for building the model. One best practice is to check the correlation between variables and independent variable.

The correlation shows that “Bill Amount” and “Payment Amount” variables are the ones with highest relationship with the Target variable. “Age” and “default” are also showing some relationship as well. Once the variables have been selected, it is time to run the model.

The best result for the regression model was given by the Random Forest Regressor algorithm with performance of 0.47 which is not what it was expected.

Since regression model did not have a good performance, classification model could give us better one. In order to transform the project from regression to classification, discretization will be used to the continuous Target variable to create bins needed for classification model.

For the classification model, the same variables “Bill Amount”, “Payment Amount”, “Age”, “default” are going to be used since the correlation between the variables does not change.

The best result for the classification model was given by the Gradient Boosting Classifier algorithm. The following table shows the accuracy with several intervals used to build the model:

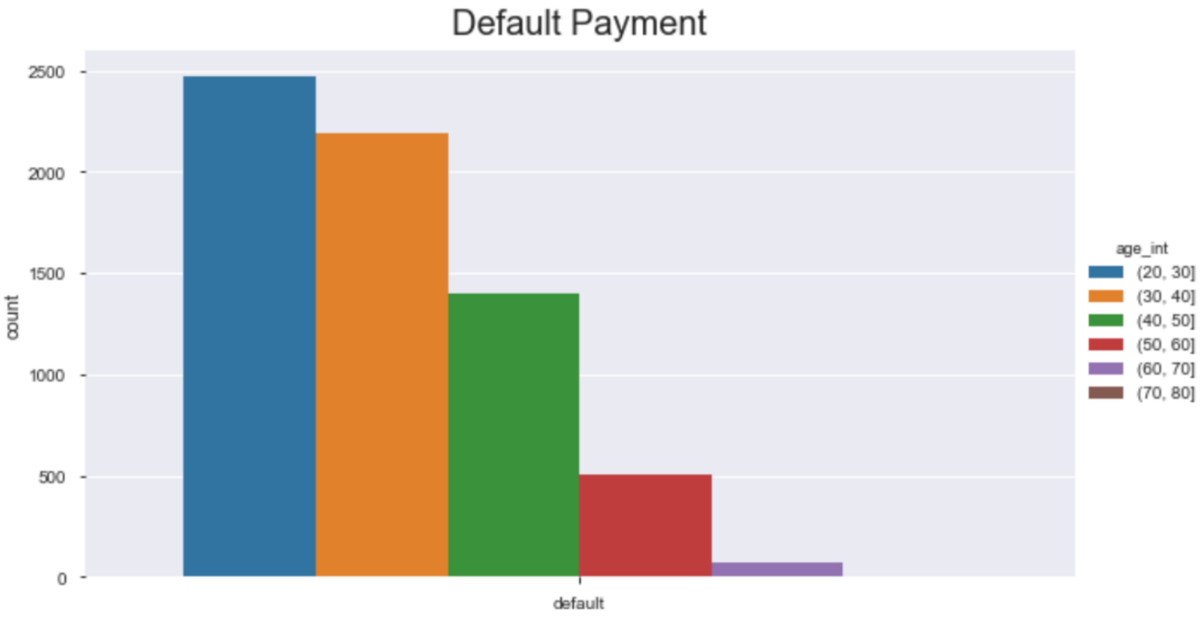
Bins	Accuracy	Range for 1st bin
4	0.811	$9,010 - $257,500
5	0.735	$9,010 - $208,000
6	**0.689**	**$9,010 - $175,000**
7	0.661	$9,010 - $151,428
8	0.631	$9,010 - $133,750
9	0.606	$9,010 - $120,000
10	0.568	$9,010 - $109,000

As shown in the table above, the accuracy of the model increases when the number of intervals (bins) decreases. However, the size of the credit limit increase when the number of intervals decreases. Therefore, it is up to the business what level of accuracy is allowed in order to have better coverage for the credit limit range.

Finally, in order to improve the accuracy of the model more data should be gathered such as credit score, salary, assets value, etc.

On a side note, additional discretization was done with the independent variables. For instance, the “Bill Amount”, and the outcome did not change, even the accuracy decrease a little bit.

Furthermore, since the range of the credit limit is wide, it is recommended to start at the lower end for young people because from the 6,636 customers who will default 70% will fall in this range (below 40 years old) as shown in the following visualization:



Another important recommendation is for the level of education. 47% of the customer's education falls at University level, and only 16% in high school. However, from the 16% in high school, 25% will default with the highest percentage, follow by university level with 24% and graduate school with 19%. There is a trend in Education level similar to the Age, the higher the level of education, the less chance he/she would default.

In summary, a model can be built to ensure that customers will pay their loans with accuracy of 0.69% or above depending on the size range of the credit limit. Furthermore, it is recommended to get more data from customers and a reduce the credit limit for younger people with lower level of education.