**“Financial Risk Analytics”**

**Abstract:**

Credit default is one of the most important problems in the banking and risk analytics industry, in current usages the credit card usage plays a key role and that makes credit cards to be default in most of the cases and need to predict such cases for the banking sector.

A Credit Card company needs to improve the ability to predict whether or not its customer will default, as well as recognise the factors that influence this probability and analyse the risk associated with the payment of credit card bills by customers.

Credit card default happens when you have become severely delinquent on your credit card payments. When a customer starts defaulting on payments of credit card bills, one is put under scanner with restrictions to prevent further defaults. These defaulters are found through observing historical data.

There are various attributes which can be used to predict default, such as demographic data (age, income, employment status, etc.), behavioural data (past loans, payment, number of times a credit payment has been delayed by the customer etc.)

This project is done to done to create an efficient machine learning model to identify a better default rate of the customers using a dataset.

This project predicts the probability of a customer defaulting on their credit card bill next month and a 30,000-sample dataset of credit card customers. We'll also describe the dataset using statistics and visualize it using a manifold learning technique.

# **Business problem statement:**

* Business Problem Understanding:

We are trying to analyse the customer defaulting on their credit card bill next month on an available dataset of 30000 samples.

* Business Objective:

Objective is to train an estimator that predicts the probability of a customer defaulting on their credit card bill the next month.

* Approach:

To perform the statistical analysis, data visualization techniques and ensemble techniques on the dataset to find the sole objective of a customer defaulting the next credit card pay cycle.

* Dataset:

A dataset of 30000 samples has been gathered for further analysis.

# **Critical Assessment of Topic Survey:**

* Find the key area, gaps identified in the topic survey where the project can add value to the customers and business.

Not able to predict the customers who will default the credit card bill.

A Credit Card company needs to improve the ability to predict whether or not its customer will default, as well as recognise the factors that influence this probability and analyse the risk associated with the payment of credit card bills by customers.

* What key gaps are you trying to solve?

To predict weather the customer will default their next credit card pay cycle beforehand depending upon the data.

# **Methodology to be Followed:**

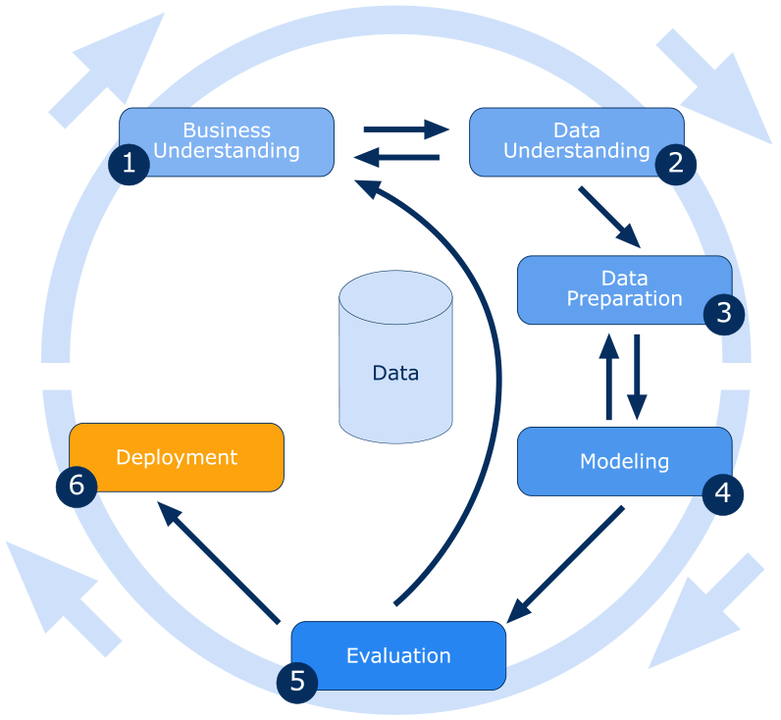


Fig 1

* Business Understanding:

Predicting the risk of a customer defaulting on their credit card bill next month.

* Data Understanding:

The proposed dataset has the payment history of particular account holders and consists of 30000 rows and 26 rows for the analysis.

* Data Preparation:

To clean the data according to the categorical features such as Gender, education and marital status as well as the continuous features such as age and credit limit.

* Modelling:

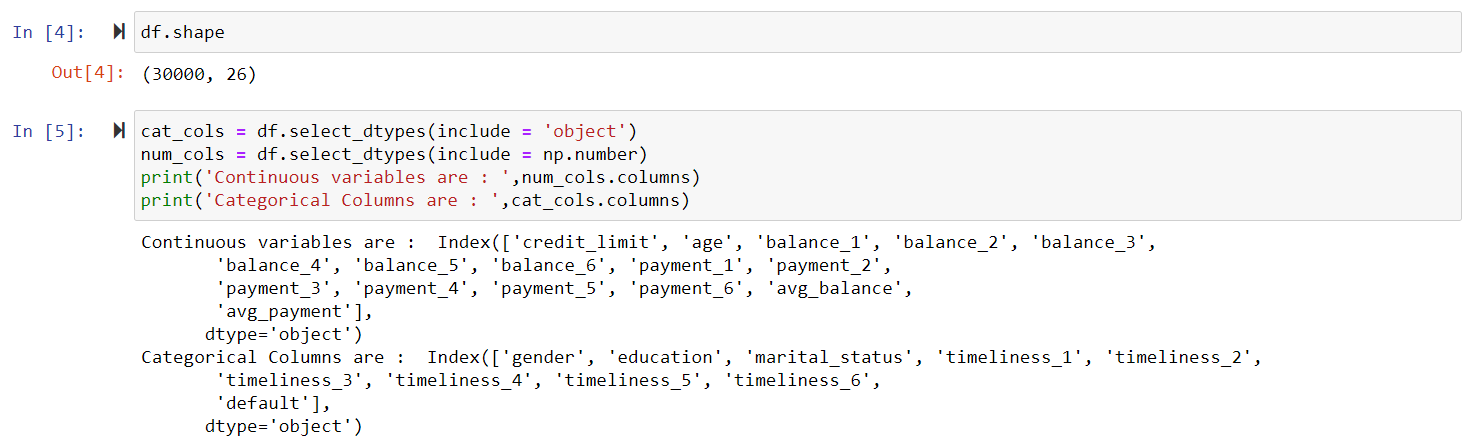
Visualize the dataset in regards with the features that are used in analysis.

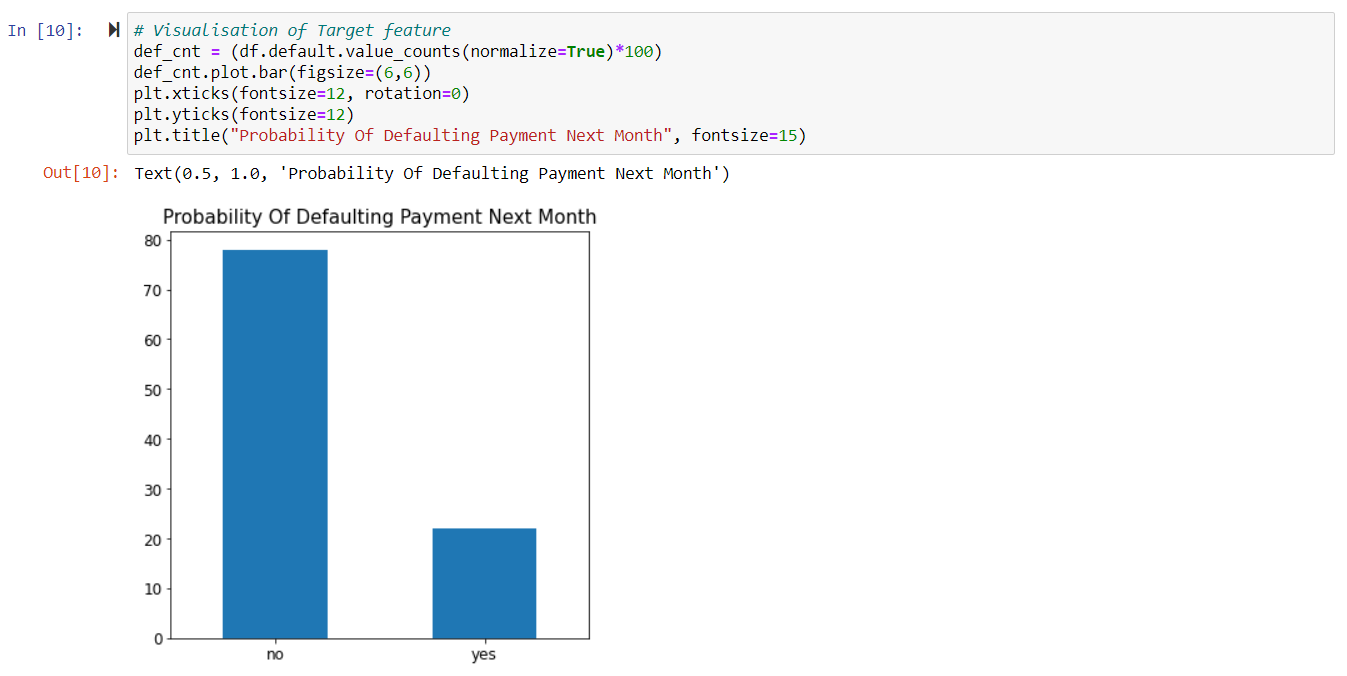
* Evaluation:

Applying the techniques of cross validation, etc.

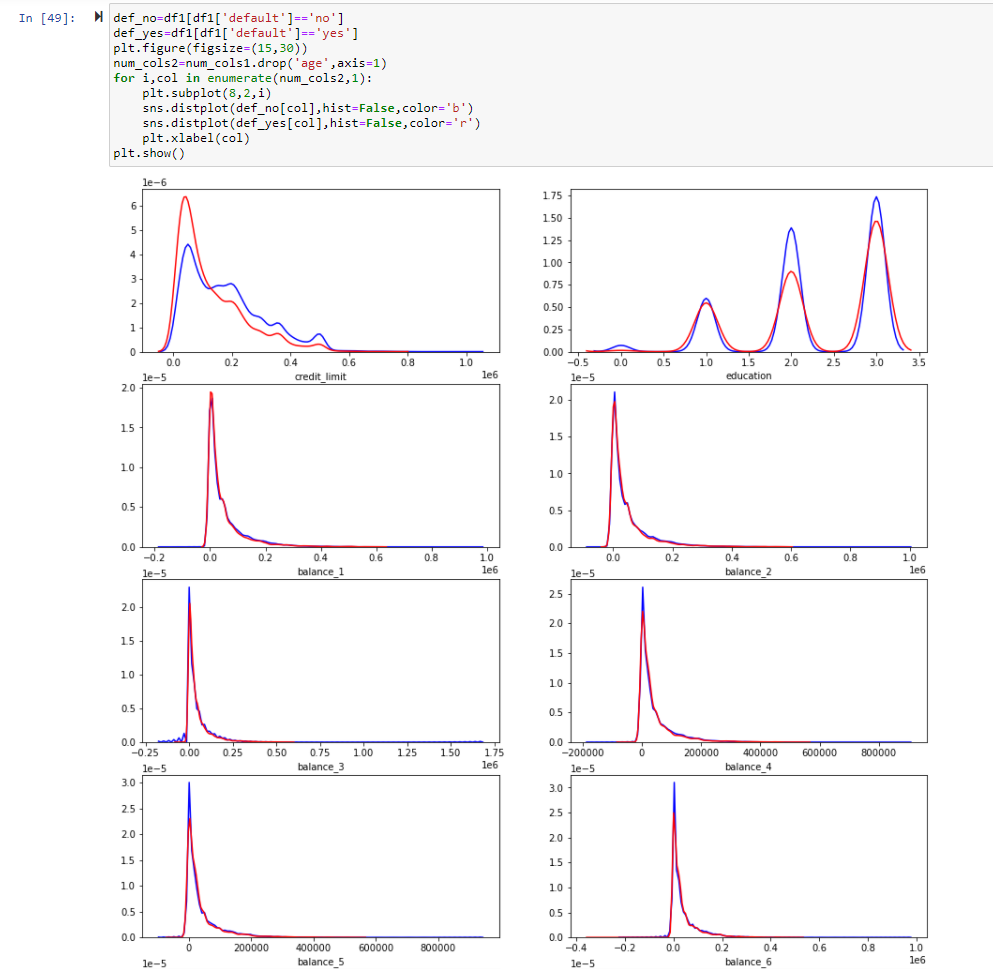
**Overview of the final process:**

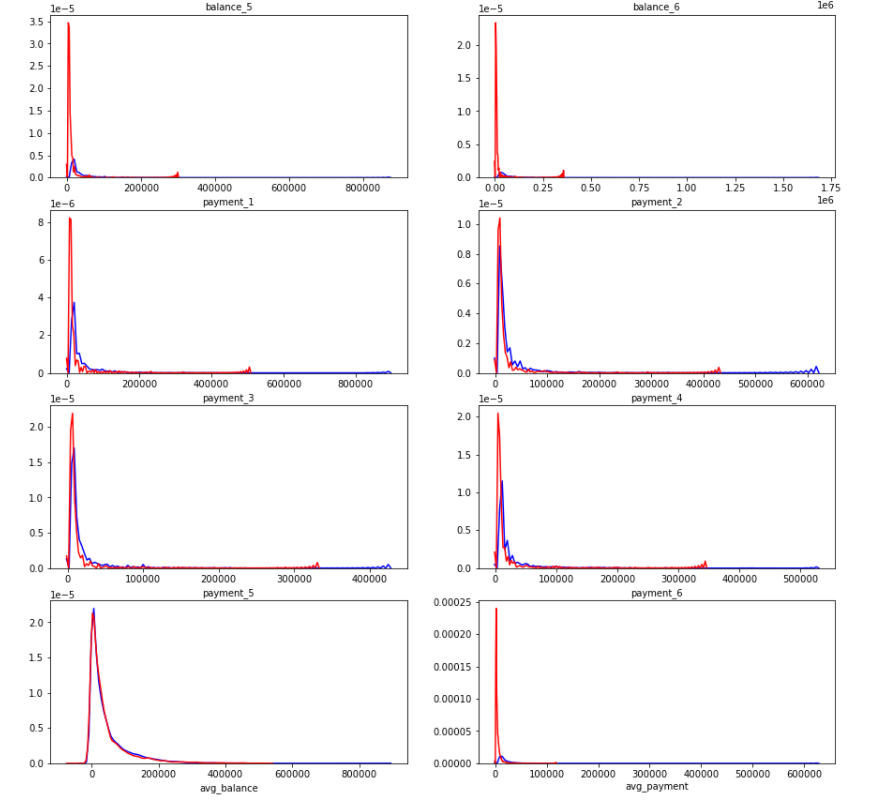
* First, we used various visualization techniques to understand the dataset.







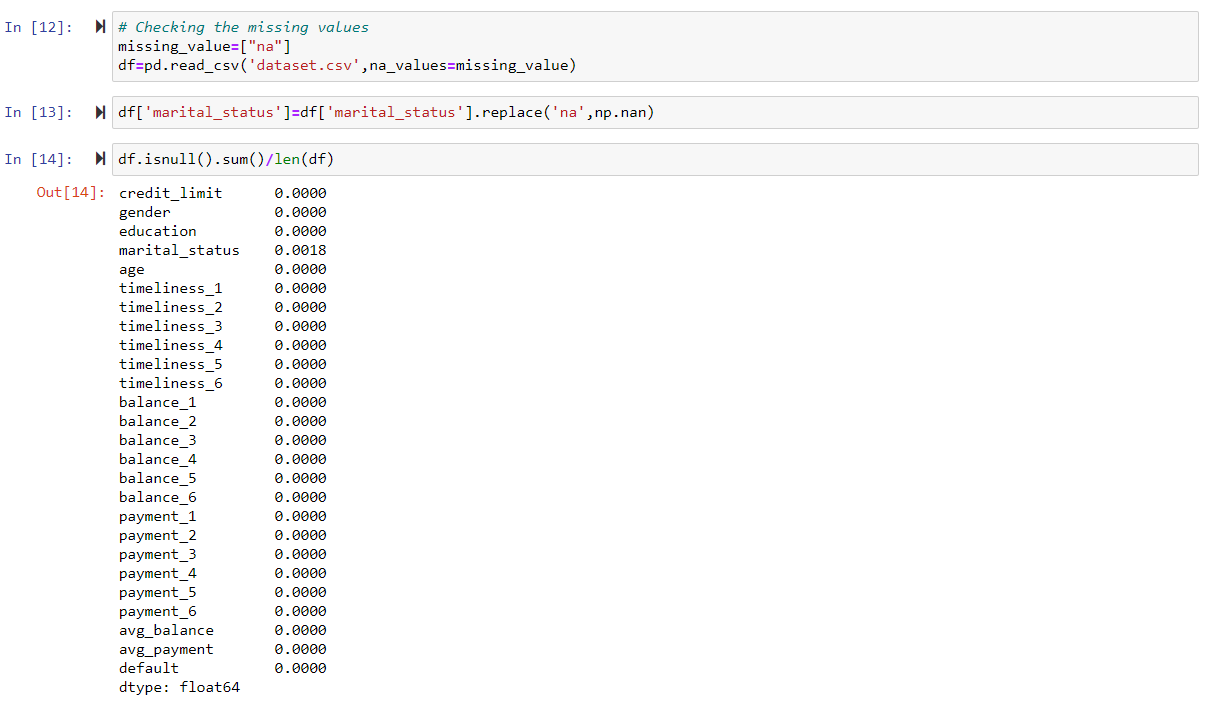




Distribution of various features with the target variable.

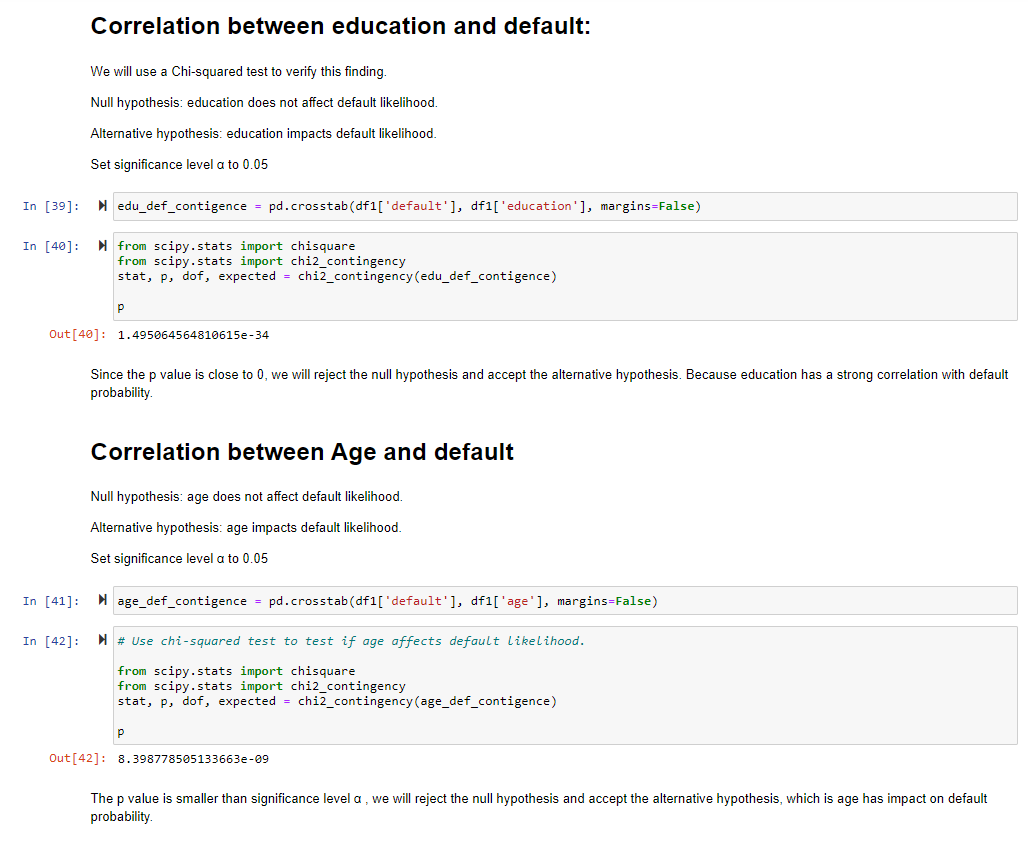
* Missing value treatment and null value treatment is done and the missing values are treated.

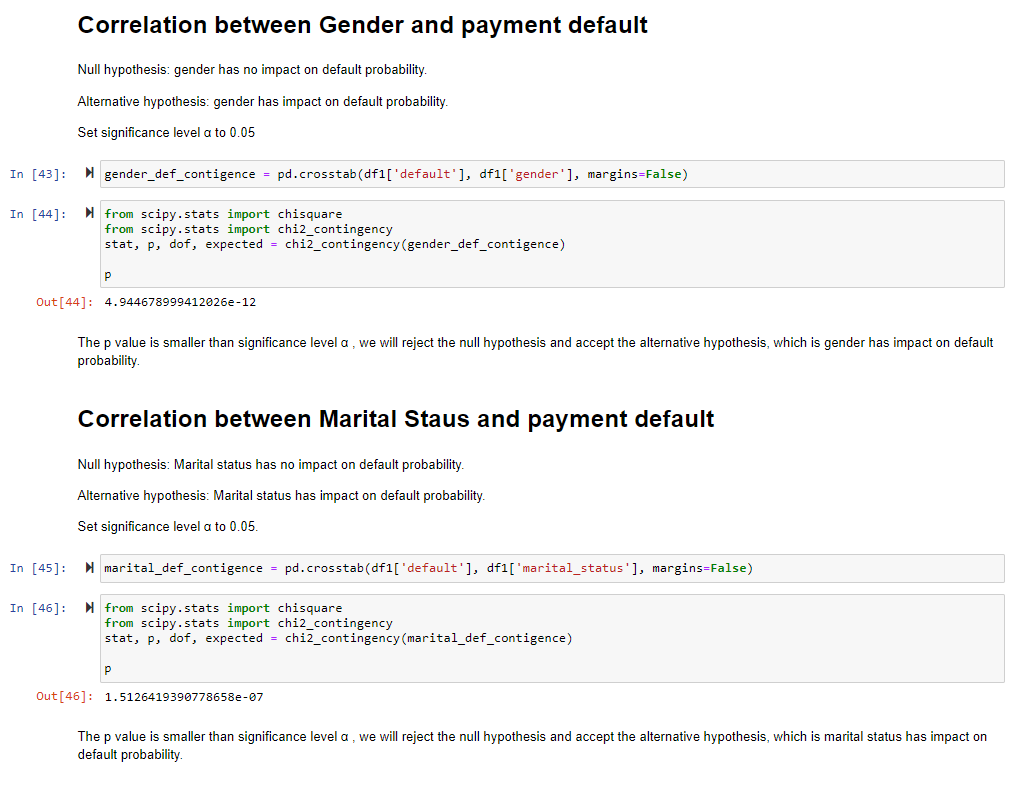


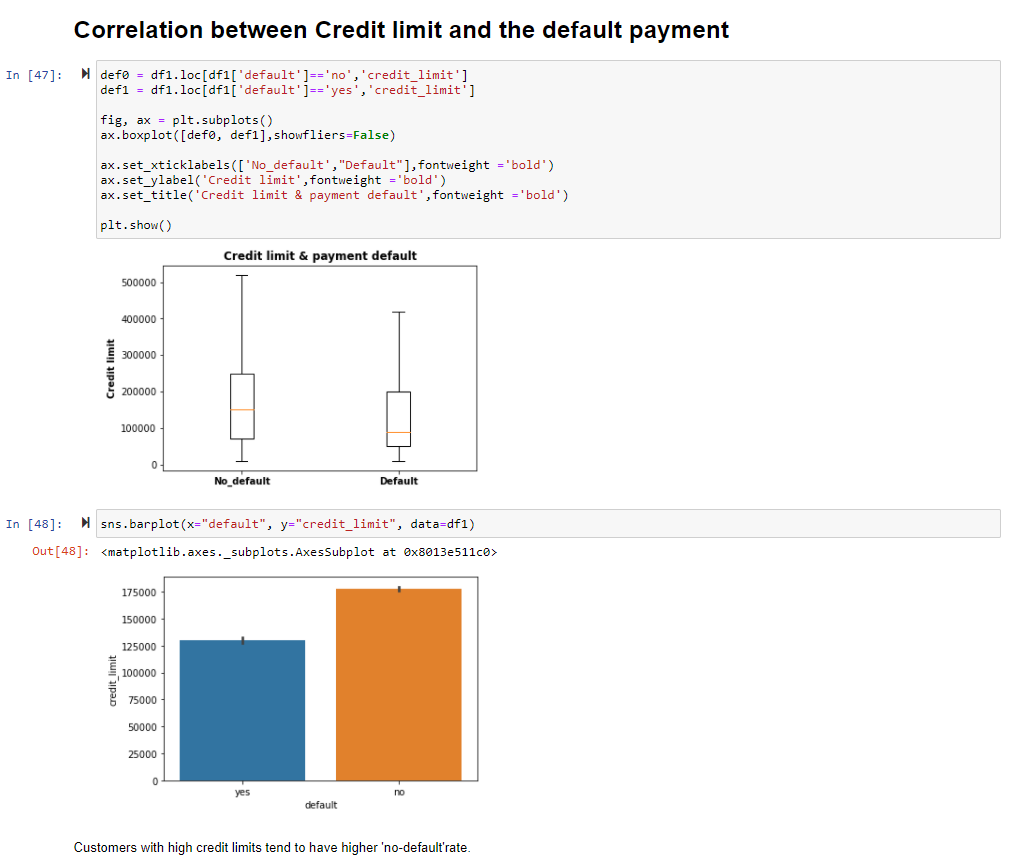


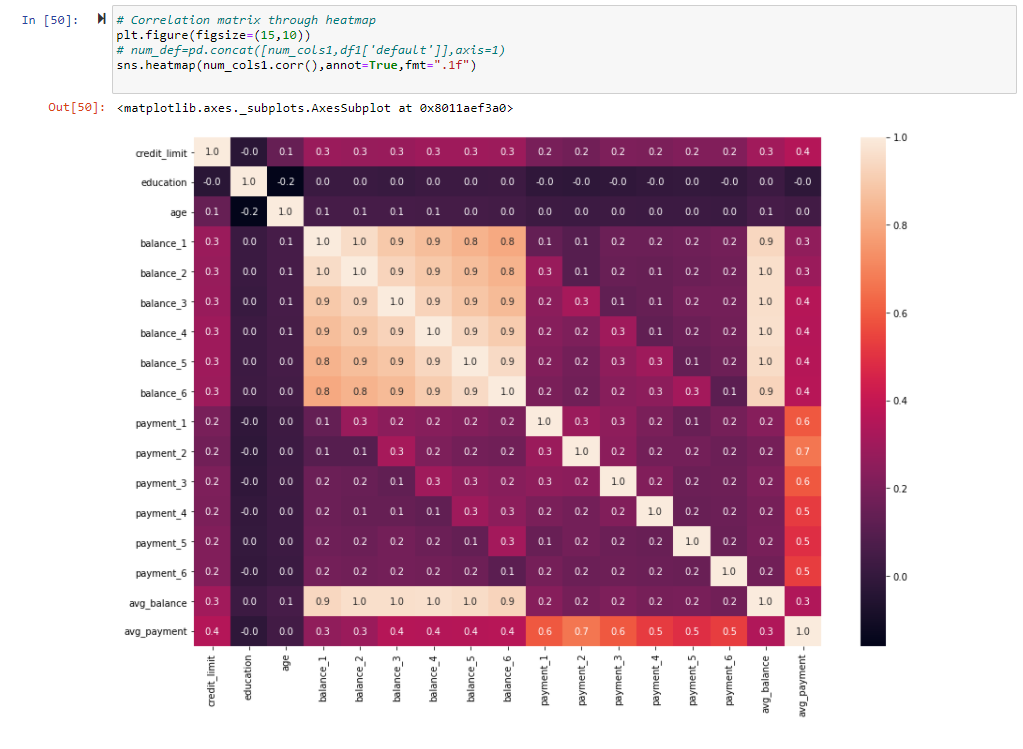


* The correlation matrix is plotted to check any correlation between various features so that necessary steps can be taken to handle this problem. Chi-square test was performed to check the impact of categorical variables on target variable.

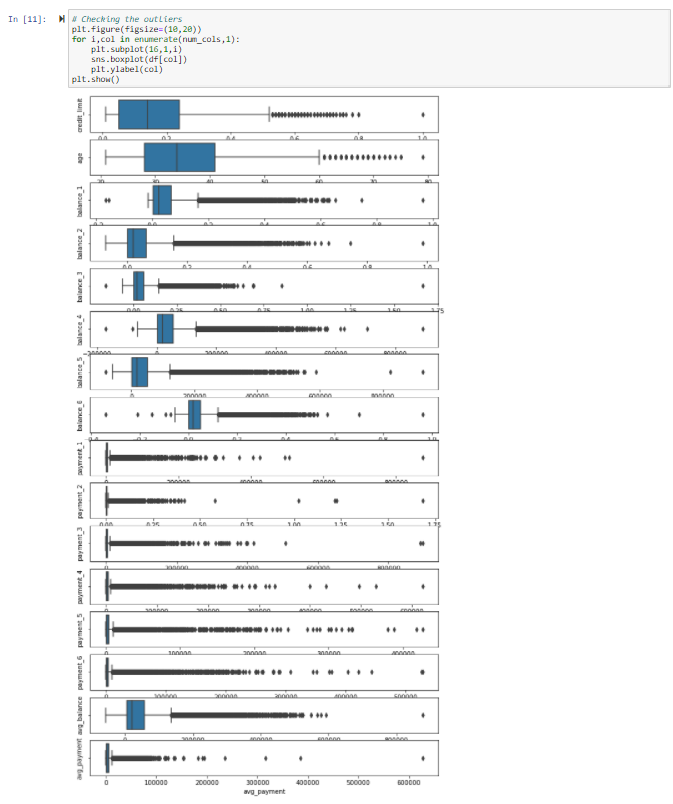


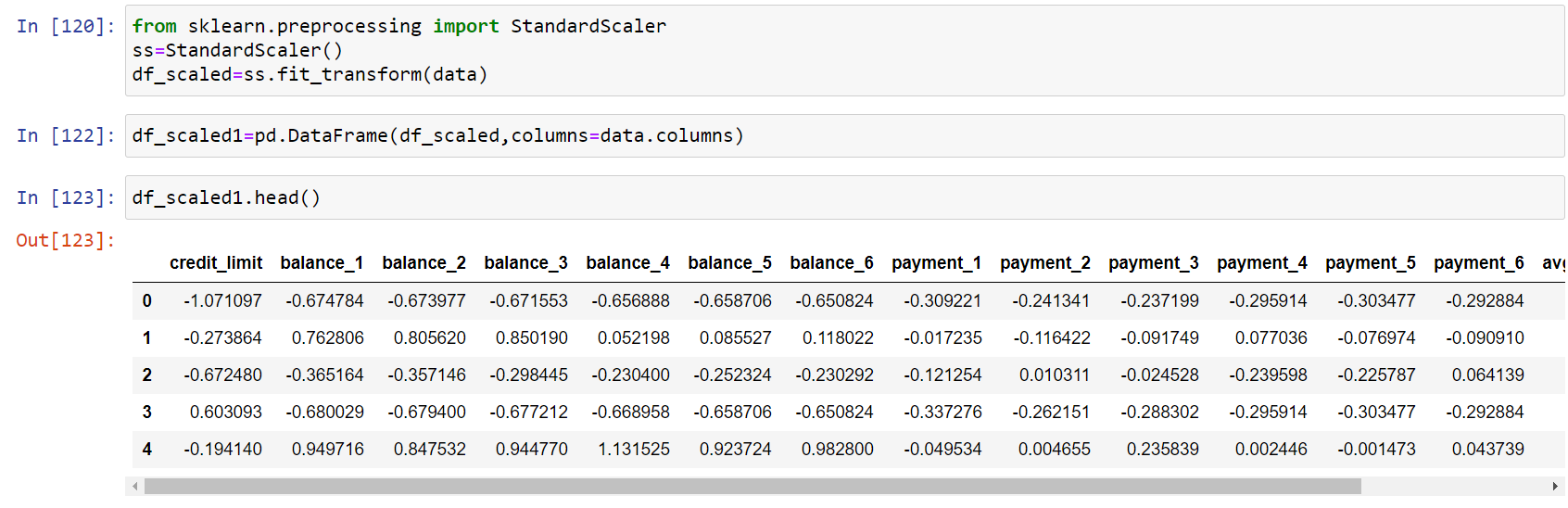






* We will check the data set to check if it needs any outlier treatment. Outliers’ proportion is huge so not removing them. Also, as the data is valid, these are extreme values. Hence, we are not treating the outliers.



* In our approach we have used various feature scaling and data transformation techniques to scale the data.
* 

**Model Evaluation:**

Describe the final model (or ensemble) in detail. What was the objective, what parameters were prominent, and how did you evaluate the success of your models(s)? A convincing explanation of the robustness of your solution will go a long way to supporting your solution

Base Models: Random Forest and Logistic Regression:

Logistic regression and Random forest algorithms are picked up for dummy models.

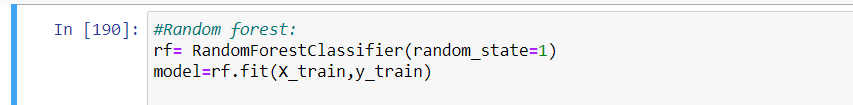
A couple of factors favourable for Random Forest for this dataset are as follows:

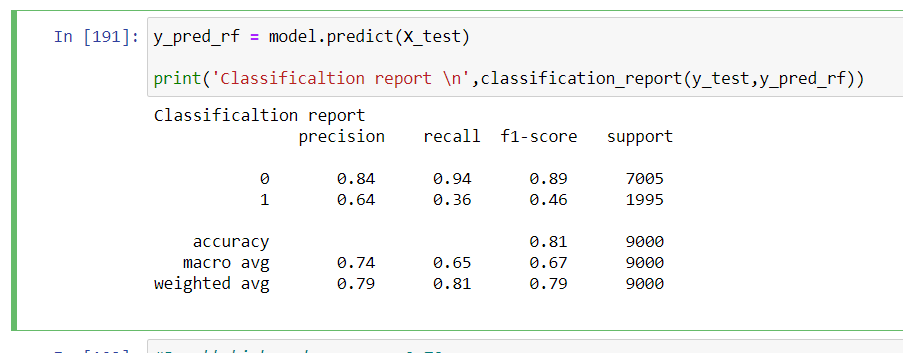
- The outlier’s proportion is huge in our dataset which is 1:2.

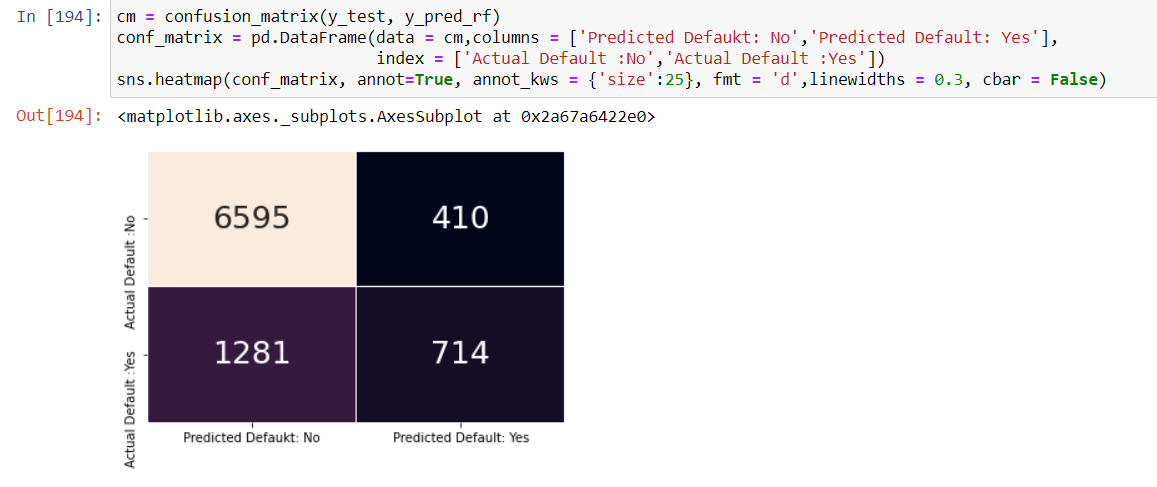
- While working on the transformation and scaling of the dataset including outliers, as our dataset contained valid negative and '0' values, we tried cube root transformation. However, the distribution of the features was getting changed.

Random Forest is robust to outliers and scaling is also not required. Therefore, prima facie Random Forest seems to be a more effective algorithm on this dataset.

Base model: Random forest:







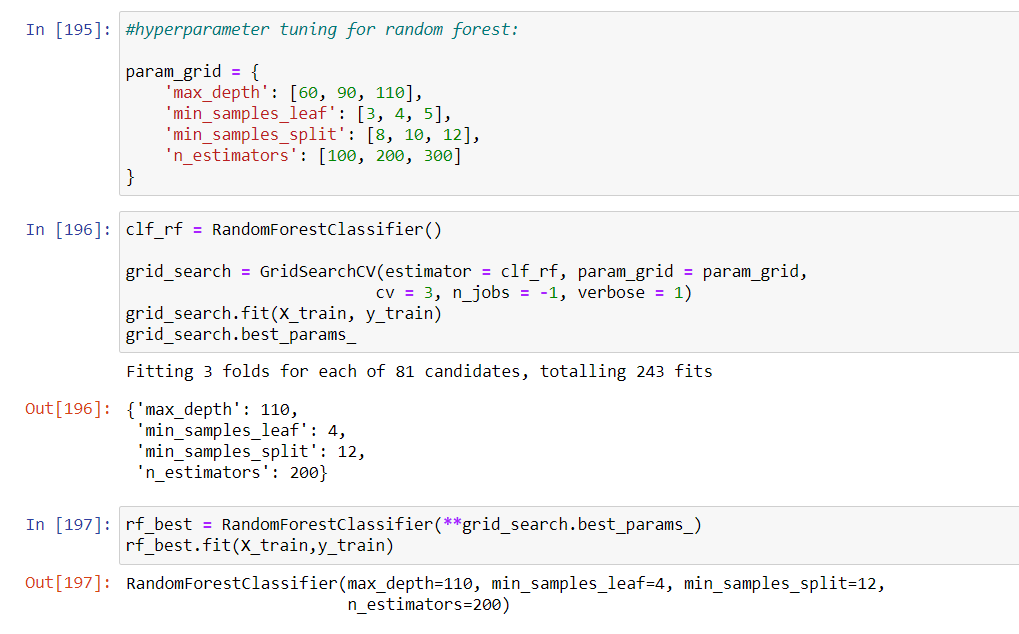
Metrics after running the base models:

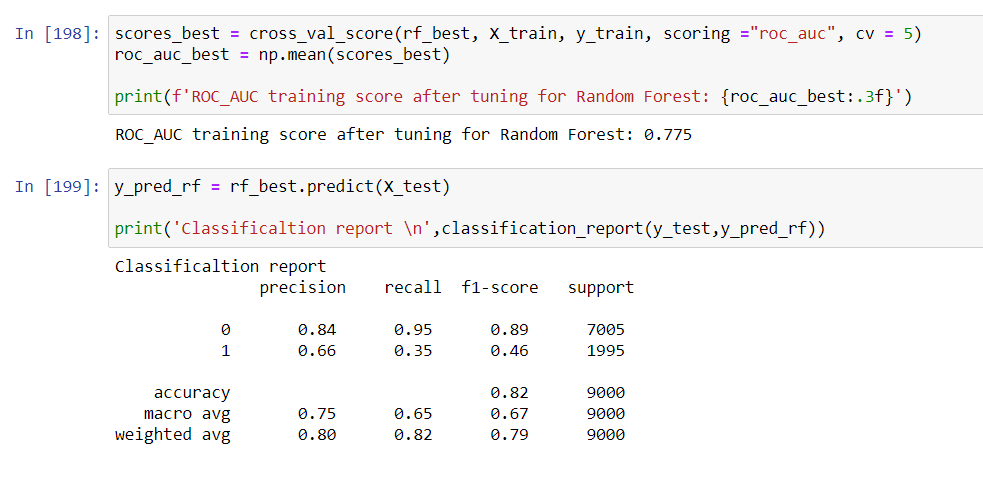
- Accuracy for both Logistic regression and Random forest is 0.81.

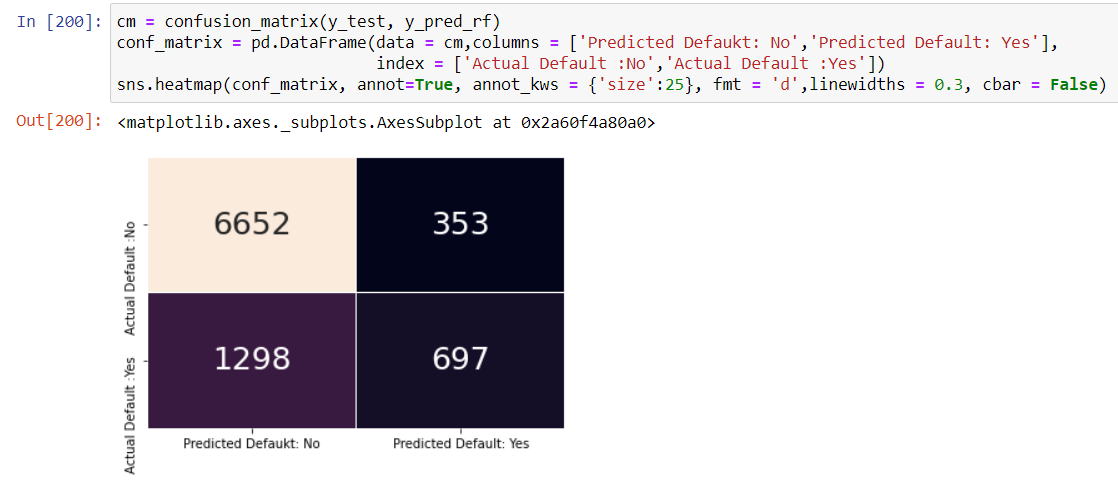
- Recall for Logistic regression is 0.97 and for Random Forest it is 0.94. However if we compare False Negative (Actual defaulter but as per model one is non-defaulter) of both models, False negative for Random Forest is much lower i.e. 1281 as compared to that of Logistic Regression which is 1508. Here as our focus is on

reducing the False Negatives, we have proceeded with Random Forest as our algorithm for the final model.

Final model: Random forest (with hyper parameter tuning):

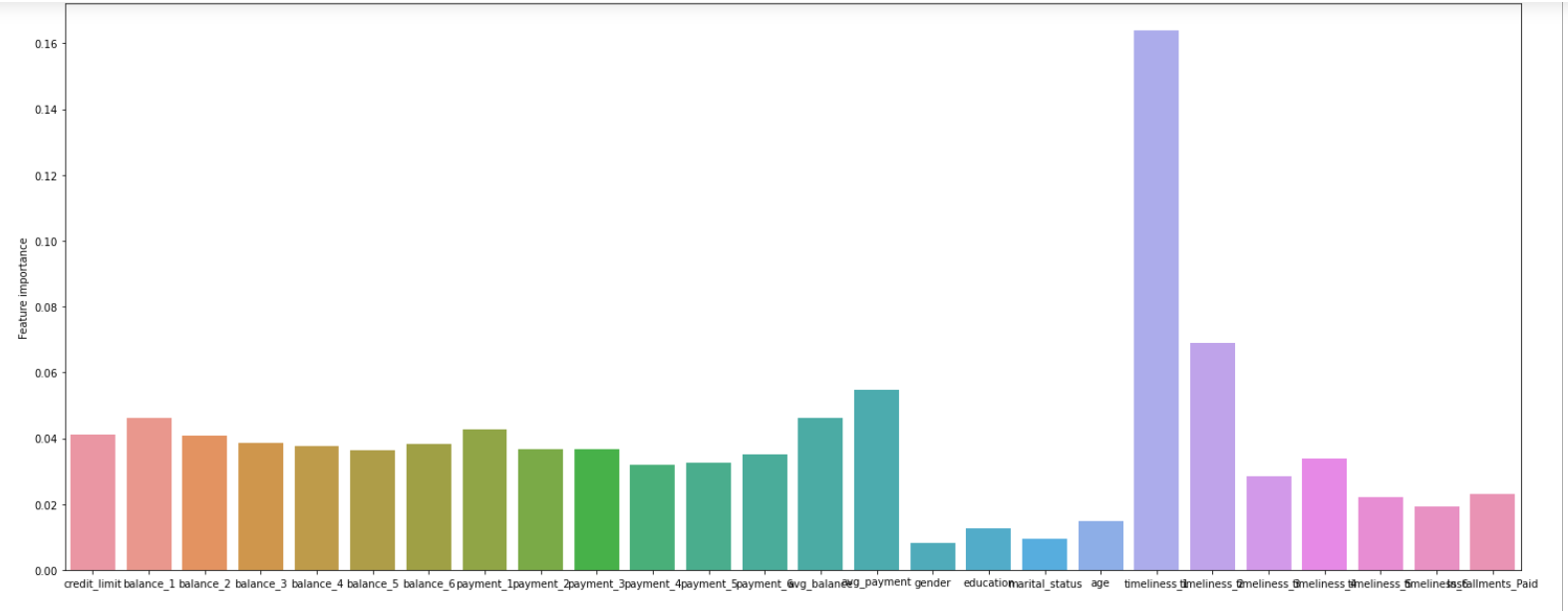






Inference: There is slight increase in the accuracy of the model after hyper parameter tuning.

Feature importance:

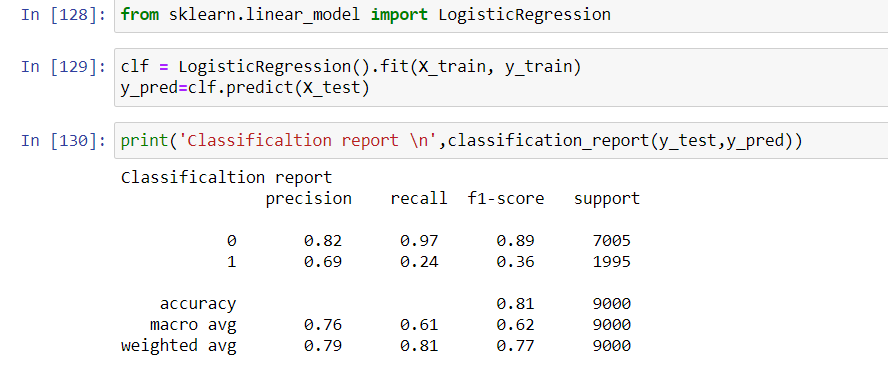


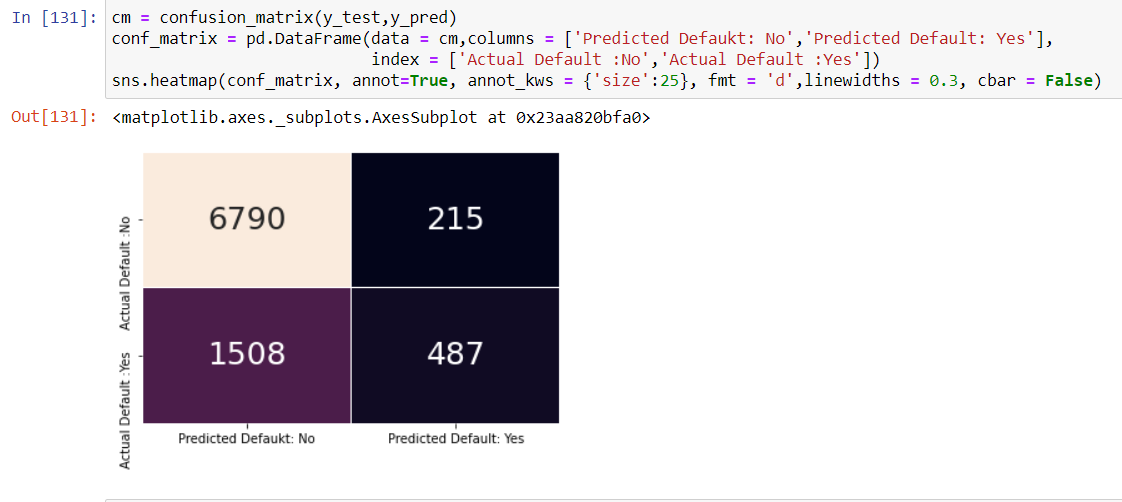
Here, From the above analysis we can see that the timeline of payment is the most significant feature to decide whether the customer will default the credit card payments or not.

Subsequently, the balance in different months significantly tells whether the customer will default the next payment or not.

**Comparison to benchmark:**

Comparison to Logistic Regression model:





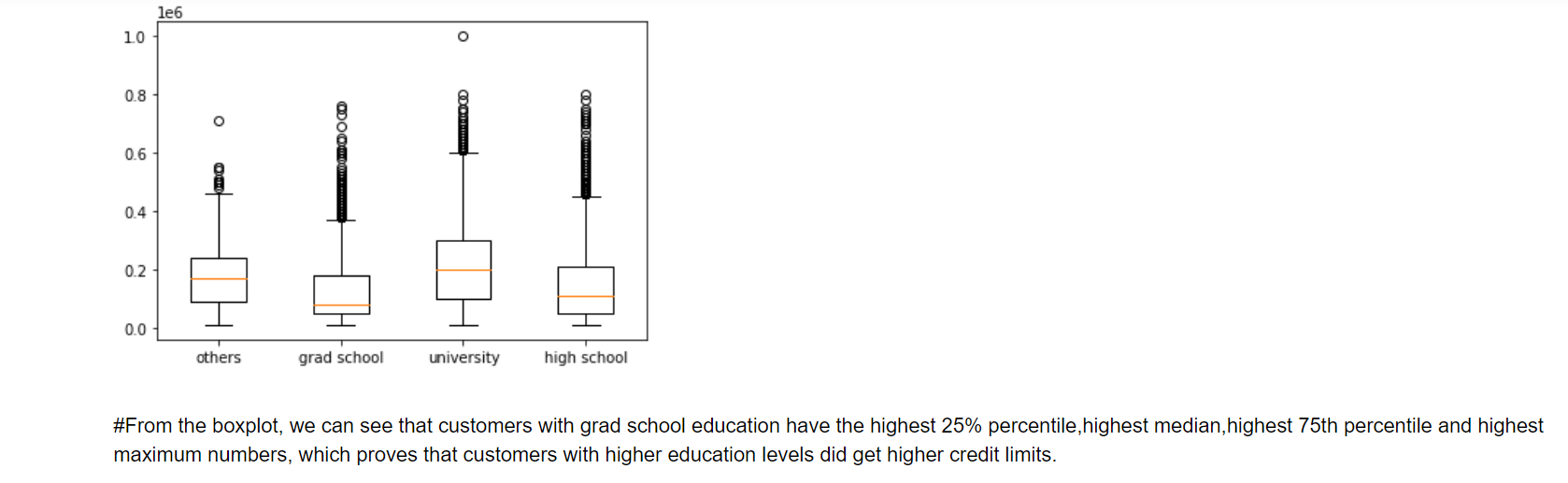
In the final model there is a significant increase in the accuracy of the model.

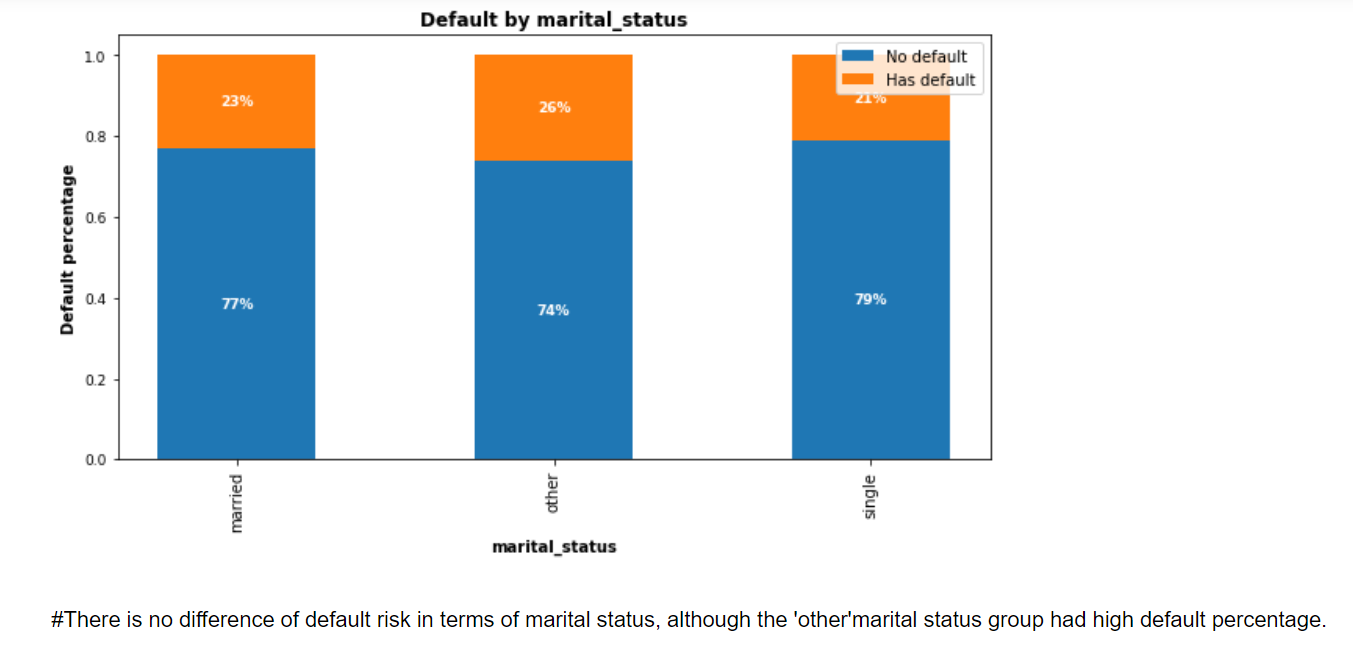
The main parameter i.e actual default vs predicted as ‘’No’’ has been reduced from 1508 to 1281. Hence increasing the reliability of the model.

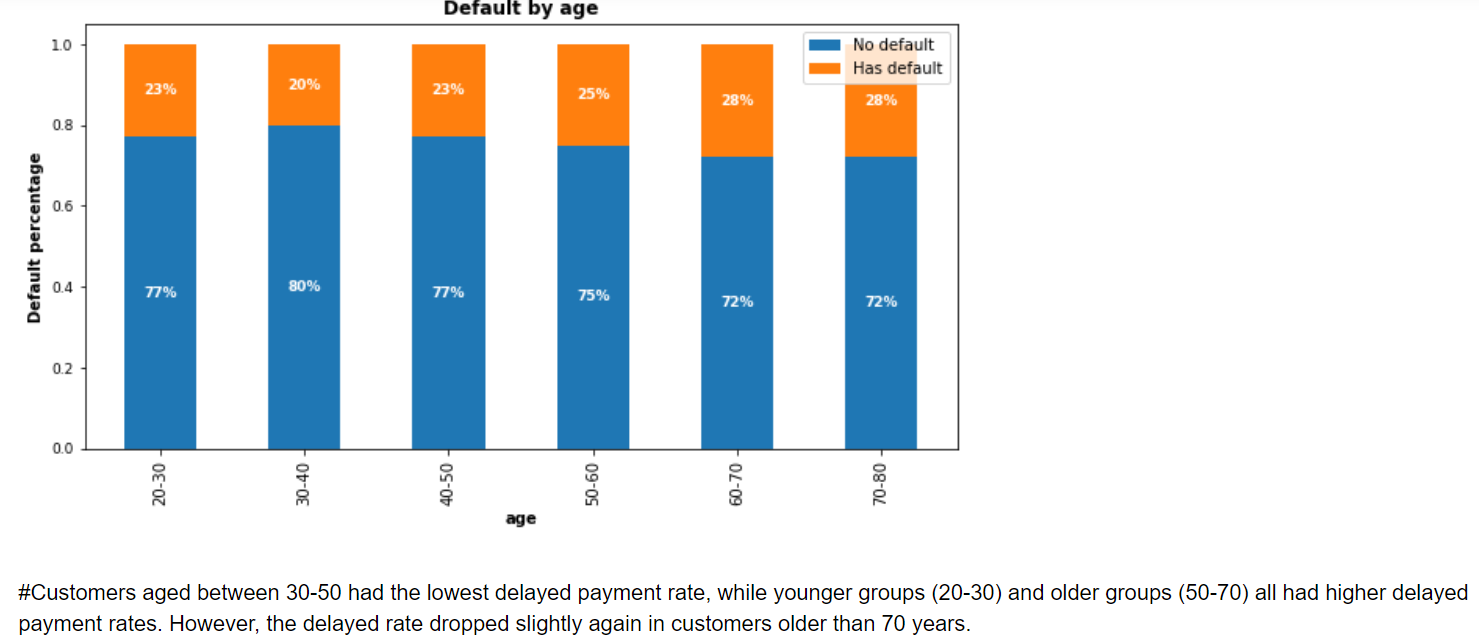
There is a significant decrease in the above parameter, since we are not treating the extreme values or the outliers, the extreme values may have affected the evaluation in logistic regression. But in random forest, due to the concept of bootstrap aggregation /majority voting might have resulted in better classification of default class and No-default class. Moreover, the random forest classifier is a tree-based algorithm and not based on distances of the data points.

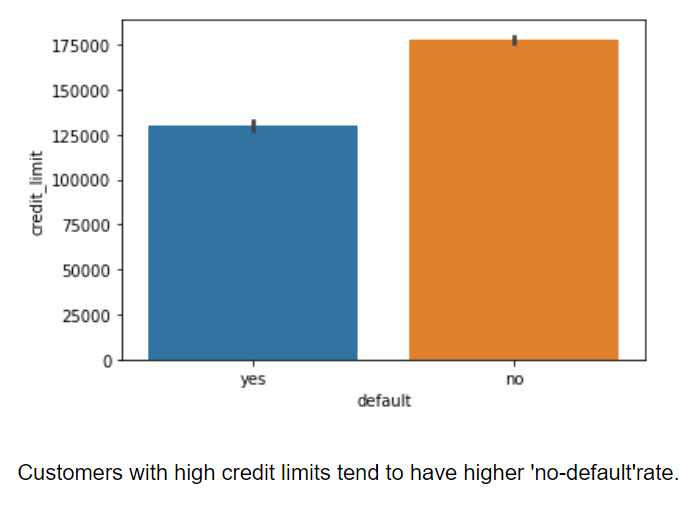
**Visualizations:**

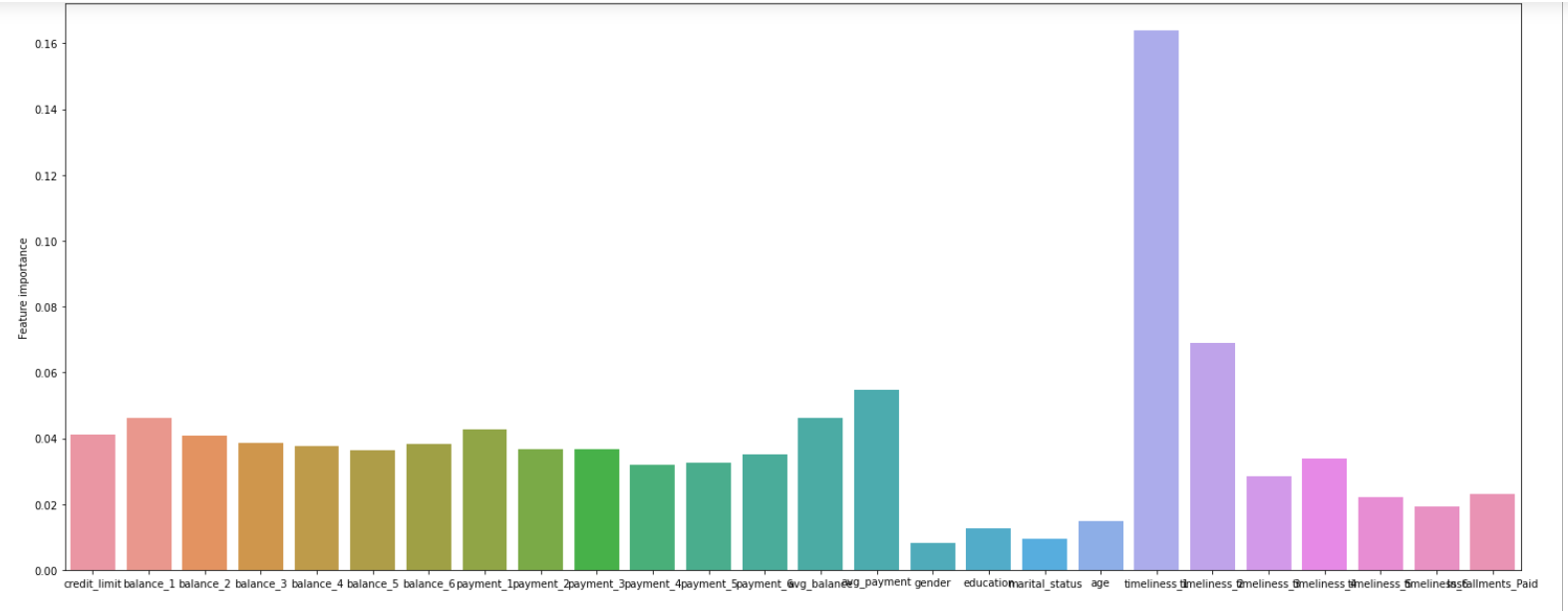
* **Various visualization plots:**











The above plot shows the significant features in accordance with their weightage impacting the target variable.

**Implications:**

The model is able to predict whether a customer will default the next payment cycle. By this the banks or the credit card organizations can predict/get informed about a customer defaulting the payment can take actions accordingly.

The model accuracy was about 82% i.e., in the majority of the situations the model is able to predict correctly. This model can inform the authorities about the customer behaviour/payments and the authorities can take business decisions effectively.

**Limitations:**

The accuracy obtained for the final model was 82% i.e. still there is a chance that the customer is defaulting the credit card payment and the model is not able to predict the same.

Case: If the credit limit is more for a particular customer and if that customer is defaulting and the model is not able to predict then there could be huge loss to the credit card companies or banks.

To enhance the solution, cascading of models could be achieved in which in each stage some data points or customers could be correctly classified. At the end there could be a representative checking the defaulters list. This can make sure that no customer or very less customers are incorrectly classified.

Estimation to the above process: There will be very few customer lists to check at the end.

**Closing Reflections:**

What have you learned from the process? What would you do differently next time?

We learned about the various parameters that are taken into consideration while allocating credit cards to the customers.

As the dataset contains huge extreme values which are valid data points, therefore we have learned how to work on such a dataset.

Here as the number of features are 26 and also the most of the features (Balance, payments, timeliness) are affiliated to each other, we could not implement the feature selection techniques. Therefore, there is a sheer possibility of the model being impacted by the curse of dimensionality due to the high number of features. Therefore, next time we would want to proceed with dimensionality reduction techniques to see the impact of them over the model performance.