**CREDIT CARD DEFAULT PREDICTION**

**SARANYA . N**

**Data Science Trainee**

**AlmaBetter , Bengaluru**

**Abstract**

Credit lending business is one of the trending businesses in this contemporary era. Banks, NBFC and other lending institutions are interested to lend money as loan to increase their profit in their business. But one of the major problems such institutions face is default of repayment of their respective loan amount from their respective customer. So for such an institution, it is very important to understand who are good customers and who are bad customers. In the present market, there are two types of customers, which are willing customers and unwilling customers. Willing customers are good for the company. Lending institutions are more interested in pumping more money so that they can have more profit. They also provide much incentive to those customers for increasing the sales. Banks and such institutions require precarious to avoid unwilling customers so that they can release such burden.

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faces by commercial banks is the risk prediction of credit clients. Recent studies mostly focus on enhancing the classifier performance for credit card default prediction rather than an interpretable model. In classification problems, an imbalanced dataset is also crucial to improve the performance of the model because most of the cases lied in one class, and only a few examples are in other categories. Traditional statistical approaches are not suitable to deal with imbalanced data. In this study, a model is developed for credit default prediction by employing various credit-related datasets. There is often a significant difference between the minimum and maximum values in different features, so Min-Max normalization is used to scale the features within one range. Data level resampling techniques are employed to overcome the problem of the data imbalance. Various under sampling and oversampling methods are used to resolve the issue of class imbalance. Different machine learning models are also employed to obtain efficient results. We developed the hypothesis of whether developed models using different machine learning techniques are significantly the same or different and whether resampling techniques significantly improves the performance of the proposed models. One-way Analysis of Variance is a hypothesis-testing technique, used to test the significance of the results.

Keywords: Machine learning, Deep learning, Feature Engineering, Accuracy, Data Mining, Validation.

**Problem statement**

This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the [K-S chart](https://www.listendata.com/2019/07/KS-Statistics-Python.html) to evaluate which customers will default on their credit card payments. Credit card debt results when a client of a credit card company  purchases an item or service through the card system. Debt accumulates  and increases via interest and penalties when the consumer does not pay  the company for the money they have spent.

**Data description**

Attribute Information:

This research employed a binary variable, default payment (Yes = 1, No = 0), as the response variable. This study reviewed the literature and used the following 23 variables as explanatory variables:

● X1: Amount of the given credit (NT dollar): it includes both the individual consumer credit and his/her family (supplementary) credit.

● X2: Gender (1 = male; 2 = female).

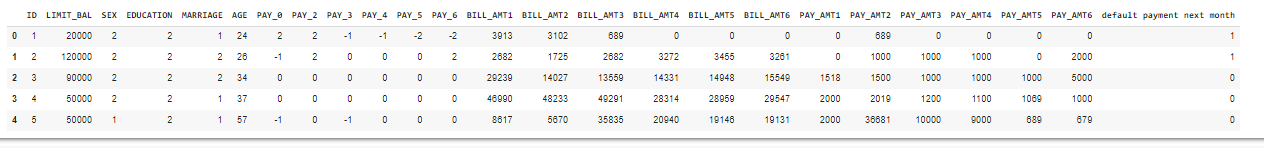
● X3: Education (1 = graduate school; 2 = university; 3 = high school; 4 = others).

● X4: Marital status (1 = married; 2 = single; 3 = others).

● X5: Age (year).

● X6 - X11: History of past payment. We tracked the past monthly payment records (from April to September, 2005) as follows: X6 = the repayment status in September, 2005; X7 = the repayment status in August, 2005; . . .;X11 = the repayment status in April, 2005. The measurement scale for the repayment status is: -1 = pay duly; 1 = payment delay for one month; 2 = payment delay for two months; . . .; 8 = payment delay for eight months; 9 = payment delay for nine months and above. ● X12-X17: Amount of bill statement (NT dollar). X12 = amount of bill statement in September, 2005; X13 = amount of bill statement in August, 2005; . . .; X17 = amount of bill statement in April, 2005. ● X18-X23: Amount of previous payment (NT dollar). X18 = amount paid in September, 2005; X19 = amount paid in August, 2005; . . .;X23 = amount paid in April, 2005.

**1.Data Preprocessing**

****

Data preprocessing is the beginning code part of a data science project. It starts with importing required libraries that are required for data preprocessing processes. Here , Matplotlib, seaborn, pandas numpy and warnings are imported as part of the prerequisite library for preprocessing the datasets. Data set is imported with the help of the read\_csv function from pandas packages. Then the datasets are investigated within the built function of python. Its investigation starts by inspecting the top 5 rows of the dataset to get overall understanding of datasets. Data set has a total of 30001 rows and 25 columns. Data types in all columns are in object d type as the first row of the dataset contains names of columns of datasets.

**2.Feature Engineering**

Feature engineering is an integral part of a machine learning project. About 60% of data scientists engaged with processing while they were building the data science project. It refers to a process of using domain knowledge to settle and transform the most relevant variables from raw material data when creating a predictive model using machine learning or statistical learning. When a raw dataset is given to a data scientist, it will not be in a properly structured manner. There will have many challenges like null values, outliers, etc... In this stage, he will solve all problems in datasets and convert datasets into machine interpretative form. When datasets are trained with a machine learning algorithm, data scientists make sure of their purity. ML model cannot train datasets that have the impurity like null value, object type, etc...So in the feature engineering step, he must make sure of the imputation of null value, conversion of object type data into numeric, etc.., so that he can train the model well. Such types of all processes are taken place in features engineering processes. Here we use some feature engineering techniques to clean datasets which are

1. Changing the column name.

2. Conversion of object columns to integers

3. Handling the missing values.

In our datasets, there are no missing values, it is not needed to take any such measures. This can be done with some libraries, those are present in python packages.

4. Dropping irrelevant values.

In the given datasets, ID columns don't have enough on predicting target columns. Even it doesn't have any relationship between them. So it is a good option to drop such types of variable, so that, computation speed will increase and error can be reduced.

**3.Explorative Data Analysis(EDA)**

Exploratory Data Analysis (EDA) is an approach to analyze the data using visual techniques. It is used to discover trends, patterns, or to check assumptions with the help of statistical summary and graphical representations. It is done with the help of a package of python by data visualization and their interpretations. In the given datasets, there are 25 feature columns and 300001 rows. Its distribution can be classified based on the following criteria

● 1. Gender

● 2. Marriage

● 3. Education

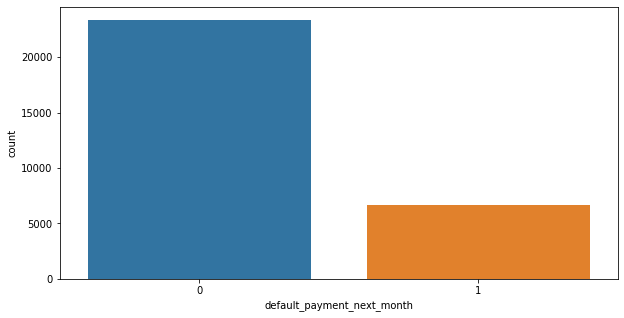
● 4. Limit Balance

● 5. Payment Status

● 6. Bill Amount

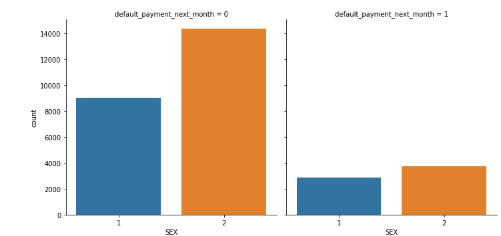
● 7. Payment Amount

**3.1 Loan Default Distributions**



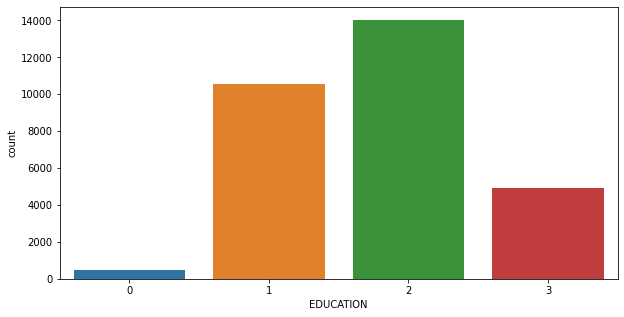
* The About 22% people are expected to default next month and 77.8% are not expected to default
* Defaulters are less than non-defaulters in the given dataset. So both the classes are not in proportion so we have to work on that to normalize data

**3.2 Gender Distribution**

****

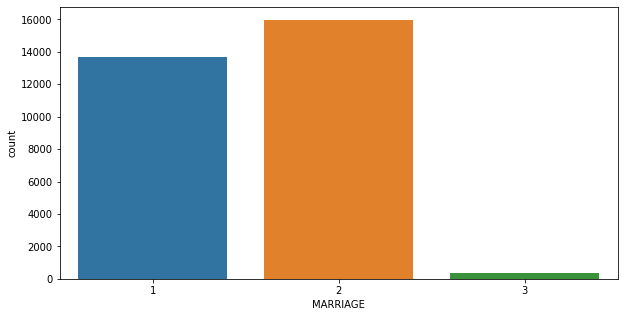
In the given dataset, there are a total of 30000 rows of data, of which 18112 are related to females. Here, 1 denotes male and 2 denotes female. The distribution of default payments and non-default are based on their respective contribution. Market share of contribution to the loan default is generally greater on male than females**.**

**3.3 Education Level**



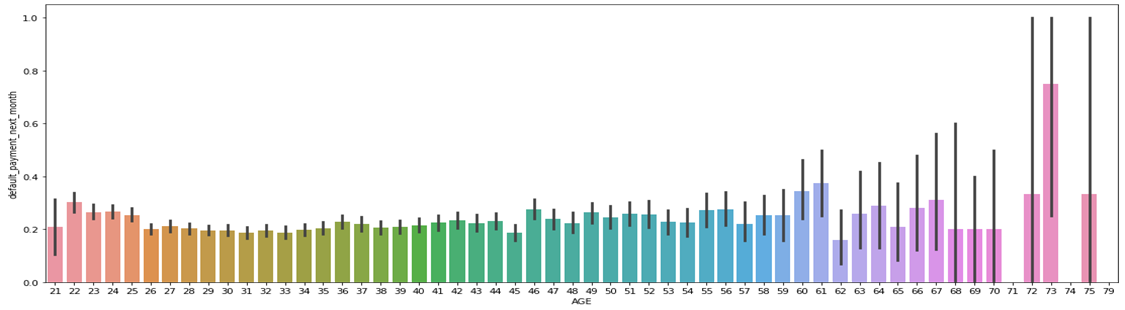
Customers who have the qualification of university degree person take high proportion than school graduate as well as high school graduates person More number of credit holders are university students i.e. around 47 percentage followed by Graduates and then High school students.

**3.4 Marriage**



The relationship of marriage in regard to default has an influence on predicting whether a person commits default or not. Marriage status can be divided on the given situation of single, married and other. Committing default has a greater chance for a married person than a single person. Share of single and married are higher than others.

**3.5 Age distribution**

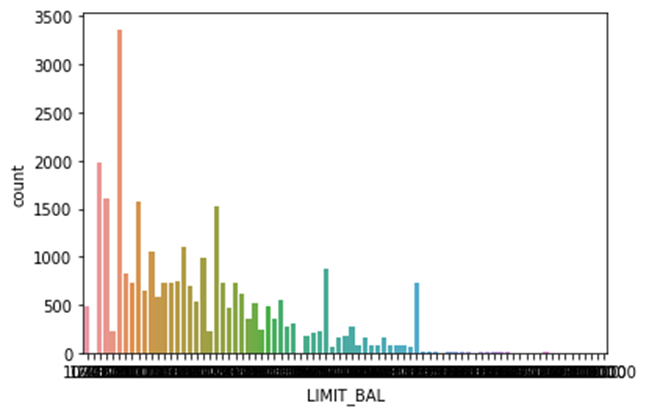


Age also has a role in whether a person commits default or not. It says the major business of credit cards is happening in ages between 26 and 35. We can make some observations that, as the age increase from 20, there is an exponential growth on having credit defaults and at age of 30, it stands at its peak. Also age above 60 years old rarely uses the credit card.

**3.6 Balance Limit**

Limit balance is the range of amounts, in which a customer can take

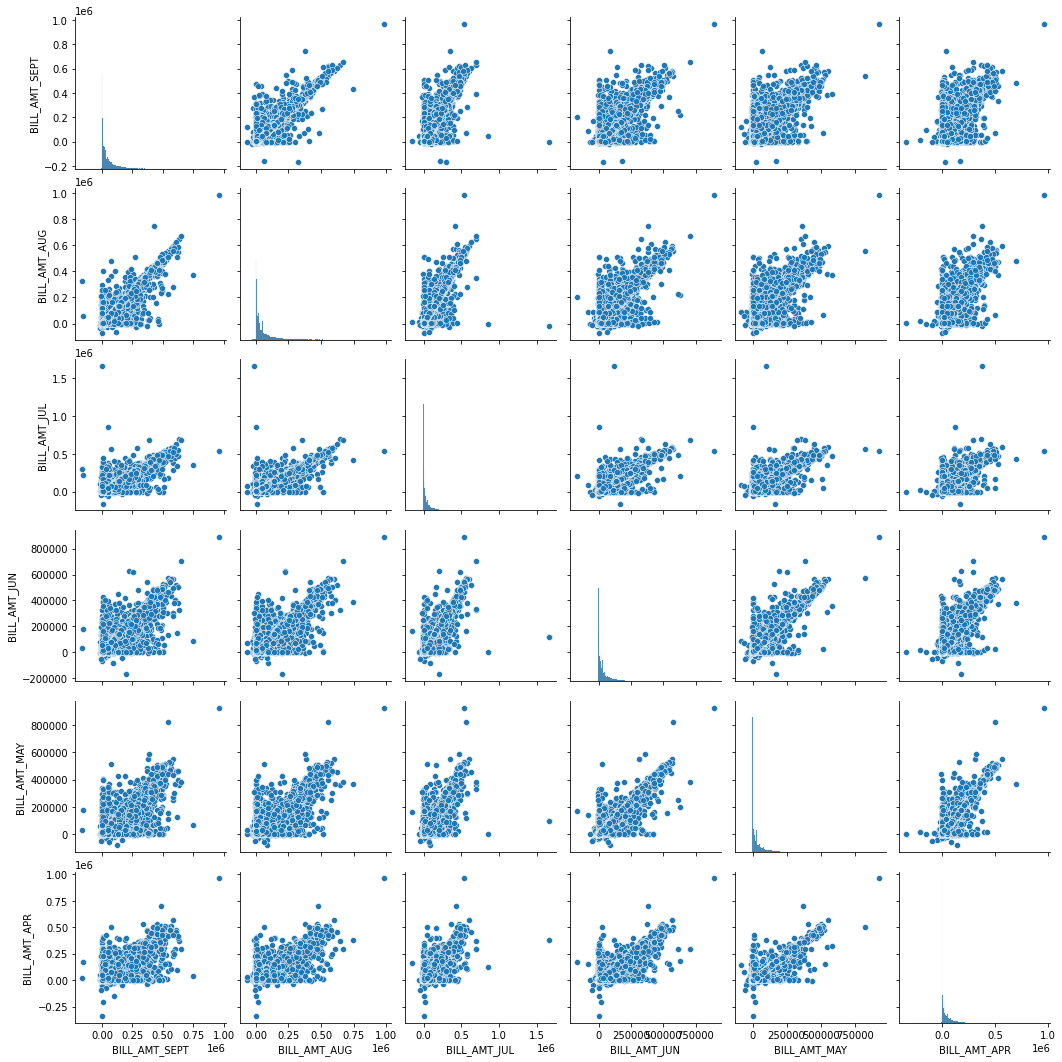
Credit from a lending institution. It includes both individual customers and companies. The credit limit varies from person to person. If a bank has a good perception of a person based on his repayment discipline, income, social status, etc., he will be led by a high margin. Banks give the high amount to such person due to high conviction on non-default and duly repayment



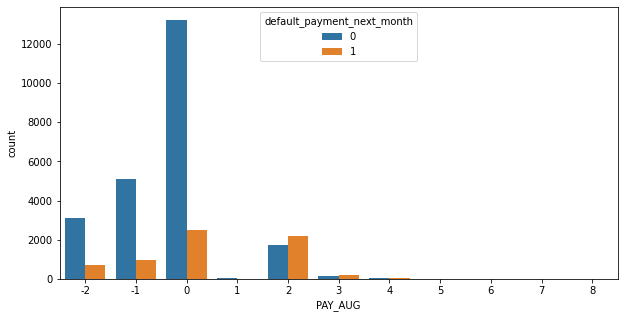
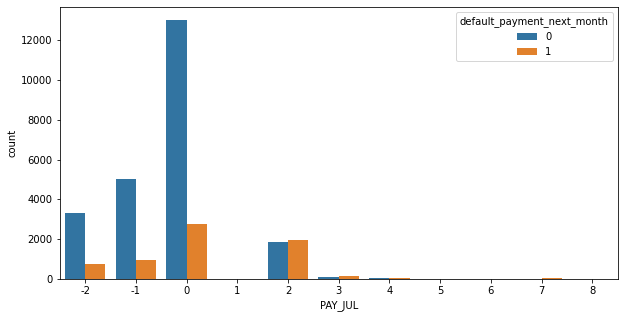
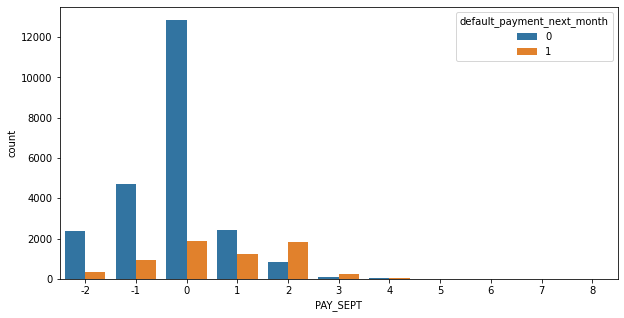
In the given datasets, its distribution also changes from person to person. The range of limit balance is between 10000 and, 1000000, and its mean is 167484. It has a standard deviation of 129747.

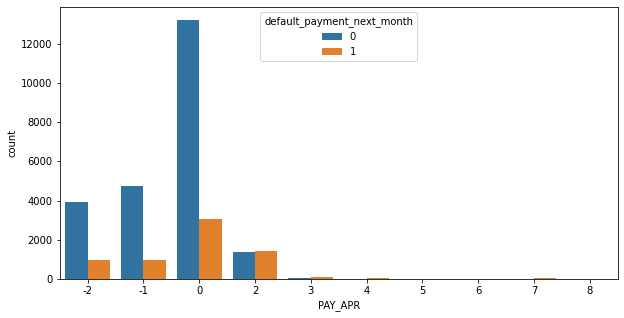
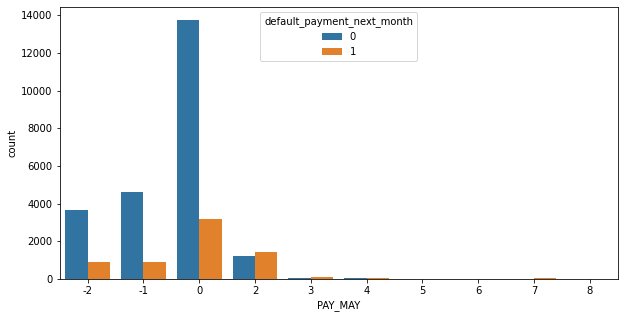
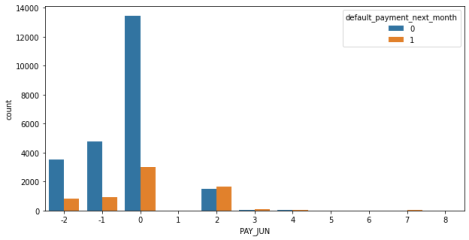
**3.7 Bill Amount**

Bill amount is the amount, due to be paid on the due date. The Bill amount is different from the payment amount, as it is the figure already paid by the customer. But in the case of bill amount, it is the amount. One of the observation, that can be derived from bill amount correlation is, that as the period elapse, its value also decreases. It shows people become lazy as times go on



**3.8 Payment Amount**

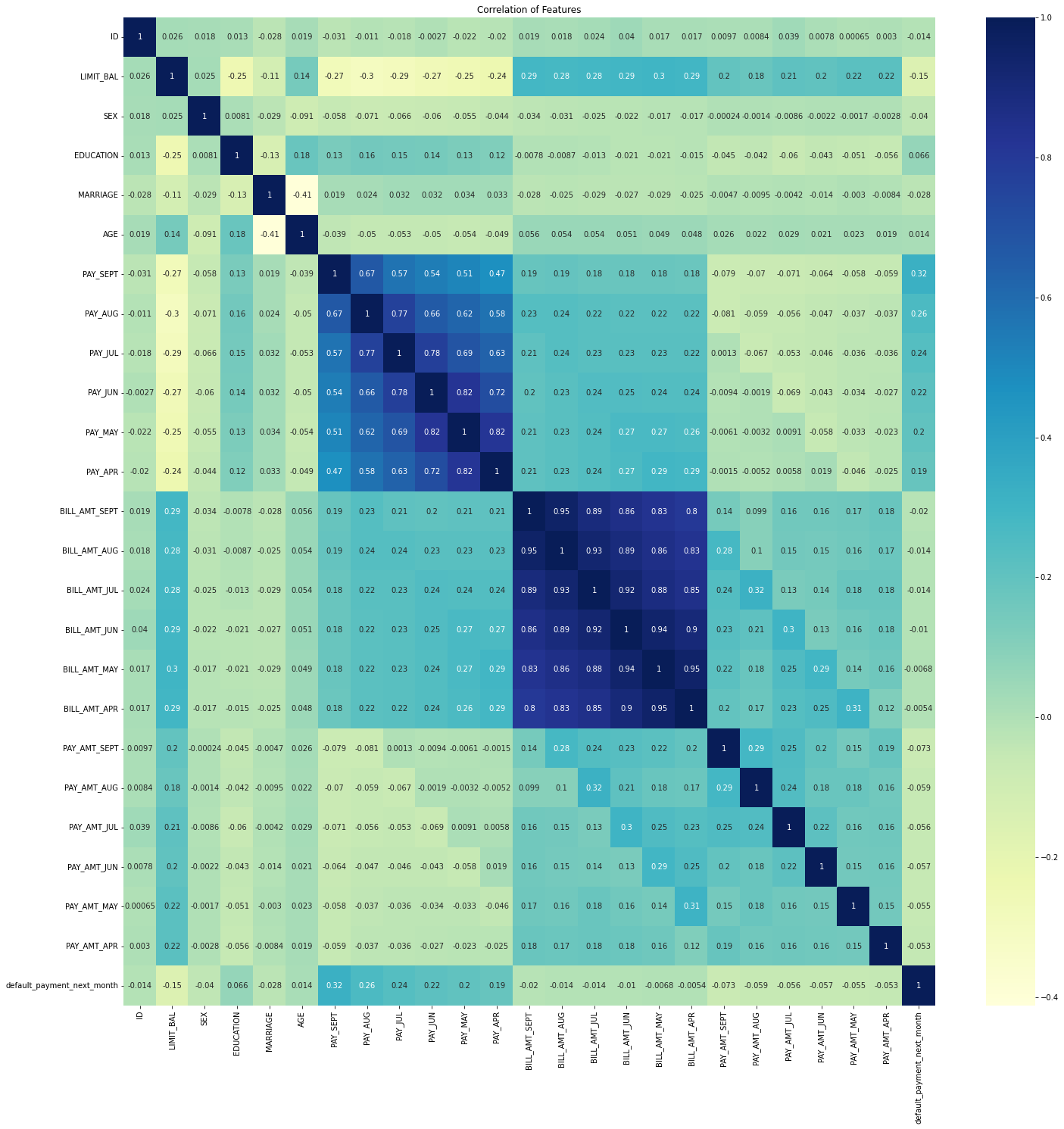
****



Amount payment is the amount of money paid by the customer on the due date. The payment amount is different from the bill amount, as it is the amount due to customers which is to be paid.

**3.9 Correlation Relationship**

Correlation is a mutual relationship or connection between two or more features. It is not necessary for correlation to be equal among some features to others. Some features may be more highly correlated to the target variable than others. Here, some are more correlated to each other. Payment status, payment amount, and bill amount are highly correlated with each other. From this correlation heat map there we can see that some of the values are negatively correlated. Like age , but we can't delete age because it is an important feature for further prediction The column ID also have some negative values it is not that much important for analysis , so we can remove it.



**4. MODEL BUILDING**

Model training is the process of fitting data into the required model after completing the data preprocessing part. This is one of the integral parts of machine learning problem-solving segments like classification and regression. This happens generally after the completion of feature engineering and explorative data analysis. In this project, it will train four of the model which are

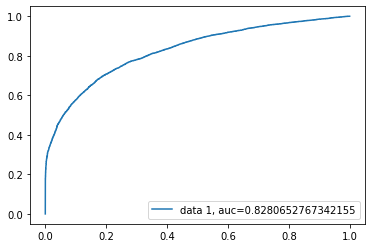
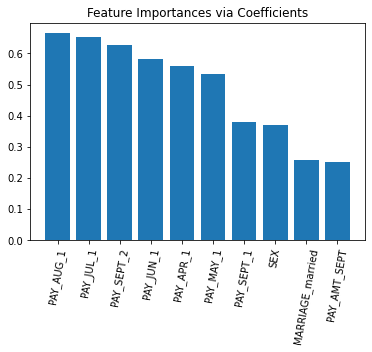
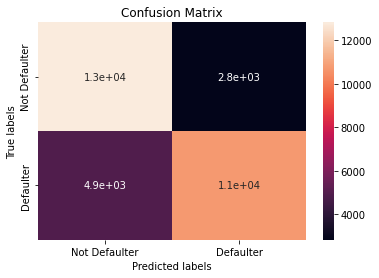
* LOGISTIC REGRESSION
* RANDOM FOREST
* SVM
* XGBOOST

**4.1 Logistic Regression**

Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of the target or dependent variable is dichotomous, which means there would be only two possible classes. Here, this technique is used to predict whether a person commits fraud or not. It has a trained model in two-phase.

**From this regression model we get the results as below:**

* The accuracy on train data is  0.7525473536269843
* The accuracy on test data is  0.7536476233707282
* *The accuracy on test data is  0.7536476233707282*
* *The precision on test data is  0.6894941634241245*
* *The recall on test data is  0.7909537271239399*
* *The f1 on test data is  0.736747280160765*
* *The roc\_score on test data is  0.7578906566654182*



**4.2 Random Forest Classifier**

As the name suggests, Random Forest is a classifier that contains a

Number of decision trees on various subsets of the given dataset and

Takes the average to improve the predictive accuracy of that dataset. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. Here, this technique is used to predict whether a person commits fraud or not.

**Following table show performance of model on different metrics of evaluation.**

* The accuracy on train data is  0.9993292234963427
* The accuracy on test data is  0.8377537124700084
* The accuracy on test data is  0.8377537124700084
* The precision on test data is  0.8075226977950714
* The recall on test data is  0.8594699061292104
* The f1 on test data is  0.8326869065133075
* The roc\_score on test data is  0.8389926270281615

**4.3 xgboost Classifier**

XG Boost is otherwise as extreme Gradient Boosting which is one of the machine learning boosting classifier models. The XG boost use plot importance() function which is a build in function to generate feature importance, which improves the performance and efficiency by algorithmic optimization and system optimization.

The accuracy on train data is 0.7859264701185038

The accuracy on test data is 0.7764087932040724

The accuracy on test data is 0.7764087932040724

The precision on test data is 0.7083009079118029

The recall on test data is 0.81996996996997

The f1 on test data is 0.7600556715379263

The roc\_score on train data is 0.7816320572370111

**4.4 Support Vector Machine**

SVC The objective of clustering is to partition a data set into groups according to some criterion in an attempt to organize data into a more meaningful form. There are many ways of achieving this goal. Clustering may proceed according to some parametric model or by grouping points according to some distance or similarity measure as in hierarchical clustering. A natural way to put cluster boundaries is in regions in data space where there is little data, i.e. In "valleys" in the probability distribution of the data. This is the path taken in support vector clustering (SVC), which is based on the support vector approach. In SVC data points are mapped from data space to a high dimensional feature space using a kernel function. In the kernel's feature space the algorithm searches for the smallest sphere that encloses the image of the data using the Support Vector Domain Description algorithm. This sphere, when mapped back to data space, forms a set of contours which enclose the data points. Those contours are then interpreted as cluster boundaries, and points enclosed by each contour are associated by SVC to the same cluster.

The accuracy on train data is 0.7525473536269843

The accuracy on test data is 0.7536476233707282

The accuracy on test data is 0.7847091628299072

The precision on test data is 0.72905317769131

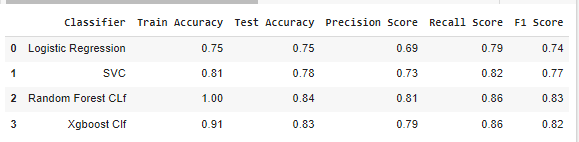
The recall on test data is 0.8203444249854057

The f1 on test data is 0.7720093393764592

The roc\_score on test data is 0.7882793428463031

**5. Model Evaluation and Selection**

Model evaluation and model selection are important parts of a data science project.

Hyper parameter tuning in the model building helps to improve model performance and can select the best among them. The same methodology has been followed in this case. 

**6.Conclusion**

**We have predicted the defaulters using multiple models in this project. We have used Logistic regression, Random forest, xgboost, SVM. We have also used gridsearchcv to tune hyper parameters.**

**1. In conclusion, we can conclude that xgboost is the best model for the present problem.**

**2. We have predicted the defaulters using multiple models in this project. We have used Logistical regression, Decision Tree, xgboost, SVM. We have also used gridsearchcv to tune hyperparameters.** **Logistic Regression, xgboost,svm, Random Forest algorithms were implemented. The important metric to compare all the algorithms in this case is ‘Recall’. As the company can’t afford to predict false negative i.e. Predict defaulter as a non-defaulter. Since, company is one, who will give to money to the customers, if, for any reason giving money to defaulter is gaining more risk to getting the investment back. Hence, here identifying false negative is important.**

**3. We have also seen the class imbalance so we did SMOTE to handle imbalance.**

**4. We did train test split and stratify the target variable.**

**5. We conclude that out of all models xgboost performed well with roc auc score of 0.832.**

**6. The best accuracy is obtained for the Random forest and xgboost classifier.**

**7. Xgboost model solves the problem with high accuracy than others. It has the precision, recall ,F!-score and ROC Score scores of 79% , 85% , 82% and 83% respectively, which is the highest among other models**.

**8. The next best models are the Random Forest classifier, it has the precision, recall ,F1-score and ROC Score scores of 80% , 85% , 83% and 83% respectively.**