**Credit Card Default Detection**

**Samara Reddy**

**Data Science Trainee,**

**Alma Better, Bangalore.**

**Contents**

1. **ABSTRACT**
2. **INTRODUCTION**
3. **LITERATURE REVIEW**
4. **OBJECTIVE**
5. **METHODOLOGY**

**4.1) Dataset Information**

**4.2) Attribute Information**

**4.3) Data Preprocessing**

**4.4) Exploratory Data Analysis**

**4.5) Processing Imbalanced Dataset**

**4.6) Hyperparameter Tuning**

**4.7) Machine Learning Models**

**4.7.1) Logistic Regression**

**4.7.2) KNN Classifier**

**4.7.3) Random Forest Classifier**

**4.7.4) Support Vector Classifier**

**4.7.4) XGBoost Classifier**

1. **EVALUATING ML MODELS**
2. **RESULTS**
3. **CONCLUSION**
4. **ACKNOWLEDGMENT**
5. **REFERENCE**
6. **ABSTRACT**

Since the 1990s, there have been significant advances in the technology space and the e-Commerce area, leading to an exponential increase in demand for cashless payment solutions. This has led to increased demand for credit cards, bringing along with it the possibility of higher credit defaults and hence higher delinquency rates, over a period of time. Also Understanding the history of clients will act as a valuable screening method for banks by providing information that can categorize clients as defaulters on a loan. Customer credit rating is a grade process where the consumer is categorized by the grade. Credit scoring model used to ascertain credit risk from new and existing customer. Credit rating is an assessment used to measure the creditworthiness of the customer. For the huge customers related data set we can use various classification techniques used in the field of data mining. The purpose of this research paper is to build a contemporary credit scoring model to forecast credit defaults for unsecured lending (credit cards), by employing machine learning techniques. Default is a keyword, used for predicting the customer who can’t repay the amount on time. Predicting future credit default accounts in advance is highly tedious task. Modern statistical techniques are usually unable to manage huge data. As much of the customer payments data available to lenders, for forecasting Credit defaults, is imbalanced (skewed), on account of a limited subset of default instances, this poses a challenge for predictive modelling. In this research, this challenge is addressed by deploying Synthetic Minority Oversampling Technique (SMOTE), a proven technique to iron out such imbalances, from a given dataset. On running the research dataset through four different machine learning models, the results indicate that the Random Forest Classifier Machine (RF Classifier model outperforms the other four classification techniques). Thus, our research indicates that the Random Forest classifier model is better equipped to deliver higher learning speeds, better efficiencies and manage larger data volumes.

1. **INTRODUCTION**

Consume now, pay later.

Credit cards are one of the most popular modes of payment for electronic transactions and make online transactions comfortable and convenient. However, since there has been an exponential expansion in credit card users over the years, banks have been determining credit risk based on an individual’s credit history. After the technology boom in the mid-1990s, companies switched to a technique that was already being used to determine credit risk and prevent defaults – namely using credit history data.

Credit risk is defined as the risk of financial loss when a borrower fails to pay the lender within a given period of time. With the rapid development of the credit cards industry, there has been a rise in credit card delinquency rates, which imposes a financial risk for the lending institutions. Credit risk is the oldest form of risk in the financial markets and has shown exponential growth in the 1990s against the backdrop of dramatic economic and technological change. In the past few years, the number of defaults has risen significantly and has cost commercial banks millions of dollars. Therefore, it has become critical that banks and lending financial institutions use robust mechanisms to forecast probabilities of credit defaults before lending. With an exponential increase in customers, many a times, credit risk must be analyzed, where customers have limited or no credit history.

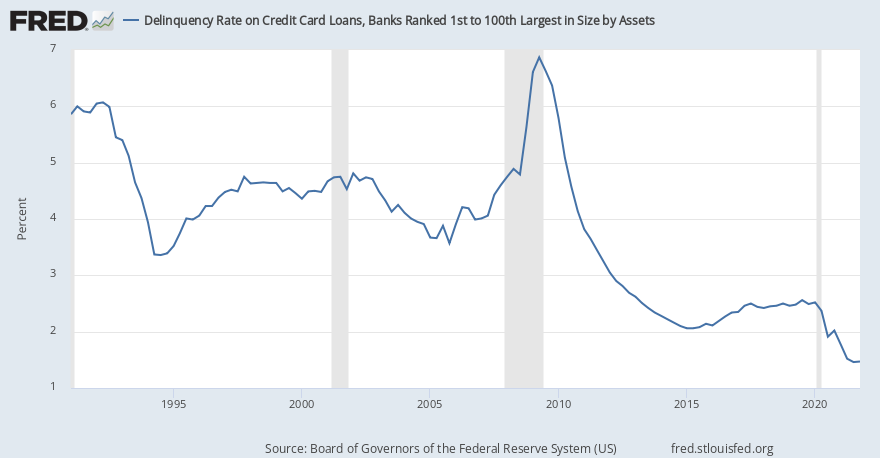
Moreover, the credit card usage database is by and large unbalanced, since majority of the customers pay their dues on time, barring a certain percentage of customers that default on payments. Machine learning algorithms are known to have proven their ability to determine the delinquency rates accurately.

Credit risk has traditionally been the greatest risk among all the risks that the banking and credit card industry are facing, and it is usually the one requiring the most capital. This can be proven by industry business reports and statistical data. Despite machine learning and big data have been adopted by the banking industry, the current applications are mainly focused on credit score predicting. The disadvantage of heavily relying on credit score is banks would miss valuable customers who come from countries that are traditionally underbanked with no credit history or new immigrants who have repaying power but lack credit history.

There has been immense development in the area of machine learning (ML) since the early 2000s. With greater access to customer data and increased computing power, credit scoring agencies are now in a better position to enable banks, by providing extensive credit analysis of their customers and prospects. Researchers are trying to determine better and efficient methods for credit risk evaluation. Financial institutions are in the process of exploring and deploying machine learning techniques that can help better decision making and enable mass customisation of product offerings. In this paper, the focus will be on evaluating and comparing popular machine learning classification models such as Logistic Regression, KNN Classifier, Random Forest Classifier, Support Vector Classifier and XGBoost Classifier to recognize patterns in the customer data (with a high degree of accuracy) for credit risk evaluation. Using publicly available datasets, machine learning classification models have been evaluated to determine credit risk defaults and deliver optimum performance; to efficiently predict delinquency rates. The same models can be deployed for automated processing of new credit card applications.

1. **LITERATURE REVIEW**

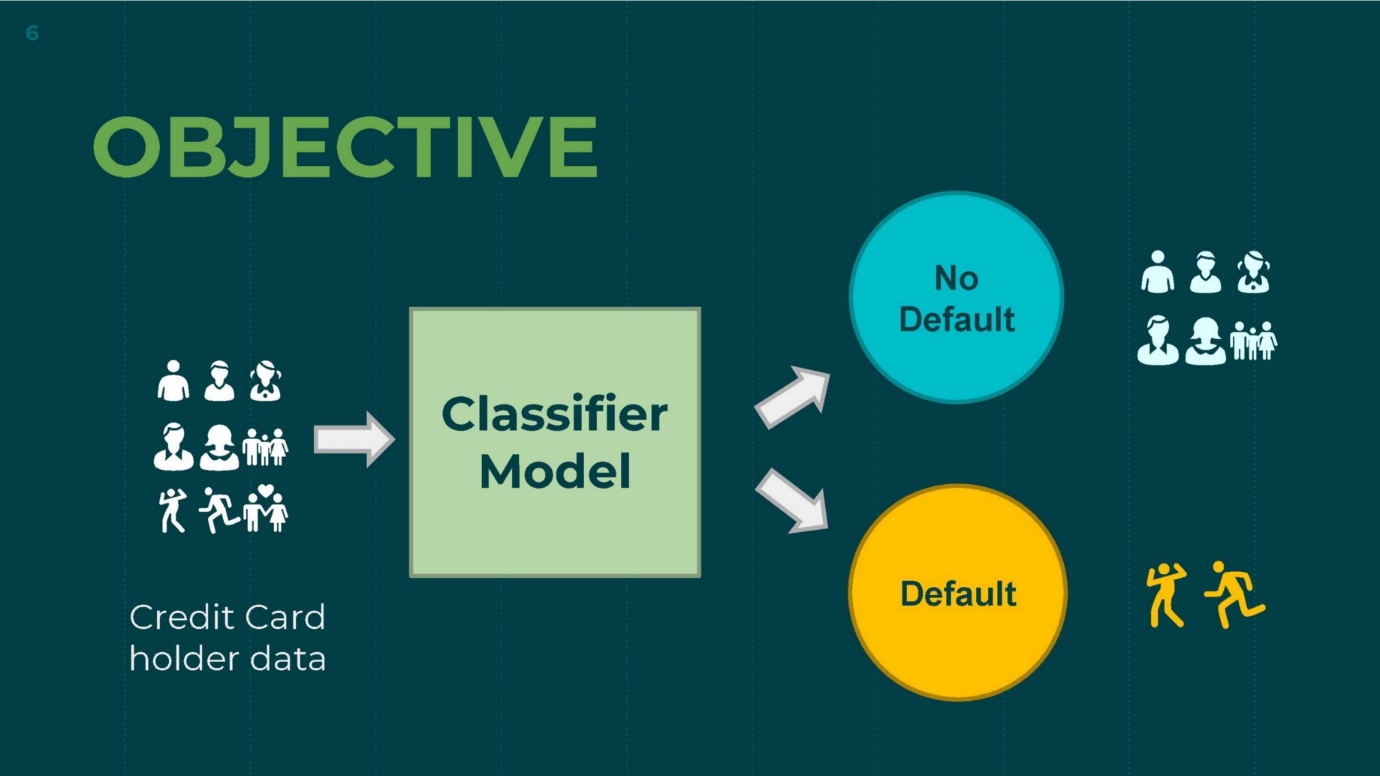
A credit card is typically issued by a commercial bank or a financial institution. It allows customers to borrow funds within a stipulated credit limit, for a given period of time, to pay for goods and services, at various points of sale, on credit in lieu of cash. These credit charges accrue, in a customers’ account as a balance, which must be squared off, on a periodic billing cycle basis, enabling customers to better manage their cash flows. Increased technological development and rise in e-Commerce has created exponential demand for payment solutions as cash alternatives. Availability of affordable credit has given a fillip to the growth of the global credit card industry. With increased deployment of unsecured credit through credit cards, delinquencies and personal bankruptcy rates also increased during the mid-1990s. As measured by the Federal Reserve Bank of New York, outstanding card balances stayed relatively flat in the years after an all-time peak in the fourth quarter of 2009 (during the financial crisis) but began to increment as the economy slowly re-bounded in the beginning of 2014.



Delinquency rates have dropped in the past year due to the pandemic and are at 1.58% as of the second quarter of 2021. Managing credit risk through prediction of credit defaults still continues to be a top priority for lenders in the unsecured lending market, in order to manage profitability and remain competitive

1. **OBJECTIVE**

The main objective of this project is to predict whether a credit card transaction is Default or not, based on the transaction amount, location and other transaction related data. It aims to track down credit card transaction data, which is done by detecting anomalies in the transaction data. Credit card fraud detection is typically implemented using an algorithm that detects any anomalies in the transaction data and notifies the cardholder (as a precautionary measure) and the bank about any suspicious transaction.



1. **METHODOLOGY**

**4.1) Dataset Information**

This data set contains the information of customers default payment. There are 30,000 different instances and 25 attributes total; each instance represents one customer, and attributes consist demographic information about the customers and their past payment history from April to September

**4.2) Attribute Information**

ID: unique identification number assigned to each customer

LIMITBAL: amount of given credit access line

SEX: gender (1 = male; 2 = female)

EDUCATION: highest degree obtained (1 = graduate school; 2 = university; 3 = high school;4 = others; 5 = unknown; 6 = unknown)

MARRIAGE: marital status (1 = married; 2 = single; 3 = others)

AGE: age in year

PAY0: monthly payment record in September

PAY2: monthly payment record in August

PAY3: monthly payment record in July

PAY4: monthly payment record in June

PAY5: monthly payment record in May

PAY6: monthly payment record in April

BILLAMT1: total amount owed in September

BILLAMT2: total amount owed in August

BILLAMT3: total amount owed in July

BILLAMT4: total amount owed in June

BILLAMT5: total amount owed in May

BILLAMT6: total amount owed in April

PAYAMT1: amount of previous payment in September

PAYAMT2: amount of previous payment in August

PAYAMT3: amount of previous payment in July

PAYAMT4: amount of previous payment in June

PAYAMT5: amount of previous payment in May

PAYAMT6: amount of previous payment in April

default payment next month: whether a customer is defaulted on next months payment or not (1 = defaulter; 0 = non-defaulter)

The columns from PAY0 to PAY6 represent the repayment status in each month from April to September: -1 indicates paying duly for one month; -2 indicates paying duly for two months; . . . ; -x indicates paying duly for x months. The positive number shows that how many months the payment has been delayed. For example, 1 means that the payment has been delayed for 1 month; 2 mean that the payment has been delayed for 2 months; . . . ; x means that the payment has been delayed for x months

The last column “**default payment next month**” will be used as the output of the final predictive model.

**4.3) Data Pre-processing**

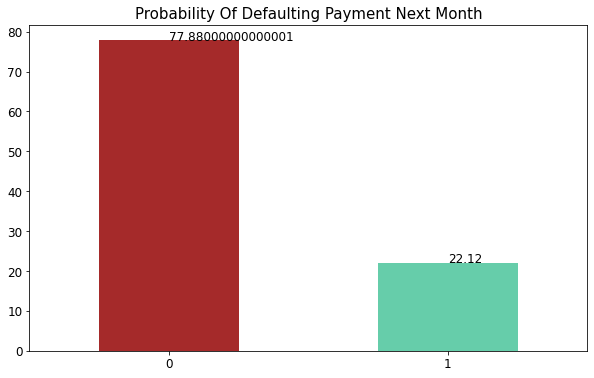
**Null Values Treatment:** The problem of missing value is quite common in many real-life datasets. Missing value can bias the results of the machine learning models and/or reduce the accuracy of the model. A Missing Value is defined as the values or data that is not stored (or not present) for some variable/s in the given dataset. Our dataset contains NO null values which may tend to our accuracy at the beginning of our project in order to get a better result.

**Encoding Categorical Variables:** Machine learning models require all input and output variables to be numeric. This means that if your data contains categorical data, you must encode it to numbers before you can fit and evaluate a model. In the project I had used three different types of Encoders.

**One Hot Encoding** - One-Hot Encoding is popular technique for treating categorical variables. It simply creates additional features based on the number of unique values in the categorical feature. Every unique value in the category will be added as a feature. One-Hot Encoding is the process of creating dummy variables.

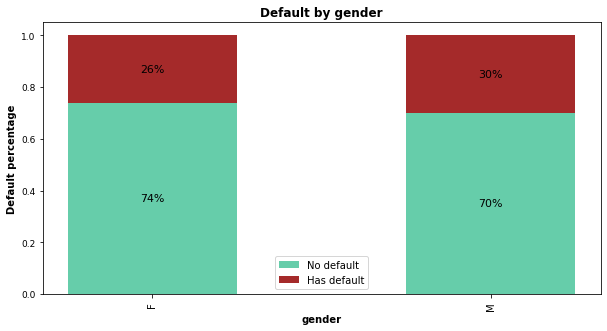
**4.4) Exploratory Data Analysis**

*Feature Analysis - The frequency of defaults*



Count of Non-default is a lot higher than default value. Non- Default data is 77.9% while Default cases are 22.1% as in dataset. Looking at data, we will need to check if it is a case of Imbalanced dataset.

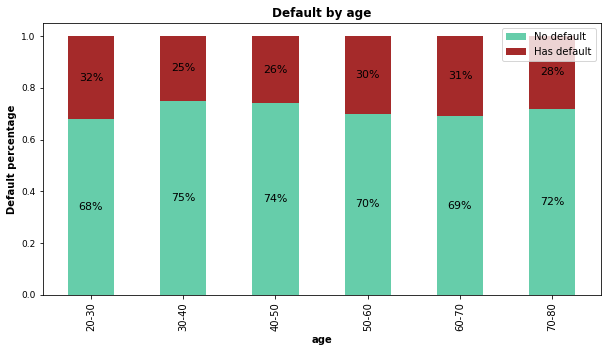
*Feature Analysis – Default by gender*



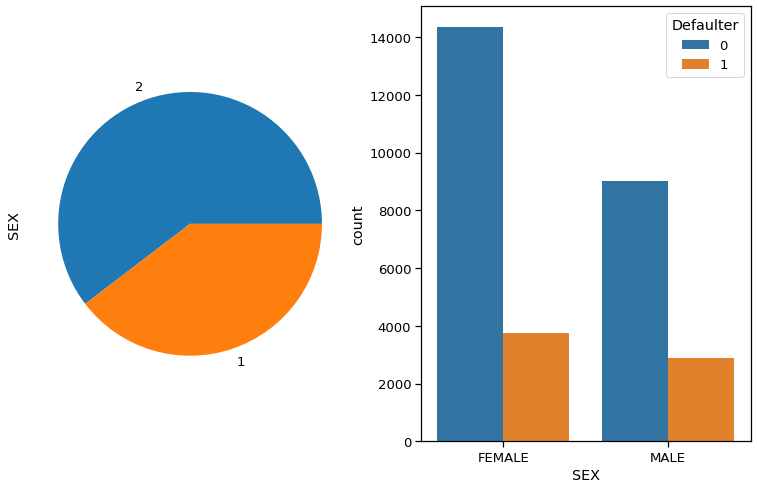
* The rate of being defaulter is comparatively higher in males with 30% of total defaulters than to 26% of female defaulter respectively.

*Feature Analysis – Default by age*

* Customers aged between 30-50 had the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all had higher delayed payment rates. However, the delayed rate dropped slightly again in customers older than 70 years.

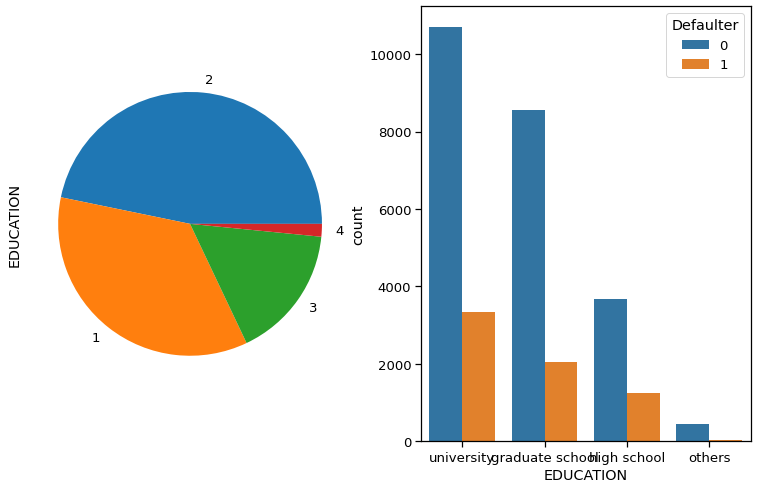


*Feature Analysis - Gender wise defaulter prediction*

**

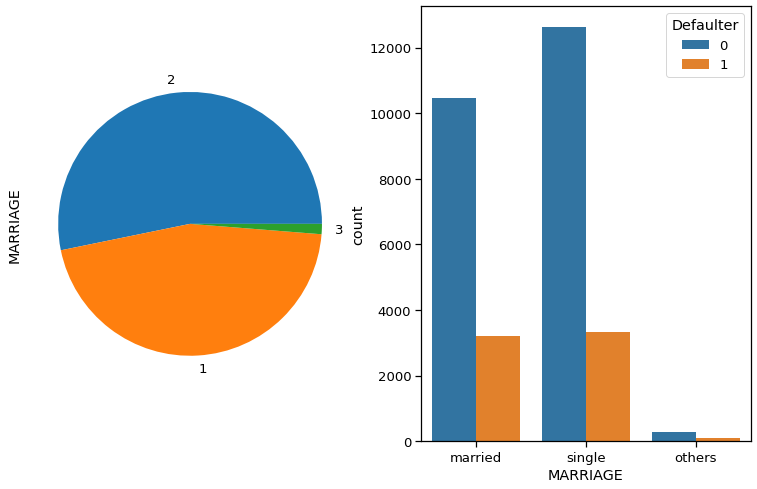
Larger percentage of females than males in default payment category.

*Feature Analysis - Education wise defaulter prediction*

**

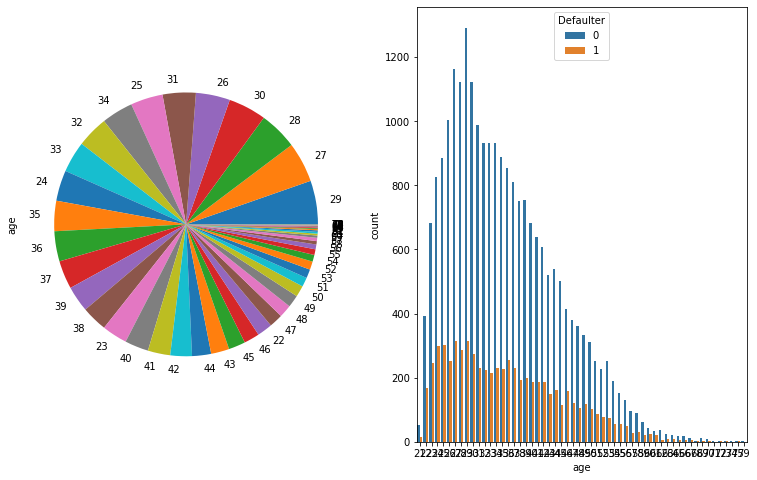
Number of defaulters have a higher possibility that the person is educated (graduated from school or university).

Feature Analysis – Marital status wise defaulter prediction

**

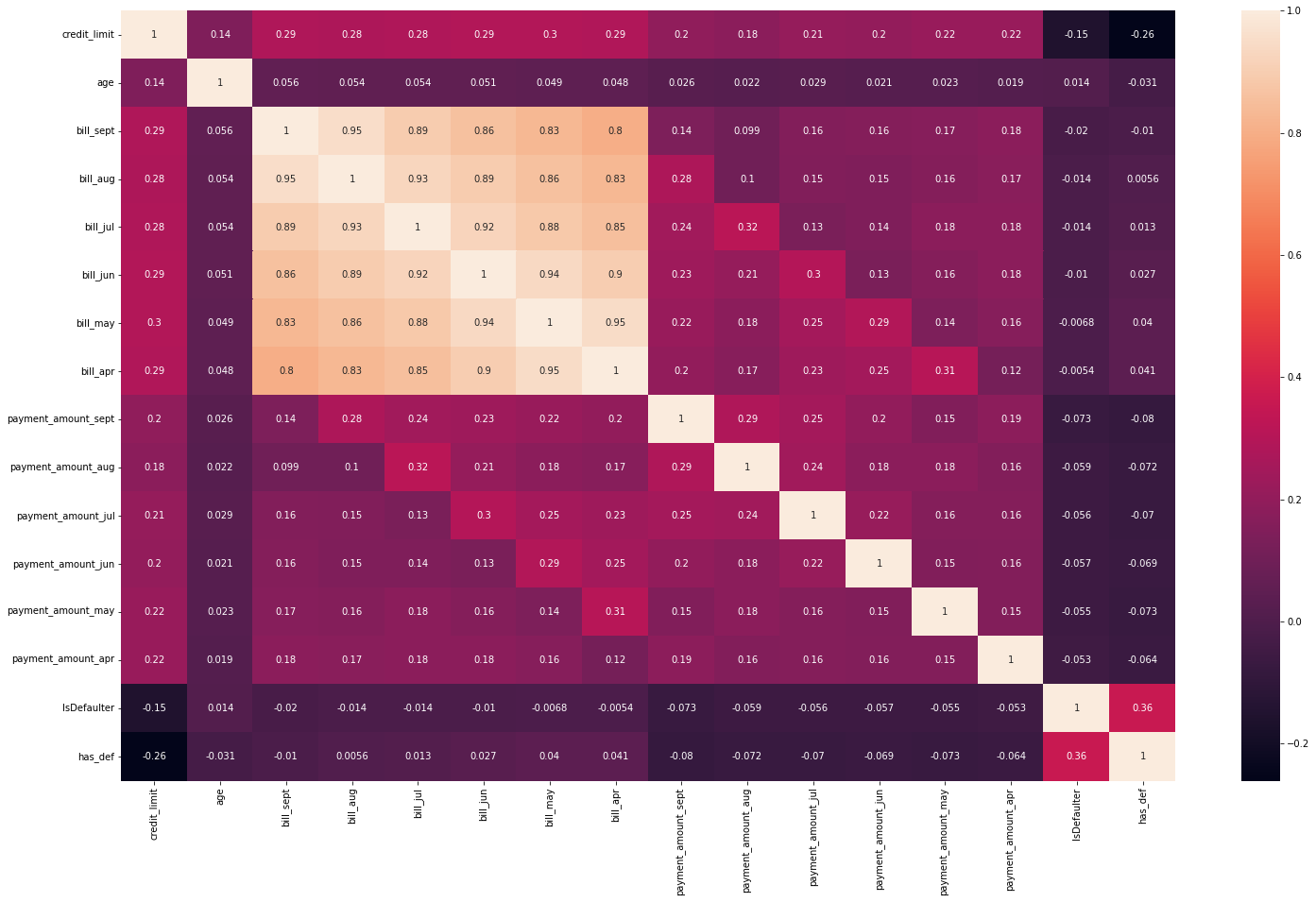
Individuals having single status have higher percentage of default than married.

Feature Analysis – AGE wise Defaulter Prediction

**

Age peaks around 28-29 years in default payment category. But there is no major relation between age and defaulter prediction.

**Correlation between parameters**



We can see that no correlation between the features so there is no need to remove or drop some features.

**4.5) Processing Imbalanced Dataset**

Imbalanced classification is primarily challenging as a predictive modelling task because of the severely skewed class distribution. This is the cause for poor performance with traditional machine learning models and evaluation metrics that assume a balanced class distribution.

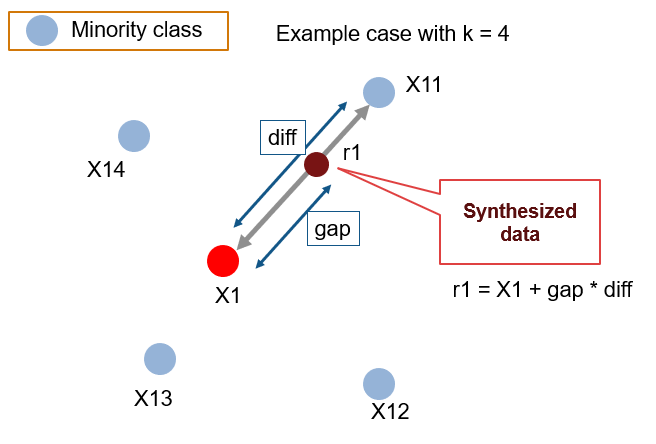
Nevertheless, there are additional properties of a classification dataset that are not only challenging for predictive modelling but also increase or compound the difficulty when modelling imbalanced datasets. For this I had used SMOTE

SMOTE (Synthetic Minority Oversampling Technique)

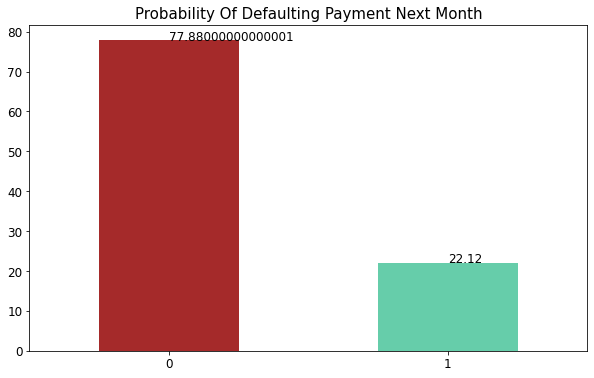
SMOTE is an oversampling technique where the synthetic samples are generated for the minority class. This algorithm helps to overcome the overfitting problem posed by random oversampling. It focuses on the feature space to generate new instances with the help of interpolation between the positive instances that lie together.

#### Working Procedure:

At first the total no. of oversampling observations, N is set up. Generally, it is selected such that the binary class distribution is 1:1. But that could be tuned down based on need. Then the iteration starts by first selecting a positive class instance at random. Next, the KNN’s (by default 5) for that instance is obtained. At last, N of these K instances is chosen to interpolate new synthetic instances. To do that, using any distance metric the difference in distance between the feature vector and its neighbors is calculated. Now, this difference is multiplied by any random value in (0,1] and is added to the previous feature vector. This is pictorially represented below:







Before SMOTE After SMOTE

**4.7) Hyperparameter Tuning**

In machine learning, hyperparameter tuning is the problem of choosing a set of optimal hyperparameters for a learning algorithm*.* Machine learning involves predicting and classifying data and to do so, you employ various machine learning models according to the dataset. Machine learning models are parameterized so that their behavior can be tuned for a given problem. These models can have many parameters and finding the best combination of parameters can be treated as a search problem. You cannot know the best value for a model hyperparameter on a given problem. You may use rules of thumb, copy values used on other issues, or search for the best value by trial and error. When a machine learning algorithm is tuned for a specific problem then essentially you are tuning the hyperparameters of the model to discover the parameters of the model that result in the most skillful predictions.

Models can have many hyperparameters and finding the best combination of parameters can be treated as a search problem. Although there are many hyperparameter optimization/tuning algorithms like grid search and Random Search etc.

GridSearchCV : Grid-searching is the process of scanning the data to configure optimal parameters for a given model. Depending on the type of model utilized, certain parameters are necessary. Grid-searching does NOT only apply to one model type. Grid-searching can be applied across machine learning to calculate the best parameters to use for any given model. Grid-Search will build a model on each parameter combination possible. It iterates through every parameter combination and stores a model for each combination. Grid Search uses a different combination of all the specified hyperparameters and their values and calculates the performance for each combination and selects the best value for the hyperparameters.

I had used GridSearch for tuning hyperparameters of respective algorithms for getting for better accuracy and to avoid overfitting in case of tree based models like Random forest and XGBoost Classifier.

**4.8) Machine Learning Models**

**Logistic Regression:** It is one of the best algorithms when it comes to handling categorical target feature. It will predict the probability of an instance belonging to the default class or not. It is the appropriate regression analysis to conduct when the dependent variable is dichotomous (binary). Like all regression analyses, the logistic regression is a predictive analysis. In other words, it is multiple regression analysis but with a dependent variable is categorical. It is called regression but performs classification based on the regression and it classifies the dependent variable into either of the classes. Logistic regression is used to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables. Logistic regression model uses sigmoid function to predict probability of positive and negative class **f(x)= 1/1+e ^(-x)**

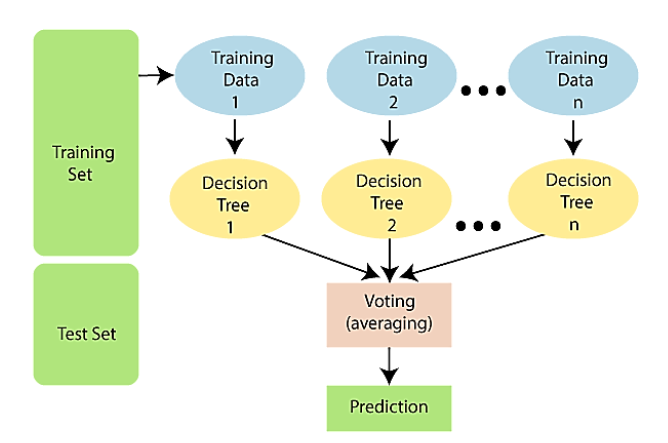
****

Decision Tree : Decision Tree is a **Supervised learning technique**that can be used for both classification and Regression problems, but mostly it is preferred for solving Classification problems. It is a tree-structured classifier, where**internal nodes represent the features of a dataset, branches represent the decision rules** and **each leaf node represents the outcome**.

In a Decision tree, there are two nodes, which are the **Decision Node** and**Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches. The decisions or the test are performed on the basis of features of the given dataset.

Random Forest Classifier: Random forest is a way of averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of overcoming over-fitting problem of individual decision tree.

In other words, random forests are an ensemble learning method for classification and regression that operate by constructing a lot of decision trees at training time and outputting the class that is the mode of the classes output by individual trees. This is used to describe data and to explain the relationship between one dependent binary variable, which is the class variable that has a categorical value like Yes/No or True/False or 0/1 and one or more nominal independent variable. It is used in measuring the probability of a binary class or response as the value of the class variable relating it with the variables used for predicting (also called predictors).

****

**Gradient boosting**: Gradient boosting algorithm is one of the most powerful algorithms in the field of machine learning. As we know that the errors in machine learning algorithms are broadly classified into two categories i.e. Bias Error and Variance Error. As gradient boosting is one of the boosting algorithms it is used to minimize bias error of the model.

Unlike, Adaboosting algorithm, the base estimator in the gradient boosting algorithm cannot be mentioned by us. The base estimator for the Gradient Boost algorithm is fixed and i.e. *Decision Stump*. Like, AdaBoost, we can tune the n\_estimator of the gradient boosting algorithm. However, if we do not mention the value of n\_estimator, the default value of n\_estimator for this algorithm is 100.

1. **EVALUATING ML MODELS**

To evaluate the performance of algorithms, we used different metrics

**1)** **Confusion matrix**: The confusion matrix is a table that summarizes how successful the classification model is at predicting examples belonging to various classes. One axis of the confusion matrix is the label that the model predicted, and the other axis is the actual label. It is a tabular representation of a classification model performance on the test set, which consists of four parameters: true positive, false

positive, true negative, and false negative

|  |  |  |
| --- | --- | --- |
|  | Predicted True | Predicted False |
| Actual True | True Positive (TP) | False Negative (FN) |
| Actual False | False Positive (FP) | True Negative (TN) |

**2) Accuracy:** Accuracy is often the most used metric representing the percentage of correctly predicted observations, either true or false. To calculate the accuracy of a model performance, the following equation can be used:

**3)Precision/Recall:**

Precision is the ratio of correct positive predictions to the overall number of positive predictions **:**

Recall is the ratio of correct positive predictions to the overall number of positive examples

**4)Area under ROC Curve(AUC)**

ROC curves use a combination of the true positive rate (the proportion of positive examples predicted correctly, defined exactly as recall) and false positive rate (the proportion of negative examples predicted incorrectly) to build up a summary picture of the classification performance

1. **RESULTS**

**1 Logistic Regression:**

Firstly, logistic regression is performed on the featured data and the classifier is evaluated using the test data. The confusion matrix showing actual and predicted are as follows,

By performing accuracy metrics, the logistic regression classifier attained

**The accuracy on train data is 0.774**

**The accuracy on test data is 0.775**

**The precision on test data is 0.705**

**The recall on test data is 0.82**

**The f1 on test data is 0.758**

**The auc on test data is 0.78**

**2 Decision Tree Classification:**

Secondly, Decision Tree Classifier is performed on the featured data and the classifier is evaluated using the test data. The confusion matrix showing actual and predicted are as follows,

**The accuracy on train data is 0.796**

**The accuracy on test data is 0.79**

**The precision on test data is 0.715**

**The recall on test data is 0.841**

**The f1 on test data is 0.773**

**The auc on test data is 0.797**

**3 Random Forest:**

And next, Random Forest Classifier is performed on the featured data and the classifier is evaluated using the test data. The confusion matrix showing actual and predicted are as follows,

**The accuracy on train data is 0.81**

**The accuracy on test data is 0.799**

**The precision on test data is 0.727**

**The recall on test data is 0.85**

**The f1 on test data is 0.783**

**The auc on test data is 0.805**

# **4 Gradient Boosting:**

At last Gradient Boosting is performed on the featured data and the classifier is evaluated using the test data. The following results were obtained

**The accuracy on train data is 0.991**

**The accuracy on test data is 0.857**

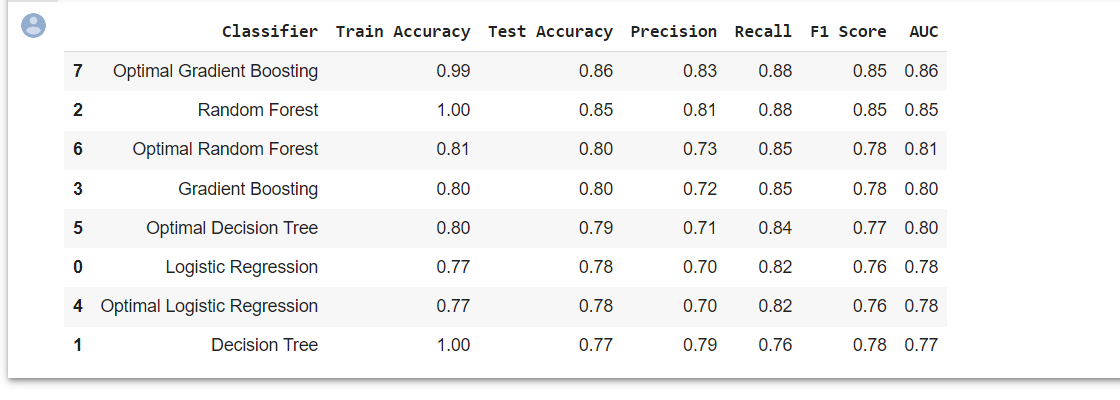
**The precision on test data is 0.832**

**The recall on test data is 0.875**

**The f1 on test data is 0.853**

**The auc on test data is 0.857**

**Table below shows the quantitative observations depicted using above classification method**

****

1. **CONCLUSION**

* No Null Values and No Duplicate Values in dataset
* About 22% people are expected to default next month and 77.8% are not expected to default
* The rate of being defaulter is comparatively higher in males with 30% of total defaulters than to 26% of female defaulter respectively.
* Customers aged between 30-50 had the lowest delayed payment rate, while younger groups (20-30) and older groups (50-70) all had higher delayed payment rates.
* Higher proportion of customers with higher bill amount but lower payment rate suggesting they are likely to do the fault
* Customers with grad school education have the highest median and highest maximum number, so we can say that people with higher education levels did get higher credit limits
* Gradient Boosting was high F1 score of 0.85 and Recall of 0.88
* Random forest has high Recall of 0.88

1. **REFERENCE**

* Top 5 Data Mining Algorithms for Classification <https://wisdomplexus.com/blogs/datamining-algorithms-classification/>
* Top 20 Python libraries for Data Science <https://builtin.com/data-science/python-libraries-data-science>
* Bellotti, T., and Jonathan, C., 2009, "Support vector machines for credit scoring and discovery of significant features," Expert Systems with Applications, 36(2), 3302-3308
* <https://www.analyticsvidhya.com/blog/2020/10/overcoming-class-imbalance-using-smote-techniques/>
* <https://machinelearningmastery.com/one-hot-encoding-for-categorical-data/>
* <https://www.analyticsvidhya.com/blog/2021/10/handling-missing-value/>
* <https://www.datacamp.com/community/tutorials/parameter-optimization-machine-learning-models>