**Data 240 Data Mining and Analytics**

**Final Project: Final-report**

Lakshmi Naga Meghana Polisetty | Sakshi Jain | Sakshi Tongia | Snehal Dashrath Karad| Vaibhav Yalla

1. **Introduction**

**1.1 Why did you choose this topic? What is your data?**

***Objective : Predict Default of Credit Card Payment***

**(https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients#)**

Credit cards are important for many financial situations. It is helpful in most financial situations, such as unplanned expenses. In addition, other advantages of using a credit card are cashback, reward points, building credit, etc. The simple principle of using this valuable tool is to use it responsibly. The responsibility of owning a credit card mainly includes paying the bills on time, which will avoid increases in interest rates, poor credit scores, decrease in credit limit, and late fees. Credit cards have introduced the convenience of carrying just a card in place of large sums of cash and also to pay for a purchase over a span of time in place of immediate full payment at the time of purchase. However, there is a possibility of defaulting the payment to the credit card lender by the customer resulting in loss.

According to the GOBanking Rates survey, about 6%, that is 14 million Americans have credit card debt over $10,000. 33% of Americans believe that it would take more than two years to clear their credit card debt. This situation has degenerated, as in the pandemic 45% of Americans have taken more credit card debt.

Using machine learning to predict credit card default will save the lenders from issuing credit cards to risky customers or to identify if a customer is going to default a payment in future and thereby can reduce their credit limit. Increase in credit-card debt acts negatively for the customer and also reduces the bottom line for the lenders. This will help the lenders in optimizing their decision, build sustainable business economics and provide a better customer experience.

***Data***

The dataset is collected from UCI Machine Learning Repository, which is sourced by I-Cheng Yeh. The dataset is aimed on the aggregated customer profiles of Taiwan with default payments from April 2005 to September 2005. The target feature is binary with a default event, which is did the customer default or not. The default event is true when the customer does not pay the bills after the latest statement date. The “Default of credit card clients Data Set” contains 23 unique features with 30,000 instances. The features include the credited amount (New Taiwan dollars), gender, education age, repayment statements in each month, etc.

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 payments. 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.

**1.2 Why does this topic need classification methods?**

As information becomes more and more available, companies will generate and collect tremendous amounts of data every day. Businesses face a major challenge when it comes to transforming information into actionable results from their databases. In data mining, large quantities of data are analyzed and explored automatically or semi-automatically in order to discover meaningful patterns and rules. In today's world, data mining is a vital part of decision support systems and plays an important role in segmenting the market, detecting fraud, and scoring behavior. According to academic studies on credit scoring, there are a wide variety of classification methods used to discriminate between good and bad borrowers. Formal statistical methods are used to classify customers into 'good' or 'bad' risk categories according to their credit score. Due to the tremendous growth in customer credit in recent years, these methods have become increasingly important. This dataset can be used to determine credit risk using a variety of classification methods.

Furthermore, for banks to control credit risk, an approach to evaluate likelihood of a default credit card payment is more effective to predict if the cardholder is ‘Risky’ or ‘Not Risky’. An accurate system to forecast the class of cardholder being a defaulter or not will help manage this issue rather efficiently allowing banks and organizations to make quick decisions. With credit risk prediction, a common problem is that there is a lot of grey area between user data expressing risk and user data expressing no risk, which results in no clear demarcation of labels/classes for data points. Generally, non-risky data points are seen to dominate risky data points. To resolve this, an efficient technique is to create an overfitted model to increase misclassification rate by misclassifying risk-free data points. Classification techniques can then be deployed to resolve this issue.

1. **Literature Review**

***Paper 1: Consumer credit-risk models via machine-learning algorithms***

In their research Khandani et al., 2010, combine credit bureau data and customer transactions. They used generalized classification and regression trees (CART) to build the model. The model predicts the possibility of credit default by customer in the forthcoming three to 12 months. The authors state that models are helpful as they are adaptive with the change in customer behavior over the credit cycles. The dataset comprises transactional, account-balance and credit bureau data for a subset of customers of a major commercial bank, with personal identification data of individuals being removed. The drawback of the data is that one customer at a time has relations with multiple banks and hence the complete credit or spending data of a customer is not available. CART model is suitable for datasets with high number of features and unlike logit and probit models where dependent variables have to fit one linear model, CART can find nonlinear relation between input variables and identify the number of independent variables that can be used. The model results show that it correctly classified 99% of the classes. Identifying the credit risk for a bank can save up to 6%-23% of the total losses they might face and hence is evident to do so.

***Paper 2: Prediction of default payment of credit card clients using Data Mining Techniques***

Subasi & Cankurt, 2019 discuss how credit risk prediction is an important problem to solve in the banking sector. In their paper, they focus on various methodologies to conduct an automated credit score assessment. This study is focused on seeing how a group of simple data mining techniques and models managed to produce high classification rates, rather than complex models. To fix the imbalanced dataset, the author uses a Synthetic Minority Over-Sampling Technique (SMOTE) technique. This was deployed to balance the training dataset, and furthermore, generated 2 additional datasets by different configurations of SMOTE. As a result, they ended up additionally using SMOTE 100%(one-pass) and SMOTE 200% (two-pass) datasets. Approaches like single classifier, multiple classifiers, non-parametric methods, etc. have been discussed. The first technique deployed for classification was C4.5 Decision Tree model. An extension to ID3, it creates maximum trees followed by pruning. Consisting of a pre-pruning parameter, it avoids the generation of nodes with fewer leaves, improving error reduction. One of the most recent techniques in credit risk analysis, the paper also emphasizes the use of random forest algorithm with ensemble bagging technique. Based on tree growing technique, it generated many splits that help give diversity to base learners. Moreover, overfitting is immensely reduced due to ensemble aggregation. As a result, Random Forest with SMOTE 200% dataset managed to obtain the highest accuracy of 89.01% followed by Decision Tree giving 80.56% accuracy. Complex models like KNN were able to give accuracy of only 80.21%. In conclusion, Random Forest and DTree work best for such a dataset along with SMOTE technique.

***Paper 3: Predictive Analytics for Default of Credit Card Clients***

In this paper, Bacova & Babic, 2021, have compared various machine learning algorithms, on the Default of credit card client’s dataset. CRISP-DM methodology was used on Taiwan's public data of credit card holders which includes 30,000 instances, and 23 different features. Features include Education, Marriage, Sex, etc. Exploratory Data Analysis (EDA) on this dataset provides important insights such as the maximum amount of given credit, the average age of cardholders, and others. The dataset was split into 70% of training data, and 30% testing data. There was no improvement in performance in splitting 60/40, or 80/20. Models Random Forest (RF), AdaBoost, XGBoost (XGB), and Gradient Boosting algorithms are applied to the dataset. A comparison of these Machine learning models is done using performance metrics. Performance metrics used in this paper are accuracy, precision, recall, and ROC curve. XGBoost has provided the best accuracy of approx. 82%, with 0.94 recall for class 0. And 0.84 precision for class 0. The limitations are the dataset is imbalanced, and no oversampling, or under-sampling techniques were used in this paper, which could have improved the performance of the models.

***Paper 4: Probability of Loan Default- Applying Data Analytics to Financial Credit Risk Prediction***

In this paper, the authors M. Paprzycki et al. compared the performance of seven classifiers, K-Nearest Neighbours (KNN), Support Vector Machines, Random Forest, Gradient Boosted Decision Trees, AdaBoost, Extreme Gradient Boosting (XGBoost), and the results of meta classifiers over two different datasets. The datasets used in this study are publicly available benchmark datasets “Give Me Some Credit” and “German Credit Dataset” both of which contain information about potential borrowers' demographic and financial state. They suggested that in the banking domain for credit scoring, Recall and the Type I error are the most important criteria for the evaluation of models hence in their study, they aimed at achieving the best recall score. Based on their experiments and evaluation criteria, Gradient Boosting Decision Trees performed the best with both a relatively high recall score (0.650) and a low Type 1 error (0.092)as compared to other models for the “Give Me Some Credit” dataset. For the “German Credit Dataset” as well, GBDT performed the best with a recall of 0.65 and a Type I error of 0.0. Their experiments to use a meta classifier for this task using two different methods, Weighted Average Recall Error (WARE) and the Weighted Average Type-1 Error (WA-T1E) did not achieve any better results than models implemented individually. Their attempts to employ a meta-classifier for this job using the Weighted Average Recall Error (WARE) and the Weighted Average Type-1 Error (WA-T1E) techniques did not perform well. The results achieved for recall and Type I error were much worse than the models implemented individually for both datasets. They concluded that these simple meta-classifiers were not a good solution to this problem and in the future, they would like to experiment with complex meta-classifiers.

***Paper 5: The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients***

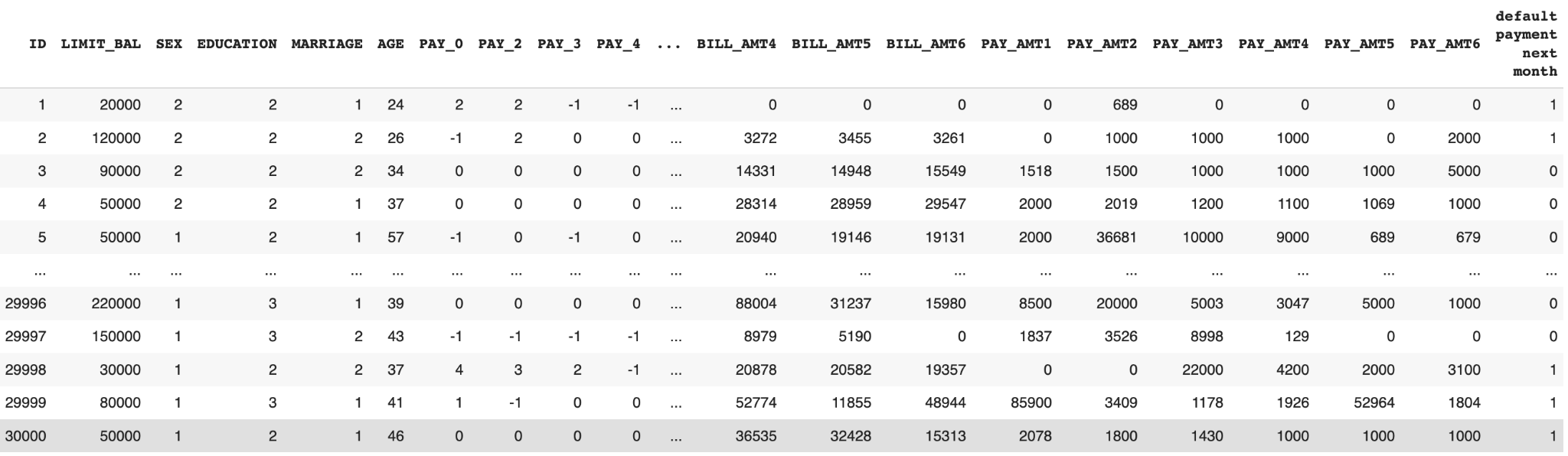
In this study, I-Cheng Yeh E Che-hui-Lien compares the predictive accuracy of the probability of default payments in Taiwanese customers using six data mining techniques. According to the paper, Taiwan's credit card issuers were facing a credit card debt crisis during the third quarter of 2006, so the banks over issued cash and credit cards to unqualified applicants in order to increase the market. Using financial information, such as business financial statements, customer transactions, and repayment records, the author mainly focuses on risk prediction. Data mining techniques were compared among six approaches to see if any were more accurate at classifying data. In order to score credit, six data mining techniques are used: discriminant analysis, logistic regression, Bayes classifier, nearest neighbor, artificial neural networks, and classification trees. Study participants were credit card holders of an important Taiwanese bank (a cash and credit card issuer), using payment data from October 2006. There are 5529 cardholders with default payments (22.12%) among the total 25,000 observations. As the response variable, this study used a binary variable - default payment (Yes = 1, No = 0). Based on error rates, K-nearest neighbor classifiers and classification trees have the lowest error rates (=0.18). Based on the area ratio, K-nearest neighbor classifiers perform best (=0.68). With the highest area ratio (=0.54) and relatively low error rate (=0.17), artificial neural networks achieve the best performance in the validation data.

***Paper 6: A conservative approach for online credit scoring***

Afshin Ashofteh and Jorge M.Bravo proposed a novel approach for online credit scoring that could be appropriate for Big Data and small data. An innovative machine learning method is presented in this research to predict the default of high-risk branches or customers using credit scoring in risk management. To form a conservative credit-scoring model and to study the impact of modeling performance on credit provider benefit, Kruskal-Wallis non-parametric statistics were used. An innovative method for identifying good loans with low false-negative rates has been presented by the authors. A computationally efficient credit-scoring model is built based on a Kruskal-Wallis nonparametric statistical analysis in two steps. Logistic regression with Ridge penalty, Random Forests, and SVM are used to learn and assess the model performance based on a two-step approach. In this paper, a sample credit dataset is used to illustrate the use of this novel time-dependent credit scoring method. According to the empirical results, the new credit scoring methodology has a low false-negative rate and a reasonable coefficient of determination. According to the paper, the model shows a high level of accuracy but is computationally less expensive. This credit scoring method was shown to be more informative and conservative in terms of classification accuracy. An AUC of 0.99 is obtained for small datasets and an AUC of 0.67 is obtained for large datasets with an 18% improvement in recall and sensitivity. As a homogeneous symmetric average aggregate measure, the authors conclude that this method is accurate and were able to renew features dynamically and weighted out the attributes as their impact factors.

***Raw Data : 3000 Rows X 25 Columns***

**Figure 1.**

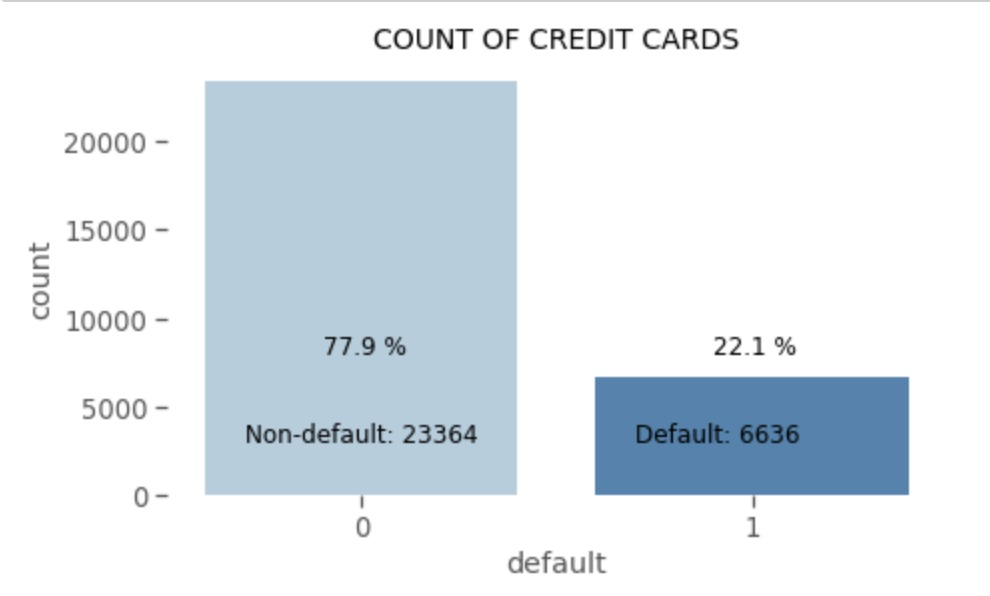


***Pre-processing Steps Performed :***

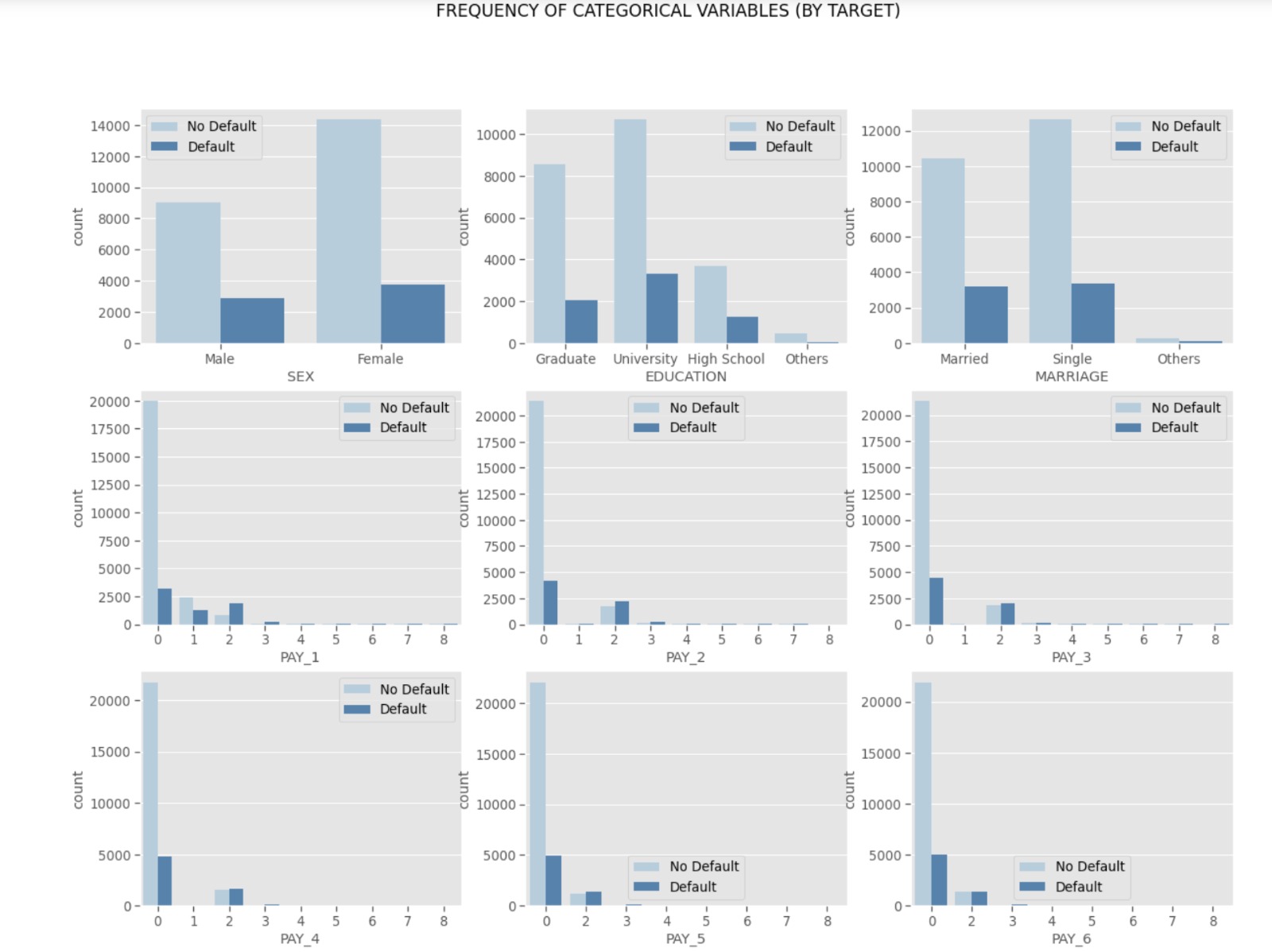
1. Replaced the column names for convenience : Target Variable changed from default.payment.next.month to Default
2. Changed PAY\_0 to PAY\_1 to have proper continuity  
   According to our findings, the PAY\_n variables indicate the number of months of delay: -1 in PAY\_n indicates thes customer "Paying Duly" however, there is no clear indication as to what does that values 0 and -2 indicate. Hence, we are adjusting the labels to 0 for Pay Duly category.
3. Changed the Values for Education column (1 = Graduate School; 2 = University; 3 = High School; 4 = Others) Any value other than 4 is changed to 4

With an Initial Exploratory Data Analysis, we have observed that the Default Data is highly imbalanced with 78% customers that have been in Non-Default and 22% of the Customers in Default Status as shown in Figure 2.

**Figure 2.**



**Figure 3.**



**3. Feature selection methods**

Feature selection is a process of selecting relevant input features from given feature list of the target problem. In a real world scenario, there can be multiple features that can be collected for a target problem. Some of those features may be highly correlated or some may be highly irrelevant to the target prediction. If we use all the given attributes, this might result in a lot of processing time and an inaccurate prediction. The aim is to select only those attributes that would be relevant to solve the machine learning problem.

In the given problem, we have 28 numbers of features. To reduce these features, we apply a series of feature selection methods to find out relevant features.

**3.1 Hypothesis testing**

Hypothesis testing is a statistical method to test if the population sample is correlated with other samples. It has the following components that need to be defined-

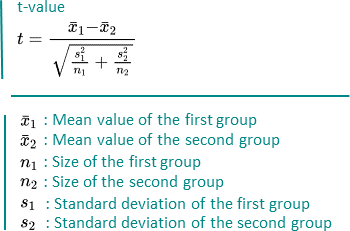
***Null Hypothesis***: It is denoted by H0, where if there is no significant difference between population, and sample data.

***Alternate Hypothesis***: The alternative hypothesis is a complete opposite to the Null Hypothesis. This statement of contrary of Null hypothesis is denoted by H1.

***Critical Value***: Higher the critical value, lower the probability of 2 samples belonging to the same distribution.

***P-value***: P-value also called as probability value, helps in supporting or rejecting the null hypothesis in hypothesis testing.  
***Degree of freedom***: The number of independent variables is called as the degree of freedom.

**Figure 4.**



***T-test***

T-Test is used to compare the means of two given samples. The implementation of this method is limited to samples smaller than 30. It assumes that the sample has a normal distribution. Additionally, one-sample or two-sample analyses are possible. Based on the number of samples, n-1 gives the degree of freedom.

Our Null Hypothesis states that;

H0 -> No difference between features and target variable, so we should remove them

H1 -> Difference between features and target variable, so we should keep them

Using 99% Confidence Interval; if p>0.01 , can't reject null hypothesis | if p<0.01, can reject null hypothesis. We apply T-Test between two classes for each feature. The features rejecting our null hypothesis will be kept, as they are most different .

**3.2 Chi-square**

Chi-square is a statistical test used to examine the differences between categorical variables from a random sample in order to judge independence between expected and observed results. There are two types of Chi tests - Chi-Square goodness of fit and test for independence. The Chi Square is done on the categorical dataset as shown in the below figure.

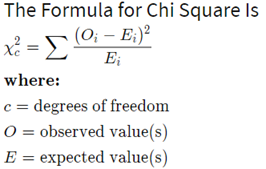
While conducting the chi-square test we have initially considered 2 hypotheses i.e the Null Hypothesis and the Alternate Hypothesis.

1. H0 (Null Hypothesis) = The 2 variables to be compared are independent.
2. H1 (Alternate Hypothesis) = The 2 variables are dependent.

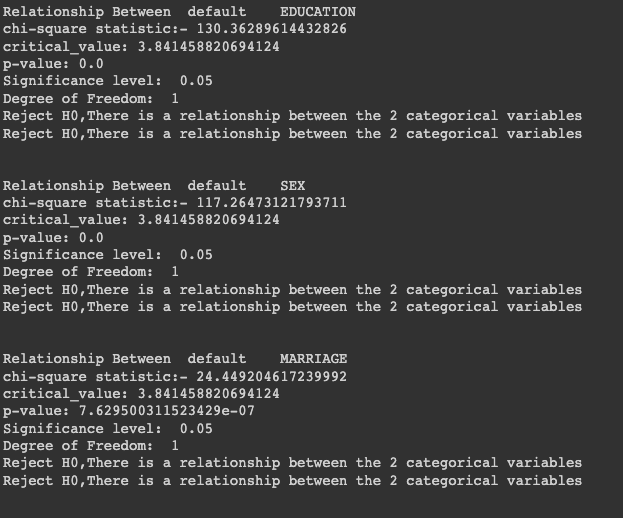
Now, if the p-value obtained after conducting the test is less than 0.05 we reject the Null hypothesis and accept the Alternate hypothesis and if the p-value is greater than 0.05 we accept the Null hypothesis and reject the Alternate hypothesis.

***Degree of freedom***: The degree of freedom in the Chi-Square test is calculated by (n-1)\*(m-1) where n and m are numbers of rows and columns respectively.

**Figure 5.**

****

**Figure 6.**

****

**3.3 Random Forest feature selection**

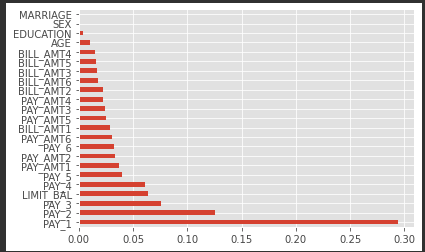
A random forest is a combination of decision trees, where the random forest has an internal thing called feature selection. Using this feature selection, we can find out what features are more important than other features based on random forest.

This feature selection in the random forest is done by the measure of impurity (Gini impurity). The features for internal nodes are selected with some criterion, which for classification tasks can be Gini impurity or information gain.

An average is calculated in the random forest, where each tree is calculated for classification and provides a decrease in accuracy. The features which are ay the top are used as important features, and the gestures which are at the bottom are shown as low feature importance features.

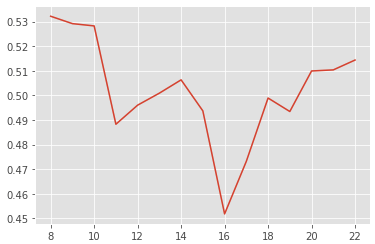
In our project, we have 23 different features, and without using our domain knowledge and other factors, we completely depended on the random forest feature selection method in this approach. As shown in Figure below, the PAY\_1 has the highest feature importance based on random forest, and the Marriage feature has the least among all other features.

**Figure 7.**



In this project, we have calculated from the eight most important features to all 23 features to know how many features are required to get the maximum score. We have calculated the AUC from the eight features to 23 features and plotted it as shown in Figure 8. From this observation, only the top eight features are more than enough to get better performance metrics, rather than having many less important features. So, we have considered the top eight important features from Random forest in the model

**Figure 8.** Calculated AUC. X-Axis is the number of features and Y-Axis is the AUC score.



The disadvantage of this process, these important features when used in a weak model would be completely useless. This feature selection method is strongly influenced by correlated features. In addition, this method is biased toward numerical data.

**3.4 Correlation Matrix**

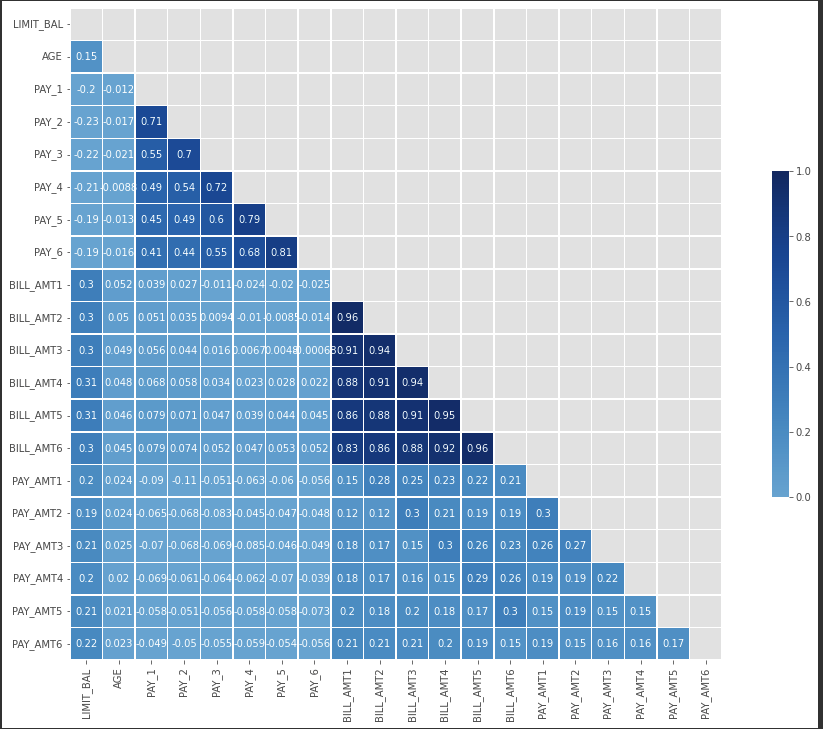
The statistical concept of correlation is frequently used to describe how nearly a linear relationship exists between two variables. If two variables are linearly dependent on each other, they will have higher correlation than the variables which are not. Thus when analyzing the data from the feature selection perspective, features which are highly correlated are more linearly dependent and hence equally affect the dependent variable. We can therefore omit one of the two features when there is a substantial correlation between two features.

Pearson Correlation Coefficient can be used with continuous variables to measure the linear relationship between variables. It has a value between -1 and 1 and measures the strength and direction of the relationship. Values between 0 and 1 signifies positive correlation, 0 signifies no correlation 0 and -1 signifies negative correlation.The chance of discovering an observation on the assumption that a certain hypothesis is correct is provided by the P-value. In order to accept or reject that hypothesis, one uses this probability.

Figure 9 shows the correlation matrix between the numerical features in our dataset. Darkest blue color shows the highly correlated features which have a correlation close to 1. Based on this matrix, all the bill amounts are highly correlated with the prior bill amounts and have values greater than 0.94 for example BILL\_AMT1 and BILL\_AMT2 have 0.96, BILL\_AMT2 and BILL\_AMT3 have 0.94, BILL\_AMT3 and BILL\_AMT4 have 0.94 etc.,

Based on the domain knowledge, we can conclude that the latest bill amount is the most important for predicting the credit card defaults as compared to the previous bills.So we decided to keep BILL\_AMT1 and discard BILL\_AMT2,BILL\_AMT3,BILL\_AMT4, BILL\_AMT5 and BILL\_AMT6.

**Figure 9.**

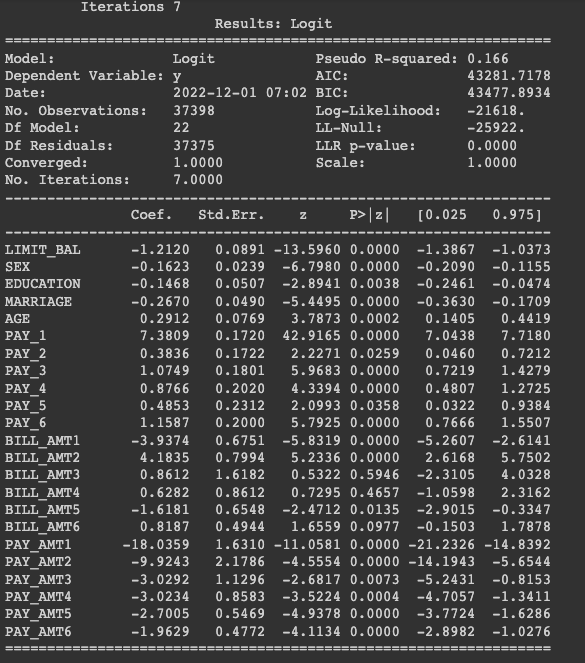
****

**3.5 Logistic regression**

In the logistic regression based model the amount of variations between features is interpreted by computing the p-value. This is an interactive process, if the initial p-value is greater than the threshold then it is removed from the dataset. If that's not the case, we add and move on the next feature till we meet convergence.

So as shown in the figure below, if the features of the p-value is less than 0.05 then the feature is more relevant, else if the feature of p-value is greater than 0.05 then the feature is discarded.

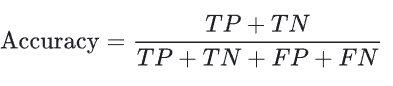
**Figure 10.**



**4. Results: Comparison between methods with different metrics**

Accuracy is one metric for evaluating the metric we have used. The accuracy is the ratio of the number of corrected predictions to the total number of predictions. Here the number of corrected predictions is the sum of true positives and true negatives. The total number of predictions is the sum of True positive, true negative, False positive, and False negative. The formula is shown below in Figure 9:

**Figure 11.**



Based on the accuracy we have calculated all the models with different combinations of feature selection methods as shown in the below table 1.

**Table 1.**Model Comparison based on Accuracy

| **Method** | **Logistic Regression** | **KNN** | **Decision Tree** | **Random Forest** | **XGBoost** |
| --- | --- | --- | --- | --- | --- |
| Without Feature Selection | 78% | 69% | 77% | 71% | 65% |
| T-Test, Chi-square | 78% | 67% | 77% | 67% | 65% |
| Correlation, Chi-square | 77% | 70% | 78% | 81.2% | 78% |
| Feature Selection by Feature Importance (Logistic Regression) | 78% | 70% | 77% | 79% | 76% |
| Feature Selection by Feature Importance (Random Forest) | 78% | 51% | 77% | 76% | 63% |

Instead of only focusing on the precision, and recall we have used F1- Score. The **F1-score** is the combination of the precision and recall into a single metric by taking their harmonic mean. The F1-score for all the models with the feature selection methods are shown in the below table 2.

**Table 2.**  Model Comparison based on F1 Score

| **Method** | **Logistic Regression** | **KNN** | **Decision Tree** | **Random Forest** | **XGBoost** |
| --- | --- | --- | --- | --- | --- |
| Without Feature Selection | 0.69 | 0.58 | 0.47 | 0.52 | 0.59 |
| T-Test, Chi-square | 0.54 | 0.37 | 0.52 | 0.28 | 0.45 |
| Correlation, Chi-square | 0.54 | 0.37 | 0.52 | 0.5 | 0.44 |
| Feature Selection by Feature Importance (Logistic Regression) | 0.54 | 0.39 | 0.52 | 0.49 | 0.45 |
| Feature Selection by Feature Importance (Random Forest) | 0.53 | 0.39 | 0.52 | 0.52 | 0.46 |

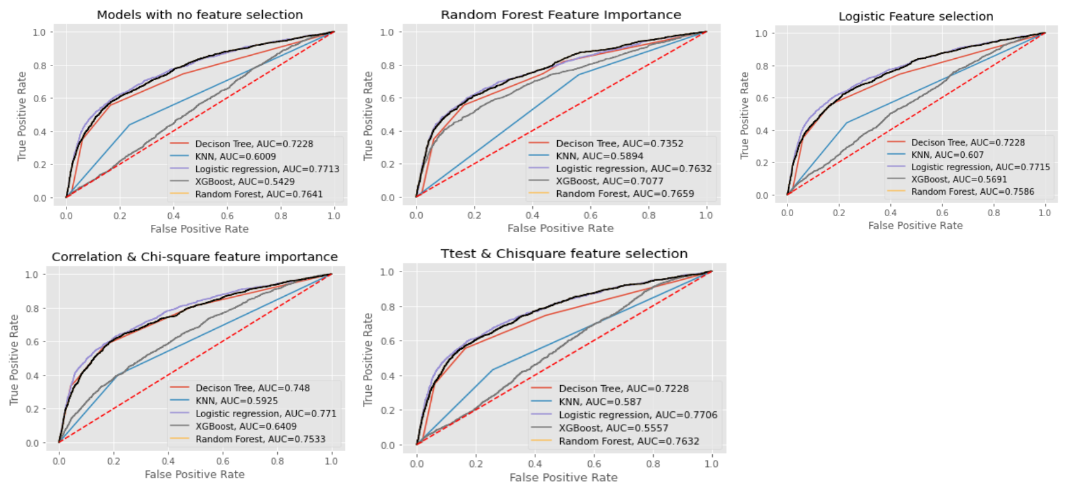
The AUC provides an aggregate measure of performance across all possible classification thresholds. The below table is based on AUC scores.

**Table 3** Model Comparison based on AUC score:

| **Method** | **Logistic Regression** | **KNN** | **Decision Tree** | **Random Forest** | **XGBoost** |
| --- | --- | --- | --- | --- | --- |
| Without Feature Selection | 0.77 | 0.60 | 0.72 | 0.59 | 0.69 |
| T-Test, Chi-square | 0.77 | 0.58 | 0.72 | 0.55 | 0.69 |
| Correlation, Chi-square | 0.77 | 0.59 | 0.74 | 0.75 | 0.73 |
| Feature Selection by Feature Importance (Logistic Regression) | 0.77 | 0.60 | 0.72 | 0.75 | 0.69 |
| Feature Selection by Feature Importance (Random Forest) | 0.76 | 0.58 | 0.73 | 0.76 | 0.73 |

The below figure provides all the models with different feature selection methods. Almost in all the models KNN had performed badly, and Random forest had performed well in all the feature selection methods.

**Figure 12.**



**5. Discussion**

**5.1 Why is one method better than other methods in your data?**

In this project, we have used different feature selection methods, which are *T-Test with Chi-Square, Correlation with Chi-Square, Feature selection using Logistics regression, and Feature selection using Random Forest.* Based on our dataset there are a few feature selection methods which had provided better results, and a few which provided less performance when used with different Machine learning models (Logistics regression, KNN, Decision Tree, Random Forest, and XGBoost).

**T-Test with Chi-Square:** A T-Test is also used to compare the mean of two given samples like the Z-test. The T-test is used on all the numerical data. In addition, Chi-Square is also used in finding better features in the categorical data.

The performance of this method is not so good because this project is using a smaller dataset. T-test works well with larger datasets.

**Correlation with Chi-square:** Chi-Square test is a statistical test which is used to find out the difference between the observed and the expected data. We can also use this test to find the correlation between categorical variables in our data. The purpose of this test is to determine if the difference between 2 categorical variables is due to chance, or if it is due to a relationship between them.

It is important to note that the variables to be compared should have only 2 categories i.e 1 and 0. The chi-square test fails to determine the correlation between variables with more than 2 categories. Displayed best performance metrics with respect to models performed.

**Logistic regression for Feature Selection:** This method is not the most optimized way of feature selection because it works on greedy algorithms. The p-value provides some value even if the feature has no relevance with the goal or target. Normally, p-value does not consider the effect of variable

**Random Forest Feature Selection Method:** This method had not performed well because of a few reasons. The dataset used in this project contains categorical data and the random forest feature selection method is completely biased toward the numerical data and high cardinality features.

Secondly, when used with weak models in our project, we achieved less accuracy and other performance metrics. This is because random forest provide important features that provide bad results for the weak models. In addition, it is also strongly influenced by correlation features.

So, from all the methods Correlation with Chi-square had performed well when compared to others as shown in below table 4.

**Table 4**

| **Method** | **Why this method?** | **Result** |
| --- | --- | --- |
| **T-Test, Chi-Square** | T-Test is also used to compare the mean of two given samples. T-test works best when samples are greater than 30.  This t-test is used on all the numerical data. Chi-square is used for all the categorical data available in the dataset | The given dataset was not that large. It is observed that t-test works better with large datasets. That’s why it might not have performed well. |
| **Correlation, Chi-Square** | We have 3 categorical features  and 23 numerical features.  Chi-square is used for feature selection of categorical features and correlation for numerical. Using the two methods together helps us in identifying the best features irrespective of being categorical or numerical. | Displayed best performance with respect to models. |
| **Feature Selection by Feature Importance (Logistic Regression)** | Logistic regression adds sparsity in the dataset.  It generates p-value for a feature based on its significance.  It can be used for categorical data, continuous data or even for a mix of both.  Hence we have used this method for feature selection | Logistic Regression works on a greedy algorithm which is why it is not the most optimized way of feature selection. Normally, p-value does not consider the effect of variable  Even if the feature is completely irrelevant, it would still generate some p-value based on significance |
| **Feature Selection by Feature Importance (Random Forest)** | It is computed based on Gini Impurity. It will provides feature importance for the Random Forest.The advantage in this method is a speed of computation, all the needed values are computed during the RF training | They are completely useless if your model is weak.  Strongly Influenced by correlation features.  They are biased towards numerical and high cardinality features. |

**5.2 Difference between all features and selected features. Why are selected features important?**

| **Method** | **Selected features** | **Discarded features** |
| --- | --- | --- |
| **T-Test** | Limit\_BAL, AGE, PAY\_1, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6, BILL\_AMT1, PAY\_AMT1, PAY\_AMT2, PAY\_AMT3, PAY\_AMT4, PAY\_AMT5, PAY\_AMT6 | BILL\_AMT2, BILL\_AMT3, BILL\_AMT4, BILL\_AMT5, BILL\_AMT6 |
| Out of the Bill amounts for 6 months, only the most recent month bill amount is selected. The repayment status and amount of previous payments for all the previous 6 months is selected. Also credit limit and age features are selected.  The Bill amounts for all the months prior to the most recent month are discarded.  The repayment status and amount of payments play a more important role as compared to the actual bill amount.  For example, for the given set of features in the dataset, even if the amount of a bill is high or low makes a very little difference if the individual is able to pay it on time.  However the pending amount for the most recent month is important to know the amount that is due for the customer. | |
| **Correlation** | PAY\_1, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6, BILL\_AMT1, PAY\_AMT1, PAY\_AMT2, PAY\_AMT3, PAY\_AMT4, PAY\_AMT5, PAY\_AMT6, Limit\_BAL, AGE | BILL\_AMT2, BILL\_AMT3, BILL\_AMT4, BILL\_AMT5, BILL\_AMT6 |
| We are finding the important features from the continuous variables. And payment status, bill amount and payment amount are important features.  Also Age and Limit\_balance have very low correlation with any of the features, hence both are selected  There was a high correlation between all the Bill amounts. Hence we decided to select only one bill amount which is for the most recent month.  Based on our understanding, the payment status of all the previous months is more important than the bill amounts | |
| **Chi-Square** | SEX, EDUCATION, MARRIAGE | NONE |
| None of the categorical features are discarded  All the categorical features are important. For education level, one of the possibilities is with higher education, the chances of a stable earning may be higher and hence level of education is an important feature.  Marital status may increase the credit limit for a household and also it is possible to make the payments in time with some help from the spouse in case required. | |
| **Feature Selection by Feature Importance (Logistic Regression)** | Limit\_BAL, SEX, EDUCATION, MARRIAGE, AGE, PAY\_1, PAY\_3, PAY\_4, PAY\_5, PAY\_6, BILL\_AMT1, BILL\_AMT2, BILL\_AMT5, PAY\_AMT1, PAY\_AMT2, PAY\_AMT3, PAY\_AMT4, PAY\_AMT5, PAY\_AMT6 | BILL\_AMT3, BILL\_AMT4, BILL\_AMT6 |
| In this method, most of the categorical features are selected as important such as mariage, age, sex etc. All the repayment status is selected, which is most important when compared to our problem.  There was a high correlation between all the Bill amounts. Hence we decided to select only one bill amount which is for the most recent month.  All the discarded features have p-value greater than 0.05. The model discards these three features because bill statements of the first two are enough for the model to predict the credit card default. | |
| **Feature Selection by Feature Importance (Random Forest)** | PAY\_1, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6, PAY\_AMT1, Limit\_BAL, BILL\_AMT1, PAY\_AMT3 | PAY\_AMT6, BILL\_AMT1, PAY\_AMT3, PAY\_AMT5, PAY\_AMT4, BILL\_AMT2, BILL\_AMT6, BILL\_AMT3, BILL\_AMT5, BILL\_AMT4, AGE, EDUCATION, SEX, MARRIAGE |
| The goal is to predict if the customer is going to pay or not in the future. So, the most important features are the repayment features (such as Pay\_1 to Pay\_6 etc). Only a very few Bill statements were considered.  Most of the Bill statements (BILL\_AMT) were also discarded because the amount of bill statements have very less importance as per Random forest model.  Here the categorical data is discarded, because this method is more biased towards numerical data. | |

**5.3 What is the meaning of your result? How to explain your result (interpretability) based on your domain knowledge and references?**

As per model results the best feature selection method for the dataset is Correlation with Chi-Square. TheFinal Selected Features are

* + PAY\_1, PAY\_2, PAY\_3, PAY\_4, PAY\_5, PAY\_6 (Repayment Status of September to April)
  + BILL\_AMT1 (Amount of Bill Statement of September)
  + PAY\_AMT1, PAY\_AMT2, PAY\_AMT3, PAY\_AMT4, PAY\_AMT5, PAY\_AMT6 (Amount of Previous Payment from September to April)
  + SEX, MARRIAGE, EDUCATION
* The best model for our dataset is **Random Forest** when it uses the features selected by **Correlation and Chi-square methods**. However we could see that XGBoost did not perform as well as Random Forest for any of the feature selection methods. We have many outliers in the BILL\_AMT\*, PAY\_AMT\* columns which consist of the data amount of the bill statement and we could not discard the outliers as it would have reduced the size of our data. One of the possible reasons may be because XGBoost is very sensitive to outliers. Also based on the domain knowledge we believe that more features are needed to improve the performance of the models
* Extracted Knowledge form the data: Repayment status, Amount of bill, Historical payment status, sex, marriage and education are important features from the given dataset.
* The most recent bill amount and payment status are relatively more important as compared to more dated bill amount and payment status.
* From the model results we could see that the accuracy ranges in between 70%-80% for all the models irrespective of feature selection methods.
* Hence, to make better prediction we believe adding more features to the dataset is important:
  + Income which signifies customers repayment capability i.e. if the customer has means/ potential to pay the amount that is due to be paid every month, and it will also be useful in defining the credit limit of the customer
  + Credit Score which signifies customers overall financial credibility across banks is required. It will tell us if the customer is credible enough to receive a credit card.

**Reference**

Ashofteh, A., & Bravo, J. M. (2021, August). A conservative approach for online credit scoring. *Expert Systems With Applications*, *176*, 114835. https://doi.org/10.1016/j.eswa.2021.114835

Bacova, A., & Babic, F. (2021, January 21). Predictive Analytics for Default of Credit Card Clients. *2021 IEEE 19th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*. <https://doi.org/10.1109/sami50585.2021.9378671>

*UCI Machine Learning Repository: default of credit card clients Data Set*. (n.d.). Retrieved October 21, 2022, from https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients

Khandani, A. E., Kim, A. J., & Lo, A. W. (2010, November). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking &Amp; Finance*, *34*(11), 2767–2787. <https://doi.org/10.1016/j.jbankfin.2010.06.001>

Subasi, A., & Cankurt, S. (2019, June). Prediction of default payment of credit card clients using Data Mining Techniques. *2019 International Engineering Conference (IEC)*. <https://doi.org/10.1109/iec47844.2019.8950597>

M. Paprzycki et al. (eds.), Intelligent Systems, Technologies and Applications, Advances in Intelligent Systems and Computing 1353, https://doi.org/10.1007/978-981-16-0730-1\_1

Yeh, I. C., & Lien, C. H. (2009, March). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems With Applications*, *36*(2), 2473–2480. https://doi.org/10.1016/j.eswa.2007.12.020