**Credit Card Default Predication**

Pravin. M, Data Science Trainee, Almabetter, Bangalore

**Abstract:**

A credit card issuer is the bank or credit union that provides the credit card and lends the money used in a transaction. A credit card network is the entity that processes each credit card transaction, handling the technical aspects of electronically moving the money around. Taiwan based credit card Company tasked at predicting the case of customers default payments. It is one of the crucial factors for company’s falls under BFSI domain. As their nature of business involves huge volumes of money which is used for giving credit to their customer base. And making profits from the returns. To increase profit from the return one should choose their customer wisely. Here comes the need for identifying valuable customers for their business purpose. This enables us build a model which is capable of identifying the most valuable customers.

**Keywords:**

Credit Card Default Prediction,

Default Prediction,

BFSI domain,

Credit Risk Analysis,

Risk Management,

Descriptive Analytics,

Exploratory Data Analysis,

Univariate Analysis,

Bivariate Analysis,

Multivariate Analysis,

Predictive Analytics,

Machine Learning,

Supervised Learning,

Classification,

Bagging,

Boosting,

Ensembles,

Hyper parameter tuning,

Randomized Search CV,

Logistic Regression,

Support Vector Classifier,

Random Forest Classifier,

Xtreme Gradient Boosting Classifier,

Confusion Matrix,

ROC\_AUC Curve,

Precision, Recall, F1 Scores,

K-S Statistic

**Table of Content**

|  |  |  |
| --- | --- | --- |
| **Sl.NO** | **CONTENTS** | **PG.NO** |
| 1 | Problem Statement | 4 |
| 2 | Business Goal | 4 |
| 3 | Introduction | 4 |
| 4 | Dataset Description | 5 |
| 5 | Data Wrangling | 6 |
| 6 | Descriptive Analytics – Exploratory Data Analysis | 7 |
| 7 | Feature Engineering | 8 |
| 8 | Predictive Analytics – ML Modelling | 9 |
| 9 | Results Comparison | 16 |
| 10 | Feature Importance | 18 |
| 11 | Conclusion | 18 |
| 12 | References | 19 |

**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.

**Business Goal**

Our Aim is to develop a model which is capable of predicting whether a customer is eligible for credit or not. 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. Because the real probability of default is unknown. To know whether our model is able to differentiate between credible and non-credible customers. We can use the K-S chart to evaluate which customers will default on their credit card payments.

**Introduction**

A credit card is a physical card that can be used to make purchases, pay bills or depending on the card, withdraw cash. A credit card is a type of unsecured borrowing under which a bank or NBFC agrees to offer you a predefined credit limit. You can make transactions up to the said limit and pay it back on the due date or convert the transaction into EMI and pay over a period of a few months. Unlike personal loans or car loans which are installment loans accounts, credit cards have a revolving credit account. This means you can borrow money on the same account again and again as you keep paying the dues. As said above, a credit card lets you borrow money (up to the given credit limit) and pay it back as and when due. When you make a purchase, the amount will be deducted from your credit limit and when you pay it back, the payment will be added back to your credit limit. This gives regular access to credit as long as you do not max out your limit. Taiwan based credit card Company tasked at predicting the case of customers default payments. It is one of the crucial factors for company’s falls under BFSI domain. As their nature of business involves huge volumes of money which is used for giving credit to their customer base. And making profits from the returns. To increase profit from the return one should choose their customer wisely. Here comes the need for identifying valuable customers for their business purpose. This enables us build a model which is capable of identifying the most valuable customers.

**Dataset 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.

**Data Wrangling**

* Importing and loading the dataset as credit\_data.
* Columns header has no names instead it is encoded with variable like X1 – X23.
* Initially, the column header is replaced with respective column names of all variables.
* And then the variable ID is set to the index of the dataset.
* Certain column names don’t have proper or chronological name for the variable. So it is need to be renamed.
* Renaming column name of PAY\_0 to PAY\_1 because the payment cycle starts from PAY\_1.
* Renaming column name of MARRIAGE to MARITAL\_STATUS.
* And renaming column name of default payment next month to DEF\_PAY\_NMO.
* For better understanding purpose the encoded values of categorical variables like SEX, MARITAL\_STATUS, DEF\_PAY\_NMO, and EDUCATION needed to encode with easily understandable names.
* Replacing values in SEX variable like 1: M, and 2: F.
* Replacing values in EDUCATION variable like 1: GS, 2: Uni, 3: HS, 4: Others.
* Replacing values in MARITAL\_STATUS variable like 1: M, 2: S, 3: O.
* Replacing values in DEF\_PAY\_NMO variable like 1: Def, 0: No\_Def.
* AGE variable is classified further into three as Young (age <= 40), Middle Aged (age > 40 & <= 60), Old Aged (age > 60).
* LIMIT\_BAL variable is classified further into three as low\_credit (Balance <= 250000), medium\_credit (Balance > 250000 & <= 500000) and high\_credit (Balance > 500000).
* And the dataset has no missing values. So there is no need for missing values imputation.

**Descriptive Analytics – Exploratory Data Analysis**

Descriptive Analytics is the examination of data or content, usually manually performed, to answer the question “What happened?” (or What is happening?), characterized by traditional business intelligence and visualizations such as pie charts, bar charts, line graphs, tables, or generated narratives. After data cleaning up. The dataset is ready for exploratory data analysis. Exploratory Data Analysis for the credit card default prediction consists of 3 sections as follows.

1. **Univariate Analysis**

It is the simplest form of analysing data. Uni means one, so in other words your data has only one variable. It doesn’t deal with causes or relationships and its major purpose is to describe. It takes data, summarizes that data and finds patterns in the data.

Here the univariate analysis involves analysing deeply into the nature of each variables specifically. It is achieved by employing data visualizing techniques. Variables involved in analysing are as follows

* LIMIT\_BAL
* SEX
* EDUCATION
* MARITAL\_STATUS
* AGE
* DEF\_PAY\_NMO and
* PAY\_1 to PAY\_2.

1. **Bivariate Analysis**

Bivariate analysis means the analysis of bivariate data. It is one of the simplest forms of statistical analysis, used to find out if there is a relationship between two sets of values. It usually involves the variables X and Y.

Here the univariate analysis involves analysing deeply into the nature of each variables specifically. It is achieved by employing data visualizing techniques. This analysis done between the dependent variable **“DEF\_PAY\_NMO”** vs all the **“Independent Variables”**. The main focus of this analysis is to find significant relationship between the dependent variable and independent variable.

1. **Multivariate Analysis**

Multivariate analysis is used to study more complex sets of data than what univariate analysis methods can handle. This type of analysis is almost always performed with software as working with even the smallest of data sets can be overwhelming by hand.

Here the univariate analysis involves analysing deeply into the nature of each variables specifically. It is achieved by employing data visualizing techniques. This analysis is focused on finding correlation between variables like

* DEF\_PAY\_NMO,
* SEX,
* EDUCATION,
* LIMIT\_BAL,
* PAY\_1 – PAY\_6,
* BILL\_AMT1 – BILL\_AMT6,
* PAY\_AMT1 – PAY\_AMT6,
* AGE, and
* MARITAL\_STATUS.

**Feature Engineering**

Feature Engineering is also known as pre-processing. It involves imputation of dataset were ever it requires. It is achieved with help of certain libraries available. In this credit\_data dataset all the categorical variables needs to be encoded which is done by means of one-hot encoding. After this the dependent variable “DEF\_PAY\_NMO” has imbalanced data with itself. Its need to be rectified with the help of Synthetic Minority Oversampling Technique (SMOTE). After imbalance treatment done, the dataset has to be split into training and testing datasets by means of train-test splitter. credit\_data dataset has some columns those are right skewed in nature to avoid any bias as well as overfitting during modelling it is need to standardized by means of standard scaler. This enables us to make the dataset in a proper format (free from missing values, imbalanced data, & etc.,) and ready for modelling.

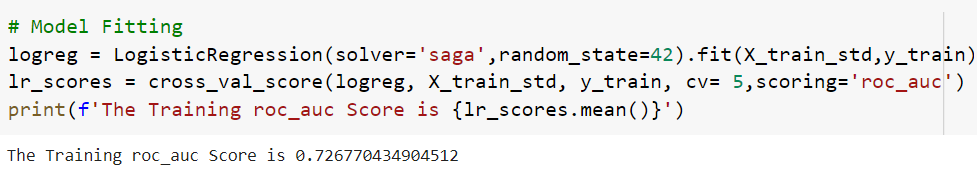
**Predictive Analytics – ML Modelling**

The term predictive analytics refers to the use of statistics and modelling techniques to make predictions about future outcomes and performance. Predictive analytics looks at current and historical data patterns to determine if those patterns are likely to emerge again. This allows businesses and investors to adjust where they use their resources to take advantage of possible future events. Predictive analysis can also be used to improve operational efficiencies and reduce risk. In BFSI domain businesses predictive analytics plays a protagonist role in their daily operations. Machine Learning algorithms are really capable of learning patterns from current and historical data. And also able to predict based on the training. This enables us to build a model based on business demand. Our Business goal is to predict the probability of credible and non-credible customers. So classification type algorithms best suits for our business goal.

* **Logistic Regression**

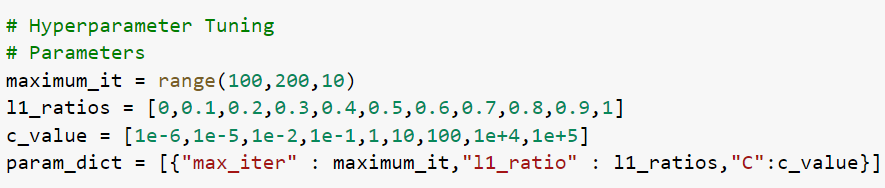
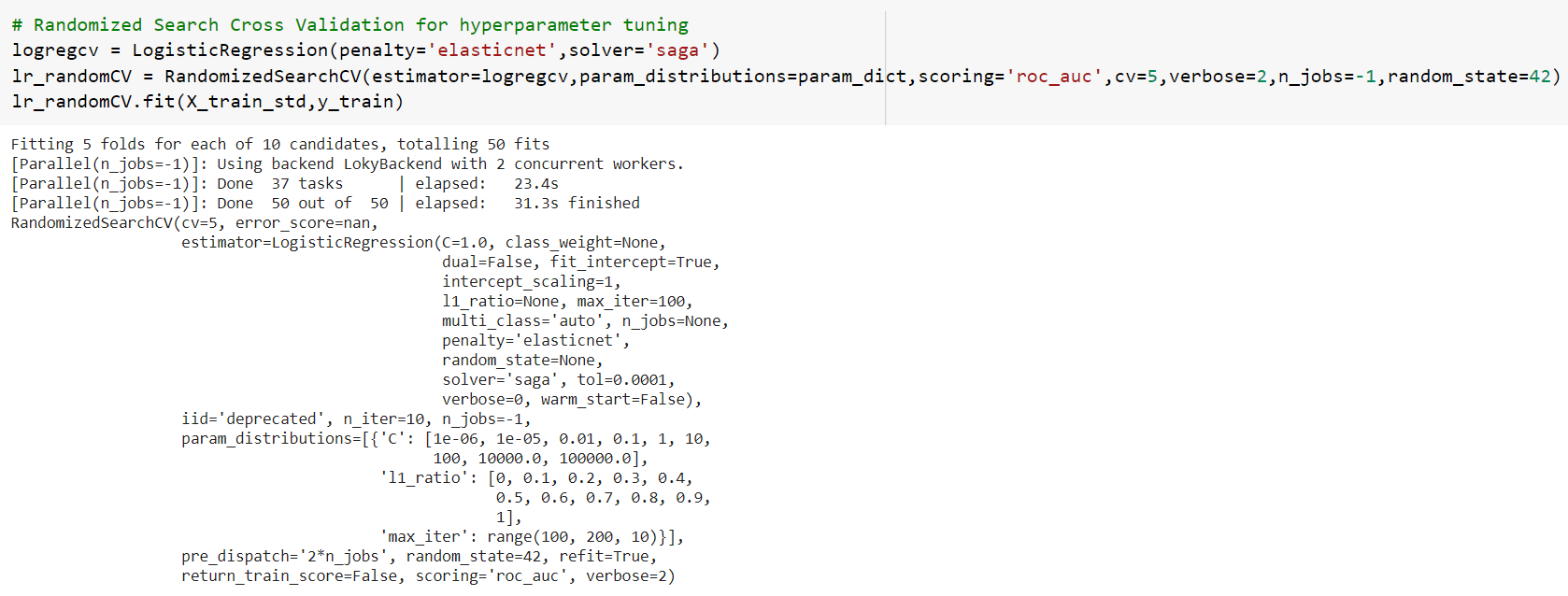
1. **Model Training**

Initially the Logistic Regression model is initiated. After that training set is fitted in the model. Then model is allowed to learn the pattern from the training dataset. After training completion. It is evaluated using anyone classification metric here we used “ROC-AUC” for evaluation. Based on evaluation score the model is improved.



1. **Hyper-parameter Tuning**

The main objective for hyper-parameter tuning is to improve model performance in testing datasets. Each model has their own hyper-parameters. Here in Logistic Regression we had focused on improving some selected hyper-parameters which might improve the overall model performance in testing dataset. The hyper-parameter tuning is done by means of Cross Validation. Here we had implemented cross validation by means of RandomizedSearchCV it helps us by reducing the effort and computational timing while performing cross validation.

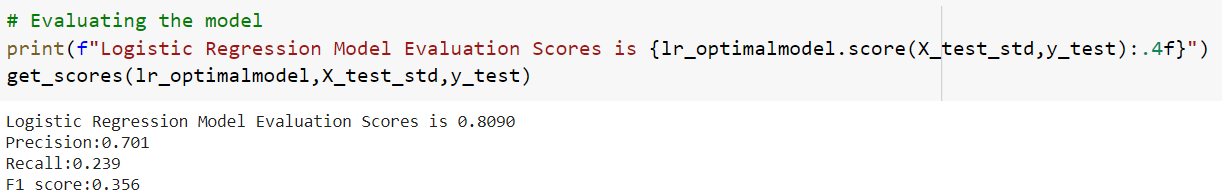
 

1. **Predicting**

After hyper-parameter tuning. The model is now allowed to make predictions on testing dataset based on its accumulated training knowledge.

1. **Evaluating / Validating the model**

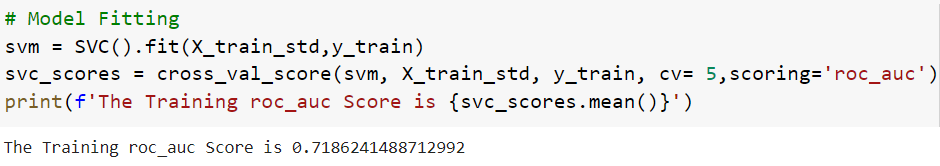
After prediction done. The Predicted results yhat are compared with original result y. And evaluation done based on classification metrics such as Accuracy, Precision, ROC-AUC, Recall, F1 Score, and K-S Statistic.



* **Support Vector Classifier**

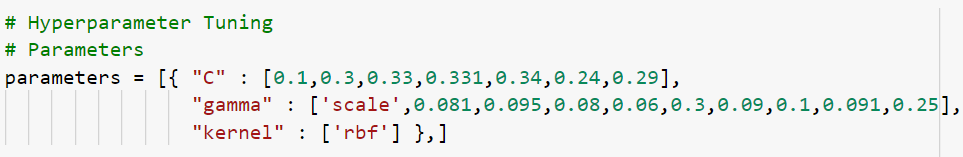
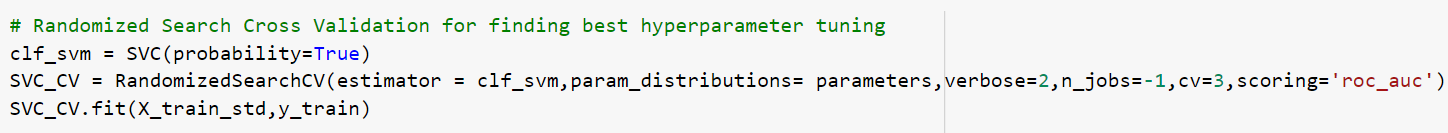
1. **Model Training**

Initially the Support Vector Classifier model is initiated. After that training set is fitted in the model. Then model is allowed to learn the pattern from the training dataset. After training completion. It is evaluated using anyone classification metric here we used “ROC-AUC” for evaluation. Based on evaluation score the model is improved.



1. **Hyper-parameter Tuning**

The main objective for hyper-parameter tuning is to improve model performance in testing datasets. Each model has their own hyper-parameters. Here in Support Vector Classifier we had focused on improving some selected hyper-parameters which might improve the overall model performance in testing dataset. The hyper-parameter tuning is done by means of Cross Validation. Here we had implemented cross validation by means of RandomizedSearchCV it helps us by reducing the effort and computational timing while performing cross validation.

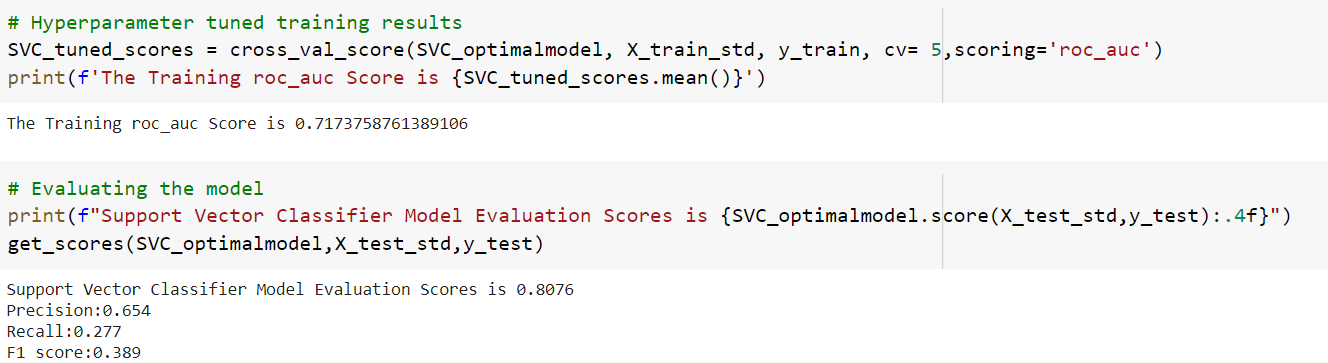
 

1. **Predicting**

After hyper-parameter tuning. The model is now allowed to make predictions on testing dataset based on its accumulated training knowledge.

1. **Evaluating / Validating the model**

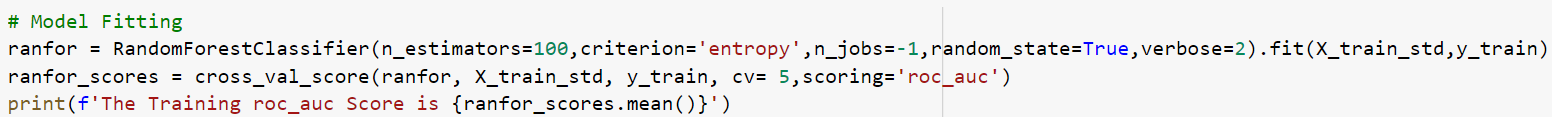
After prediction done. The Predicted results yhat are compared with original result y. And evaluation done based on classification metrics such as Accuracy, Precision, ROC-AUC, Recall, F1 Score, and K-S Statistic.



* **Random Forest Classifier**

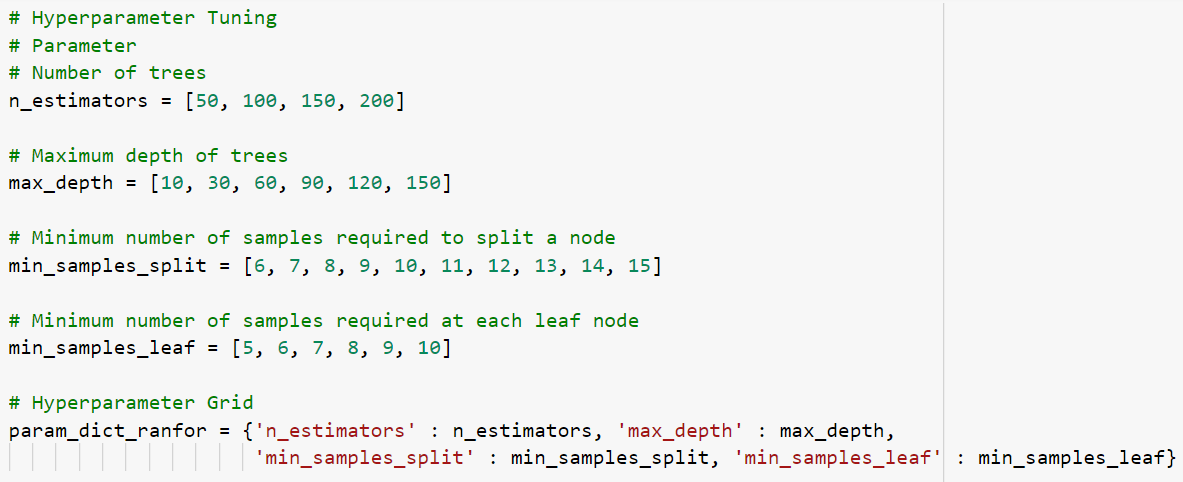
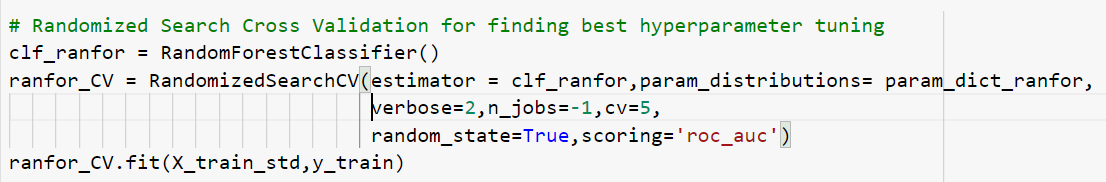
1. **Model Training**

Initially the Random Forest Classifier model is initiated. After that training set is fitted in the model. Then model is allowed to learn the pattern from the training dataset. After training completion. It is evaluated using anyone classification metric here we used “ROC-AUC” for evaluation. Based on evaluation score the model is improved.

1. **Hyper-parameter Tuning**

The main objective for hyper-parameter tuning is to improve model performance in testing datasets. Each model has their own hyper-parameters. Here in Random Forest Classifier we had focused on improving some selected hyper-parameters which might improve the overall model performance in testing dataset. The hyper-parameter tuning is done by means of Cross Validation. Here we had implemented cross validation by means of RandomizedSearchCV it helps us by reducing the effort and computational timing while performing cross validation.

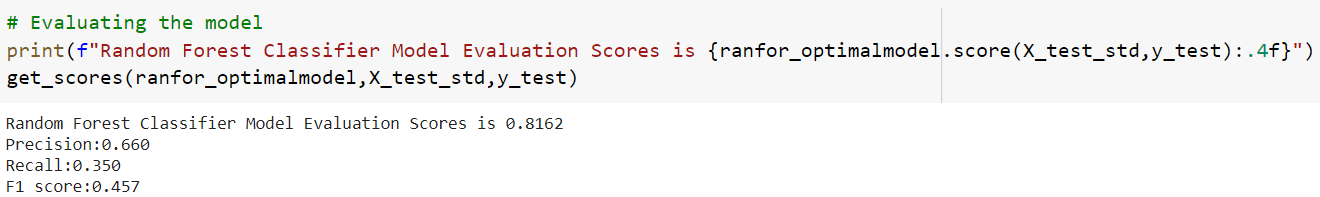
 

1. **Predicting**

After hyper-parameter tuning. The model is now allowed to make predictions on testing dataset based on its accumulated training knowledge.

1. **Evaluating / Validating the model**

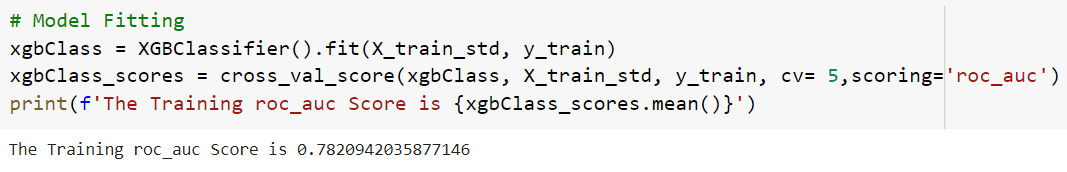
After prediction done. The Predicted results yhat are compared with original result y. And evaluation done based on classification metrics such as Accuracy, Precision, ROC-AUC, Recall, F1 Score, and K-S Statistic.



* **XGBoost Classifier**

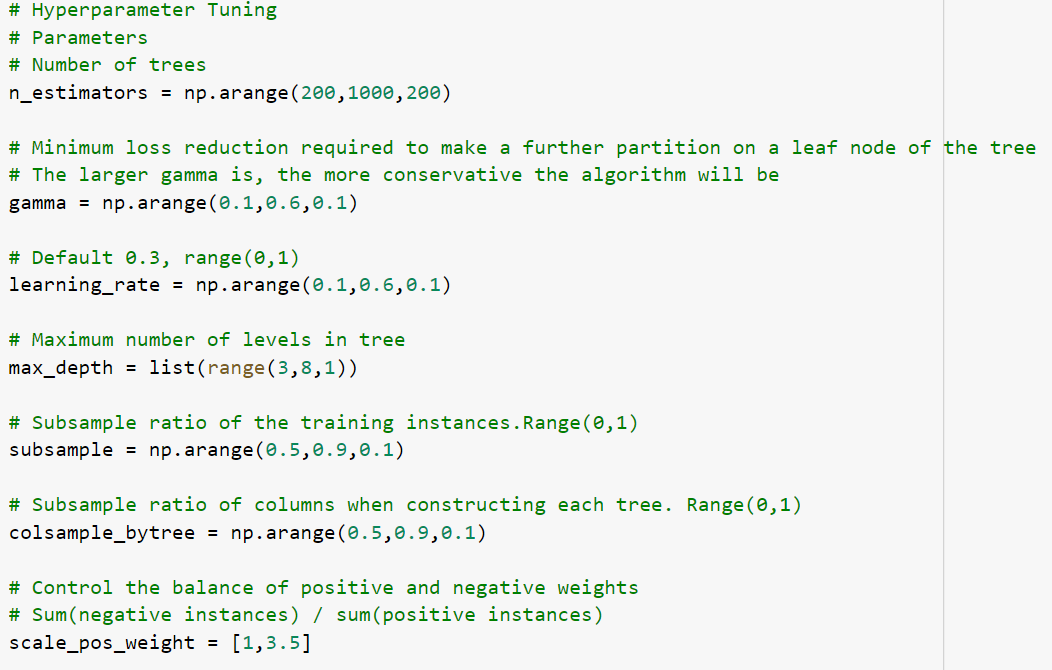
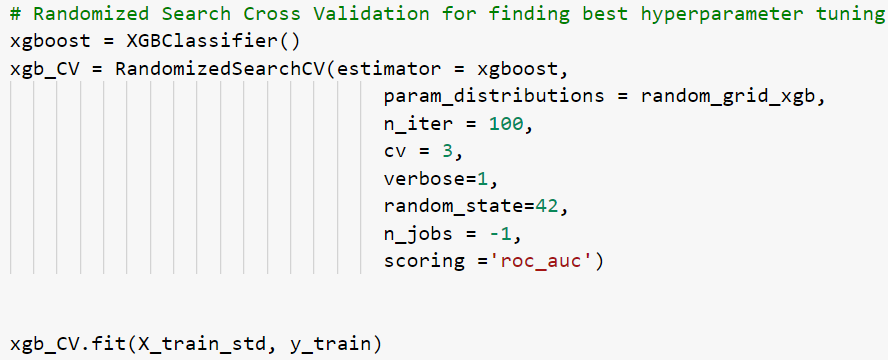
1. **Model Training**

Initially the Xtreme Gradient Boosting model is initiated. After that training set is fitted in the model. Then model is allowed to learn the pattern from the training dataset. After training completion. It is evaluated using anyone classification metric here we used “ROC-AUC” for evaluation. Based on evaluation score the model is improved.



1. **Hyper-parameter Tuning**

The main objective for hyper-parameter tuning is to improve model performance in testing datasets. Each model has their own hyper-parameters. Here in XGBoost Classifier we had focused on improving some selected hyper-parameters which might improve the overall model performance in testing dataset. The hyper-parameter tuning is done by means of Cross Validation. Here we had implemented cross validation by means of RandomizedSearchCV it helps us by reducing the effort and computational timing while performing cross validation.

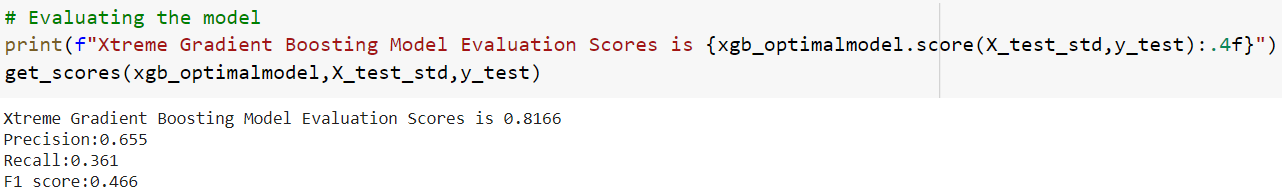
 

1. **Predicting**

After hyper-parameter tuning. The model is now allowed to make predictions on testing dataset based on its accumulated training knowledge.

1. **Evaluating / Validating the model**

After prediction done. The Predicted results yhat are compared with original result y. And evaluation done based on classification metrics such as Accuracy, Precision, ROC-AUC, Recall, F1 Score, and K-S Statistic.



**Results Comparison**

* **Confusion Matrix**

A confusion matrix is a technique for summarizing the performance of a classification algorithm. Classification accuracy alone can be misleading if you have an unequal number of observations in each class or if you have more than two classes in your dataset. Calculating a confusion matrix can give you a better idea of what your classification model is getting right and what types of errors it is making.

* **ROC-AUC Curve Comparison**

ROC - AUC curve is a performance measurement for the classification problems at various threshold settings. ROC is a probability curve and AUC represents the degree or measure of separability. It tells how much the model is capable of distinguishing between classes. Higher the AUC, the better the model is at predicting 0 classes as 0 and 1 classes as 1.

* **Precision Recall Curve Comparison**

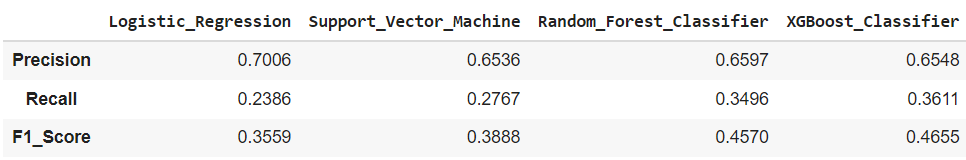
The precision-recall curve shows the trade-off between precision and recall for different threshold. A high area under the curve represents both high recall and high precision, where high precision relates to a low false positive rate, and high recall relates to a low false negative rate. High scores for both show that the classifier is returning accurate results (high precision), as well as returning a majority of all positive results (high recall).

* **K-S Statistic Comparison**

Kolmogorov-Smirnov (KS) Statistics is one of the most important metrics used for validating predictive models. It is widely used in BFSI domain. It compares the two cumulative distributions and returns the maximum difference between them. It is a non-parametric test which means you don't need to test any assumption related to the distribution of data. In KS Test, Null hypothesis states null both cumulative distributions are similar. Rejecting the null hypothesis means cumulative distributions are different. In simple words, it helps us to understand how well our predictive model is able to discriminate between events and non-events.

* **Overall Performance Comparison**

It involves comparing Precision, Recall, and F1 Score of each model. It also helps us to take right decision in choosing a best model with optimal overall performance.



* **Modelling Inference**

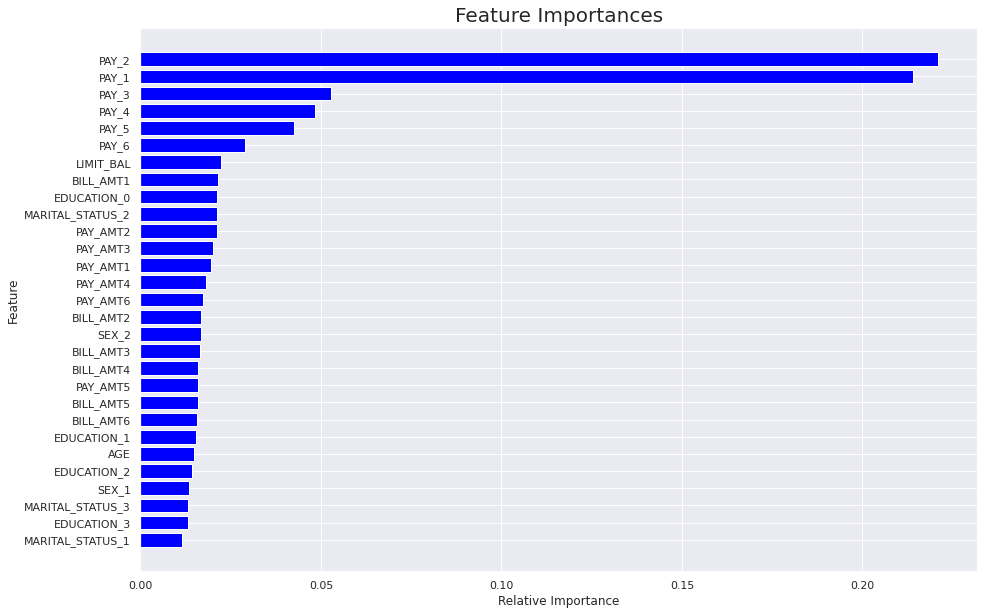
Based on KS Statistic and other metrics like Precision, Recall, F1 Score and ROC - AUC Curves. The Good Model is selected based on following criteria.

1. KS Statistic - To select a model based on KS Statistic it should have KS value above 40 in top 3 decile. **"XGBoost Classifier have KS = 42.6% in 3rd decile"**
2. ROC-AUC - To select a model based on ROC-AUC it should have highest score among all other model. **"XGBoost Classifier have ROC-AUC = 81%"**
3. Precision - Here the Business Nature lies under BFSI domain so mostly recall is given higher priority.
4. Recall - For BFSI domain problem recall value should be higher.

**"XGBoost has Recall = 36%"**

**Feature Importance**

After Results Comparison. Feature Importance is done. So that we can able to categorize all the variables based on their importance in contributing for Predicted outcomes.

****

**Conclusion**

* PAY\_2 and PAY\_1 are the most recent 2 months payment status and they are the strongest predictors of future payment default risk.
* If business demanded a model with high precision score then – Logistic Regression would be best fit model.
* If business demanded a model with high recall score then – XGBoost Classifier would be best fit model.
* Overall Best Performer based on ROC – AUC, Precision, Recall, F1 Score and KS Statistic then – XGBoost Classifier is the best model.

**References:**

* Paisabazzar - [www.paisabazaar.com](http://www.paisabazaar.com)
* Investopedia - [www.investopedia.com](http://www.investopedia.com)
* Listendata - [www.listendata.com](http://www.listendata.com)
* Machinelearningmastery – <http://www.machinelearningmastery.com>
* Towards Data Science – <www.towardsdatascience.com>