Predicting Credit Card defaults through classification algorithms of Machine Learning

Abstract

Credit Card, since the day of its existence, revolutionized the contemporary short-term credit market. Credit cards have been one of the most thriving financial services by the banks over the past years. However, with the rise in the number of credit card users, banks have been experiencing an escalating credit card default rate. This paper is an attempt to address this problem of predicting Credit Card defaults based on the previous month's payment history and records using Machine Learning techniques on python. One of the most common methods to predict credit card defaults is Binary Logistic Regression.

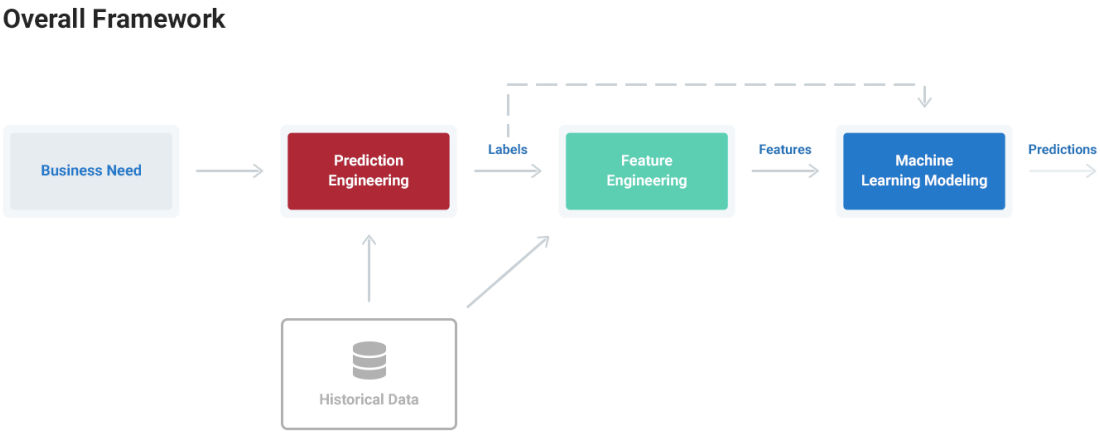
This paper provides a prediction accuracy comparison among various classification models of Machine learning. The three Machine learning algorithms used here are Binary Logistic Regression, Random forest, and XGBoosting (Extreme Gradient Boosting). These algorithms are run on the data set of credit card clients in the USA for 6 months. We observed the highest AUC value of 0.78 for XGBoosting and AUC value of 0.758 for the Random Forest method.

Introduction

The science of Machine Learning has become widely popular and influential in the past few years. Presently, almost every organization whether, government, corporate houses, Tech firms, startups or, research institutions heavily, rely on Machine learning algorithms and methods. Traditionally, software engineering combined human- created *rules* with *data* to **create answers to a problem**. Instead, machine learning uses *data* and *answers* to **discover the rules behind a problem**. (Chollet, 2017). [1]

Machine Learning is a scientific study that enables computer systems to execute tasks without giving a set of specific instructions or commands. Machine Learning works on algorithms to construct mathematical models based on training data in order to make predictions or decisions without being explicitly programming to do so. In layman terms, basically what machine learning

does is that it uses historical data and recognizes the patterns and trends in that data and then estimate predictions and develops the basis for the decision. This feature of Machine Learning makes it more adept for fostering solutions that are flexible to changing business parameters.
The general machine learning framework is outlined below:



There are different types of Machine Learning Algorithms based on the type of input or output data or the task they are designed to achieve.
They can be comprised into four major learning algorithms

a. Supervised Learning
b. Unsupervised Learning
c. Semi-supervised Learning
d. Reinforcement Learning

The Prediction output from a Supervised Machine Learning Model can be either a classification output or a regression output. A classification output is the one that ends up being categorical from a finite set. For ex. [Yes, No] or [low, medium, high].
A regression output is the one where output has a numerical value within a range. For ex. Real no.
In this paper, we have only focused on the classification output as we aim to predict whether a credit card user will default or not default in next month's bill payment. Thus, our prediction output can be classified into two categories, either default or not default.
The classification algorithms mostly run under the concept of linear separability i.e. different data points can be separated by a line known as "**Decision Boundaries**". The most common types of

classification algorithms are Logistic Regression, Support Vector Machines, Decision Trees, Neural Networks, Learning Vector Quantization.

With the continuous expansion in credit market and mounting debts, banks face severe risks of default; Thus the prediction of monthly EMI payments, credit card bills, etc. serves an essential purpose in determining the profitability of the banks. Thus, this estimation of likelihood of whether a client or account holder will default or not next month falls under classification problem. Hence, we have used three different kinds of classification algorithms (**Random Forest**, **XGBoosting** & **Logistic Regression**) to forecast credit card defaults.

Data and Methodology

The data used for this fraud prediction model mainly belongs to real-time transaction history of randomly selected users of a known credit card company whose name cannot be disclosed.

This dataset contains information on default payments, demographic factors, credit data, history of payment, and bill statements of credit card clients in the USA from April 2019 to September 2019[2]. The data is a table of 3000 rows and 25 columns of variables, including gender, bill amount and repayment status. The data used were already labeled by the bank as default or no default. The objective is to identify the most accurate model among the three classification models using binary logistic regression, random forest, and XGBoost. The data acquired is already preprocessed as there were no missing values based on the original variables, various transformations were conducted to the data in accordance with distribution, such as log transformation, data discretization or standardization to create more derivatives variables. We have explored the data by investigating density plots of various features, and then different predictive models were created based on that.

The 25 variables used in data are explained as follows:

- **ID**: ID of each client

- **LIMIT_BAL**: Amount of given credit in NT dollars (includes individual and family/supplementary credit
- **SEX**: Gender (1=male, 2=female)
- **EDUCATION**: (1=graduate school, 2=university, 3=high school, 4=others, 5=unknown, 6=unknown)
- **MARRIAGE**: Marital status (1=married, 2=single, 3=others)
- **PAY_0**: Repayment status in September, 2019 (-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)
- **PAY_2**: Repayment status in August, 2019 (scale same as above)
- **PAY_3**: Repayment status in July, 2019 (scale same as above)
- **PAY_4**: Repayment status in June, 2019 (scale same as above)
- **PAY_5**: Repayment status in May, 2019 (scale same as above)
- **PAY_6**: Repayment status in April, 2019 (scale same as above)
- **BILL_AMT1**: Amount of bill statement in September, 2019 (NT dollar)
- **BILL_AMT2**: Amount of bill statement in August, 2019 (NT dollar)
- **BILL_AMT3**: Amount of bill statement in July, 2019 (NT dollar)
- **BILL_AMT4**: Amount of bill statement in June, 2019 (NT dollar)
- **BILL_AMT5**: Amount of bill statement in May, 2019 (NT dollar)
- **BILL_AMT6**: Amount of bill statement in April, 2019 (NT dollar)
- **PAY_AMT1**: Amount of previous payment in September, 2019 (NT dollar)
- **PAY_AMT2**: Amount of previous payment in August, 2019 (NT dollar)
- **PAY_AMT3**: Amount of previous payment in July, 2019 (NT dollar)
- **PAY_AMT4**: Amount of previous payment in June, 2019 (NT dollar)
- **PAY_AMT5**: Amount of previous payment in May, 2019 (NT dollar)
- **PAY_AMT6**: Amount of previous payment in April, 2005 (NT dollar)
- **default. payment. Next. month**: Default payment (1=yes, 0=no)

Description of Data

- There are 3000 unique credit card clients with an average credit card limit of 167484 with a standard deviation of 129747 approx. maximum value being 1M.
- The average age of the company's credit card clients is 35.5 years and S.D of 9.2 with education level mostly being graduate and university level. Most of the clients are either married or single, other being less frequent.
- As the value 0 for default payment means 'not default' and value 1 means 'default', the mean of 0.221 means that there are 22.1% of credit card contracts that will default next month

Methods

Random Forest classifier

In machine learning, there are a plethora of classification techniques that help us to predict or forecast binary business decisions efficiently. One of the techniques like the random forest is a part of Ensemble learning, which is the process by which multiple models, such as classifiers or experts, are strategically generated and combined to solve a particular computational intelligence problem. Ensemble learning is primarily used to improve the (classification, prediction, function approximation, etc.) performance of a model, or reduce the likelihood of an unfortunate selection of a poor one.

Random forest, like its name implies, consists of a large number of individual decision trees that operate as an ensemble. Each individual tree in the random forest spits out a class prediction, and the class with the most votes becomes our model's prediction.

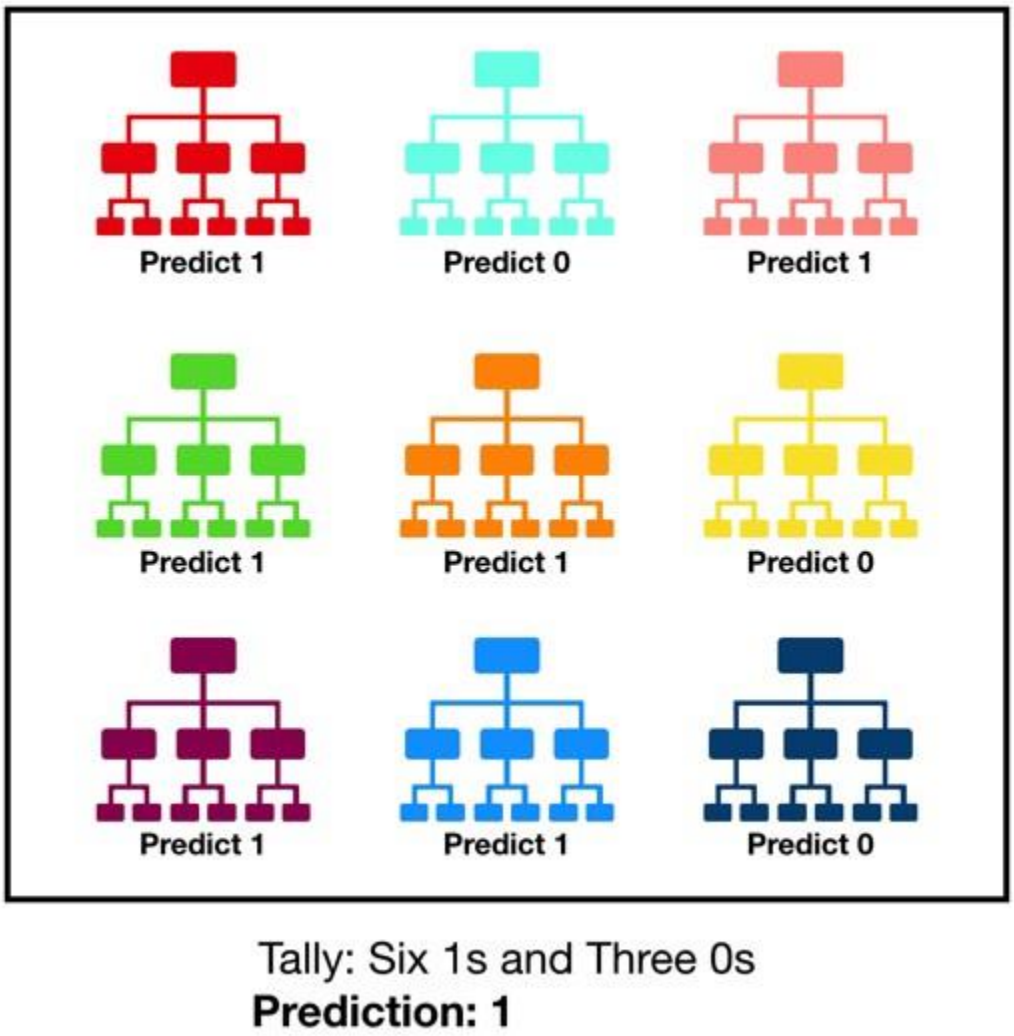
The underlying logic behind is a simple but powerful one - "The wisdom of crowds". It overcomes the weakness of a single decision tree which is often prone to overfitting or high variance by combining lots of small and uncorrelated decision trees models operating as a committee which the majority vote for decision.

To ensure that the models are uncorrelated, random forest uses two methods

Bootstrap Aggregation (Bagging) - which allows each individual tree to randomly sample from the dataset with replacement, resulting in different trees. In bagging, if we have a sample of size

N, we are still feeding each tree a training set of size N (unless specified otherwise). But instead of the original training data, we take a random sample of size N with replacement.

Feature Randomness- In a normal decision tree, when it is time to split a node, we consider every possible feature and pick the one that produces the most separation by calculating Gini impurity or information gain between the observations in the left node vs. those in the right node. In contrast, each tree in a random forest can pick only from a random subset of features. This forces even more variation amongst the trees in the model and ultimately results in lower correlation across trees and more diversification.[3]



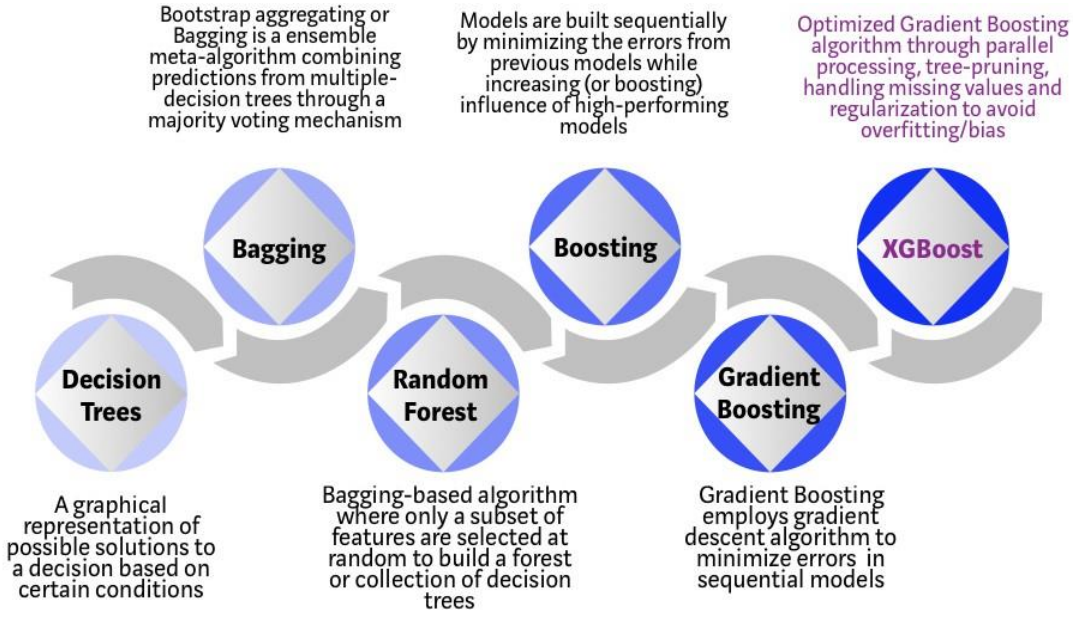
Logistic regression

Logistic regression is similar to linear regression except the outcome or regressand is a binary or dichotomous variable. It can be used on various datasets where the regressand is qualitative in nature. and we do not have to restrict ourselves to dichotomous outcomes; it can be trichotomous. In general, we can have multiple-category response variables.

Given a set of independent variables, the output of the estimated logistic regression (the sum of the products of the independent variables with the corresponding regression coefficients) can be used to assess the probability an observation belongs to one of the classes. Specifically, the regression output can be transformed into a probability of belonging to, say, class 1 for each observation. The estimated probability that a validation observation belongs to class 1 (e.g., the estimated probability that the customer defaults) for the first few validation observations.

For our data, we are using Binary Logistic Regression that predicts the value for dependent variable either '1' if the person is likely to default his/her next month's credit card bill payment or '0' otherwise.

XGBoost



XGBoost is a decision tree-based ensemble machine learning algorithm that uses a gradient boosting framework. Prediction problems involving unstructured small to medium tabular form of data gradient boosting is considered the best as it delivers results with highest accuracy.
The algorithm has following features:

1. A wide range of applications: Can be used to solve regression, classification, ranking, and user-defined prediction problems.
2. Portability: Runs smoothly on Windows, Linux, and OS X.
3. Languages: Supports all major programming languages including C++, Python, R, Java, Scala, and Julia.
4. Cloud Integration: Supports AWS, Azure, and Yarn clusters and works well with Flink, Spark, and other ecosystems[4].

Though gradient boosting and XGBoost are both ensemble tree methods that apply the principle of boosting weak learners using gradient descent loss function, XGBoost takes improvise upon it though system optimization and algorithmic enhancements.
XGBoost is an optimization technique that provides the best combination of software and hardware that yield superior results using fewer computing resources in the least amount of time.

Results and Analysis

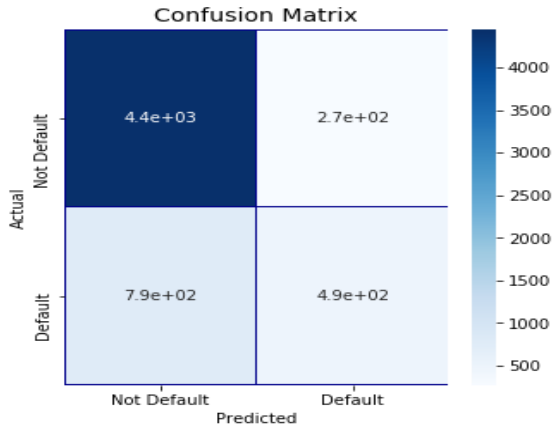
Classification is generally estimated in terms of the confusion matrix and the ROC curve. The confusion matrix gives the overall correct classification for the learning performed while ROC curve generates a graphical description of the true-positive versus false-positive rate. For more detailed analysis and comparison accuracy, Recall and Precision score is also estimated below.

Confusion matrix

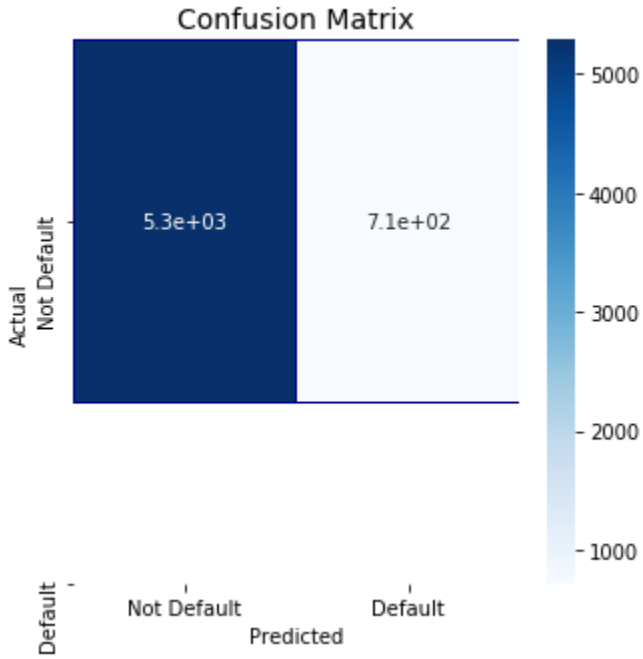
Confusion matrix is a technique for effectiveness of the performance measurement of our model. It is mostly a 2 x 2 matrix of varied combinations of actual and predicted values. It's used to measure very important aspects of classification models such as recall, Precision, accuracy, Specificity and ROC-AUC curve.

The Confusion matrix for all the three models are given below

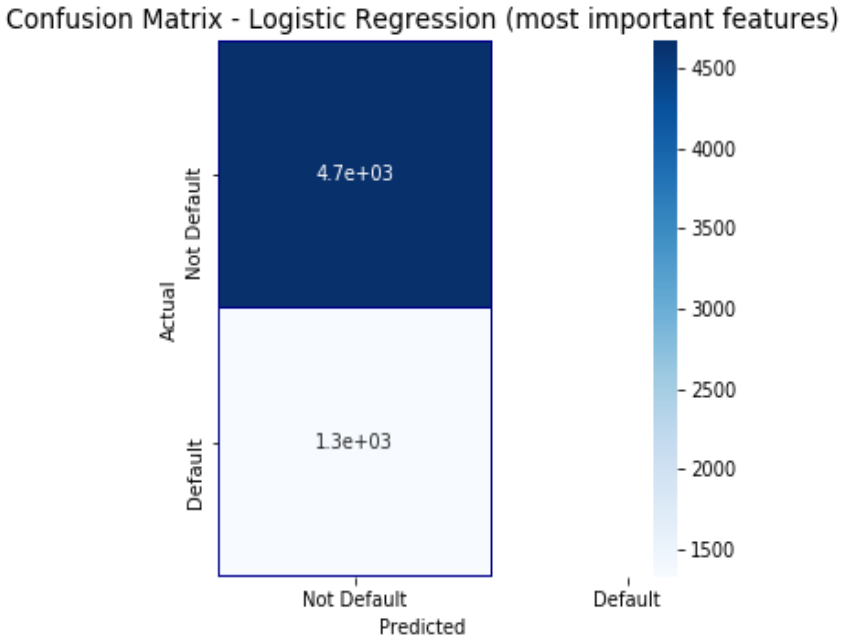
1. **Random Forest**



2. **XGBoosting**



3. Binary Logistic Regression



Given below is a sumaries table of results of all the three models

Model	Recall Score	Precision Score	Accuracy Score	F-Score	AUC
Random Forest	0.66	0.74	82.25%	0.69	0.758
XGBoosting	0.66	0.76	82.68%	0.69	0.78
Binary Logistic Regression	0.50	0.39	77.88%	0.44	0.638

- **Accuracy score**.

$$Accuracy\ Score = \frac{Number\ of\ Correct\ Prediction}{Total\ Number\ of\ Cases}$$

i.e. out of all the cases, how many does the model predicted accurately.

For binary classification, accuracy can also be calculated in terms of positives and negatives as follows:

$$Accuracy\ Score = \frac{True\ Positives + True\ Negatives}{True\ Positive + True\ Negatives + False\ Postives + False\ Negatives}$$

True positive = correctly classified or detected.

False positive = incorrectly classified or detected.

False negative: incorrectly rejected.

True negative: correctly rejected.

- **Recall Score**

$$Recall\ Score = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

i.e. out of all actual default, how many of them are predicted correctly by model.

- **Precision Score**

$$Precision\ Score = \frac{True\ Positive}{True\ Positive + False\ Positive}$$

i.e. the out of all the cases, how many does the model predicted accurately.

- **F-measure/ F1 score**

$$F - measure = \frac{2 * Recall * Precision}{Recall + Precision}$$

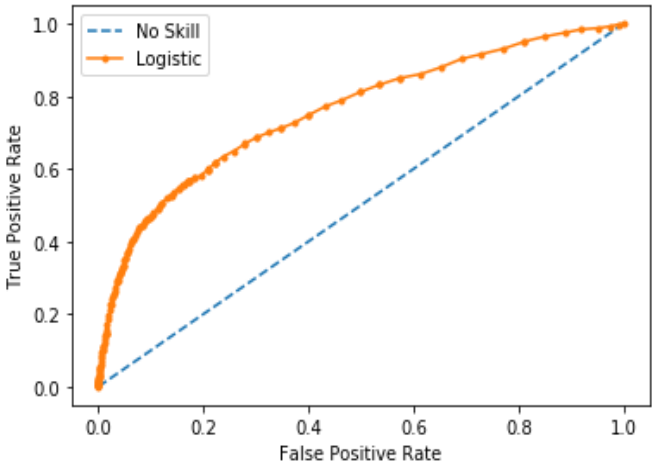
- **ROC-AUC curve**

ROC-AUC (**Area Under the Receiver operating Characteristics**) curve is one of the most significant evaluation methods of performance measurement at various threshold settings. As the value of the AUC increases, our model becomes a better predictor of the data. The model's separability capability is judged by AUC value. If the value is near to the 1 which means it has

good measure of separability where as a model that gives AUC value near to 0 is not considered a good fit for the prediction of our data.

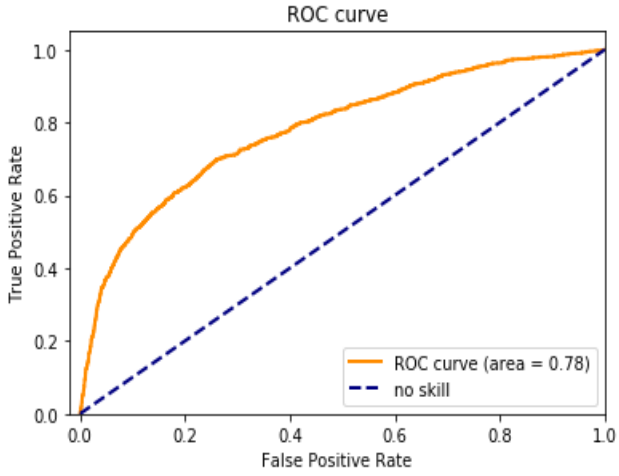
The **ROC-AUC** curve for our model is as follows

1. **Random Forest**



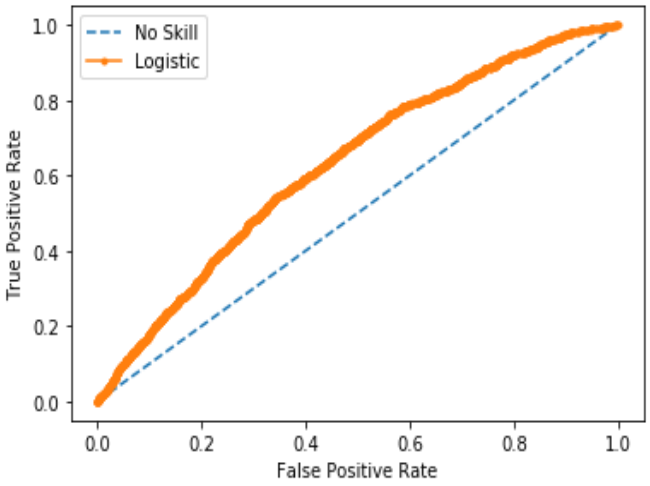
Area Under the curve (AUC): 0.758

2. **<u>XGBoosting</u>**



Area Under the Curve (AUC): 0.78

3. Binary Logistic regression



Area Under the Curve (AUC): 0.638

Conclusion

Credit card frauds are becoming more and more widespread these days. Moreover, in today's culture of less saving and more spending, loan repayment defaults, credit card frauds, banking frauds are prevalent in this society. Therefore, this project was pertinent to come up with the most accurate model that can be improvised upon by banks and can be useful to enhance their risk monitoring system, to come up with a more accurate risk premium for the customer. Machine learning is already assisting many industries to improve their product marketing, risk management in an automatic, scientific, and effective manner.

In this project, we explored three major classification models, namely Random Forest Classifier, XGBoost, and Logistic Regression. We could not include artificial neural networks, which are more useful for big data and handle image and audio files. The models have feuded with monthly bill amounts, delayed payments behavior along with gender, and education to build the default predicting mechanism. We present this work to demonstrate the wonders that machine learning can do.

The results are compared based on different accuracy measures which was shown in the table above. The traditional Binary Logistic Regression model achieved least accuracy of 77.88% The highest level of accuracy was achieved by XGBoosting of 82.68% whereas Random Forest was also proved to be

nearly as accurate as XGBoosting with the accuracy score of 82.25%. One of the most important criteria of performance evaluation is AUC value and the XGBoosting achieved the highest AUC value of 0.78. The logic behind the ROC curve is A skillful model will give a greater probability to a randomly chosen real positive occurrence than a negative occurrence on average. This is what we mean when we say that the model has skill.it means it can differentiate between positives and negatives. Generally, skillful models are represented by curves that bow up to the top left of the plot.

A no-skill model will not be able to discriminate between the classes and would predict a random class or a constant class in all cases. A model with no skill is represented at the point (0.5, 0.5). A model with no skill at each threshold is represented by a diagonal line from the bottom left of the plot to the top right and has an AUC of 0.5. Besides, ROC precision and recall are important to see if the model is predicting more class 0(true negatives) or class 1(true positives).

Clearly, XGBoost outperformed all other models but was close to random forest classifier for this specific case as both come under Ensemble Learning technique but XGBoost is Extreme gradient boosting technique is dominating in the field of machine learning because of its performance, speed and accuracy.

Through XGBoost feature importance was estimated to see which feature of the data has most influence over target which is default next month it was seen that bill amount in sept month (BILL_AMT1) and repayment status in same month has most decisional power over repayment in the month of October which is very intuitive. Because of the same reason logistic regression was built on top three features which can be changed over and experimented to see change in results.

References

[1] Gavin Edwards, 2018; Machine Learning: An introduction, https://towardsdatascience.com/machine-learning-an-introduction-23b84d51e6d0

[2] Data source: https://www.kaggle.com/gpreda/default-of-credit-card-clients-predictive-models

[3] Tony Yiu, 2019; Understanding Random Forest, https://towardsdatascience.com/understanding-random-forest-58381e0602d2

[4] Vishal Morde, 2019; XGBoost Algorithm: Long May She Reign! : https://towardsdatascience.com/https-medium-com-vishalmorde-xgboost-algorithm-long-she-may-rein-edd9f99be63d