Credit Card Default Prediction

Iqbal Babwane,
Sameer Ansari,
Lukman Haider
Data Science Trainees,
AlmaBetter, Mumbai

1. Abstract:

Financial threats are displaying a trend about the credit risk of commercial banks as the incredible improvement in the financial industry has arisen. In this way, one of the biggest threats faces by commercial banks is the risk prediction of credit clients. Recent studies mostly focus on enhancing the classifier performance for credit card default prediction rather than an interpretable model. In classification problems, an imbalanced dataset is also crucial to improve the performance of the model because most of the cases lied in one class, and only a few examples are in other categories. Traditional statistical approaches are not suitable to deal with imbalanced data. There is often a significant difference between the minimum and maximum values in different features, so Min-Max normalization is used to scale the features within one range. Data level resampling techniques are employed to overcome the problem of the data imbalance. Various under sampling and oversampling methods are used to resolve the issue of class imbalance. Different machine learning models are also employed to obtain efficient results. This model will help commercial banks, financial organizations, loan institutes, and other decision-makers to predict the loan defaulter earlier.

2. Introduction

The rapid growth in E-Commerce industry has led to an exponential increase in the use of credit cards for online purchases and consequently they have been surging in the fraud related to it. In recent years, for banks has become very difficult for detecting the fraud in credit card system. Machine learning plays a vital role for detecting the credit card fraud in the transactions. For predicting these transactions banks make use of various machine learning methodologies, past data has been collected and new features are been used for enhancing the predictive power. The performance of fraud detecting in credit card transactions is greatly affected by the sampling approach on data-set, selection of variables and detection techniques used. The performance of the techniques is evaluated for different variables based on sensitivity, specificity, accuracy and error rate. The main idea is by analyzing the customer data and by combining machine-learning algorithm to identify the default credit card user. 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.

This project possesses various contributions in the domain of credit risk prediction.

1) First, latest dataset has been used to build a machine learning model for credit risk prediction.
2) Second, the data imbalance problem has been explored by comparing the different resampling techniques and evaluate the performance that which the resampling technique has given effective results with a machine learning classifier.
3) Limited work was done on resampling techniques for data balancing in this domain because only a few resampling techniques were employed and also obtained less efficient results.
4) Lastly, the interpretable model is also deployed on the web to ease the different stakeholders. This model will help commercial banks, financial organizations, loan institutes, and other decision-makers to predict the credit defaulter earlier.

3. Problem Statement

This project is aimed at predicting the case of customers default payments in Taiwan. From the perspective of risk management, the result of predictive accuracy of the estimated probability of default will be more valuable than the binary result of classification - credible or not credible clients. We can use the K-S chart to evaluate which customers will default on their credit card payments.

4. Data Description

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.

5. Data Cleaning

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset. When combining multiple data sources, there are many opportunities for data to be duplicated or mislabeled. If data is incorrect, outcomes and algorithms are unreliable, even though they may look correct. There is no one absolute way to prescribe the exact steps in the data cleaning process because the processes will vary from dataset to dataset. But it is crucial to establish a template for your data cleaning process so you know you are doing it the right way every time.

Data cleaning means fixing bad data in your data set.

Bad data could be:

- Empty cells
- Data in wrong format
- Wrong data
- Duplicates

6. Data Visualization

Data visualization is the discipline of trying to understand data by placing it in a visual context so that patterns, trends, and correlations that might not otherwise be detected can be exposed.

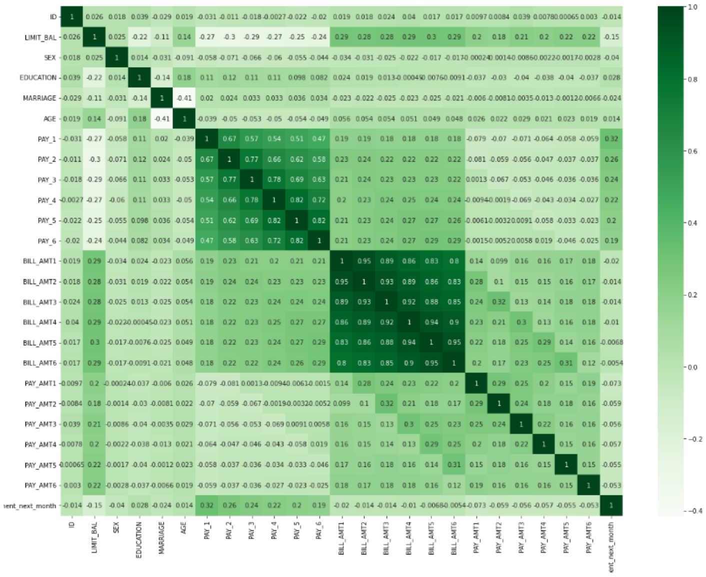
Python offers multiple great graphing libraries packed with lots of different features. Whether you want to create interactive or highly customized plots, Python has an excellent library for you.

To get a little overview, here are a few popular plotting libraries:

- Matplotlib: low level, provides lots of freedom
- Pandas Visualization: easy to use interface, built on Matplotlib
- Seaborn: high-level interface, great default styles

Observation:

Here many features are correlated with each other, but we can’t delete those features. Because it contains the past transaction details of the customers.



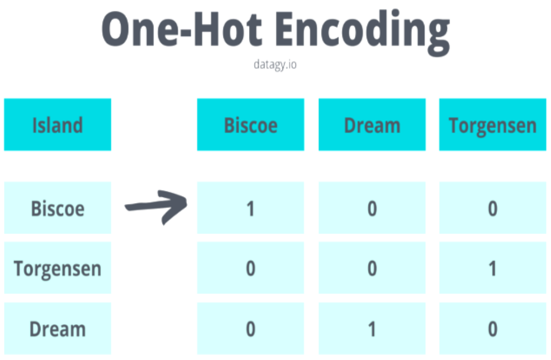
7. Feature Engineering

Feature engineering is the act of converting raw observation into desired features using statistical or machine learning approaches. Feature engineering refers to manipulation- addition, deletion, combination, mutation of our dataset to improve machine learning model training, leading to better performance and greater accuracy. Effective feature engineering is based on sound knowledge of business problem and the available data sources.

i. One hot encoder data

One-Hot encoding is used in machine learning as a method to quantify categorical data.

One-hot encoding approach eliminates the order but it causes the number of columns to expand vastly. So, for columns with more unique values try using other techniques like Label Encoding



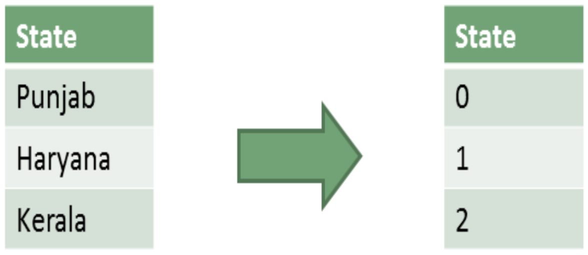
One-Hot Encoding

datagy.io

Island		Biscoe	Dream	Torgensen
Biscoe	→	1	0	0
Torgensen		0	0	1
Dream		0	1	0

ii. Label Encoder

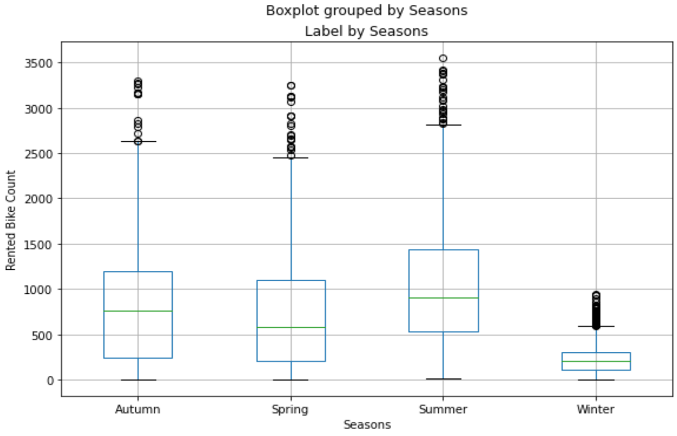
Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.



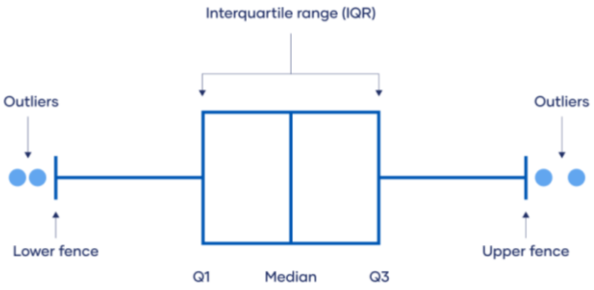
State		State
Punjab	→	0
Haryana		1
Kerala		2

8. Outlier

Outliers is a data point in the dataset that differs significantly from the other data or observation. The thing to remember that, not all outliers are the same. Some have a strong influence, some not at all. Some are valid and important data values. Some are simply errors or noise. Many parametric statistics like mean, correlations, and every statistic based on these is sensitive to



Analysis of outlier



- **Outlier detection**

  We use following methods to detect Outlier using Interquartile Range.

Square root:

The square root method is typically used when your data is moderately skewed. Now using the square root (e.g., sqrt(x)) is a transformation that has a moderate effect on distribution shape. It is generally used to reduce right skewed data. Finally, the square root can be applied on zero values and is most commonly used on counted data. Square Root Transformation: Transform the values from y to √y.

Log Transformation:

The logarithmic is a strong transformation that has a major
effect on distribution shape. This technique is, as the square root method, octenyl used for reducing right skewness. Worth noting, however, is that it cannot be applied to zero or negative values. Log Transformation: Transform the values from y to log(y).

Cube root transformation:

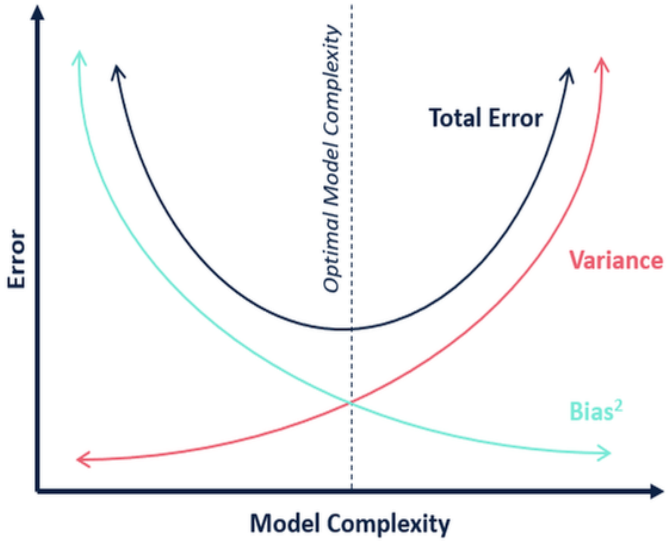
Cube root transformation involves converting x to $x13$. This is a fairly strong transformation with a substantial effect on distribution shape: but is weaker than the logarithm. It can be applied to negative and zero values too. Negatively skewed data Cube Root Transformation: Transform the values from y to $y^{1/3}$

Overfitting: So, what is overfitting? Well, to put it in more simple terms it's when we built a model that is too complex that it matches the training data "too closely" or we can say that the model has started to learn not only the signal, but also the noise in the data. The result of this is that our model will do well on the training data, but won't generalize to out-of-sample data, data that we have not seen before.

Bias-Variance tradeoff: When we discuss prediction models, prediction errors can be decomposed into two main subcomponents we care about: error due to "bias" and error due to "variance". Understanding these two types of error can help us diagnose model results and avoid the mistake of over/under fitting. A typical graph of discussing this is shown below:

Bias: The red line, measures how far off in general our models' predictions are from the correct value. Thus, as our model gets more and more complex, we will become more and more accurate about our predictions (Error steadily decreases).

Variance: The cyan line, measures how different can our model be from one to another, as we're looking at different possible data sets. If the estimated model will vary dramatically from one data set to the other, then we will have very erratic predictions, because our prediction will be extremely sensitive to what data set, we obtain. As the complexity of our model rises, variance becomes our primary concern.



9. Fitting different models

i. Logistic Regression

ii. Decision trees

iii. KNN - K-Nearest Neighbor

iv. Random Forest

v. SVC – Support Vector Classification

vi. Gradient Boosting

Logistic Regression:

Logistic regression is a classification algorithm that predicts the probability of an outcome that can only have two values (i.e., a dichotomy). A logistic regression produces a logistic curve, which is limited to values between 0 and 1. Logistic regression models the probability that each input belongs to a particular category.

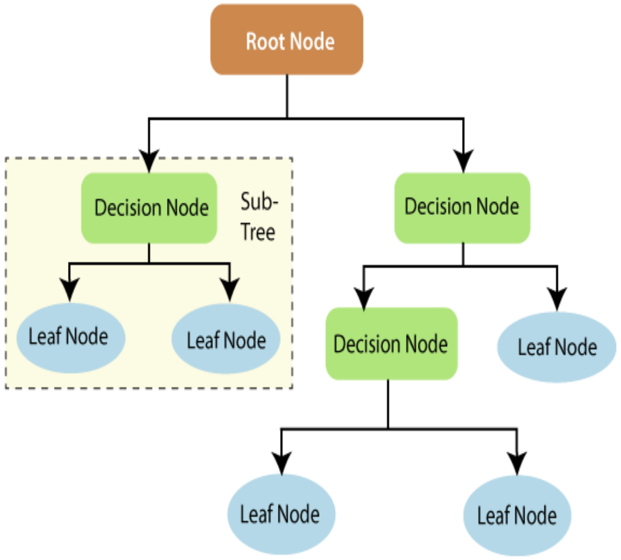
Logistic regression is an excellent tool to know for classification problems, which are problems where the output value that we wish to predict only takes on only a small number of discrete values. Here we'll focus on the binary classification problem, where the output can take on only two distinct classes.

In Logistic Regression, the log-odds of a categorical response being "true" (1) is modeled as a linear combination of the features:

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

Decision trees:

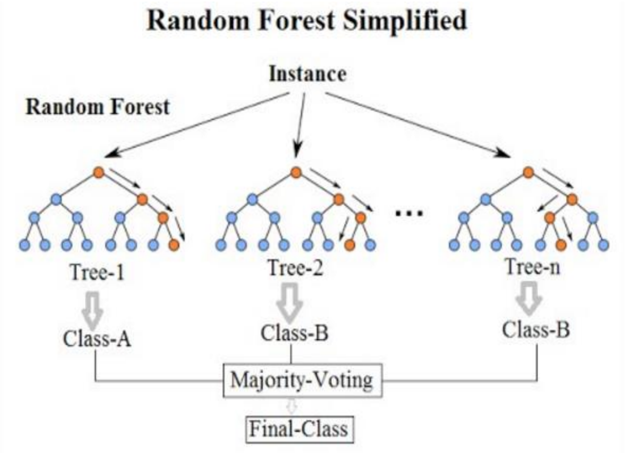
Decision Tree is a supervised learning method used in data mining for classification and regression methods. It is a tree that helps us in decision-making purposes. It separates a data set into smaller subsets, and at the same time, the decision tree is steadily developed. The final tree is a tree with the decision nodes and leaf nodes.



Random Forest:

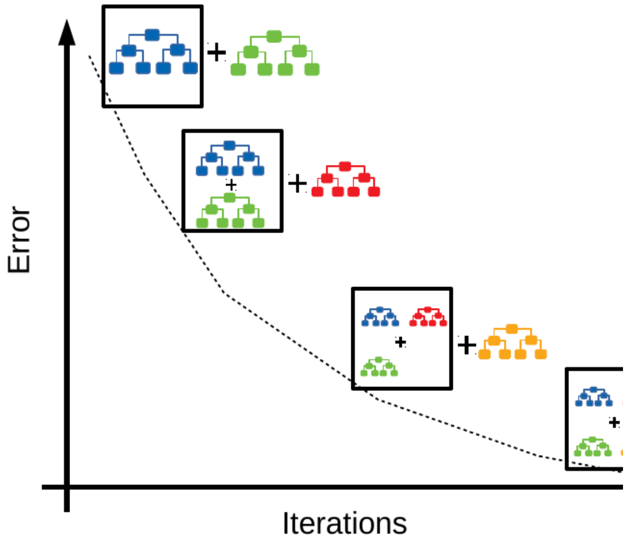
Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique**.**

“Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset.”

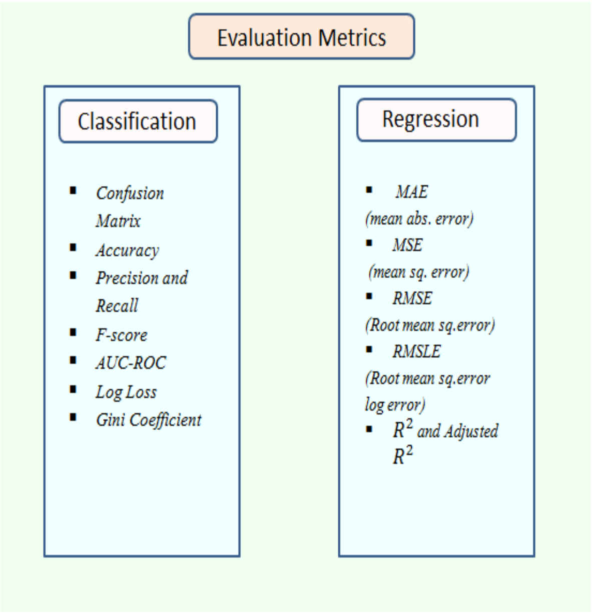


Gradient Boosting:

Gradient Boosting algorithm is used to generate an ensemble model by combining the weak learners or weak predictive models. Gradient boosting algorithm can be used to train models for both regression and classification problem. Gradient Boosting Regression algorithm is used to fit the model which predicts the continuous value.



10. Model Evaluation:



Evaluation Metrics

Classification

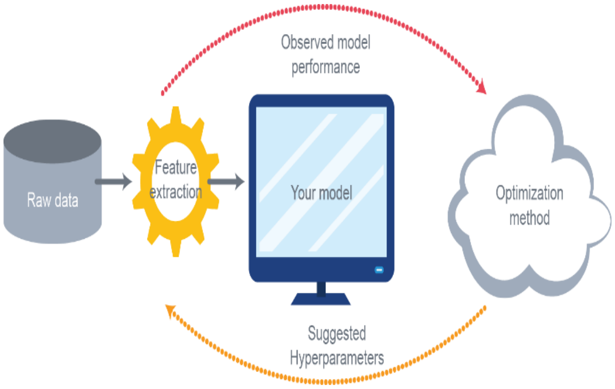
- *Confusion Matrix*
- *Accuracy*
- *Precision and Recall*
- *F-score*
- *AUC-ROC*
- *Log Loss*
- *Gini Coefficient*

Regression

- *MAE (mean abs. error)*
- *MSE (mean sq. error)*
- *RMSE (Root mean sq.error)*
- *RMSLE (Root mean sq.error log error)*
- *R^2 and Adjusted R^2*

Hyper parameter tuning:

A Machine Learning model is defined as a mathematical model with a number of parameters that need to be learned from the data. By training a model with existing data, we are able to fit the model parameters. However, there is another kind of parameter, known as Hyperparameters, that cannot be directly learned from the regular training process. They are usually fixed before the actual training process begins. These parameters express important properties of the model such as its complexity or how fast it should learn.



Hyperparameters are those parameters that are explicitly defined by the user to control the learning process. Some key points for model parameters are as follows:

- These are usually defined manually by the machine learning engineer.
- One cannot know the exact best value for hyperparameters for the given problem. The best value can be determined either by the rule of thumb or by trial and error.

- Some examples of Hyperparameters are the learning rate for training a neural network, K in the KNN algorithm.

Grid Search CV:

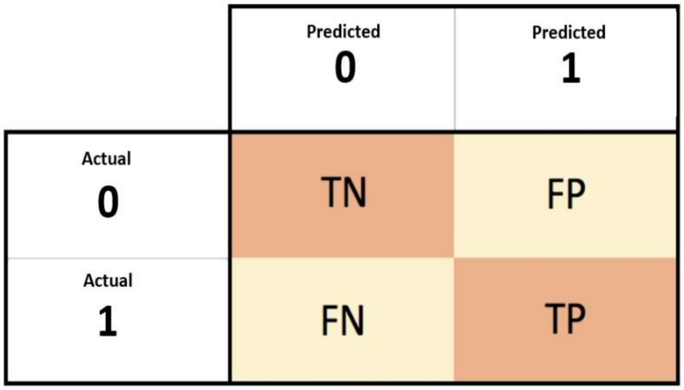
The Grid Search Method considers some hyperparameter combinations and selects the one returning a lower error score. This method is specifically useful when there are only some hyperparameters in order to optimize. However, it is outperformed by other weighted-random search methods when the Machine Learning model grows in complexity.

Grid Search is an optimization algorithm that allows us to select the best parameters to optimize the issue from a list of parameter choices we are providing, thus automating the 'trial-and-error' method. Although we can apply it to multiple optimization issues; however, it is most commonly known for its utilization in machine learning in order to obtain the parameters at which the model provides the best accuracy.

Randomized Search CV:

In Random Search, the hyperparameters are chosen at random within a range of values that it can assume. The advantage of this method is that there is a greater chance of finding regions of the cost minimization space with more suitable hyperparameters, since the choice for each iteration is random. The disadvantage of this method is that the combination of hyperparameters is beyond the scientist’s control.

Evaluation Metrics:



	Predicted 0	Predicted 1
Actual 0	TN	FP
Actual 1	FN	TP

Accuracy: Accuracy will require two inputs (i) actual class labels (ii) predicted class labels. To get the class labels from probabilities (these probabilities will be probabilities of getting a HIT), you can take a threshold of 0.5. Any probability above 0.5 will be labelled as class 1 and anything less than 0.5 will be labelled as class 0.

Precision: Precision for a label is defined as the number of true positives divided by the number of predicted positives. Report precision in percentages.

Recall: Recall for a label is defined as the number of true positives divided by the total number of actual positives. Report recalls in percentages.

F1-Score: This is defined as the harmonic mean of precision and recall.

Log Loss: The loss function for linear regression is squared loss. The loss function for logistic regression is Log Loss, which is defined as follows:

$$\text{Log Loss} = \sum_{(x,y)\in D} -y\log(y') - (1-y)\log(1-y')$$

Where:

- (x, y) ϵ D is the data set containing many labelled examples, which are (x, y) pairs.

- y is the label in a labelled example. Since this is logistic regression, every value of y must either be 0 or 1.
- y' is the predicted value (somewhere between 0 and 1), given the set of features in x.

AUC-ROC: The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve.

11. Conclusion:

This study focused on predicting Credit Card Default Prediction using given dataset. Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Regressor, KNN and SVC are used to predict. This statistical data analysis shows interesting outcomes in prediction method and also in an exploratory analysis.

hence the prediction from the logistic model was very low. Best predictions are obtained with an **SVC** model with a **Recall** score for train is **82%** and test score is **80%.**

12. References:

i. https://stackoverflow.com/

ii. https://www.almabetter.com/

iii. https://www.w3schools.com/

iv. https://www.geeksforgeeks.org/machine-learning/