INDIAN INSTITUTE OF INFORMATION TECHNOLOGY,BHAGALPUR

Data Science Project Report on Data Analysis and Prediction Using Machine Learning

GROUP – 12

Anil Kumar – 180101005
Karan Chaudhary – 180101022
Nikhil Kumar – 180102022
Sandeep Kumar – 180101040
Suraj Kumar – 180102040
Ujjwal Kumar - 180102043

Contents

INTRODUCTION

For this project we will be exploring publicly available data from Lending Club connects people who need money (borrowers) with people who have money (investors). Hopefully, as an investor you would want to invest in people who showed a profile of having a high probability of paying you back. We will try to create a model that will help predict this (www.lendingclub.com).

DATA ATTRIBUTES

- **Credit.Policy:** if the customer meets the credit underwriting criteria of LendingClub.com, and 0.
- **Purpose:** The purpose of the loan (takes values "credit_card", "debt_consolidation", "educational", "major_purchase", "small_business", and "all_other").
- **int.rate:** The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be more risky are assigned higher interest rates.
- **Fico:** The FICO credit score of the borrower.
- **dti:** The debt-to-income ratio of the borrower (amount of debt divided by annual income).
- **days.with.cr.line:** The number of days the borrower has had a credit line.
- **revol.bal:** The borrower's revolving balance (amount unpaid at the end ofthe credit card billing cycle).
- **revol.util:** The borrower's revolving line utilization (the amount of thecredit line used relative to total credit available).
- **inq.last.6mths:** The borrower's number of inquiries by creditors in the last6 month.
- **delinq.2yrs:** The number of times the borrower had been 30+ days past due.
- **pub.rec:** The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

PROBLEM DEFINTION

Situation | Problem definition: This project is all about predicting whether a borrower will return the money to the investor or not on the basis of borrower profile which contains almost thirteen different attributes like interest rate, fico score, the purpose of taking the loan, and many more. Once we predicted that the borrower is viable. Then the company will do the disbursement of the loan.

Task: Now the task was to come up with an efficient machine learning model which will result in maximum accuracy on a given data set.

GRAPH EXPLORATRY ANALYSIS

It's also known as Exploratory Data analysis (EDA). It refers to the critical process of performing Initial investigations on data. It's used to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations. EDA is primarily used to see what data can reveal beyond the formal modeling or hypothesis testing task and provides a provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate. Originally developed by American mathematician John Tukey in the 1970s, EDA techniques continue to be a widely used method in the data discovery process today.

CHARACTERISTICS

- They are not structured studies.
- It is usually low cost, interactive and open ended.
- It is time-consuming research and it needs patience and has risks associated with it.
- There are no set of rules to carry out the research properly, as they are flexible, broad and scattered.
- Such research usually produces qualitative data, however in certain cases quantitative data can generalize for a larger sample through use of surveys and experiments.

ADVANTAGES

- Improves understanding of variables by extracting averages, mean, min. and max Values etc.
- Discover errors, outliers, and missing values in the data.
- Identify pattern by Visualizing data in graphs such as box plots, scatter plots, and histograms.
- The researcher has a lot of flexibility and can adapt to changes as the research progresses.
- It enables the researcher understand at an early stage, if the topic is worth investing the time and resources and if it is worth pursuing.

DISADVANTAGES

- The main disadvantage of exploratory research is that they provide qualitative data. Interpretation of Such information can be Judgmental and biased.
- Even though it can point you in the right direction towards what is the answer, it is usually inconclusive.
- Most of the times, exploratory research involves a smaller sample, hence the results cannot be accurately interpreted for a generalized population.
- Many a times, if the data is being collected through secondary research, then there is a chance of that Data being old and is not updated.
- EDA does not effective when we deal with high-dimensional data.

HISTOGRAM

A histogram is a graphical representation of the distribution of a dataset. Although its appearance is similar to that of a standard bar graph, instead of making comparisons between different items or categories or showing trends over time, a histogram is a plot that show us the underlying frequency distribution or the Probability distribution of a single continuous numerical variable.

Histograms are two-dimensional plots with two axes; the vertical axis is a frequency axis whilst the horizontal axis is divided into a range of numeric values (intervals or **bins**) or time intervals.

CHARACTERISTICS

- A histogram is used to display continuous data in a categorical form.
- In a histogram, there are no gaps between the bars, unlike a bar graph.
- The width of the bins is equal.
- A histogram is designed to provide quick apprehension of many kinds of information, including the mean, minimum and maximum values of the information plotted on the chart.
- The frequency of each bin is shown by the area of vertical rectangular bars. Each bar covers a range of continuous numeric values of the variable under study. The vertical axis shows frequency values derived from counts for each bin.

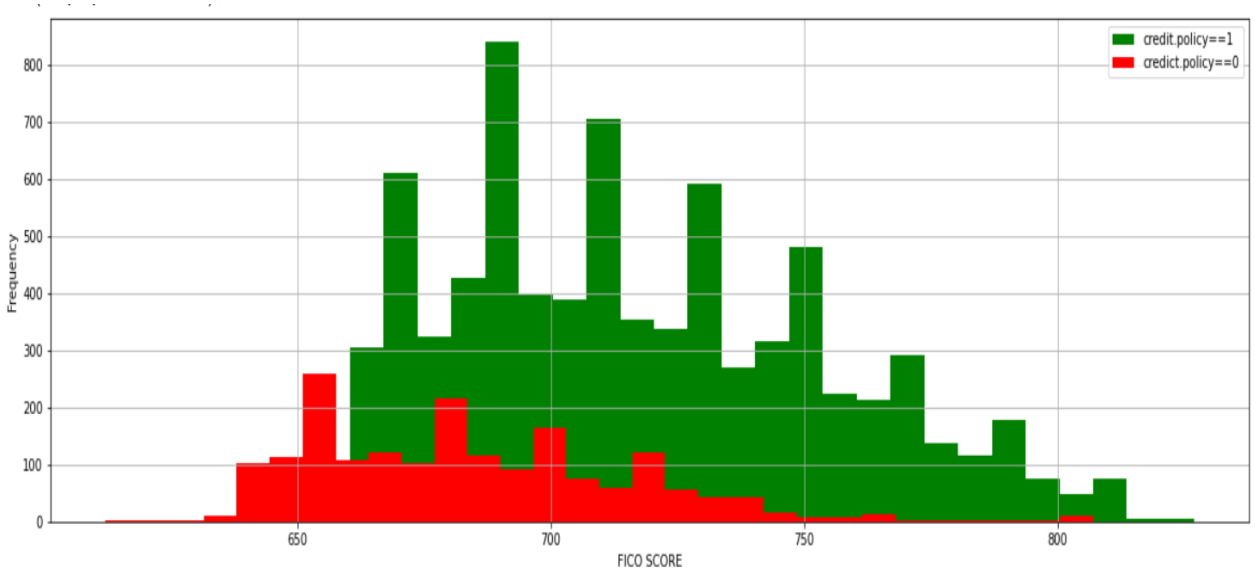
ADVANTAGES

- It displays the large amount of data graphically which are difficult to interpret in tabular form.
- It shows the occurrence frequency of various data points in a set of data.
- It shows the process of central tendency (centre of process) and helps in calculating process capability.
- The main advantages of a histogram are its simplicity and versatility. It can be used in many different situations to offer an insightful look at frequency distribution.
- It can also help to detect any unusual observations (outliers) or any gaps in the data.

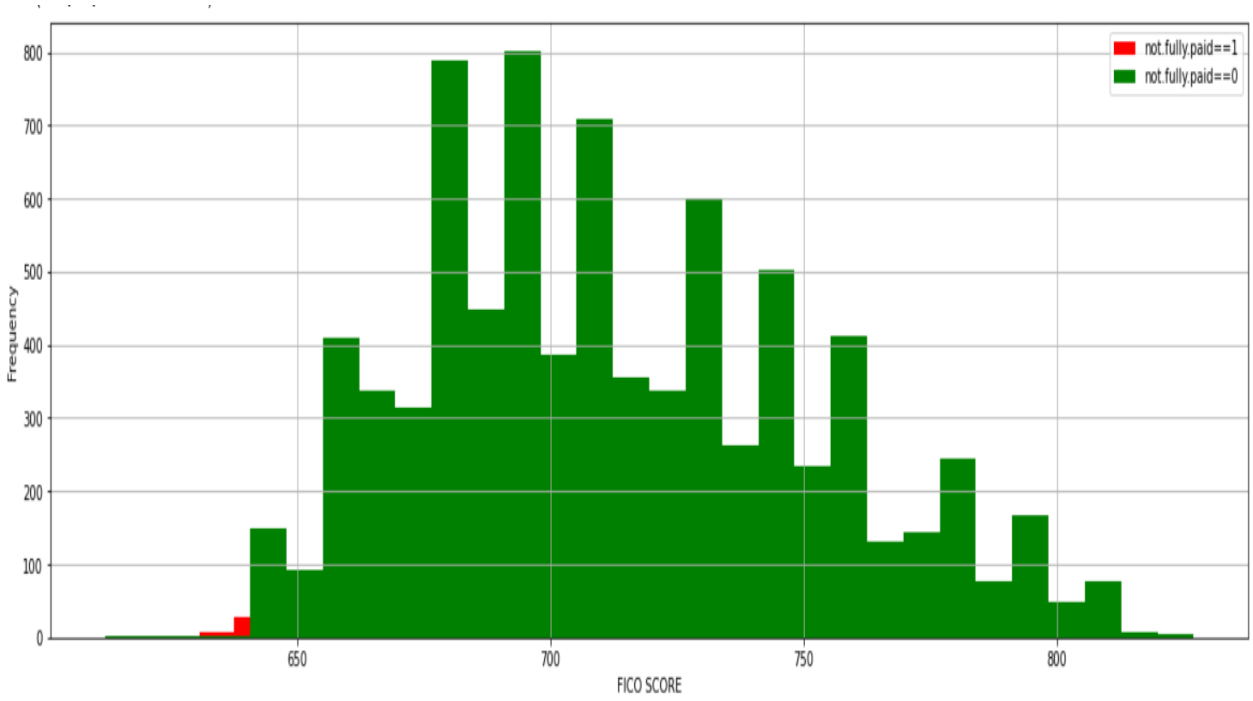
DISADVANTAGES

- It cannot read exact values because data is grouped into categories.
- More difficulty comes in comparing two data sets by using histogram.
- It can be used only with continuous data.
- A histogram can present data that is misleading. For example, using too many blocks can make analysis difficult, while too few can leave out important data.
- Histograms are based on two sets of data, but to analyse certain types of statistical data, more than two sets of data are necessary. So, it fails in this criterion.

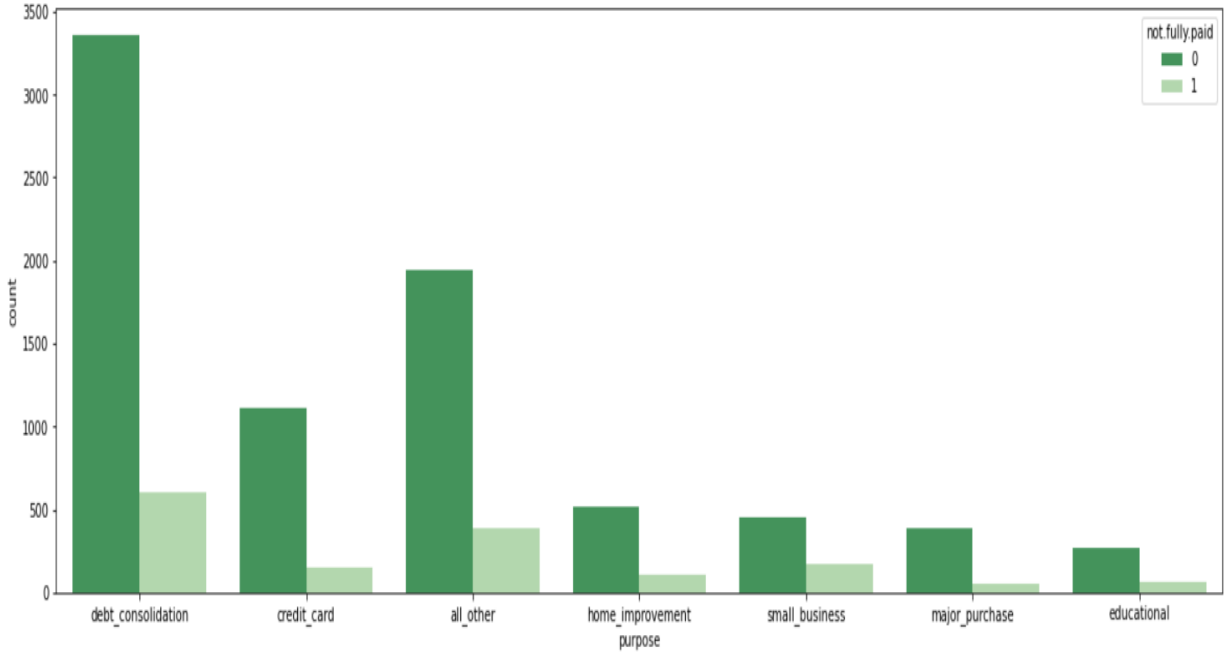
- Histogram of two **FICO** distributions on top of each other, one for each **credit. policy** outcome.



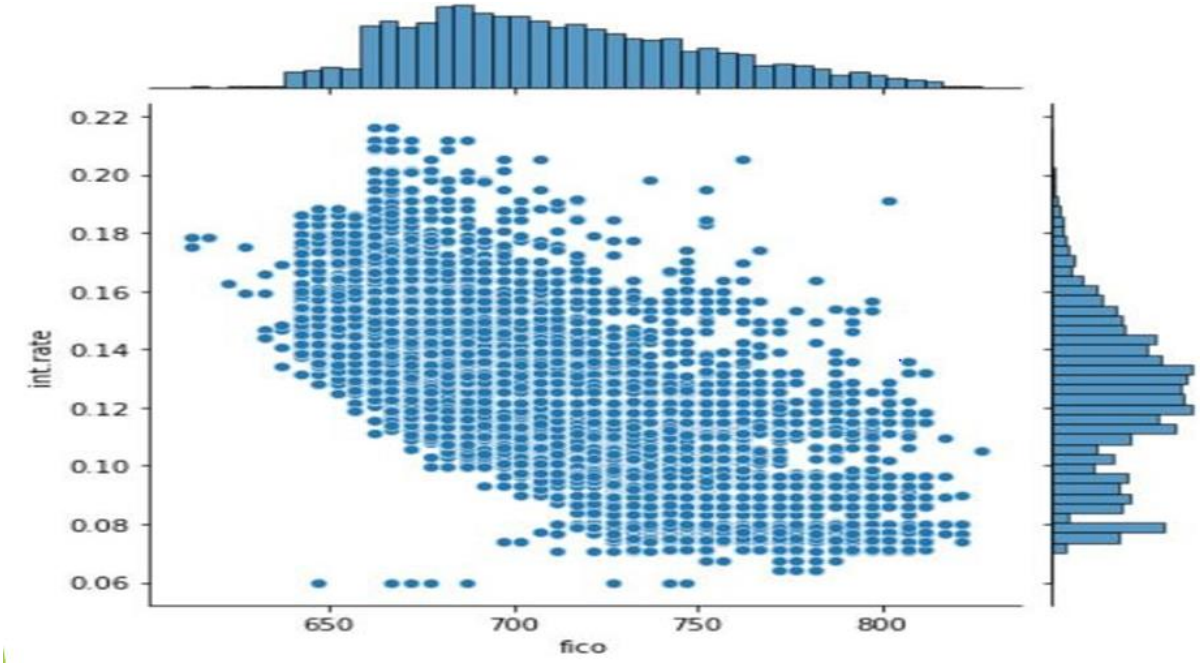
- Histogram of two FICO distributions on top of each other, one for each not.fully.paid.



- Count plot between purpose and not.fully.paid



- Join plot between FICO score and Interest rate



OBSERVATION OF GRAPH EXPLORATORY ANALYSIS

We have seen in **graph exploratory analysis** that the data distribution is not so random. Hence, we can use both **linear** as well as **non-linear** classification model to predict the result.

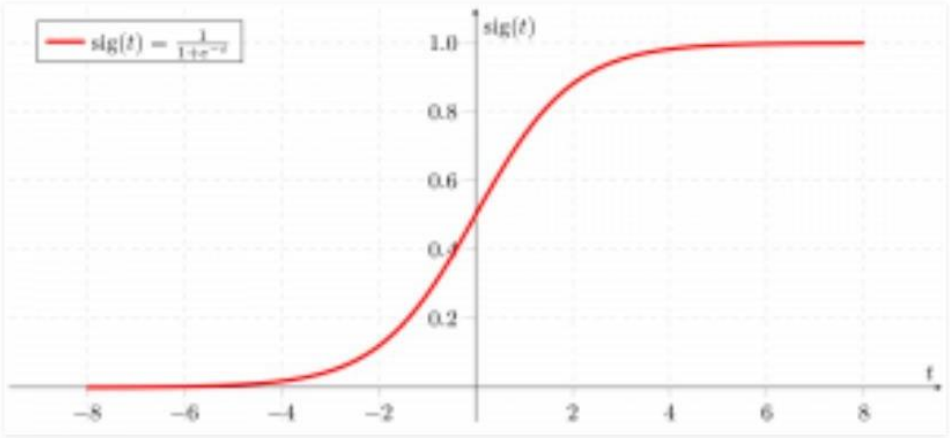
We will be using **logistic regression** as our **linear** model where as in case **non-linear** model we will be using two different types of models to predict our result meticulously.

- Tree Model – Decision tree classifier
- Ensemble Model – Random Forest classifier

LOGISTIC REGRESSION

- Logistic regression is a supervised learning classification algorithm used to predict the probability of a target variable. The nature of target or dependent variable is dichotomous, which means there would be only two possible classes.

$$g(z) = \frac{1}{1+e^{-z}}$$



- In simple words, the dependent variable is binary in nature having data coded as either 1 (stands for success/yes) or 0 (stands for failure/no).
- Mathematically, a logistic regression model predicts P(Y=1) as a function of X. It is one of the simplest ML algorithms that can be used for various classification problems such as spam detection, Diabetes prediction, cancer detection etc.

LOGISTIC REGRESSION ALGORITHM

The logistic function is defined as: transformed = 1 / (1 + e^-x), Where 'e' is the numerical constant Euler's number and x is the input, we plug into the function. All of the inputs get transformed into the range [0, 1] and that the smallest negative numbers resulted in values close to zero and the larger positive numbers resulted in values close to one. The logistic regression has coefficients just like linear regression

for example:

output = b0 + b1*x1, where b0, b1 are coefficients.

- First, we calculate a prediction using the current values of the coefficients HΘ(x), gives the predicted value, but it's not accurate, so we find the error by using cost function J(Θ).

$$J(\theta) = -\frac{1}{m}\sum\left[y^{(i)}\log(h\theta(x(i))) + \left(1-y^{(i)}\right)\log(1-h\theta(x(i)))\right]$$

$$h\theta(X) = \frac{1}{1+e^{-(\beta_0+\beta_1 X)}}$$

- Calculate new coefficient values based on the error in the prediction. The main purpose of doing it is to reduce the error between predicted value and actual value. So, we try to find such value of coefficients which makes cost function tend to 0, and we do it by using gradient descent.

$$\theta j := \theta j - \alpha \frac{\partial}{\partial \theta j} J(\theta)$$

- The process is repeated until the model is accurate enough (e.g., error drops to some desirable level) or for a fixed number iteration.

Want $\min_\theta J(\theta)$:
Repeat {
$\theta_j := \theta_j - \alpha \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)}) x_j^{(i)}$
(simultaneously update all θ_j)
}

LOGISTIC REGRESSION ASSUMPTIONS

- In case of binary logistic regression, the target variables must be binary always and the desired outcome is represented by the factor level 1.
- There should not be any multi-collinearity in the model, which means the independent variables must be independent of each other.
- We must include meaningful variables in our model.
- We should choose a large sample size for logistic regression.

SIGMOID FUNCTION

- The sigmoid function is a mathematical function used to map the predicted values to probabilities.
- The value of the logistic regression must be between 0 and 1, which cannot go beyond this limit, so it forms a curve like the "S" form. The S-form curve is called the Sigmoid function or the logistic function.
- It maps any real value into another value within a range of 0 and 1.In logistic regression, we use the concept of the threshold value, which defines the probability either 0 or 1.

ADVANTAGES

These are the advantages of logistic regression:

- Logistic regression is easier to implement, interpret and very efficient to train.
- It makes no assumption about distributions of classes in feature space.
- It can easily extend to multiple classes (multinomial regression) and a natural probabilistic view of class predictions.
- It not only provides measure of how appropriate a predictor (coefficient size) is, but also its direction of association (positive or negative).

DISADVANTAGES

These are the disadvantages of logistic regression:

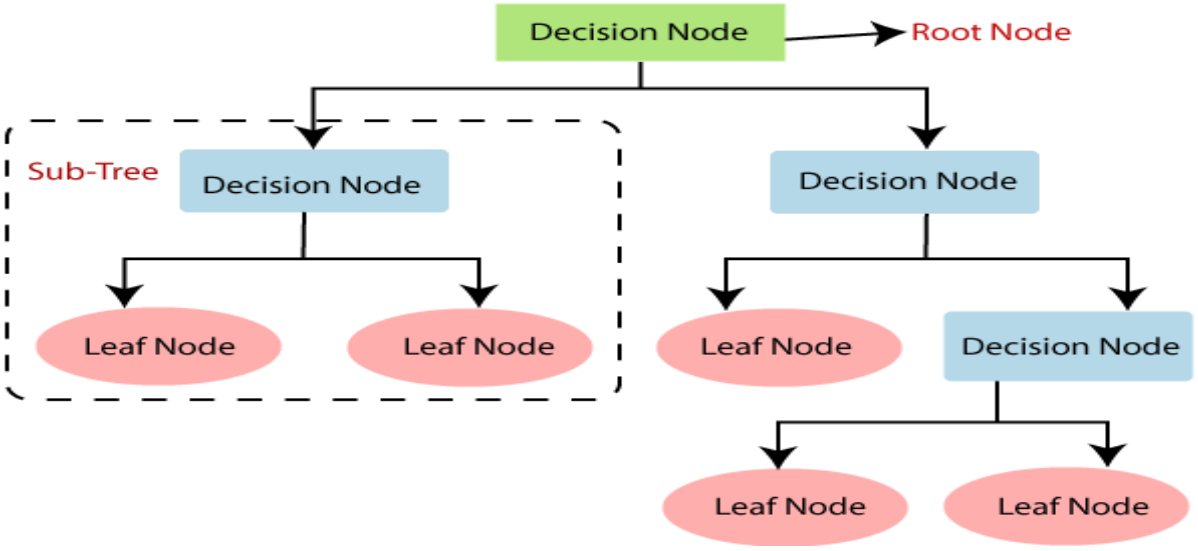
In logistic regression, we use the concept of the threshold value, which defines the probability either 0 or 1. It constructs linear boundaries.

The major limitation of Logistic Regression is the assumption of linearity between the dependent variable and the independent variables.

DECISION TREE CLASSIFIER

- Decision tree builds classification or regression models in the form of a tree structure. It breaks down a data set into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. Leaf node represents a classification or decision.

- In a Decision tree, there are two nodes, which are the **Decision Node** and **Leaf Node.** Decision nodes are used to make any decision and have multiple branches, whereas Leaf nodes are the output of those decisions and do not contain any further branches.
- the decision nodes are where the data is split.
- The decisions or the test are performed on the basis of features of the given dataset.
- It is called a decision tree because, similar to a tree, it starts with the root node, which expands on further branches and constructs a tree-like structure.
- It is a graphical representation for getting all the possible solutions to a problem/decision based on given conditions.



DECISION TREE ALGORITHM

The algorithm starts from the root node of the tree. This algorithm compares the values of root attribute with the record (real dataset) attribute and, based on the comparison, follows the branch and jumps to the next node. For the next node, the algorithm again compares the attribute value with the other sub-nodes and move further. It continues the process until it reaches the leaf node of the tree.

- Begin the tree with the root node, says S, which contains the complete dataset.
- Find the best attribute in the dataset using Attribute Selection Measure (ASM).
- Divide the S into subsets that contains possible values for the best attributes.
- Generate the decision tree node, which contains the best attribute.
- Recursively make new decision trees using the subsets of the dataset created in step -3. Continue this process until a stage is reached where we cannot further classify the nodes and called the final node as a leaf node.

DECISION TREE ASSUMPTIONS

- At the beginning, we consider the whole training set as the root.
- Feature values are preferred to be categorical. If the values are continuous then they are discretized prior to building the model.
- On the basis of attribute values records are distributed recursively.
- We use statistical methods for ordering attributes as root or the internal node.

DECISION TREE TERMINOLOGIES

- Root Node: Root node is from where the decision tree starts. It represents the entire dataset, which further gets divided into two or more homogeneous sets.
- Leaf Node: Leaf nodes are the final output node, and the tree cannot be segregated further after getting a leaf node.
- Splitting: Splitting is the process of dividing the decision node/root node into sub-nodes according to the given conditions.
- Branch/Sub Tree: A tree formed by splitting the tree.
- Pruning: Pruning is the process of removing the unwanted branches from the tree.

ADVANTAGES

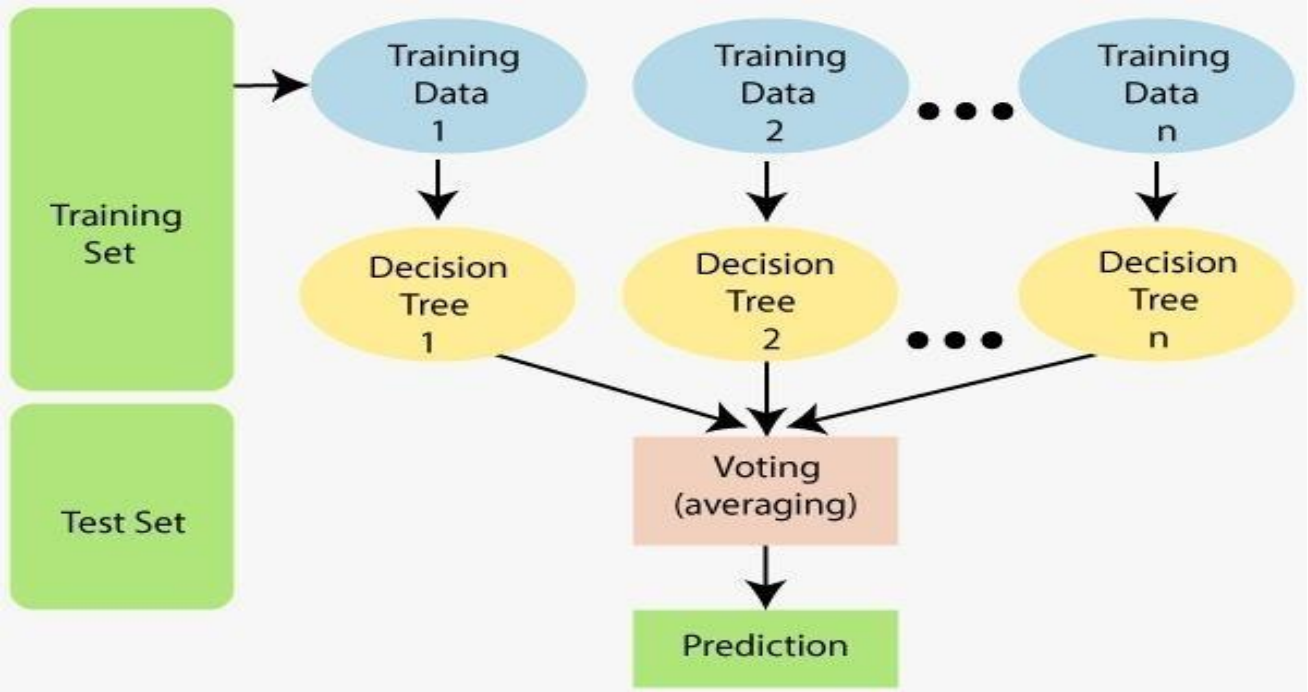
- Compared to other algorithms decision trees requires less effort for data preparation during pre-processing.
- Missing values in the data also do NOT affect the process of building a decision tree to any considerable extent.
- A Decision tree model is very intuitive and easy to explain to technical teams as well as stakeholders.
- A decision tree does not require normalization of data.
- There is less data cleaning required once the variables have been created. Cases of missing values and outliers have less significance on the decision tree's data.

DISADVANTAGES

- A small change in the data can cause a large change in the structure of the decision tree causing instability.
- For a Decision tree sometimes, calculation can go far more complex compared to other algorithms.
- Decision tree often involves higher time to train the model.
- It's less effective in predicting the outcome of a continuous variable.

RANDOM FOREST CLASSIFIER

- Random forest is a supervised learning algorithm. Random forest builds multiple decision trees and merges them together to get a more accurate and stable prediction.
- It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model.
- It 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. Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output.
- The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.



RANDOM FOREST ALGORITHM

Random Forest works in two-phase first is to create the random forest by combining N decision tree, and second is to make predictions for each tree created in the first phase.

It adds additional randomness to the model, while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model.

- Select random K data points from the training set.
- Build the decision trees associated with the selected data points (Subsets).
- Choose the number N for decision trees that you want to build.

  Repeat Step 1 & 2.
- For new data points, find the predictions of each decision tree, and assign the new data points to the category that wins the majority votes.

RANDOM FOREST ASSUMPTIONS

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random Forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.

ADVANTAGES

- It reduces overfitting in decision trees and helps to improve the accuracy.
- It is flexible to both classification and regression problems.
- It works well with both categorical and continuous values.
- Random Forest algorithm is very stable. Even if a new data point is introduced in the dataset, the overall algorithm is not affected much since the new data may impact one tree, but it is very hard for it to impact all the trees.
- It is capable of handling large datasets with high dimensionality.

DISADVANTAGES

- It requires much computational power as well as resources as it builds numerous trees to combine their outputs.
- It also requires much time for training as it combines a lot of decision trees to determine the class.
- Due to the ensemble of decision trees, it also suffers interpretability and fails to determine the significance of each variable.

IMPLEMENTATION STEPS

These are the implementation steps:

- Importing Libraries
- Mounting Google Drive
- Importing Data
- Exploratory Data Analysis
- Encoding The String Data Column into Integer Column
- Splitting The Dataset into The Training Set and TestSet
- Feature Scaling
- Training The Logistic Regression Model on The Training Set
- Prediction of Result using Logistic Regression
- Building Confusion Matrix of Logistic Regression
- Accuracy of Logistic regression
- Training The Decision Tree Model on The Training Set
- Prediction of Result using Decision Tree Model
- Building Confusion Matrix of Decision Tree
- Accuracy of Decision Tree Model
- Training The Random Forest Classification Model On The Training Set
- Prediction of Result using Random Forest Model
- Building The Confusion Matrix of Random Forest Model
- Accuracy of Random Forest Model

CONCLUSION

Here is the performance of each model.

Logistic Regression: 83.84 (false negative :379)

Decision Tree: 73.86 (false negative: 293)

Random Forest: 84.00 (false negative: 366)

Result: Eventually we got the maximum accuracy of **84%** with the help of the Random Forest Model.

REQUIREMENTS OF SOFTWARE AND HARDWARE

TECHNOLOGY

- Python
- Machine Learning
- Data Science

LIBRARIES

- NumPy
- Pandas
- Sklearn
- Matplotlib
- Seaborn

SOFTWARE USED

- Microsoft Office
- Google Colab
- Windows 10