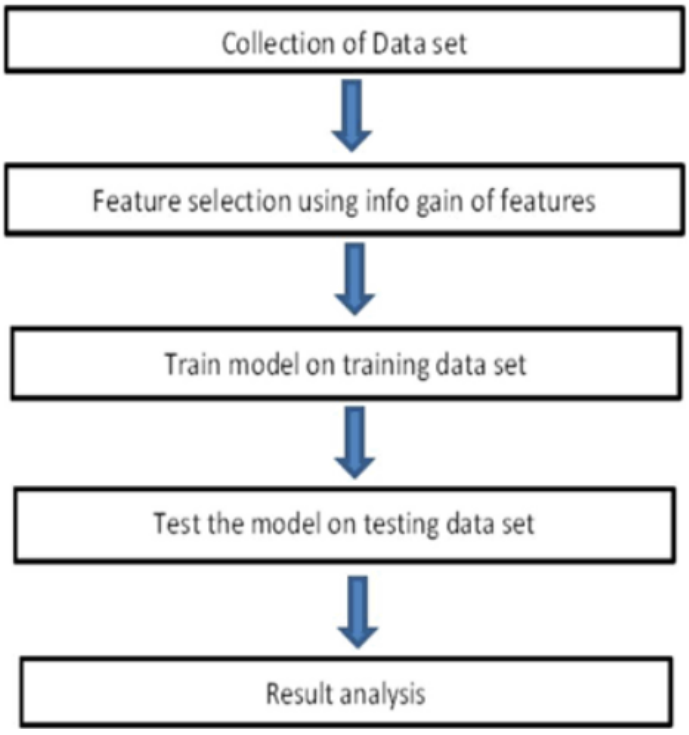
Loan Status Prediction Performance Insights

INTRODUCTION:

The two most pressing issues in the banking sector are: 1) How risky is the borrower? 2) Should we lend to the borrower given the risk? The response to the first question dictates the borrower's interest rate. Interest rate, among other things (such as time value of money), tests the riskiness of the borrower, i.e. the higher the interest rate, the riskier the borrower. We will then decide whether the applicant is suitable for the loan based on the interest rate. Lenders (investors) make loans to creditors in return for the guarantee of interest-bearing repayment. That is, the lender only makes a return (interest) if the borrower repays the loan. However, whether he or she does not repay the loan, the lender loses money. Banks make loans to customers in exchange for the guarantee of repayment. Some would default on their debts, unable to repay them for a number of reasons. The bank retains insurance to minimize the possibility of failure in the case of a default. The insured sum can cover the whole loan amount or just a portion of it. Banking processes use manual procedures to determine whether or not a borrower is suitable for a loan based on results. Manual procedures were mostly effective, but they were insufficient when there were a large number of loan applications. At that time, making a decision would take a long time. As a result, the loan prediction machine learning model can be used to assess a customer's loan status and build strategies.

METHODOLOGY:



MACHINE LEARNING CONCEPT:

Four machine learning models have been used for the prediction of loan approvals. Below are the description of the models used

LOGISTIC REGRESSION:

This is a classification algorithm which uses a logistic function to predict binary outcome (True/False, 0/1, Yes/No) given an independent variable. The aim of this model is to find a relationship between features and probability of particular outcome. The logistic function used is a logit function which is a log of odds in the favor of the event. Logit function develops a s-shaped curve with the probability estimate similar to a step function.

DECESION TREE:

This is a supervised machine learning algorithm mostly used for classification problems. All features should be discretized in this model, so that the population can be split into two or more homogeneous sets or subsets. This model uses a different algorithm to split a node into two or more sub-nodes. With the creation of more sub-nodes, homogeneity and purity of the nodes increases with respect to the dependent variable.

RAMDOM FOREST:

This is a tree based ensemble model which helps in improving the accuracy of the model It combines a large number of Decision trees to build a powerful predicting model. It takes a random sample of rows and features of each individual tree to prepare a decision tree model. Final prediction class is either the mode of all the predictors or the mean of all the predictors.

XGBOOST:

This algorithm only works with the quantitative variable. It is a gradient boosting algorithm which forms strong rules for the model by boosting weak learners to a strong learner. It is a fast and efficient algorithm which recently dominated machine learning because of its high performance and speed.

LIBRARIES FOR DATA ANALYSIS:

The models are implemented using Python 3.10 with listed libraries:

PANDAS:

Pandas is a Python package to work with structured and time series data.The data from various file formats such as csv, json, sql etc can be imported using Pandas. It is a powerful open source tool used for data analysis and data manipulation operations such as data cleaning, merging, selecting as well wrangling

SEABORN:

Seaborn is a python library for building graphs to visualise data. It provides integration with pandas. This open source tool helps in defining the data by mapping the data on the informative and interactive plots. Each element of the plots gives meaningful information about the data.

SKLEARN:

This python library is helpful for building machine learning and statistical models such as clustering, classification, regression etc. Though it can be used for reading, manipulating and summarizing the data as well, better libraries are there to perform these functions.

UNDERERSTANDING THE DATASET:

The machine learning model is trained using the training data set. Every new applicant details filled at the time of application form acts as a test data set. On the basis of the training data sets, the model will predict whether a loan would be approved or not. We have 13 features in total out of which we have 12 independent variables and 1 dependent variable i.e. Loan_Status in train dataset and 12 independent variables in test dataset. The Loan_ID, Gender, Married, Dependents, Education, Self_Employed, Property_Area, Loan_Status are all categorical.

EXPLORATORY DATA ANALYSIS UNIVARIATE VISUAL ANALYSIS TARGET VARIABLE – LOAN STATUS:

We will start first with an independent variable which is our target variable as well. We will analyse this categorical variable.

PREDICTOR VARIABLES:

There are 3 types of Independent Variables: Categorical, Ordinal & Numerical.

CATRGORICAL FEATURE:

- Gender
- Marital Status
- Employment Type
- Credit History

- 80% of loan applicants are male in the training dataset.
- Nearly 70% are married
- About 75% of loan applicants are graduates
- Nearly 85–90% loan applicants are self-employed
- The loan has been approved for more than 65% of applicants.

ORDINAL FEATURE

- Number of Dependents
- Education Level
- Property or Area Background

Our Visual Analysis below

- Almost 58% of the applicants have no dependents.
- Highest number of applicants are from Semi Urban areas, followed by urban areas.

NUMERIC FEATURE

- The Applicant's Income
- The Co-Applicant's Income

DATA CLEANING:

CONVERTING CATEGORICAL TO NUMERIC VARIABLES:

We drop the bins which we created for the exploration part. And Changing the '3+' in dependent variables to 3 makes it a numerical variable. Similarly we also convert the target variable's categories into 0 and 1 so that we can find it's correlation with numerical variables. All the more reason to as Algorithms like Logistic Regression only work with numeric values as input

MISSING VALUE IMPUTATION:

The total amount of missing or corrupt data in our data. To fix this, we replace missing categorical variables with it's mode and missing numerical variables with it's mean.

CREATING NEW MODELS:

As we can see p value for only credit history is less than 0.05. Hence we will remove all other independent variables and create new model and check its accuracy.

MODEL 1:

While finding correlation between independent variables, we found that applicant income and loan amount have good correlation so we will remove one of the variable from the data and then create a model

Accuracy of Model 1 is 81 %

MODEL 2:

Creating third model with two columns credit history and co applicant income.

Accuracy of Model 2 is Testing the *model by splitting data into 70% for train and 30 % for test.*

CONCLUSION:

- We did Exploratory data Analysis on the features of this dataset and saw how each feature is distributed
- We analysed each variable to check if data is cleaned and normally distributed.
- We cleaned the data and removed NA values
- We also generated hypothesis to prove an association among the Independent variables and the Target variable. And based on the results, we assumed whether or not there is an association.
- We calculated correlation between independent variables and found that applicant income and loan amount have significant relation.
- Finally, we got a model with co applicant income and credit history as independent variable with highest accuracy.
- We tested the data and got the accuracy of 83 %