EXTERNSHIP REPORT

ON

Loan status prediction using exploratory data analysis

Submitted by:

Team no:010

Team name:Dreamer's

Jafreen Sana nousheen

Samala Saipranathi

Tingilkar Sneha

Dandi Suma

Abstract:

• Loan prediction is a very common real-life problem that each retail bank faces at least once in its lifetime. If done correctly, it can save a lot of man hours at the end of a retail bank.

• Customers first apply for a home loan after that company validates the customer eligibility for the loan. The loan eligibility process (real time) can be automated based on customer details provided through for example filling an online application form. These details can be Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and many more. Hence the goal is to identify the customer segments that are eligible and in fact germane for loan amounts so that they can be specifically targeted. In this first Phase, some preliminary operations were carried out to understand all the features clearly.

• Visualization of all the variables for instance Married, to determine whether the married tend to repay the loans back or the unmarried was done. Similarly Gender, Self Employed, Applicant Income, Credit History, and on Independent Variables like Education etc... plotting Box and Bar Plots.

• All the variables were Normalized Before Plotting and cross checked if it is Normal using graphs. Some of the Bivariate Analysis Include Loan_Status and Education etc. Loan_status is our Target Variable.

• Data cleaning includes dealing with missing values Mean (without outliers) and Mode (with Outliers) Replacement Method. Also Correlation is visualized using Heatmap.

• In the Upcoming Phases, the plan is to progress further by performing feature engineering, Model Building using regression etc..

Introduction:

The term banking can be defined as receiving and protecting money that is deposited by the individual or the entities. This also includes lending money to the people which will be repaid within the given time. Banking sector is regulated in most of the countries as it is the important factor in determining the financial stability of the country. The provision of banking regulation act allows public to obtain loans. Loans are good sum of money borrowed for a period and expected to be paid back at given interest rate. The purpose of the loan can be anything based on the customer requirements. Loans are broadly divided as open-ended and close-ended loans. Open-ended loans are the loans for which the client has approval for a specific amount. Examples of open-end loans are credit cards and a home equity line of credit (HELOC). Close-ended loans decreases with each payment. In other words, it is a legal term that cannot be modified by the borrower. Personal loans, mortgages, auto payments, instalment loan and student loans are the most common examples of close-ended loans. Secured or collateral loan are those loans that are protected by an asset. Houses, Vehicles, Savings accounts are the personal properties used to secure the loan.

Unsecured loans are also known as personal or signature loans. Here the lender believes that the borrower can repay the loan based on financial resources possessed by the borrower. Liquidity risk is the risk that arises from the lackof marketability of an investment that cannot be bought or sold quickly enough to prevent or minimize a loss. The interest rate risk is the risk in which the interest rates priced on loans will be too low to earn the bank money.

Literature survey:

In [1] the researchers analyse the data set using data mining technique. Data mining procedure provides a great vision in loan prediction systems, since this will promptly distinguish the customers who are able to repay the loan amount within a period. Algorithms like “J48 algorithm”, “Bayes net”, Naive Bayes” are used. On applying these algorithms to the datasets, it was shown that “J48 algorithm” has high accuracy (correct percent) of78.3784% which provides the banker to decide whether the loan can be given to the costumer or not.

In paper [2], “loan prediction using Ensemble technique”, used “Tree model”, “Random forest”, “svmmodel” and combined the above three models as Ensemble model. A prototype has been discussed in paper [2] so that the banking sectors can agree/reject the loan request from their customers. The main method used is real coded genetic algorithms. The combined algorithms from the ensemble model, loan prediction can be done in an easier way. It is found that tree algorithm provides high accuracy of 81.25%.

In paper [3], using R-language, an improved risk prediction clustering algorithm is used to find the bad loan customers since probability of default (PD) is the critical step for the customers who comes for a bank loan. So, a frame work for finding PD in the data set is provided by data mining technique. R- Language has the technique called as KNN (K nearest neighbour) algorithm and it is used for performing multiple imputation calculation when there are missing values seen in the data set.

The paper [4] had used tree model. It helps to find whether the banking sector people will be able to overcome the loan problem with their customers. It provides a high accuracy of 80.87%.

The paper [5] uses decision tree induction algorithm and found that the algorithm finds a best way to evaluate the credit risk. To avoid the credit risk, bankers holds the technique called as “credit score”, where it helps the lenders to keep note on who are the applicants who will able to repay the amount or probability of going into the default risks. The input given for credit evaluation was customer data, WEKA software, cibil score. The methodology used in prediction system was problem and data understanding, data filtering, system modelling and finally system evaluation. This was done on the banks existing dataset containing 1140 records and 24 attributes. At last the system was tested and helps the bankers to make a correct decision on whether to accept or reject the loan approval.

The paper [6] used predictive model technique and descriptive model technique to predict the loan approval in banks. In predictive model technique, classification and regression were used and in descriptive model technique clustering and association were

used. Classifiers also implement several algorithms like naive Bayes, kNN algorithms of R language and regressors implements several algorithms like decision trees, neural networks, etc., To undergo this prediction analysis,out of all these algorithms, naive Bayes produces a most accurate classifier and the algorithms like decision tree, neural network, K-NN algorithms will be more accurate regressors. The main goal of the paper is to predict the loan classification based on the type of loan, loan applicant and the assets (property) that loan applicant holds. It was found that the decision tree algorithm gave an improved accuracy of almost 85% on doing the analysis

Loan applicant data analysis:

Whenever the bank makes decision to give loan to any customers then it automatically exposes itself to several financial risks. It is necessary for the bank to be aware of the clients applying for the loan. This problem motivates to do an EDA on the given dataset and thus analysing the nature of the customer. The dataset that uses EDA undergoes the process of normalisation, missing value treatment, choosing essential columns using filtering, deriving new columns, identifying the target variables and visualising the data in the graphical format. Python is used for easy and efficient processing of data. This paper used the pandas library available in Python to process and extract information from the given dataset. The processed data is converted into appropriate graphs for better visualisation of the results and for better understanding.For obtaining the graph Matplot library is used.

A. Annual Income Vs Purpose Of Loan

In this Figure 1, the X axis represents the purpose of loan i.e. the purpose for which the loan is applied. Debt consolidation, home improvement is some of the purposes. High,moderate and low represents the annual income of people who fall in the range as below. Low represents the annual income of people between therange of minimum to 10 lakhs and Moderate represents the annual income of people between 10 lakhs and 25 lakhs and High represents the annual income of people above 25lakhs. By these criteria, a new column called Category is derived

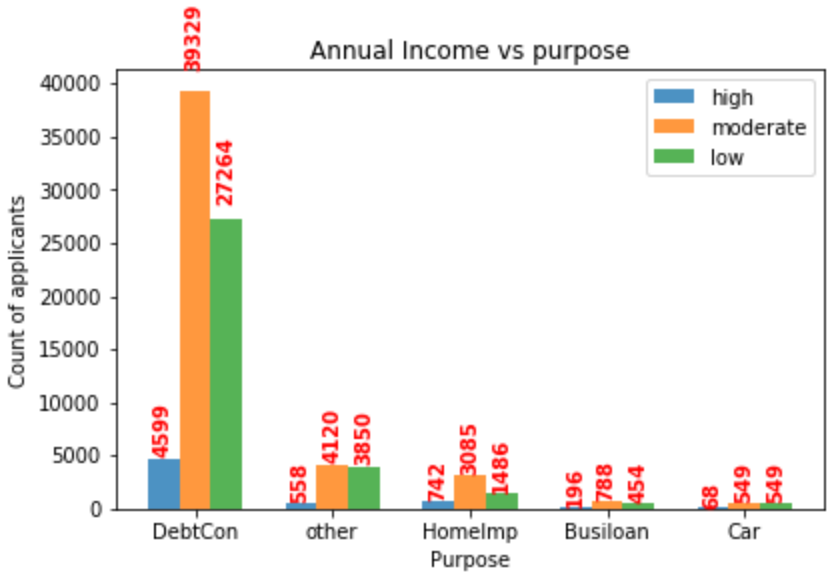


Fig 1. Annual Income vs purpose

Thus, grouping the Category that is high, moderate and low.

Inference from the Figure 1 is as follows:

· People in moderate category seek loan in the higher numbers.

· The field debt consolidation shows the highest distribution.

· Low and moderately categorized applicants try for other purpose and car loans equally.

B. Trust Customer Classification

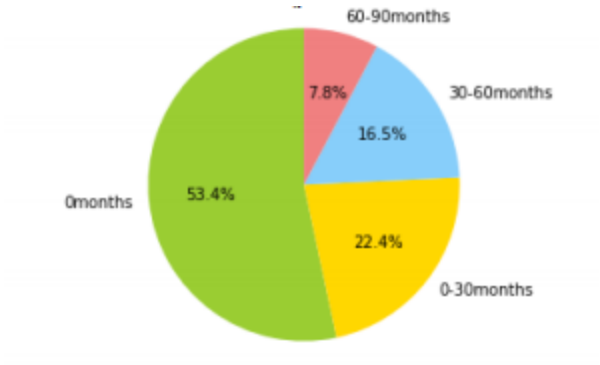


Fig 2. Trust Customers

From the Figure 2 it is inferred as follows:

· There are many customers who does not have delinquents, has applied for the loan which intimates or indirectly conveys that the applicant has some chances to get approval of the loan as the applicant have no delinquents. The result is about 53.3% of applicants.
· And it is also inferred that the number of people applied for loan gradually decreases with the increase in the delinquent months. This shows that, the applicant has minimum chances to get the approval of loans

C. Loan Term Vs Delinquent Months

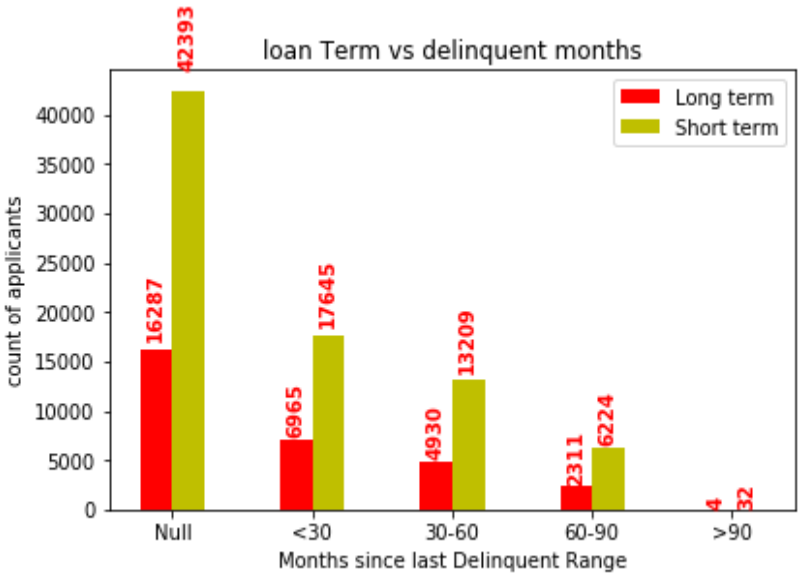


Fig 3. loan term vs delinquent months

· Figure 3 deals with the customers who can pay the loan within term period against customers who cannot repay their monthly depts within the particular term.
From the Figure 3 it can be concluded that:

· This analysis can find a higher number of customers who are able to repay without deliquiates and for short term.

· Almost all applicants who are even delinquent more than 90 months prefers only shortterm.

· Applicants who delinquent more than 90 months are less in number and it indirectly conveys that their loan will never be sanctioned and if its yes, the applicant will not be able to pay it back

D. Loan Term Vs Credit Category

From the categories poor, fair, good, very good and undefined, the value the credit category is found, as applicants without credit score fall into the category undefined,people have credit score between 300-850 between these there are some categories such as credit score between 300 and 579 falls in poor, credit score between 580 and 669 falls in category fair and the credit score between 670 and 739 is good and above this is considered to be very good.

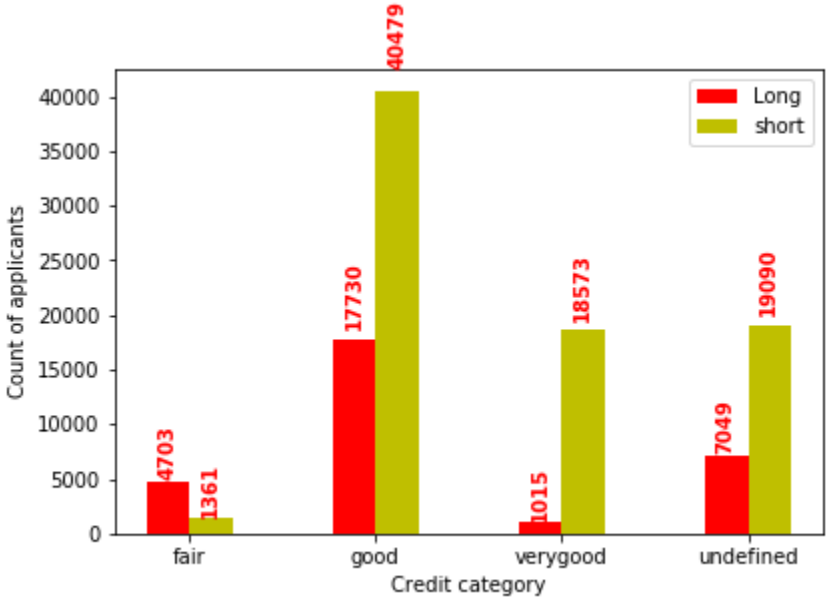


Fig 4. Loan term vs Credit category

Figure 4 spectacles the repayment period of the loan versus credit score under various categories by grouping the derived column credit category and loan term.Figure 4 have deduced the following:

· Customers with good and very good credit score prefer for short term payback period in contradiction to customers with fair credit score.

· People applying loan for first time prefer short term Loans because lender doesn't do a credit check so that the applicant avail loans easily.

E. Loan Term Vs Years In Current Job

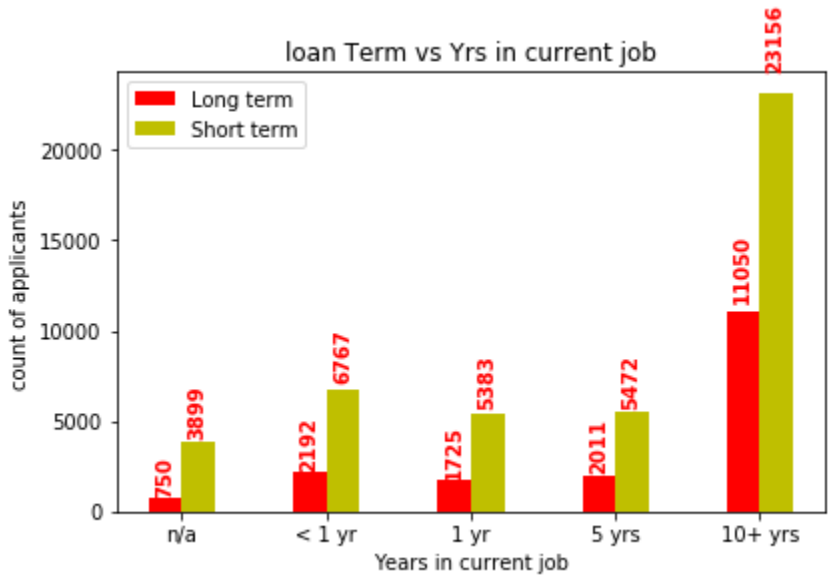


Fig 5. Loan term vs Years in current job

Figure 5 displays the count of applicants who have various years of experience in current job against the term period of repayment of loans.

From the Figure 5 it is concluded that:

· Applicants who have various years of experience in the same job claim the loan for a short period of time.

· Also, the applicants who are freshers claim the loan for a short period of time.

· This also infers that, long term loans are borrowed by people who are yet to start their own business and can repay only after this business set well and brings profit.

· Long term goals carry a greater risk to the money lenders and thus not very easily approved by the banks.

F. Loan Payment Chances Vs Home Ownership

Loan payment chances which have been classified into canpay, maypay and not payable.From Current credit balance, subtracting the monthly debt of that person, one can find whether the person would be able to repay the loan or not.

By subtracting, if applicants have less than 50,000 balance, the applicant is considered in the not payable category and applicants having balance between 50,000 and 3 lakhs

are considered to be may pay category and above than that the applicant are considered to be can pay category.
From the Figure 6 it is concluded that,

· People living in rent home fall under not payable category among ownership credentials

· Applicants who have kept home in mortgage apply loan in highest numbers.

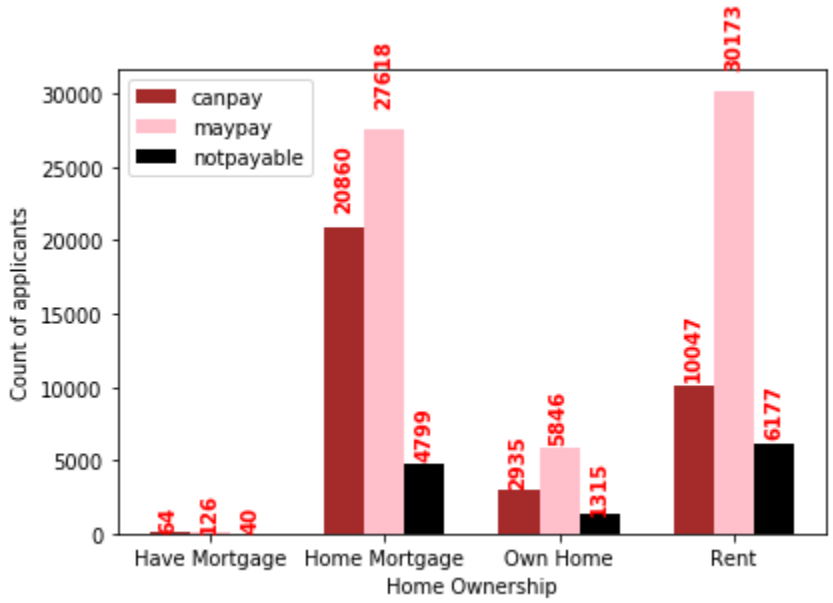


Fig 6. Loan payment chances vs Home ownership

Decision Trees :

The basic algorithm of decision tree requires all attributes or features should be discretized .
The knowledge depicted in decision tree can represented in the form of IF-THEN rules.

One of the most easy and famous classification algorithm is Decision Tree Algorithm .This algorithm helps interpreting and understanding better.

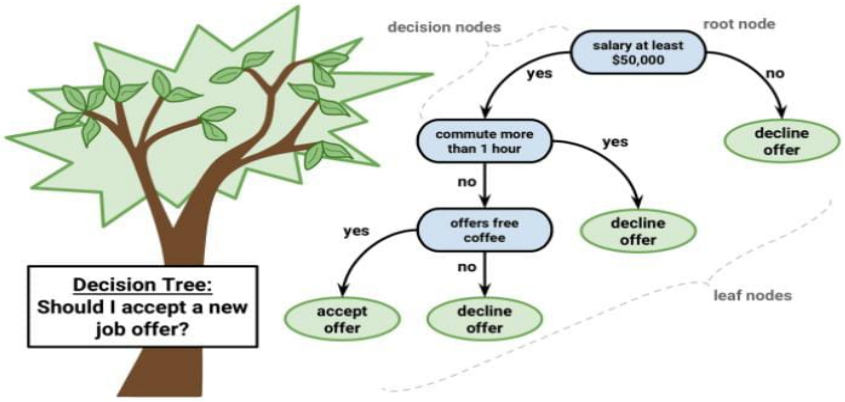
Decision tree algorithms is one of the supervised learning algorithms. The decision treealgorithm is capable of solving both, classification and regression problems, which distinguishes it from rest of the supervised learning algorithms.

The main objective of using this algorithm is to predict the value/class of the target variable by learning some decision rules.

For predictions using this algorithm, we have to begin with the root node .The value of records attribute and root node are compared .This comparison tells what will be the next node that needs to be followed.

Every node acts as a test case for some other attribute, and every edge running down from a node corresponds to a probable answer of that test case. In the starting the entire train dataset is considered as the root of the decision tree.

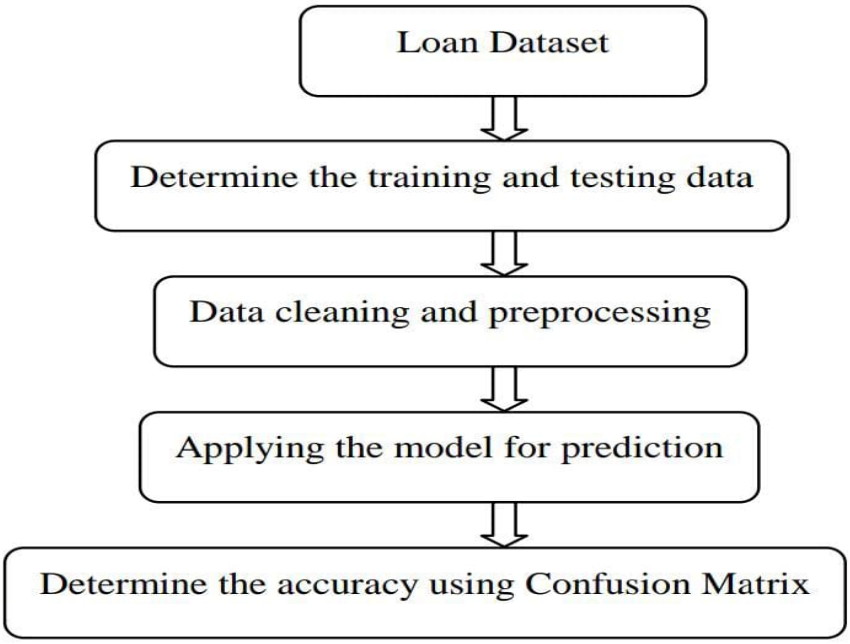
Before building this model the continuous variables are converted into categorical.



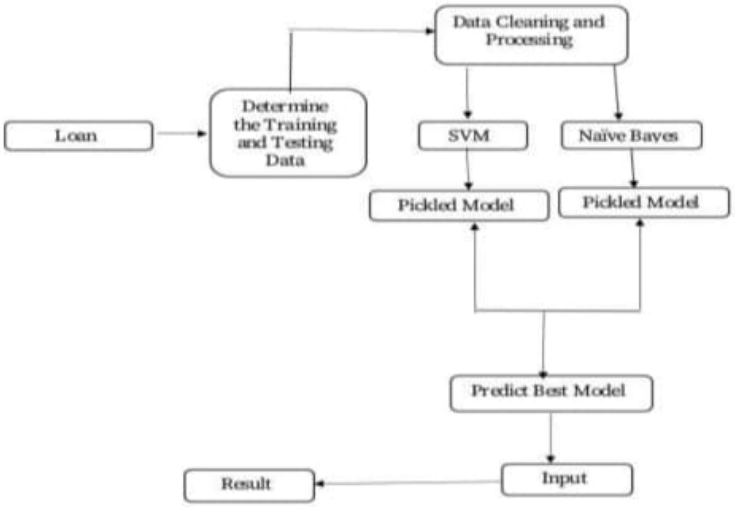
Proposed Method

The architecture of the proposed model is shown in flow chart . The major objective of this project is to derive patterns from the datasets which are used for the loansanctioning process and create a model based on the patterns derived in the previous step.

Classification data mining algorithms are used to filter outthe probable loan defaulters from the list. For analysis purposes, essential inputs like
gender, age, marital status, residential status, job, income, loan expectation, existing client, accountbalance, total debt, etc., are collected and used to find the
appropriate attributes.



System Architecture:



Problem Statement:

There is a company named Dream Housing Finance that deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. However doing this manually takes a lot of time. Hence it wants to automate the loan eligibility process (real time) based on customer information

So the final thing is to identify the factors/ customer segments that are eligible for taking loan. How will the company benefit if we give the customer segments is the immediate question that arises. The solution is ….Banks would give loans to only those customers that are eligible so that they can be assured of getting the money back.

Hence the more accurate we are in predicting the eligible customers the more beneficial it would be for the Dream Housing Finance Company.

Type of problem:

Loan_Status is yes or no. So this can be solved by any of the classification techniques like

- Logistic Regression .
- Decision Tree Algorithm.
- Random Forest Technique.

I have mentioned only few. We will be dealing with each of techniques later in this blog.

Data description:

- The data set provided in the above link is of Dream Housing Finance company that deals with granting home loans in urban, semi-urban, and rural areas.
- It consists of the following variables: Gender, Marital Status, Education, Number of Dependents, Income of the applicant, Income of the Co-applicant, Loan Amount(thousands), Credit History, Unique Loan ID, Self-Employed Status, Term of the Loan Amount(months), Property Area, Loan Status(If Loan is approved then Y else N).
- The train csv data contains the Loan Status variable (Target variable) and other variables used for training our model.
- The test csv data doesn't contain the Loan Status variable (Target variable), as the model will be applied to this data to predict the Loan Status.
- Factors affecting the loan acceptance decision:

- All the lenders have their criteria, but primarily these factors are considered while approving a loan application:-
- **Household Income:** Higher household income increase the chance of loan approval.
- **Credit History:** Applicants with a history of timely debt repayments have a higher probability of getting their applications approved.
- **Loan Amount:** Lower loan amount results in a higher approval rate.
- **Debt payments:** Applicants with a lower monthly debt payment have a higher probability of getting their application approved.
- **Education:** Higher qualifications of an individual increase the likelihood of getting a high-paying job. Therefore, result in a higher income and increases the chances of loan approval.
- **Dependents:** Less number of dependents in the family increases the chances of loan approval.

Dataset Information:

Dream Housing Finance company deals in all home loans. They have presence across all urban, semi urban and rural areas. Customer first apply for home loan after that company validates the customer eligibility for loan. Company wants to automate the loan eligibility process (real time) based on customer detail provided while filling online application form. These details are Gender, Marital Status, Education, Number of Dependents, Income, Loan Amount, Credit History and others. To automate this process, they have given a problem to identify the customers segments, those are eligible for loan amount so that they can specifically target these customers.

This is a standard supervised classification task.A classification problem where we have to predict whether a loan would be approved or not. Below is the dataset attributes with description.

Description about the Data Columns:

There are 2 data sets that are given. One is training data and one is testing data. It's very useful to know about the data columns before getting in to the actual problem for avoiding confusion at a later state. Now let us understand the data columns (that has been already given by the company itself) first so that we will get a glance.

Variable	Description
Loan_ID	Unique Loan ID
Gender	Male/ Female
Married	Applicant married (Y/N)
Dependents	Number of dependents
Education	Applicant Education (Graduate/ Under Graduate)
Self_Employed	Self employed (Y/N)
ApplicantIncome	Applicant income
CoapplicantIncome	Coapplicant income
LoanAmount	Loan amount in thousands
Loan_Amount_Term	Term of loan in months
Credit_History	credit history meets guidelines
Property_Area	Urban/ Semi Urban/ Rural
Loan_Status	Loan approved (Y/N)

There are altogether 13 columns in our data set. Of them Loan_Status is the response variable and rest all are the variables /factors that decide the approval of the loan or not.

Now let us look in to the each variable and can make some assumptions.(It's just assumptions right, there is no harm in just assuming few statements)

- Loan ID -> As the name suggests each person should have a unique loan ID.
- Gender -> In general it is male or female. No offence for not including the third gender.
- Married -> Applicant who is married is represented by Y and not married is represented as N. The information regarding whether the applicant who is married

is divorced or not has not been provided. So we don't need to worry regarding all these.

- Dependents -> the number of people dependent on the applicant who has taken loan has been provided.

- Education -> It is either non -graduate or graduate. The assumption I can make is " The probability of clearing the loan amount would be higher if the applicant is a graduate".

- Self_Employed -> As the name suggests Self Employed means , he/she is employed for himself/herself only. So freelancer or having a own business might come in this category. An applicant who is self employed is represented by Y and the one who is not is represented by N.

- Applicant Income -> Applicant Income suggests the income by Applicant.So the general assumption that i can make would be "The one who earns more have a high probability of clearing loan amount and would be highly eligible for loan "

- Co Applicant income -> this represents the income of co-applicant. I can also assume that " If co applicant income is higher , the probability of being eligible would be higher "
- Loan Amount -> This amount represents the loan amount in thousands. One assumption I can make is that " If Loan amount is higher , the probability of repaying would be lesser and vice versa"

- Loan_Amount_Term -> This represents the number of months required to repay the loan.

- Credit_History -> When I googled it , I got this information. A credit history is a record of a borrower's responsible repayment of debts. It suggests → 1 denotes that the credit history is good and 0 otherwise.

- Property_Area -> The area where they belong to is my general assumption as nothing more is told. Here it can be three types. Urban or Semi Urban or Rural

- Loan_Status -> If the applicant is eligible for loan it's yes represented by Y else it's no represented by N.

Exploratory Data Analysis:

Well don't get to worry about the fancy names like exploratory data analysis and all. By looking at the columns description in the above paragraph, we can make many assumptions like

- The one whose salary is more can have a greater chance of loan approval.

- The one who is graduate has a better chance of loan approval.
- Married people would have a upper hand than unmarried people for loan approval .
- The applicant who has less number of dependents have a high probability for loan approval.
- The lesser the loan amount the higher the chance for getting loan.

Advantages :

- Loan Prediction is very helpful for employ of banks as well as for applicant also.
- The aim of this project is to provide quick, immediate and easy way to choose the deserving applicants.
- It can provide special advantages to the bank.
- This helps all other departments to carried out other formalities.

Disadvantages :

- The disadvantage of this model is that it emphasize different ways to each factor but in real life sometime loan can be approved on the basis of single strong only, which is not possible through this system.

Applications:

Marketing:

- Businesses can use decision trees to enhance the accuracy of their promotionalcampaigns by observing the performance of their competitors' products and services.

- Decision trees can help in audience segmentation and support businesses in producing better-targeted advertisements that have higher conversion rates.

Retention of Customers:

- Companies use decision trees for customer retention through analyzing their behaviors and releasing new offers or products to suit those behaviors. By using decision tree models, companies can figure out the satisfaction levels of their customers as well.

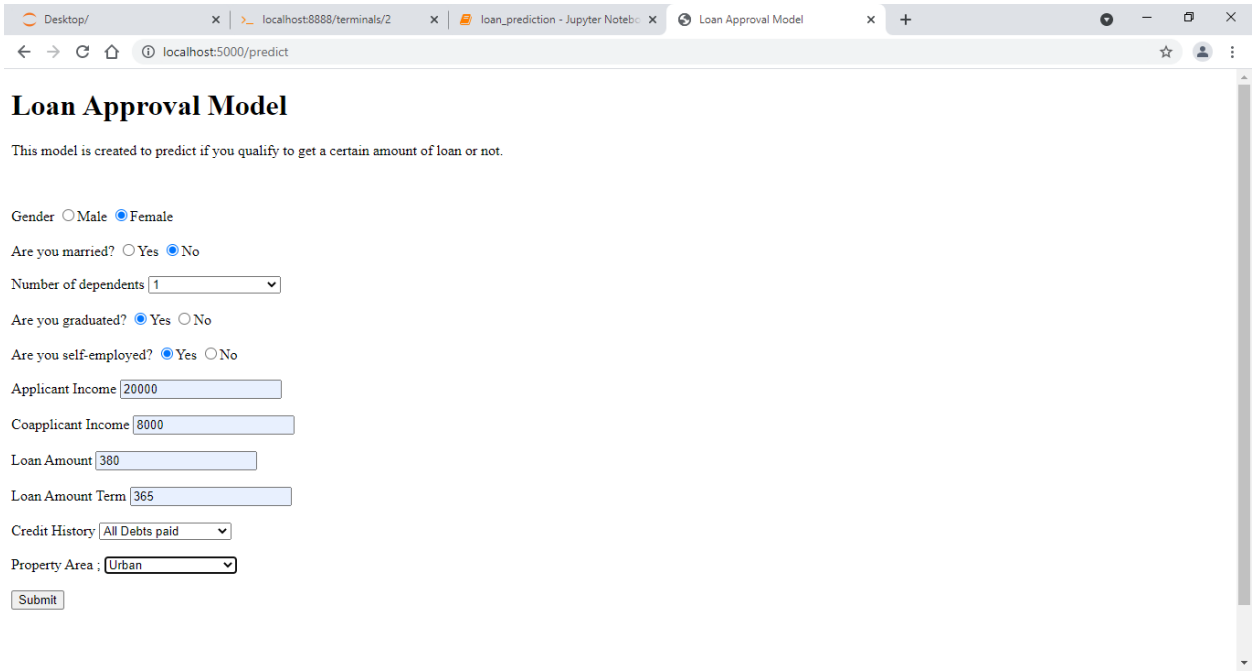
Diagnosis of Diseases and Ailments:

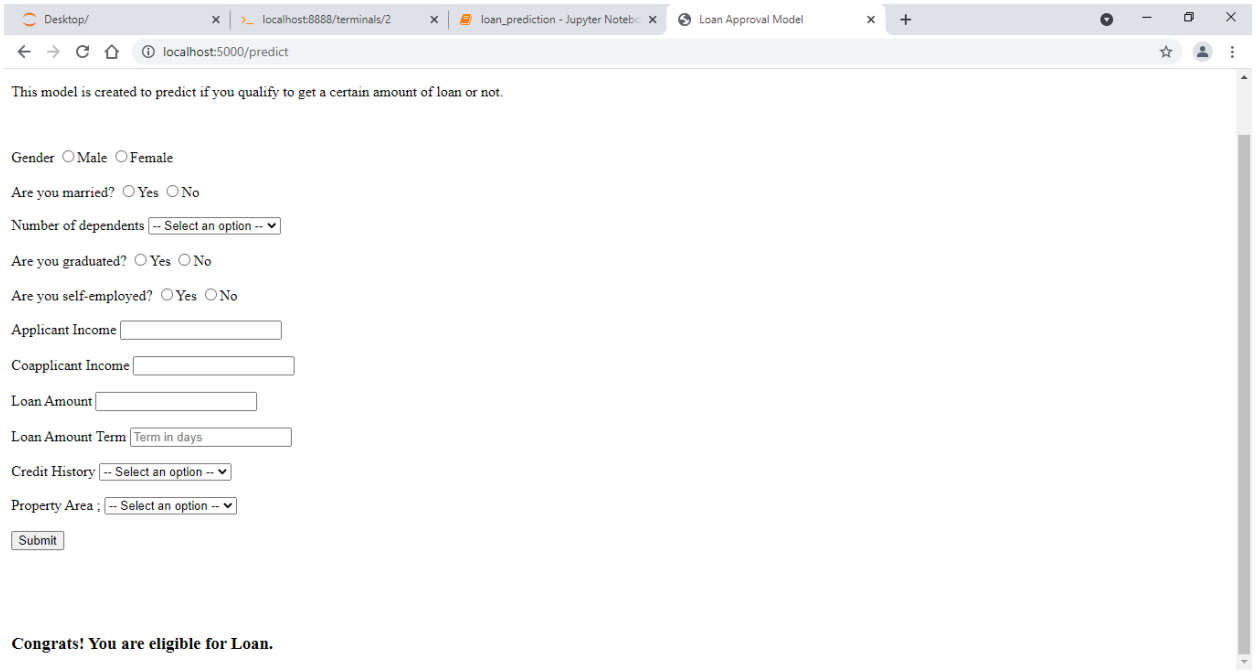
- Decision trees can help physicians and medical professionals in identifying patients that are at a higher risk of developing serious (or preventable) conditions such as diabetes or dementia. The ability of decision trees to narrow down possibilities according to specific variables is quite helpful in such cases.

Detection of Frauds:

- Companies can prevent fraud by using decision trees to identify fraudulent behavior beforehand. It can save companies a lot of resources, including time and money.

Ouput:





Conclusion:

We have proposed an efficient and reliable bridge for loan status prediction , when there is urgent need for people to take loan it may not be possible for people to communicate with each and every bank status for that loan status prediction can fulfill their requirements in short time span. So , that it can overcome the rate. Thus the proposed website can help people who is need of anytime and anywhere. This website is very useful for smart City and smart nation purpose

Future work:

This paper work can be extended to higher level in future. Predictive model for loans that uses machine learning algorithms, where the results from each graph of the paper can be taken as individual criteria for the machine learning algorithm.

References: https://github.com/smartinternz02/SI-GuidedProject-4195-1626238230

https://youtu.be/Nbh41J1gbwA

https://youtu.be/DL4CMPp249s