Loan Application Status Prediction Project

Introduction –

In this article, I am going to solve the Loan Application Status Prediction problem dataset in a very detailed manner. This is a classification problem in which I have to classify whether the loan will be approved or not. The classification refers to a predictive modelling problem where a class label is predicted for a given example of input data.

Problem Statement –

In the given dataset it includes details of applicants who have applied for loan. The dataset includes details like Gender, Marital Status, Education, credit history, loan amount, their income, dependents etc. To start this process, they were given a problem to identify the applicants' details, those who are eligible for loan amount so that they can specifically target these customers. It is a classification problem, from the given information about the applicant's I have to predict whether the they'll be to pay the loan or not.

I will start by exploratory data analysis, then pre-processing, and finally I will be testing different models such as Logistic regression and decision trees, KNeighbors Classifier model, SVM model, etc.

The given dataset consists of the following rows:

Independent Variables:

- Loan ID - The individual having unique Loan ID

- Gender - Male or Female

- Married - Applicant is married or not (Yes/No)

- Dependents - Number of dependents

- Education - Applicant Education (Graduate/Under Graduate)

- Self Employed - Self-employed (Yes/No)

- Applicant Income - Applicant income

- Coapplicant Income - Coapplicant income

- Loan Amount - Loan amount

- Loan Amount Term - Term of loan in months

- Credit History - credit history guidelines (yes/no)

- Property Area - Urban/ Semi Urban/ Rural

Dependent Variable (Target Variable):

- Loan Status - Loan approved or not (Yes/No) this is the target variable

Data Analysis –

I will be importing the necessary libraries and load the data. I will be using matplotlib for visualisation and pandas for data manipulation. I shall download the given dataset from here –

https://github.com/dsrscientist/DSData/blob/master/loan_prediction.csv

I will be importing the necessary libraries and load the data.

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project/v/1&cellId=3

Then I will be uploading the given dataset.

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1&cellId=4-5

Checking the size of the dataset.

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1&cellId=6

In the given dataset I have 614 rows and 13 columns including target columns.

EDA Conclusion –

I can see that there is some missing data, so I can further explore this using the pandas statistical describe function.

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1&cellId=11

Checking the null values in the given dataset.

https://jovian.ai/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1#C14

There are some variables have missing values thus I will have to deal with given data and also there seems to be some outliers for the Applicant Income, Coapplicant income and Loan Amount. Thus, also see that about 84% applicants having a credit history. Because the mean of Credit History field is 0.84 and also it has either 1 for having a credit history or 0 for not.

It would be interesting to study the distribution of the numerical variables mainly the Applicant income and the loan amount. To do this I will be using matplotlib/seaborn for the data visualization.

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1&cellId=37-38

The people with better education should normally have a higher income, I shall check that by plotting the education level against the income. The distributions are quite similar but I can see that the graduates have more outliers which means that the people with higher income are most likely well educated.

Another interesting variable is credit history, in that I observe how it affects the Loan Status of the applicants. The people with a huge credit history a way more likely to pay their loan. This means that credit history will be an influential variable in this dataset.

Pre-processing the Dataset –

Dropping the unwanted column in the given dataset.

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1&cellId=16-20

To try out different models I shall create a function that is takes in a model. The accuracy which means that using the model on the train set and measuring the error of the same dataset. And also, I use a technique cross validation which splits randomly data into train and test set, trains the model using the train set and validates it with the test set and takes the average error. The latter method gives a better idea on how the model performs in real life.

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1&cellId=46-55

Building Machine Learning Models –

Now I can test different models –

Classifying by using Logistic regression model –

This is a classification algorithm it is uses as 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.

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1&cellId=57-65

Classifying by using KNeighbors Classifier model –

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1&cellId=67-73

Classifying by using Random Forest Classifier model -

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.

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1&cellId=80-86

Classifying by using Decision Tree Classifier model -

This is a supervised machine learning algorithm mostly used for classification problems. All features in these discretized in this model. The homogeneity and purity of the nodes increases with respect to the dependent variable.

https://jovian.ml/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1&cellId=74-80

I have used multiple algorithms for training the dataset purposes like Decision Tree, Random Forest, SVC, Logistic Regression, etc. Among all the algorithms logistic regression performs best on the validation of the given dataset with an accuracy score is **70.7%.**

Saving the given model –

https://jovian.ai/admanenilima19/loan-application-status-prediction-final-project-e1950/v/1#C92

Conclusion –

I did the EDA on the features of this dataset and saw how each feature is distributed.

I analysed each variable to check for the data is cleaned and normally distributed.

I cleaned the data and removed all NA values.

I also prove that an association between the independent and the target variables.

Also calculated the relation between the applicant income and the loan amount have the significant relation.

I also found that credit history is crating more impact on the loan giving decisions.

The accuracy of Logistic Regression is: 70.17 %.

The accuracy of KNN is: 62.68 %.

The accuracy of SVM is: 58.53 %.

The accuracy of Decision Tree Classifier is: 53.63 %.

The accuracy of Random Forest Classification is: 58.53 %.