Loan prediction classification problem.

This document covers the steps and procedures towards addressing the loan prediction machine learning problem. This covers steps from problem identification to model deployment.

Problem definition:

A housing finance company deals in home loans both in the urban, semi-urban, and rural areas. These customers first apply for a home loan and after that a company validates their eligibility for a loan. The bank wants to automate this process (real-time based) for customers filling in an online application form. Details such as age, gender, marital status, education level, number of dependents, applicant income, co-applicant income, loan amount, credit history, property area, and loan term amount.

The task is to automate this process by identifying the customer segments who are eligible for the loan based on information such as gender, marital status, education status among.

Data understanding and preprocessing:

The data was obtained from *kaggle.com* and stored in csv format was loaded using the *pandas* data frame to enable getting an overview of the data. For purposes of exploratory data analysis, the *pandas profiling package* was used to make the process faster. The observations from the profiling report were;

- There were 13 variables and 381 observations.
- 75 cells had missing values
- No duplicates were found

For purposes of data cleaning, the Loan_ID column was removed because it wasn't in anyway relevant for loan analysis. For variables such as Gender, Dependents, Self_Employed, and Credit_History, the missing values were replaced by mode being that they are categorical variables and for the Loan_Amount_Term, mean was the best replacement metric it being a numerical variable. The categorical variables were later hot encoded (changed to numerical from string) to make data ready for machine learning tasks.

Model building process.

The input (entire features apart from Loan_Status) and target attributes (Loan_Status) were separated and then the input data scaled (normalized) to avoid data leakage.

The data was then split into the training sets and testing set at a 0.2 split ratio which implies 20% was for testing and 80% for training.

With the nature of our dataset, 4 classification algorithms were tested; *random forest, xgboost, SVC, and decision tree classifier.*

Model evaluation.

The xgboost model was more accurate with an accuracy score of 82% on the testing set implying that it would return a correct loan status 82 percent of the time when information is fed into the model.

Conclusions.

It was observed also that the most important factor towards loan status prediction was credit history since financially it influences one paying ability.

A simple interface using *gradio* was designed to enable users to only feed in the required information and status of yes or no returned.