This project proposes a solution to the loan prediction problem imposed on https://datahack.analyticsvidhya.com/contest/practice-problem-loan-prediction-iii/.

The Dataset

The training set (train_ctrUa4K.csv) is composed by 614 rows, each of them containing information about home loans. Goal of the classification is to identify the customers who are eligible for loan amount so that they can specifically target these customers. A brief explanation of the available features is provided in the subsequent table.

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)

Feature Engineering - Filling NA values

a) Train set

At first, the “Loan_ID” column is dropped because it can’t provide useful information for the prediction process. Then the missing values of ‘Gender’, ‘Married’, ‘Dependents’, ‘Self_Employed’, ‘LoanAmount’, ‘Loan_Amount_Term’ and ‘Credit_History’ columns are filled by using different tactics:
- ‘Married’ -> the rows that have missing value in this column are only 3, so they are dropped.
- ‘Credit_History’ -> it is filled based on ‘Loan_Status’ meaning that if the ‘Loan_Status’ has 'Y' then the ‘Credit_History’ will be '1' else it will be '0'.

- ‘Dependents’ -> it is filled based on ‘Married’ as we use the mean number of dependent members of the married (1 member) or the mean number of dependent members of the not married (0 members).
- ‘LoanAmount’ -> it is filled based on ‘Education’ and 'Self_Employed' as we use the mean value of ‘LoanAmount’ for the following cases: graduate and not self employed, graduate and self employed, not graduate and not self employed and last, not graduate and not self employed.
-‘Gender’, ‘Self_Employed’, ‘Loan_Amount_Term’ -> these columns are filled by using KNN meaning that we treat each column as the target feature and all the other columns are used for training a KNN model.

Also, there is the problem of conversion of non numeric columns to numeric. The ‘Married’, ‘Education’, ‘Self_Employed’ and ‘ Loan_Status’ are simply converted by replacing the original values (‘Yes/No’, ‘Graduate/Not Graduate’, ‘Y/N’, ‘Y/N’ respectively) with 1 and 0. For ‘Gender’, ‘Dependents’ and ‘Property_Area’ columns we create a new column for each value (‘Male/Female’, ‘0/1/2/3+’, ‘Urban/Semi Urban/Rural’), so 9 new columns are created and the original ‘Gender’, ‘Dependents’ and ‘Property_Area’ are dropped. In these 9 new columns we put 1 and 0 depending on the values of the original 3 columns.

Last but not least, the numeric and fully completed (no NA values) ‘ApplicantIncome’ and ‘CoapplicantIncome’ columns are summed into one single column called ‘Total_Income’.

b) Test set

The test set is treated exactly as the train set. However, in this set there isn’t a completed ‘Loan_Status’ column as it’s the target column that our model has to fill. Consequently the ‘Credit_History’ column is filled differently because in train set its completion is based on ‘Loan_Status’. So, in test set the ‘Credit_History’ is filled by using KNN.

Algorithms

In this project, the performance of 5 common classification algorithms is tested: Decision Trees, Random Forest, KNN (K-Nearest Neighbor), SVM (Support Vector Machines) and Logistic Regression.

Evaluation and Metric

In order to determine which algorithm performs the best the train set is splitted into a new train and test set (70% - 30%). So, 5 models are implemented.

According to the competition statement , the metric that is used in order to evaluate the performance of every model is accuracy.