Data Science Capstone Project

By Muhammad Daniel Haikel

Agenda

Problem Statement

Data Set

Methodology

Classification Findings

Conclusion

Problem Statement

- To predict whether a customer will default on their loan based on the features found in the data set available.

Data Set

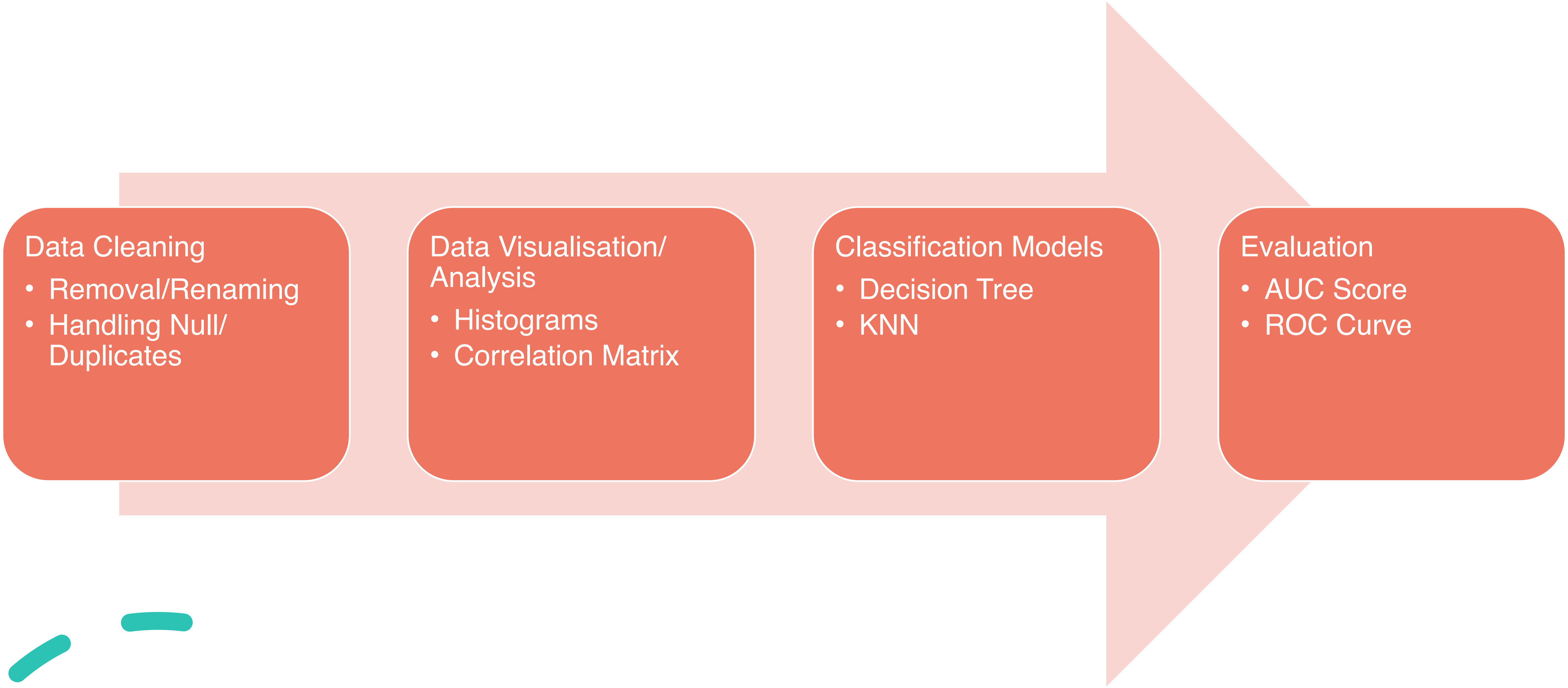
- Loan_ID
- Gender
- Married
- Dependents
- Education
- Self_Employed
- ApplicantIncome
- CoapplicantIncome
- LoanAmount
- Loan_Amount_Term
- Credit_History
- Property_Area
- Loan_Status

There are 13 columns of data available that can be used for analysis.

Loan_ID was dropped as it does not influence the loan status of an individual.

The target value is Loan_Status and the rest of the data set are used as features in the models used later.

Methodology



Data Cleaning

- Removal/Renaming
- Handling Null/ Duplicates

Data Visualisation/ Analysis

- Histograms
- Correlation Matrix

Classification Models

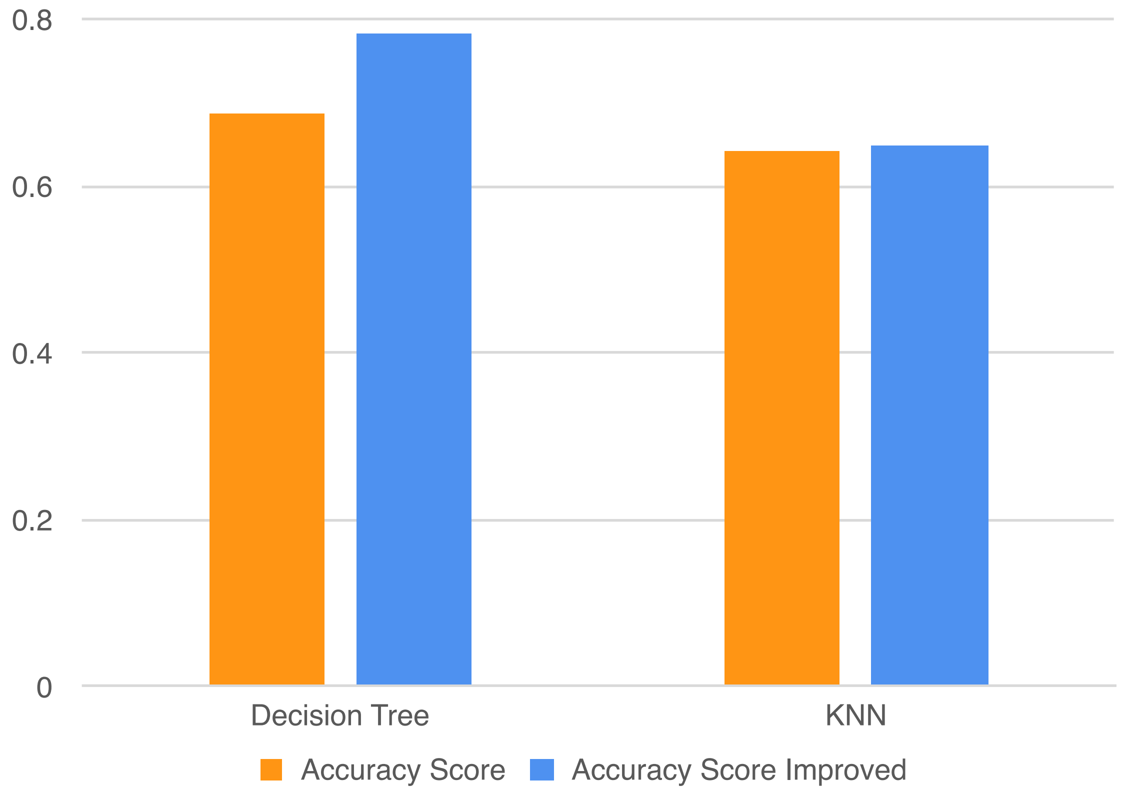
- Decision Tree
- KNN

Evaluation

- AUC Score
- ROC Curve

Classification Findings

Applied Decision Tree Classifier and KNN algorithm to predict whether a customer will default on their loan based on their features against their loan status. Accuracy score for both classification models are as above and are slightly better after tuning. The Decision Tree Classifier model performed better for this dataset.



Conclusion

Correlation and collinearity of the features in this data set is very low with the highest correlation with Loan Status is Credit History. With the collinearity being low within the features, the data set might be unclean or sample size being too small.