Bhuvana Chandrika Kothapalli

Spandan Pal

Mansi Sharma

John Hwang

Serena Song

****

Data Science Programming Final Project

Group 18

**Description:**

In this project, we are going to study a loan approval dataset to understand which factors affect the loan approval of a candidate. The dataset includes various pieces of information that play a role in assessing whether applicants are eligible for requested loans. Using this data, we will try to conduct exploratory analysis to determine the important factors affecting the loan approval. We will also try to design models to learn patterns and relationships from historical data and use them to make predictions about future loan approval outcomes.

The analysis of this dataset aids in addressing issues faced by loan applicants and financial institutions. It helps reveal key factors impacting loan approval, enabling applicants to understand their financial situation better and helps institutions optimize their risk assessment models with this data, potentially lowering default rates. Moreover, this can aid in the automation of the loan approval process which would not only streamline operations but also lead to better customer experience. In summary, this dataset aids in comprehending and predicting loan approval, optimizing risk assessment and enhancing financial planning.

**Exploratory Analysis of the Data:**

We have a total of 4269 records present in the Loan Approval dataset. The dataset contains fields such as loan amount, income, asset valuations (commercial/ residential/ luxury/ bank), CIBIL score, loan term etc. The details for these predictors can be seen in Exhibit 1.

When checked for data quality, the dataset did not reveal any null records. Following this, we plotted a correlation heatmap for all the numerical predictors in our dataset. This can be seen in Exhibit 2. During this time, we see that there is a significant correlation between income\_annum, loan\_amount along with other fields like luxury\_assets\_value and bank\_asset\_value. Due to this, we would remove some of the columns from our analysis dataset in the future steps, to avoid multicollinearity and to reduce complexity of the model.

Following this, we investigate the relationship between the predictors and the variable to be predicted i.e. loan status. This can be seen in Exhibit 3. Here, we notice that for the fields no\_of\_dependents, income\_annum, loan\_amount, residential\_assets\_value, commercial\_assets\_value, luxury\_assets\_value, bank\_asset\_value, we do not see any variation in the distribution indicating that these variables do not impact loan status. At the same time, we see that as loan\_term increases, the chances of rejection increases. Moreover, we see that as the Cibil score crosses a threshold, the chances of approval increase significantly.

Following this, we investigate the relationship between the categorical variables and the loan status. This can be seen in Exhibit 4. Here, we see that education or self\_employed do not have any variation with respect to loan status, which indicates that the fields do not have a strong impact on loan status.

We tried using log transformation on the numerical fields to see but, we do observe any change in the relationships between the predictors. Combining the asset variables also does not give us any fruitful result.

**Solution and Insights:**

Fields like commercial\_asset\_value, residential\_asset\_value, loan\_amount have a skewed distribution. To handle this, we create equal buckets for the fields. We also convert the categorical fields into new binary columns using one-hot encoding. Moreover, we remove the fields which we do not need, like load\_id. We also remove fields like luxury\_assets\_value, bank\_asset\_value to reduce complexity of the model. At last, we scale the data using the StandardScaler function.

**Models:**

We split our data into a 30-70 test-train split which we will use for the following models: Logistic Regression, K-nearest Neighbors, Decision Trees, Random Forest, Gradient Boosting and Neural Network.

Logistic Regression：Since this is a classification problem, we first used Logistic Regression on our data. Here, we get an accuracy of 0.93, precision of 0.94, recall of 0.93 and an AUC score of 0.935. These can be seen in Exhibit 5.

K-nearest Neighbors：A Grid Search Cross Validation was performed to determine the optimal hyperparameters for the KNN model. This process identified that the model performed best with 20 neighbors and distance-based weighting. The KNN model was trained using the optimal parameters on the training data. The trained model was then used to make predictions on the test data. The model's performance was assessed using metrics like precision, recall, and F1-score. Additionally, an ROC curve was plotted, and the AUC (.8997) was calculated to provide a comprehensive view of the model's performance. This can be seen in Exhibit 6.

Decision Tree：A Grid Search Cross Validation was performed to determine that the model performed best with a maximum of 6 depths. A Decision Tree model was trained using the optimal hyperparameters on the training data, which was then used to make predictions on the test data.

The model's performance was assessed using metrics like precision, recall, and F1-score. Additionally, an ROC curve was plotted, and the AUC of .9347 was calculated, showing that a decision tree fits the loan dataset better than KNN (which had an AUC of .8404). The runtime of the decision tree was also comparable with the KNN model (.5179 vs .5241 respectively), showing better prediction for less computational cost. This can be seen in Exhibit 7.

Random Forest：A Grid Search Cross Validation was performed to determine that the model performed best with a depth of 30 and 150 estimators. A Random Forest was trained using the optimal hyperparameters on the training data, which was then used to make predictions on the test data.

The model's performance was assessed using metrics like precision, recall, and F1-score. Additionally, an ROC curve was plotted, and the AUC of .9344 was calculated, showing that the random forest model has similar prediction power to the decision tree. However, the random forest model’s runtime of 1.4266 seconds was much worse than the decision tree model. This result showed that the random forest model was computationally heavier than the previous models we have fit. This can be seen in Exhibit 8.

Gradient Boosting: A Grid Search Cross Validation was performed to determine that the model performed best with a depth of 4 , 150 estimators and a learning rate of 0.01. A gradient boosting model was trained using the obtained parameters on the training data, which was then used to make predictions on the test data. The model's performance was assessed using metrics like precision, recall, and F1-score. Additionally, an ROC curve was plotted, and the AUC of .940 was calculated, showing the highest accuracy till now. However, the runtime of the boosting model of 1.88 seconds was much worse than random forest. This can be seen in Exhibit 9.

Neural Network: For Neural Network, we construct a Keras neural network for binary classification comprising an input layer, two ReLU-activated hidden layers with 64 and 32 neurons respectively, and a sigmoid-activated output layer. It's trained on data using the RMSprop optimizer. Post-training, predictions made on the test are evaluated, with their performance metrics provided in a classification report. The model's discriminative ability is visually represented through an ROC curve and quantified by the AUC score. Finally, the training duration is displayed, providing insights into the model's efficiency. The Neural Network achieves the top AUC of .982 but also takes the longest to run compared to other models. This can be seen in Exhibit 10.

**Conclusion:**

The CIBIL Score, commonly known as the Credit Score, stands out as the primary predictor for loan approval, with scores over 540 greatly enhancing approval chances. Likewise, the Loan Term inversely affects approval rates, meaning longer loan terms decrease approval likelihood. Interestingly, while education and self-employment factors seemed inconsequential during initial analysis, the machine learning model indicates their significance, likely due to interplay with other variables. This can be seen in Exhibit 11.

Neural Network emerges as the most accurate model for this dataset. Using it with only significant predictors yields an AUC score of .978, just slightly below the .982 achieved when using all predictors, confirming minimal accuracy loss when focusing on significant indicators.

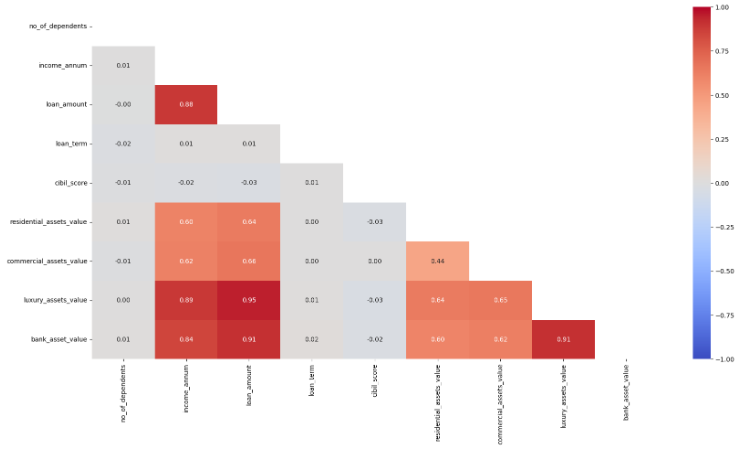
**APPENDIX**

**Exhibit 1 :** Details of all the variables present in the Loan Approval Dataset

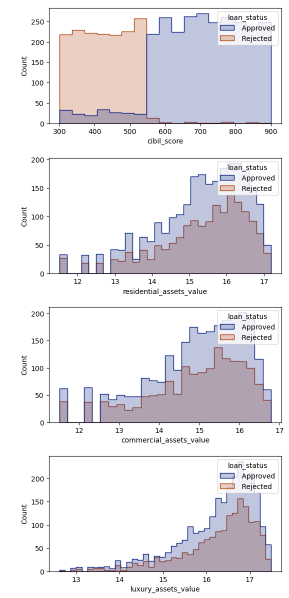
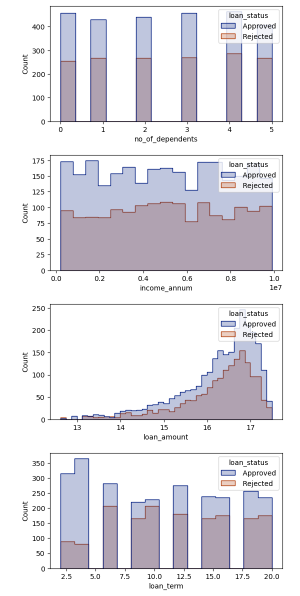
Variable Description

| loan\_id | Unique ID of the Loan |
| --- | --- |
| no\_of\_dependents | Number of dependent members related to the applicant |
| education | Educational Status of the Applicant (Graduate/Non Graduate) |
| self\_employed | Self Employment Status of the Applicant (Yes/No) |
| income\_annum | Income Per Annum of the Applicant |
| loan\_amount | Request Amount For the Loan |
| loan\_term | The number of years applicable for the loan |
| cibil\_score | The credit score of the applicant |
| residential\_assets\_value | Valuation of the residential assets of the applicant |
| commercial\_assets\_value | Valuation of the commercial assets of the applicant |
| luxury\_assets\_value | Valuation of the luxury assets of the applicant |
| bank\_asset\_value | Valuation of the bank assets of the applicant |
| loan\_status | Loan Approval Status (Approved/Rejected) |

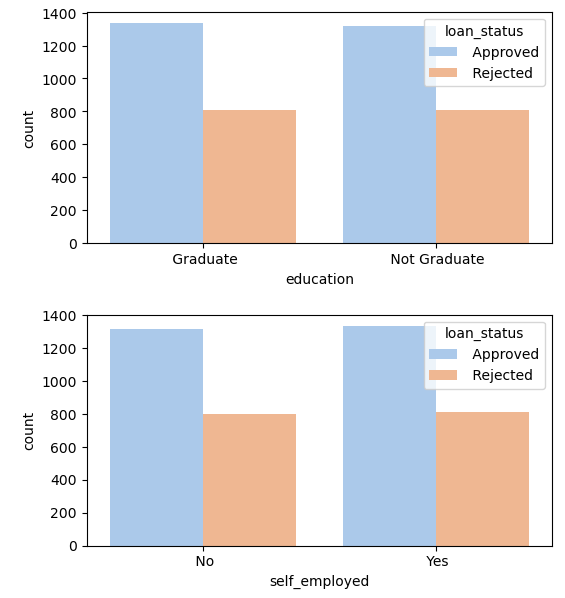
**Exhibit 2 :** Correlation Heatmap of all the Numerical Attributes



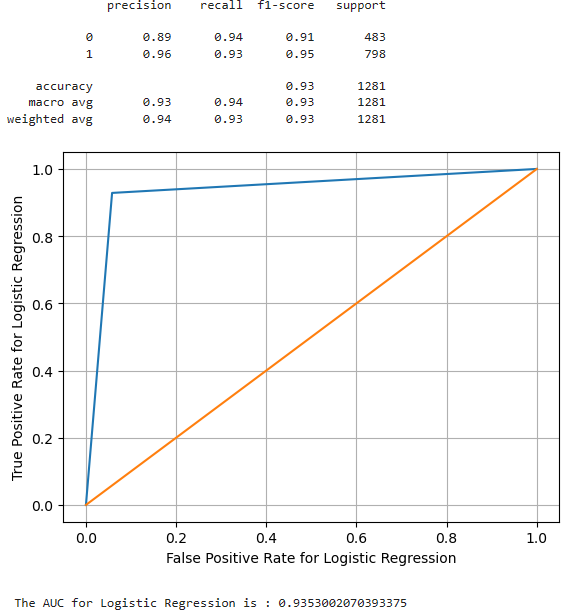
**Exhibit 3 :** Histogram of all the numerical variables with respect to loan status



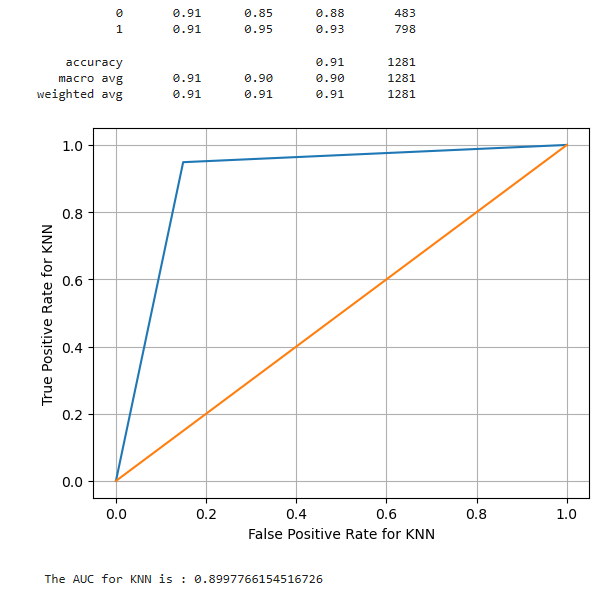
**Exhibit 4 :** Count of all the categorical variables with respect to loan status



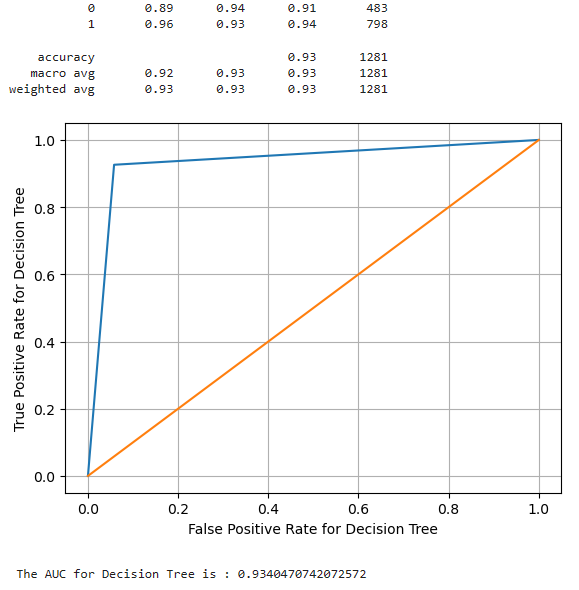
**Exhibit 5 :** Logistic Regression



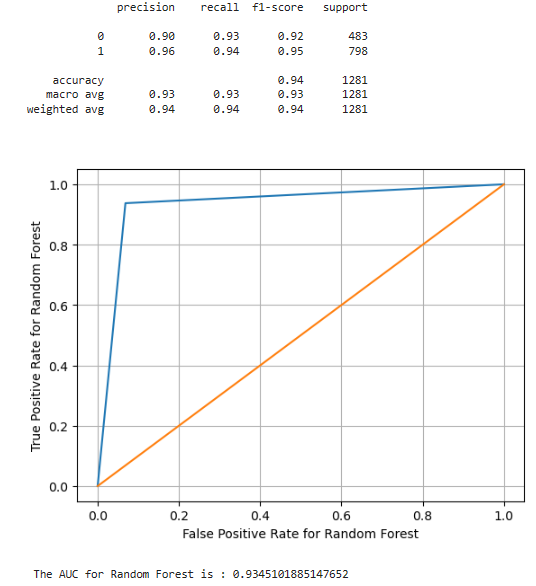
**Exhibit 6 :** KNN



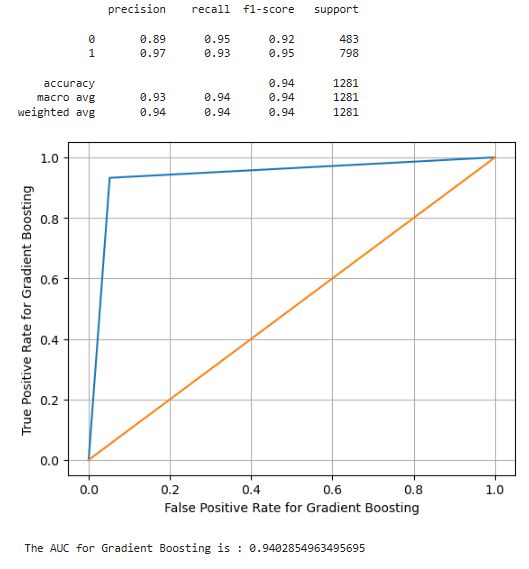
**Exhibit 7 :** Decision Tree



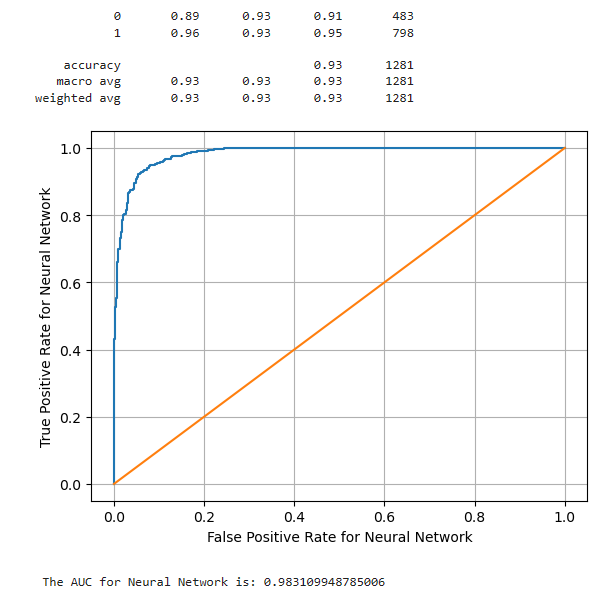
**Exhibit 8:** Random Forest



**Exhibit 9 :** Gradient Boosting



**Exhibit 10 :** Neural Network



**Exhibit 11 :** Feature Importance

