Capstone Project (Loan Approval)

Daniel Chan

Problem Statement

To predict whether a customer is eligible for loan approval, given customer detail provided while filling online application form regarding custoemr background and credit history.

*Dataset taken from:

https://www.kaggle.com/ninzaami/loan-predication

Data Preprocessing

- 13 columns x 614 rows of data
- **Target Column**: ‘Loan_Status’
- **Numerical Columns** included: ‘ApplicantIncome', 'CoapplicantIncome', 'LoanAmount', 'Loan_Amount_Term',
- **Categorical Columns** included: ‘Loan_ID', 'Gender', 'Married', 'Dependents', 'Education', 'Self_Employed', 'Credit_History', 'Property_Area', 'Loan_Status'

Data Preprocessing

- Cleaning numerical columns:
  - Outliers for numerical columns (using z score>3 as a threshold) were removed
  - Missing values were filled using median
  - After train-test split, feature scaling was utilized to build a better machine learning model later
- Cleaning categorical columns
  - Missing values were filled using mode
  - Label Encoder used to transform data

EDA

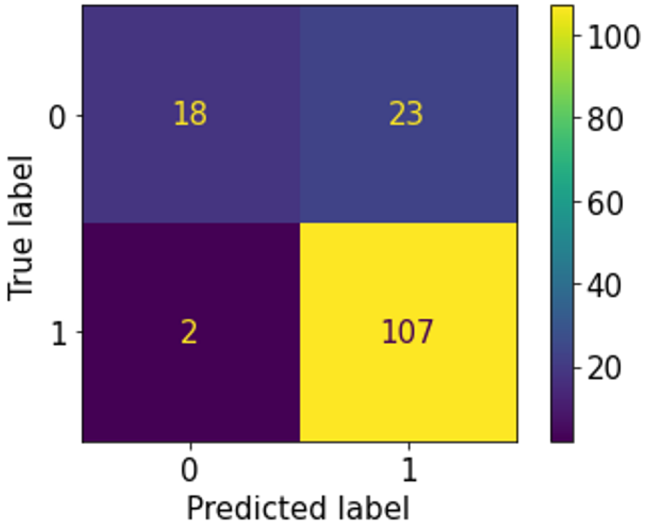
	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History
count	600.000000	600.000000	600.000000	600.000000	600.000000
mean	4870.668333	1421.538200	141.301667	343.000000	0.856667
std	3380.099718	1684.082008	74.952416	64.089357	0.350705
min	150.000000	0.000000	9.000000	12.000000	0.000000
25%	2880.500000	0.000000	100.000000	360.000000	1.000000
50%	3768.500000	1188.500000	126.500000	360.000000	1.000000
75%	5704.250000	2253.250000	160.250000	360.000000	1.000000
max	20833.000000	8980.000000	650.000000	480.000000	1.000000

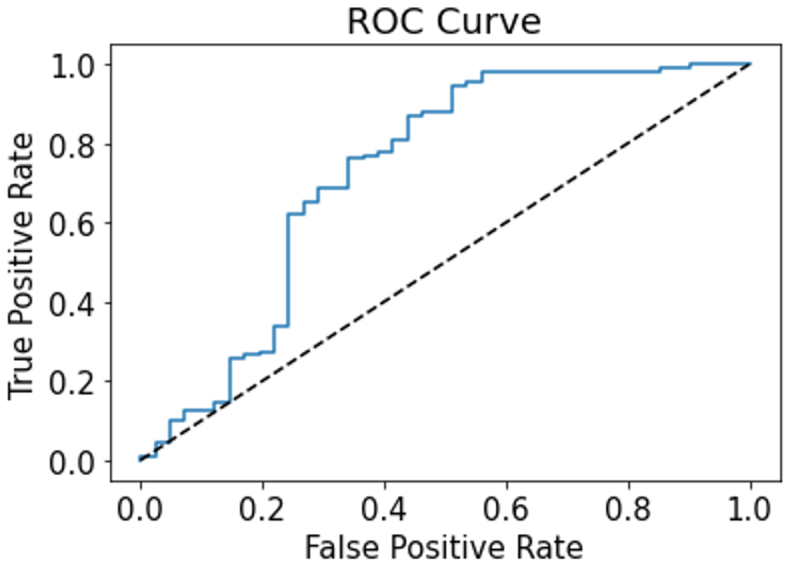
EDA

- Applicant Income was skewed left, with most (75%) of applicants earning <$5704
- Of the data…
  - 81.5% were from male customers
  - 65.2% were married customers
  - 59.3% had 0 dependents
  - 77.7% were graduates
  - 86.7% were not self-employed
  - 85.7% had a credit history
  - 69.2% had their loan approved
- Distribution amongst customers' property areas was roughly equal
- Customer's Income had a high correlation with Loan Amount (0.605)

Model accuracy and evaluation: Logistic Regression

- **Accuracy: 0.83, AUC = 0.710** (closer to 1 is better)
- False Positive Rate: 23/150=15.33%



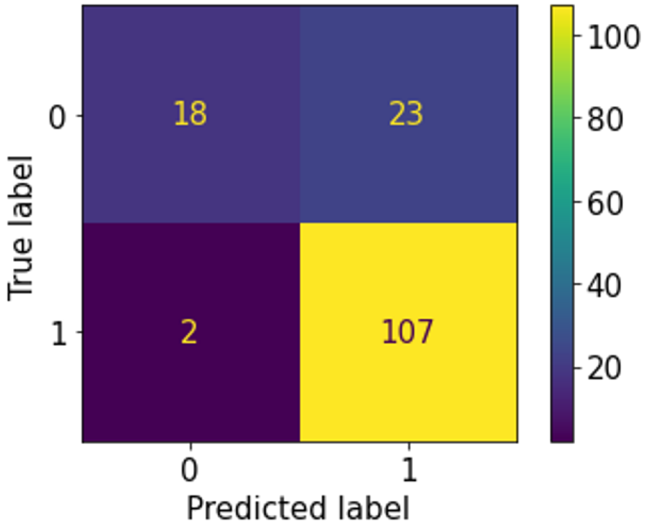


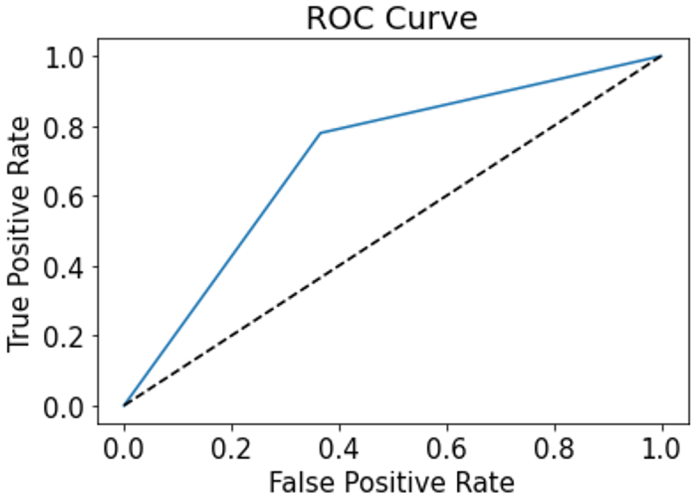
Logistic Regression (Hyperparameter Tuning Attempt)

- Attempts to adjust model parameters were unsuccessful in achieving higher accuracy

Model accuracy and evaluation: Decision Tree

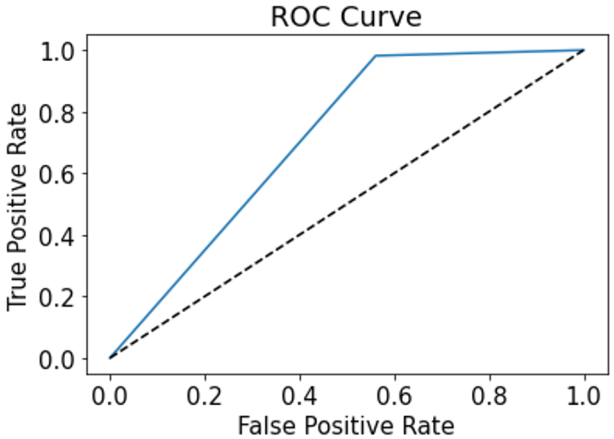
- **Accuracy: 0.76, AUC = 0.707** (closer to 1 is better)
- False Positive Rate: 23/150=15.33%

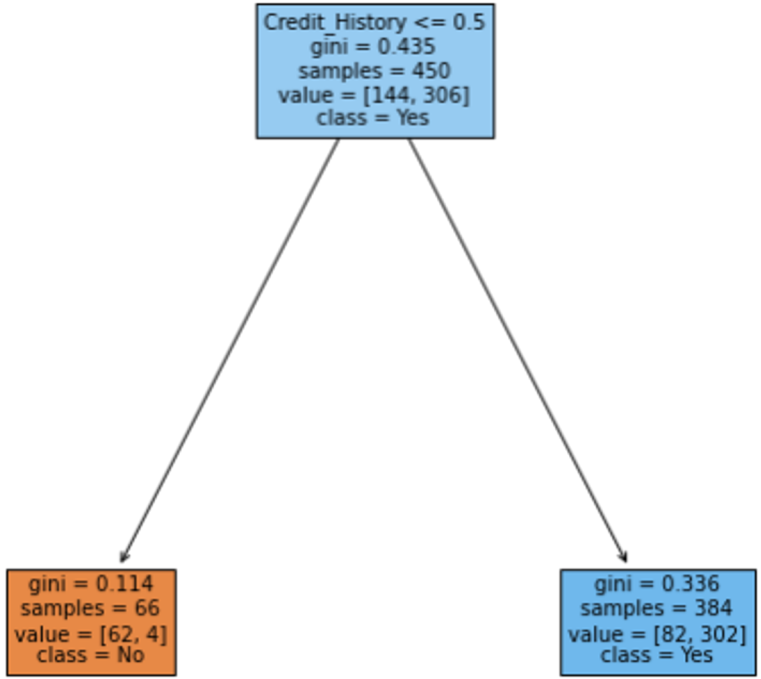




Decision Tree (Hyperparameter Tuning Attempt)

- **Adjusted parameters in attempt to find higher accuracy**
- **Highest Accuracy for best model found: 0.83, AUC = 0.707** (closer to 1 is better)





Suggested Model

- While the parameter-adjusted decision tree model achieved an accuracy score similar to the logistic regression model (accuracy = 0.83 for both), the decision tree model would **not** be recommended as the model's prediction relies on one key variable: customer's **Credit History**
  - Using on a model that relies on only one key variable would fail to make use of other variables containing potentially useful customer data
  - Relying on only one variable for prediction brings up practical issues too (e.g.missing data)
  - Furthermore…
    - Credit History data was heavily skewed (85.7% of customers had a credit history while only 14.3% did not)
    - Credit History data also had the most missing values out of all variables (n=49), which was then filled with the median value i.e. having a credit history

Suggested Model

- Meanwhile, the logistic regression model factors in multiple variables of customer data into its regression equation to churn out predictions

- Hence, the logistic regression model would be the recommended model for loan approval prediction

Dataset Limitations and Future Directions

- A more representative sample of customer data, eg for certain customer demographics that were heavily skewed in the current dataset (e.g. Gender, Married, Graduates, Credit History), may enable the creation of a more accurate machine learning model across demographics
- Target Variable (“Loan_Status’’ was not equally distributed which may affect modelling
- Alternative prediction models that were not evaluated in this project can be explored (e.g. K-neighbours, Random Forest)