



Credit Risk Assessment of Borrowers & Machine Learning Model for Approval for Loan Application

Problem Statement

Default in loan repayment by borrowers **increases the liquidity risk** of financial institutions. If severe, banks may run into problem meeting short-term debt obligations and face insolvency.

To reduce the bank's vulnerability to failure due to bad debts, it is necessary to:

- i. Understand the type of loans that are susceptible to being defaulted, and
- ii. Understand the profile of the borrowers who defaulted payment by analysing the:
 - (a) Borrowing histories; and
 - (b) Backgroundof past loan applicants.

A machine learning model can be developed to predict the likelihood of an applicant failing to repay loan and reject the application.

Overview of Dataset

The dataset consists of the profile of the borrower (e.g. age, income, employment length), and the type and quantum of loan taken.

Features

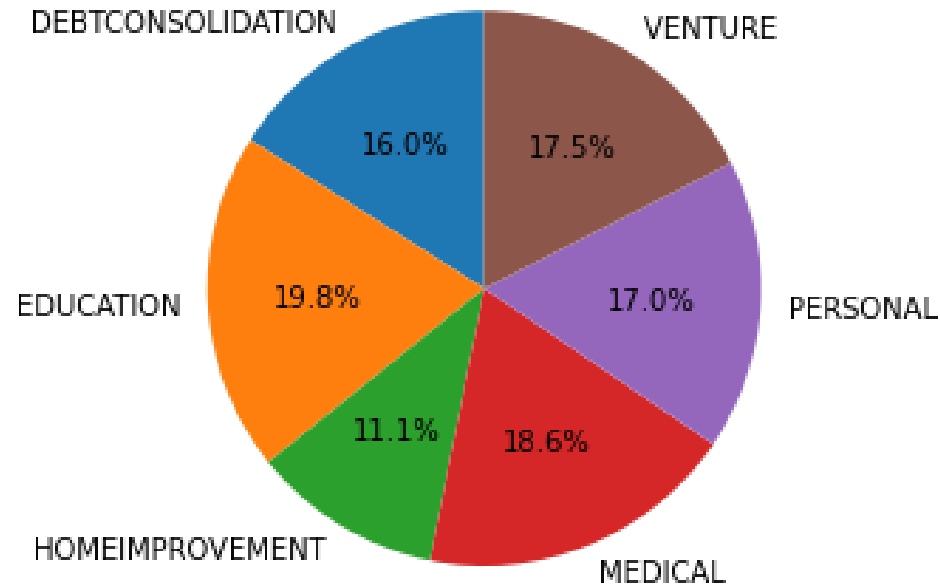
1. Age
2. Annual Income
3. Home Ownership
4. Employment Length
5. Purpose of Loan
6. Loan Grade
7. Loan Amount
8. Interest Rate
9. Loan Status (0 = non-default; 1 = default)
10. Percentage of Income
11. Historical Default
12. Credit History Length

Predictions of the likelihood of default of payment by loan applicants will be made using the K Nearest Neighbor, Logistic Regression and Decision Tree model.

I. Overview of Loan Demand and Loan Status

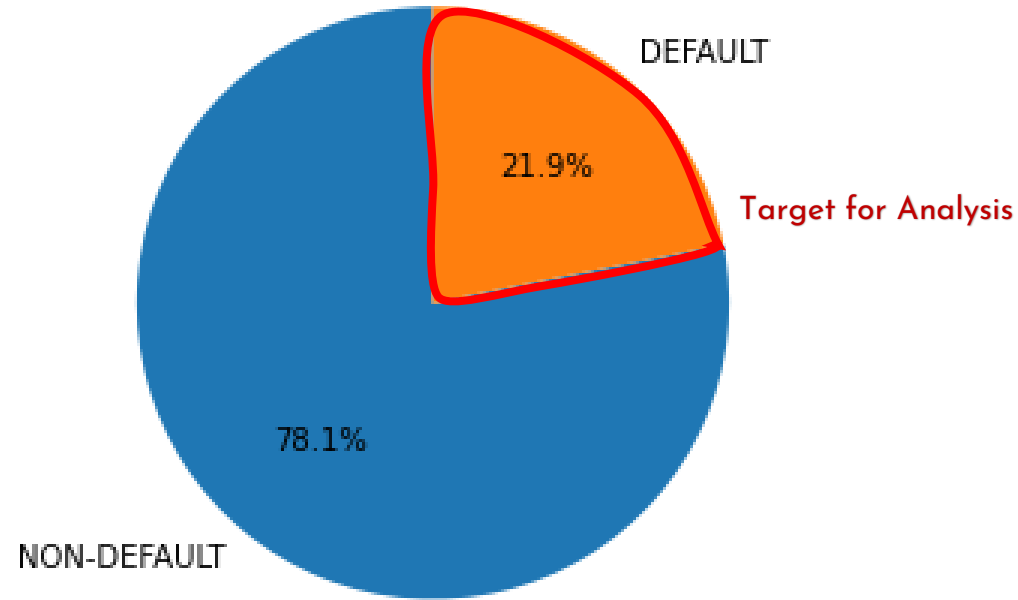


Demand for Loan



Insight: The demand for loan is relatively equal across all categories, with the exception of 'Home Improvement'.

Proportion of Borrowers Defaulting Repayments

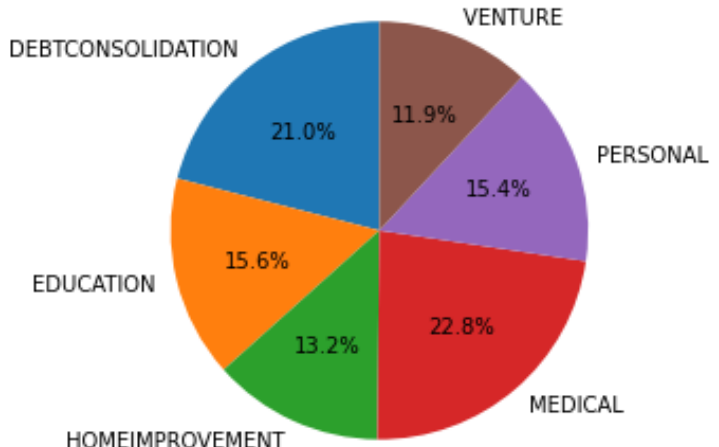
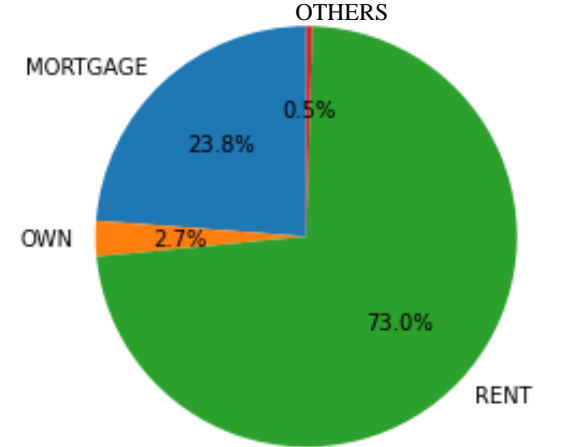


Insight: 21.9% of the borrowers had defaulted repayments on their loan.

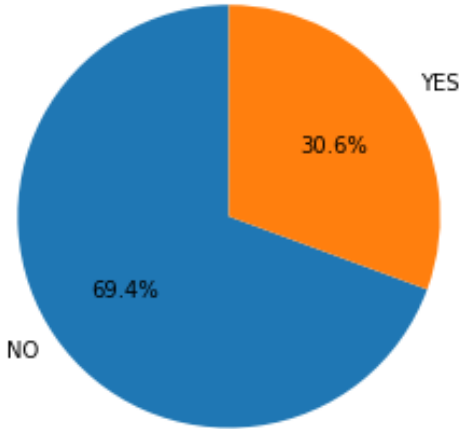
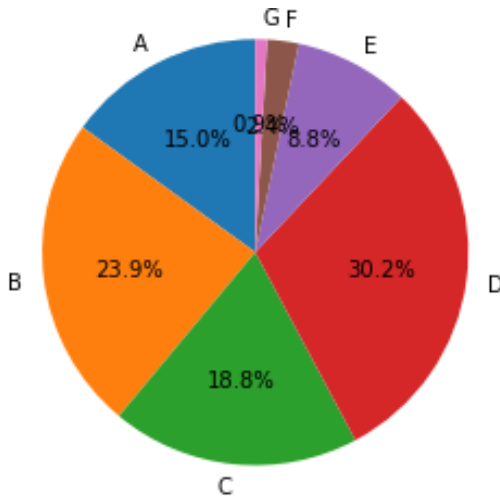
II. Understanding the Profile of Borrowers who Defaulted Repayments



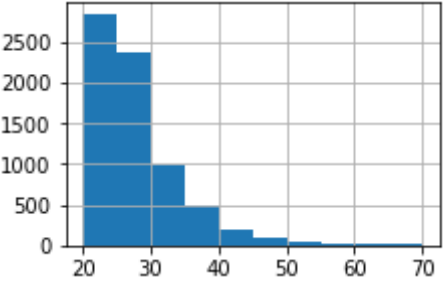
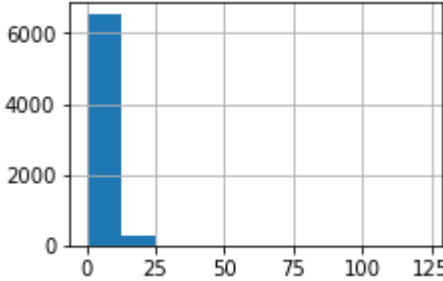
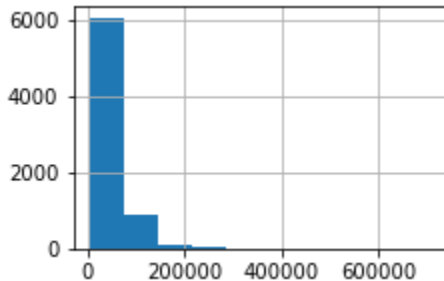
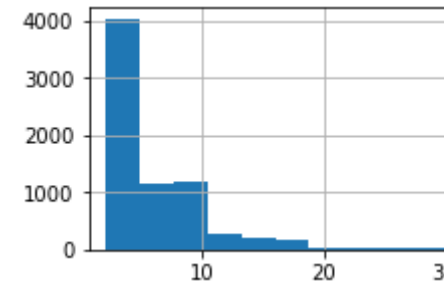
Categorical Data Review

Purpose of Loan	Home Ownership																								
 <table border="1"> <caption>Purpose of Loan Data</caption> <thead> <tr> <th>Purpose</th> <th>Percentage</th> </tr> </thead> <tbody> <tr> <td>MEDICAL</td> <td>22.8%</td> </tr> <tr> <td>DEBTCONSOLIDATION</td> <td>21.0%</td> </tr> <tr> <td>PERSONAL</td> <td>15.4%</td> </tr> <tr> <td>EDUCATION</td> <td>15.6%</td> </tr> <tr> <td>HOMEIMPROVEMENT</td> <td>13.2%</td> </tr> <tr> <td>VENTURE</td> <td>11.9%</td> </tr> </tbody> </table>	Purpose	Percentage	MEDICAL	22.8%	DEBTCONSOLIDATION	21.0%	PERSONAL	15.4%	EDUCATION	15.6%	HOMEIMPROVEMENT	13.2%	VENTURE	11.9%	 <table border="1"> <caption>Home Ownership Data</caption> <thead> <tr> <th>Ownership Type</th> <th>Percentage</th> </tr> </thead> <tbody> <tr> <td>RENT</td> <td>73.0%</td> </tr> <tr> <td>MORTGAGE</td> <td>23.8%</td> </tr> <tr> <td>OWN</td> <td>2.7%</td> </tr> <tr> <td>OTHERS</td> <td>0.5%</td> </tr> </tbody> </table>	Ownership Type	Percentage	RENT	73.0%	MORTGAGE	23.8%	OWN	2.7%	OTHERS	0.5%
Purpose	Percentage																								
MEDICAL	22.8%																								
DEBTCONSOLIDATION	21.0%																								
PERSONAL	15.4%																								
EDUCATION	15.6%																								
HOMEIMPROVEMENT	13.2%																								
VENTURE	11.9%																								
Ownership Type	Percentage																								
RENT	73.0%																								
MORTGAGE	23.8%																								
OWN	2.7%																								
OTHERS	0.5%																								
<p>Insight: Close to half (<u>43.8%</u>) of the borrowers who defaulted loan repayment took up medical loan or consolidated their unsecured loan into one account.</p>	<p>Insight: Most of the borrowers who defaulted loan repayment stayed in rental houses.</p>																								

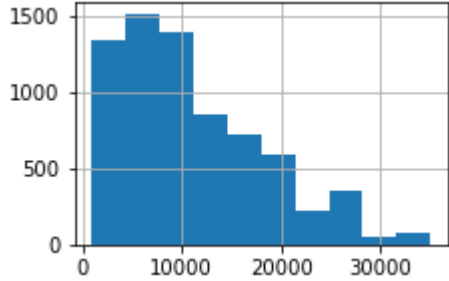
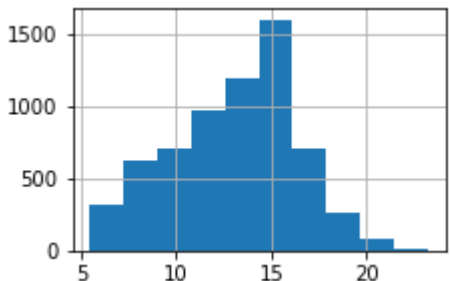
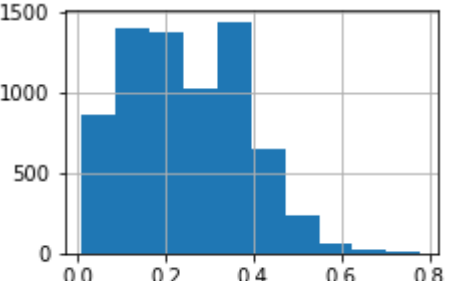
Categorical Data Review

History of Defaulting Repayment	Loan Grade																						
 <table border="1"><thead><tr><th>Category</th><th>Percentage</th></tr></thead><tbody><tr><td>NO</td><td>69.4%</td></tr><tr><td>YES</td><td>30.6%</td></tr></tbody></table>	Category	Percentage	NO	69.4%	YES	30.6%	 <table border="1"><thead><tr><th>Grade</th><th>Percentage</th></tr></thead><tbody><tr><td>A</td><td>15.0%</td></tr><tr><td>B</td><td>23.9%</td></tr><tr><td>C</td><td>18.8%</td></tr><tr><td>D</td><td>30.2%</td></tr><tr><td>E</td><td>8.8%</td></tr><tr><td>F</td><td>0.2%</td></tr><tr><td>GF</td><td>0.1%</td></tr></tbody></table>	Grade	Percentage	A	15.0%	B	23.9%	C	18.8%	D	30.2%	E	8.8%	F	0.2%	GF	0.1%
Category	Percentage																						
NO	69.4%																						
YES	30.6%																						
Grade	Percentage																						
A	15.0%																						
B	23.9%																						
C	18.8%																						
D	30.2%																						
E	8.8%																						
F	0.2%																						
GF	0.1%																						
<p>Insight: Close to <u>70%</u> of the borrowers who defaulted loan repayment did not have historical default history. This could be one of the key reasons contributing to the good to moderate loan grade assigned (72.9% holding load grade of B, C, and D).</p>																							

Numerical Data Review

Age	Employment Length	Annual Income	Credit History Length
 <p>A histogram showing the distribution of age. The x-axis ranges from 20 to 70 with major ticks every 10 units. The y-axis ranges from 0 to 2500 with major ticks every 500 units. The distribution is right-skewed, with the highest frequency (approx. 2800) in the 20-25 age bin, followed by a sharp decline.</p>	 <p>A histogram showing the distribution of employment length. The x-axis ranges from 0 to 125 with major ticks every 25 units. The y-axis ranges from 0 to 6000 with major ticks every 2000 units. The distribution is highly right-skewed, with a peak frequency of over 6000 in the 0-5 bin.</p>	 <p>A histogram showing the distribution of annual income. The x-axis ranges from 0 to 600,000 with major ticks every 200,000 units. The y-axis ranges from 0 to 6000 with major ticks every 2000 units. The distribution is highly right-skewed, with a peak frequency of over 6000 in the 0-\$50,000 bin.</p>	 <p>A histogram showing the distribution of credit history length. The x-axis ranges from 0 to 30 with major ticks every 10 units. The y-axis ranges from 0 to 4000 with major ticks every 1000 units. The distribution is highly right-skewed, with a peak frequency of over 4000 in the 0-5 bin.</p>
<p>Insight: Most of the borrowers who defaulted loan repayment are between 20 and 30 years old.</p>	<p>Insight: Most of the borrowers who defaulted loan repayment are in the workforce for less than 12.5 years.</p>	<p>Insight: Most of the borrowers who defaulted loan repayment are drawing an annual income of between \$4,000 and \$6666.67.</p>	<p>Insight: Most of the borrowers who defaulted loan repayment and history of defaulting payment defaulted loan for between 2 and 3.33 years.</p>

Numerical Data Review

Loan Amount	Interest Rate	Percent Income
 <p>A histogram showing the distribution of loan amounts. The x-axis ranges from 0 to 30,000 with major ticks at 0, 10,000, 20,000, and 30,000. The y-axis ranges from 0 to 1,500 with major ticks at 0, 500, 1,000, and 1,500. The distribution is right-skewed, with the highest frequency (around 1,500) occurring between 5,000 and 10,000, and the frequency decreasing as the loan amount increases.</p>	 <p>A histogram showing the distribution of interest rates. The x-axis ranges from 5 to 20 with major ticks at 5, 10, 15, and 20. The y-axis ranges from 0 to 1,500 with major ticks at 0, 500, 1,000, and 1,500. The distribution is roughly bell-shaped and centered around 15%, with the highest frequency (around 1,500) occurring between 15% and 17.5%.</p>	 <p>A histogram showing the distribution of percent income. The x-axis ranges from 0.0 to 0.8 with major ticks at 0.0, 0.2, 0.4, 0.6, and 0.8. The y-axis ranges from 0 to 1,500 with major ticks at 0, 500, 1,000, and 1,500. The distribution is multimodal, with peaks around 0.1 (frequency ~1,400), 0.2 (frequency ~1,400), and 0.35 (frequency ~1,400), and a significant drop in frequency between 0.2 and 0.35.</p>
<p>Insight: Most of the borrowers who defaulted loan repayment took up small loan, slightly exceeding \$10,000.</p>	<p>Trend: The number of borrowers who defaulted loan repayment increases as interest rate increases (peaks at about 15%).</p>	<p>Insight: The proportion of income taken up by the loan of most borrowers who defaulted loan repayment stood at 10%, 20%, and 40%.</p>

III. Understanding the Correlation Between Different Attributes



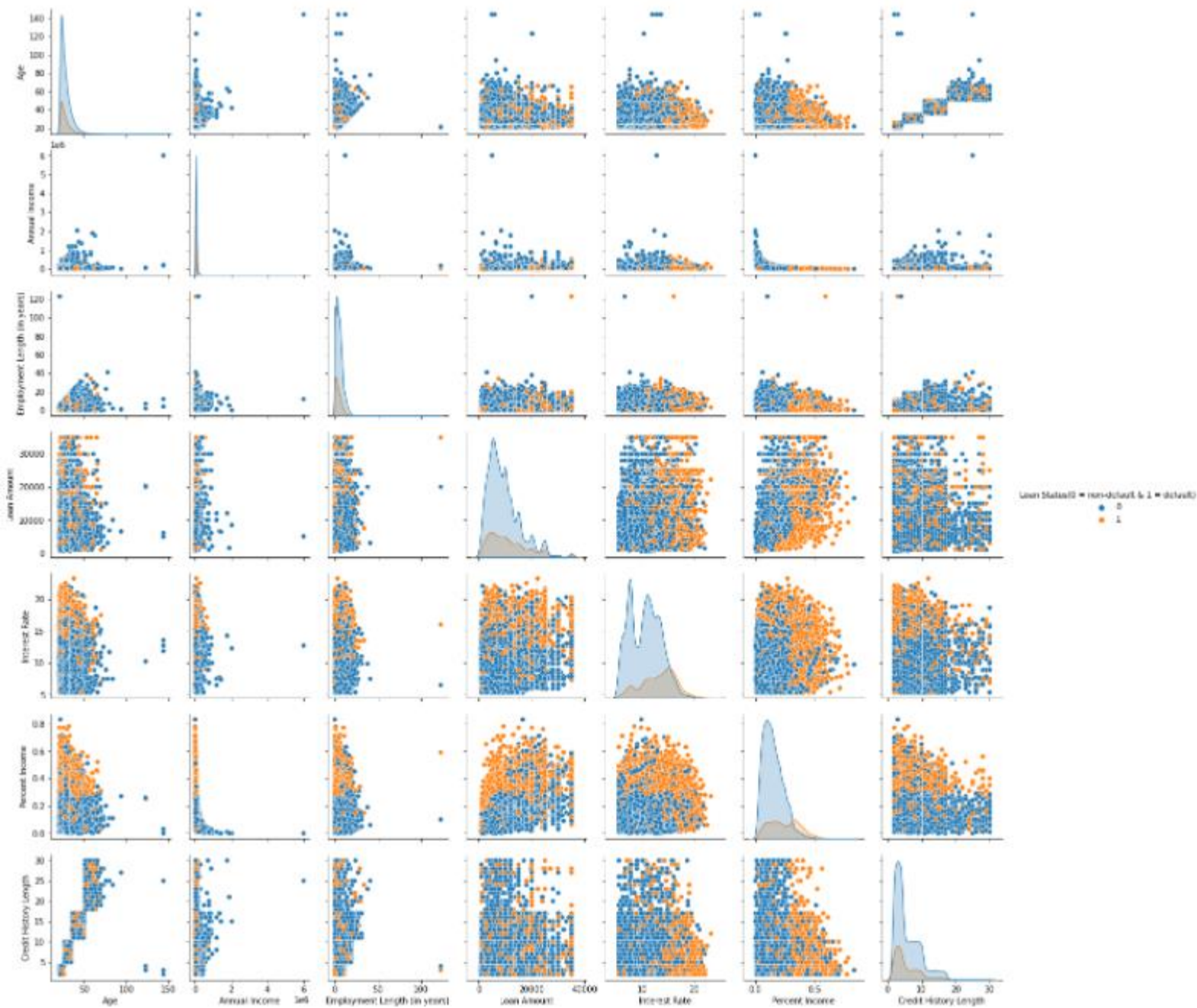
Heatmap

Attributes with Close Correlation

1. Age & Credit History Length: 0.86
2. Percent Income and Loan Amount: 0.57
3. Percent Income and Loan Status: 0.38
4. Interest Rate and Loan Status: 0.34
5. Loan Amount and Annual Income: 0.27

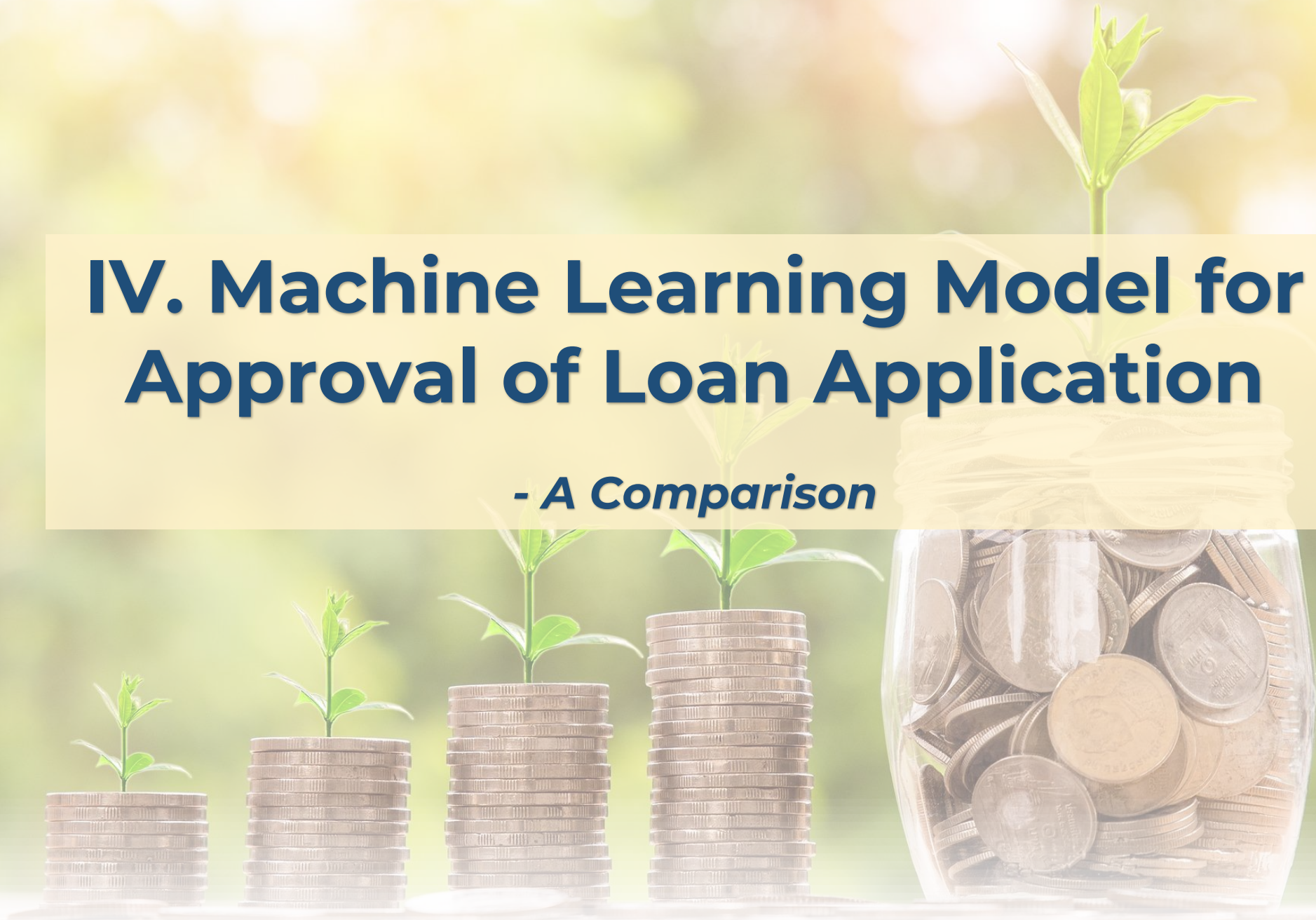


Summary



IV. Machine Learning Model for Approval of Loan Application

- A Comparison

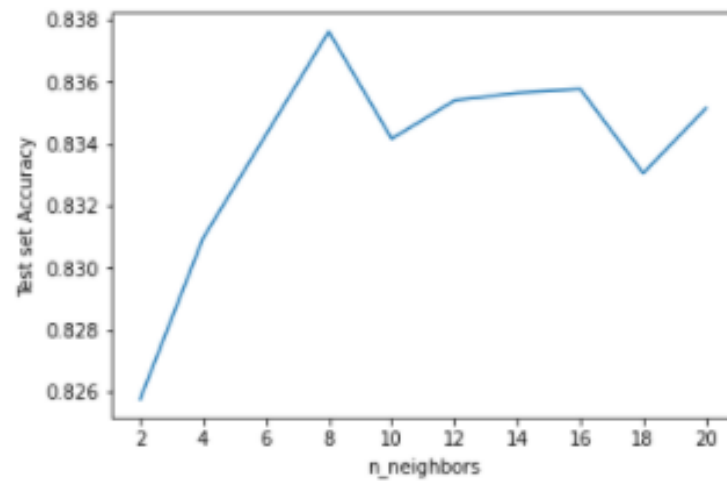


Logistic Regression

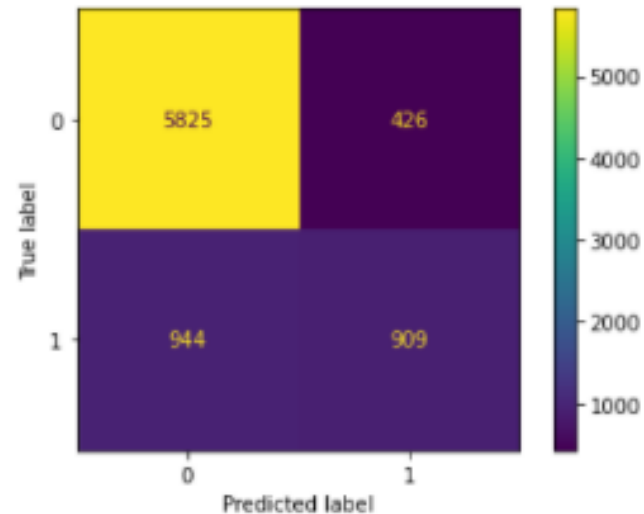
Accuracy Score: 0.817

K Nearest Neighbors

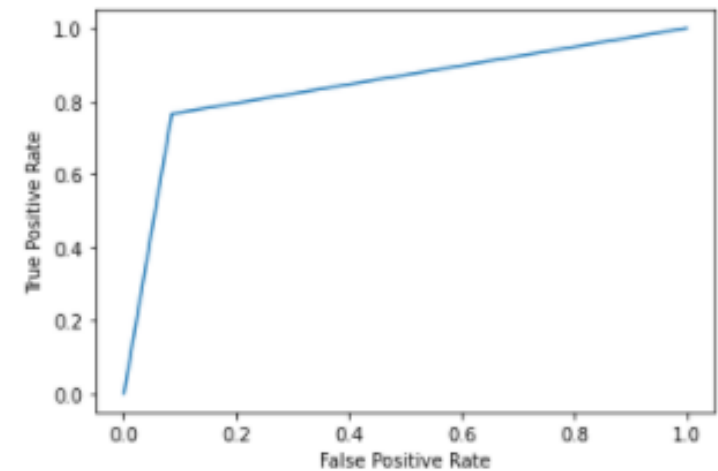
Accuracy Score: 0.831



Confusion Matrix



AUC Score: 0.711

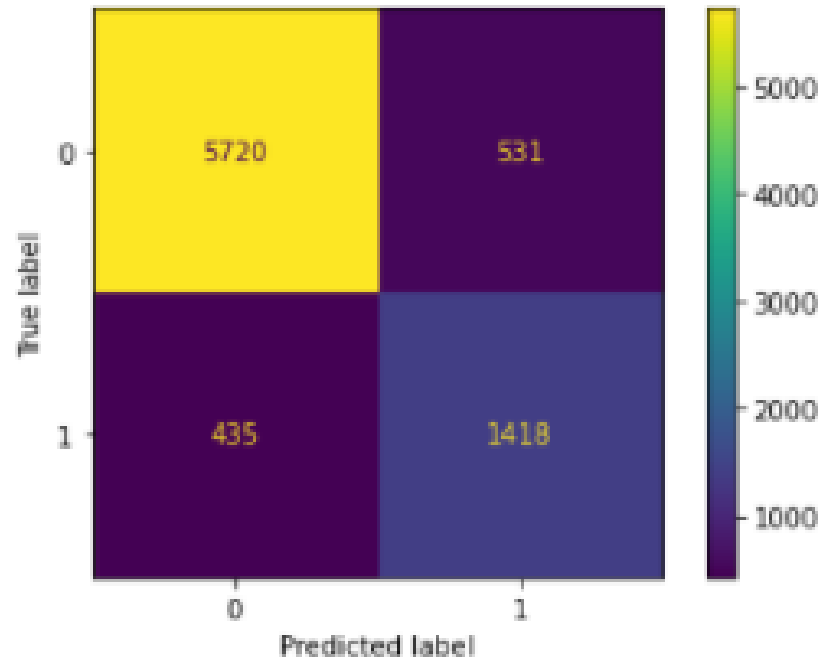




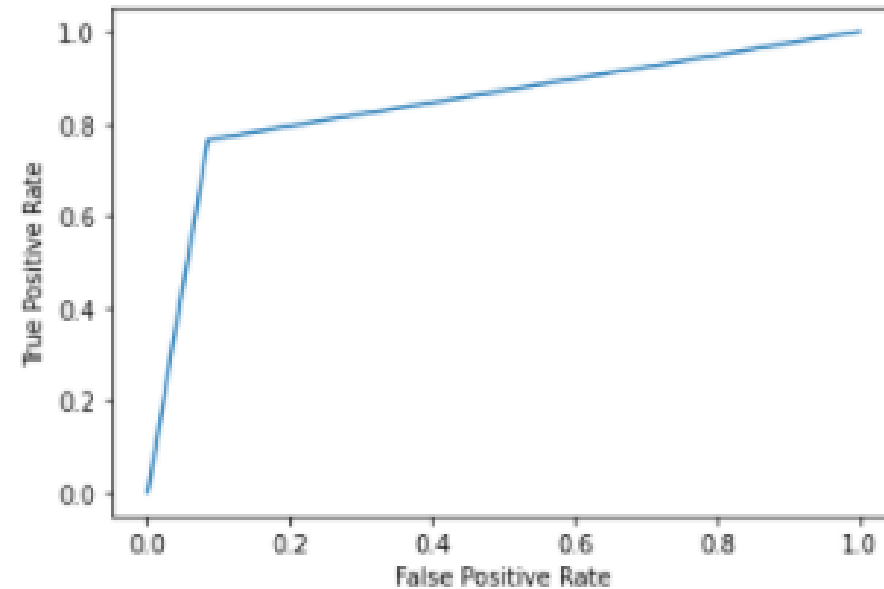
Decision Tree Classifier

Accuracy Score: 0.881

Confusion Matrix



AUC Score: 0.84



Conclusion : Decision Tree Classifier provides the best accuracy score amongst all the machine learning models.



V. Selecting the Best Parameters for Decision Tree Classifier Model

Best Estimators

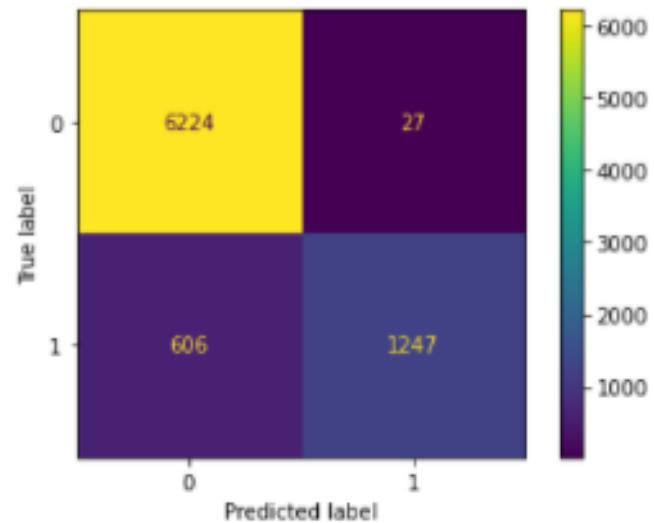
Best Predictors:

- (a) max_depth=6;
- (b) min_samples_leaf=2;
- (c) min_samples_split=30; and

Accuracy Score: 0.921

AUC Score: 0.834

Confusion Matrix



Suggestion

- Consider expanding the number of predictors to improve the accuracy of prediction. Additional information that may be collected include: (a) total expenditure (b) total savings (c) other outstanding loans (type) (d) total quantum of outstanding loan (c) total investments
- Continue to feed new data to the ML model to ensure that it remains up to date and relevant against the evolving operating environment
- Implement computational expensive machine learning classification framework such as XGBoostClassifier or Neural Network to produce better prediction outcome.

