LOAN APPROVAL LOGISTIC REGRESSION MODEL

Loading Dataset

```
Dataset Loaded Successfully
Shape: (4269, 13)

Columns: ['loan_id', ' no_of_dependents', ' education', ' self_employed', ' income_annum', ' loan_amount', ' loan_term', ' cibil_score', ' residential_
assets_value', ' commercial_assets_value', ' luxury_assets_value', ' bank_asset_value', ' loan_status']

Missing Values:
 loan_id                      0
 no_of_dependents            0
 education                   0
 self_employed               0
 income_annum                0
 loan_amount                 0
 loan_term                   0
 cibil_score                 0
 residential_assets_value    0
 commercial_assets_value     0
 luxury_assets_value         0
 bank_asset_value            0
 loan_status                 0
dtype: int64
```

- The dataset given was loaded to know the shape of the data, the columns, and find if there are any null values in the dataset.
- The result indicated 4269 rows with 13 columns, There were no missing values in the dataset.

Data Types

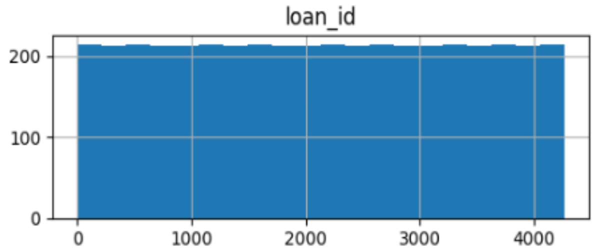
```
Data Types:

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 4269 entries, 0 to 4268
Data columns (total 13 columns):
 #   Column                     Non-Null Count  Dtype
---  ------                     --------------  -----
 0   loan_id                    4269 non-null   int64
 1    no_of_dependents          4269 non-null   int64
 2    education                 4269 non-null   object
 3    self_employed             4269 non-null   object
 4    income_annum              4269 non-null   int64
 5    loan_amount               4269 non-null   int64
 6    loan_term                 4269 non-null   int64
 7    cibil_score               4269 non-null   int64
 8    residential_assets_value  4269 non-null   int64
 9    commercial_assets_value   4269 non-null   int64
 10   luxury_assets_value       4269 non-null   int64
 11   bank_asset_value          4269 non-null   int64
 12   loan_status               4269 non-null   object
dtypes: int64(10), object(3)
memory usage: 433.7+ KB
None
```

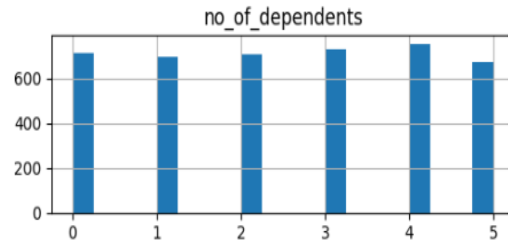
- The data types were checked to determine the numeric variables and the needed categorical variables for the modeling.

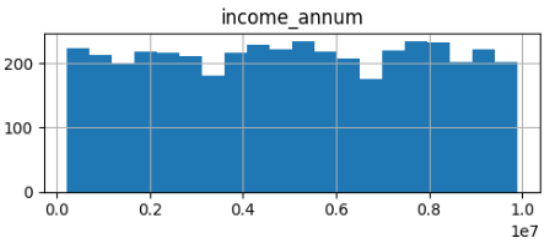
```
Sample Records:
   loan_id  no_of_dependents      education  self_employed   income_annum  \
0        1                 2       Graduate             No        9600000
1        2                 0   Not Graduate            Yes        4100000
2        3                 3       Graduate             No        9100000
3        4                 3       Graduate             No        8200000
4        5                 5   Not Graduate            Yes        9800000

    loan_amount   loan_term   cibil_score   residential_assets_value  \
0      29900000          12           778                    2400000
1      12200000           8           417                    2700000
2      29700000          20           506                    7100000
3      30700000           8           467                   18200000
4      24200000          20           382                   12400000

    commercial_assets_value   luxury_assets_value   bank_asset_value  \
0                  17600000              22700000            8000000
1                   2200000               8800000            3300000
2                   4500000              33300000           12800000
3                   3300000              23300000            7900000
4                   8200000              29400000            5000000

   loan_status
0     Approved
1     Rejected
2     Rejected
3     Rejected
4     Rejected
```

- The picture shows the details of the dataset, of previous loan application status of the applied customers.

Distribution of numeric variables

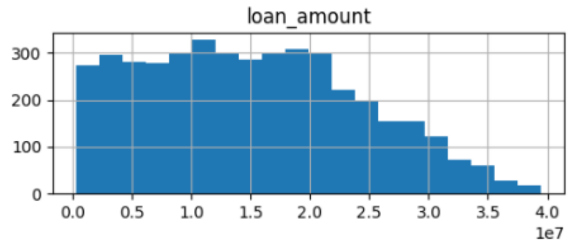


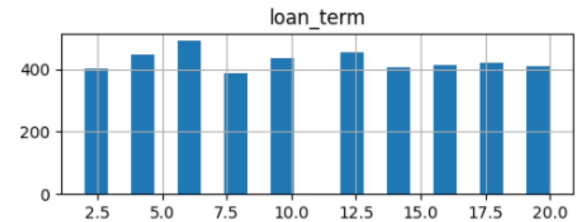


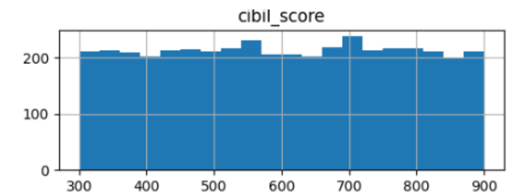


- Visual Exploratory Data Analysis was carried out on the data to know what influences the approval and rejection of previous applications. The following were deduced:

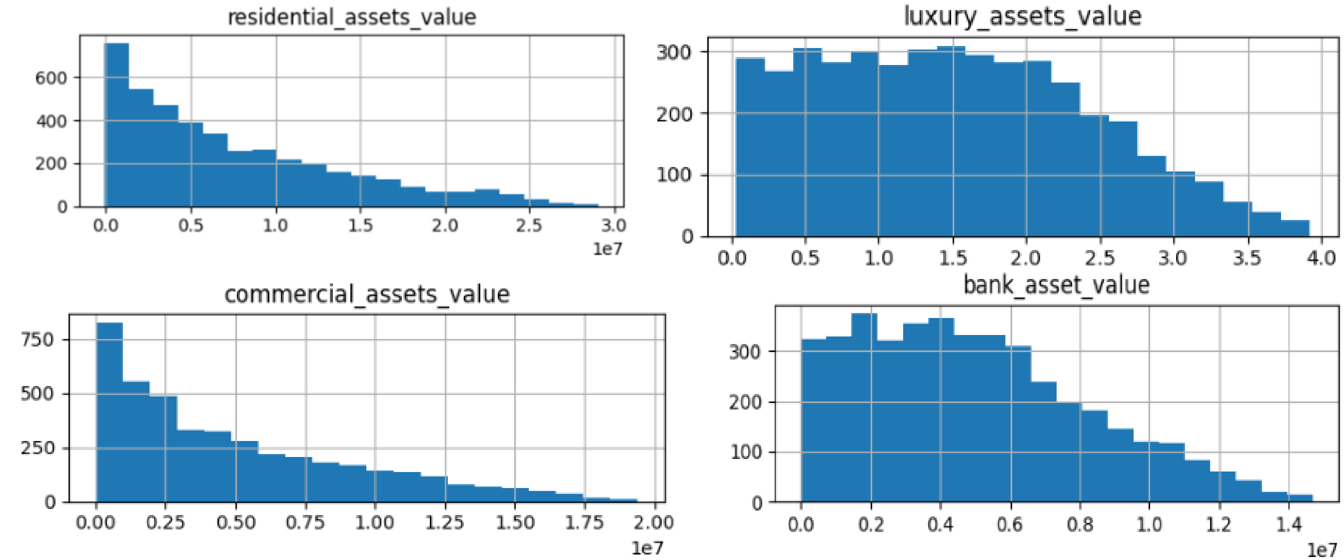
  - The distribution for loan_id is a uniform, flat histogram. This is expected, as loan_id is a unique identifier assigned to each loan and does not have a meaniful numerical distribution.
  - No_of_dependents: This variable also shows a relatively uniform distribution. The frequency is similar for each category of dependents, from 0 to 5, suggesting that the number of dependents is evenly distributed across the loans in the dataset.
  - Income_annum: The distribution of income_annum appears to be slightly right-skewed, meaning that most of the loans are associated with lower annual incomes, with a long tail of observations for higher incomes.





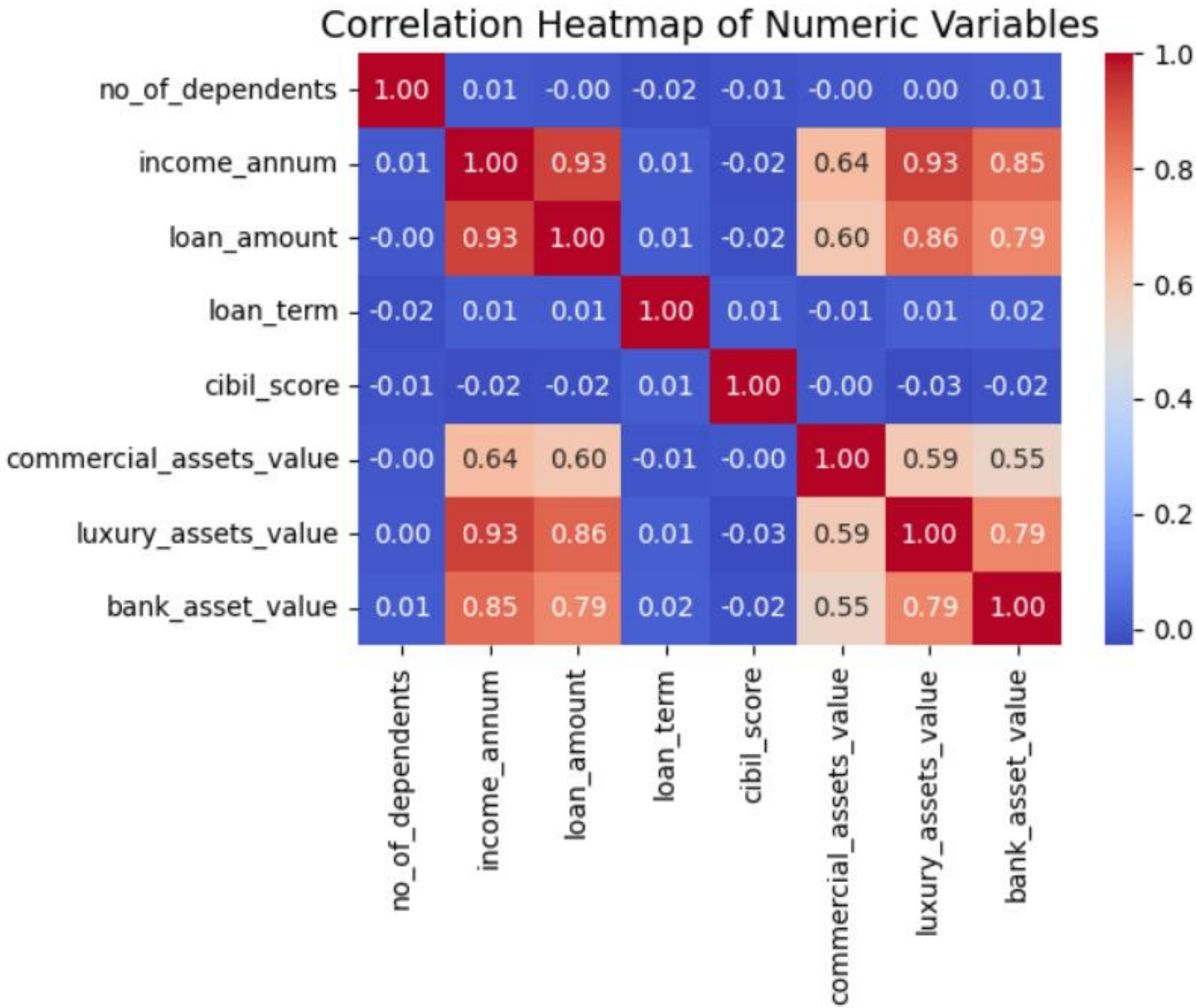


- Loan_amount: Similar to income_annum, the loan_amount distribution is right-skewed. The majority of loan amounts are concentrated at the lower end of the spectrum, with fewer loans for very large amounts.
- Loan_term: The distribution for loan_term is fairly uniform, suggesting that loan terms (in years or months) are distributed quite evenly across the different loan categories shown.
- Cibil_score: This distribution appears to be relatively uniform, with most CIBIL scores falling between 300 and 900. A CIBIL score is a credit rating that reflects a borrower's creditworthiness. The even spread indicates a wide range of credit scores in the dataset.



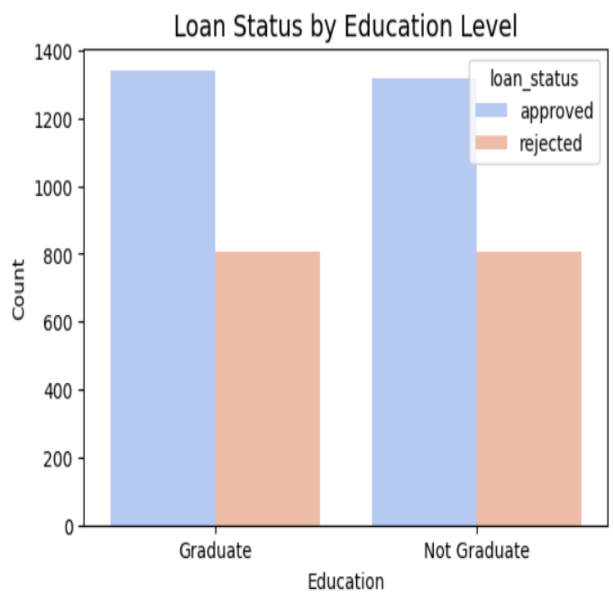
Residential_assets_value, Commercial_assets_value, Luxury_assets_value, and Bank_asset_value:

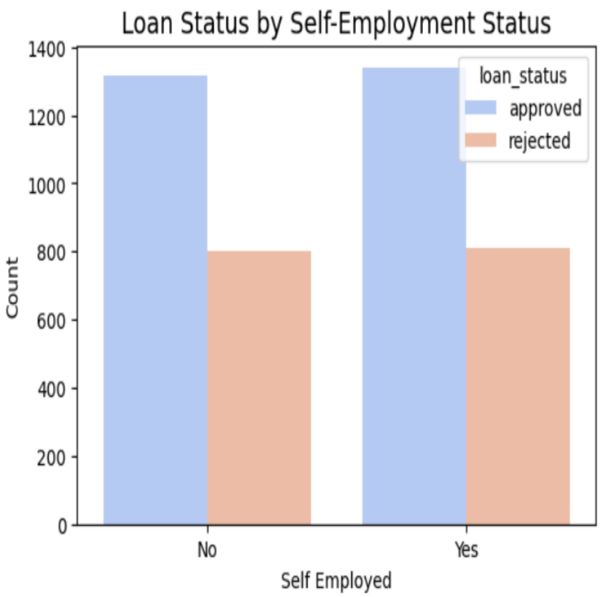
All these asset value variables show a heavily right-skewed distribution. The vast majority of individuals in the dataset have asset values at the lower end of the range, with very few individuals possessing extremely high asset values.


Correlation Heatmap of Numeric Variables
no_of_dependents
income_annum
loan_amount
loan_term
cibil_score
commercial_assets_value
luxury_assets_value
bank_asset_value
1.00 0.01 -0.00 -0.02 -0.01 -0.00 0.00 0.01
0.01 1.00 0.93 0.01 -0.02 0.64 0.93 0.85
-0.00 0.93 1.00 0.01 -0.02 0.60 0.86 0.79
-0.02 0.01 0.01 1.00 0.01 -0.01 0.01 0.02
-0.01 -0.02 -0.02 0.01 1.00 -0.00 -0.03 -0.02
-0.00 0.64 0.60 -0.01 -0.00 1.00 0.59 0.55
0.00 0.93 0.86 0.01 -0.03 0.59 1.00 0.79
0.01 0.85 0.79 0.02 -0.02 0.55 0.79 1.00
1.0
0.8
0.6
0.4
0.2
0.0

Interpretation of the specific heatmap:

- By examining the image, several strong correlations can be identified:
  - Income_annum and loan_amount have a strong positive correlation of 0.93. This suggests that as a person's annual income increases, the loan amount they take out also tends to increase.
  - Loan_amount and income_annum also have a strong positive correlation, shown by the value of 0.93. This is the same relationship, as the matrix is symmetrical.
  - Income_annum and luxury_assets_value have a strong positive correlation of 0.93, indicating a relationship between higher income and the value of luxury assets.
  - Income_annum and bank_asset_value have a strong positive correlation of 0.85.
  - Loan_amount and luxury_assets_value have a strong positive correlation of 0.86.
  - Loan_amount and bank_asset_value have a strong positive correlation of 0.79.





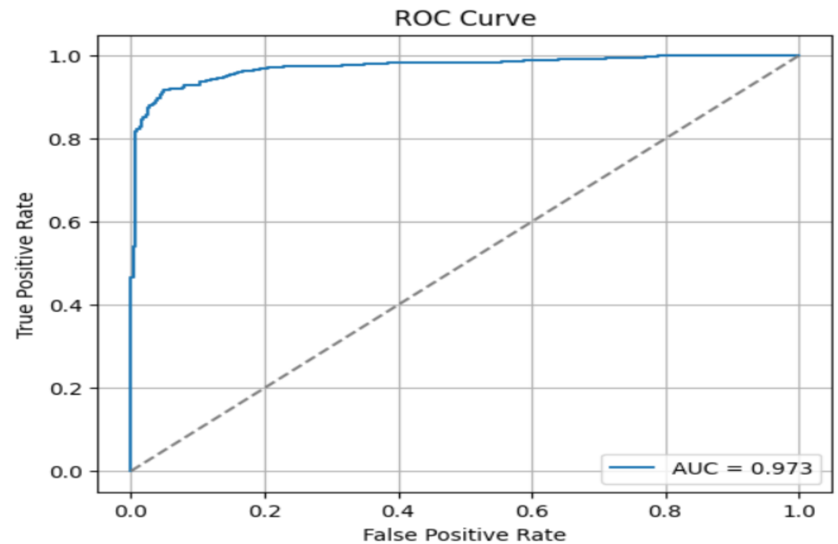
- Loan Status by Educational Level and Self Employment Status reveal that the level of education or the employment status of the applicant does not influence the approval or rejection of a loan.

```
===== Loan Status Distribution =====
loan_status
Approved    62.22
Rejected    37.78
Name: proportion, dtype: float64

===== Loan Status by Education =====
loan_status   Approved  Rejected
education
Graduate         62.45     37.55
Not Graduate     61.98     38.02

===== Loan Status by Self-Employment =====
loan_status    Approved  Rejected
self_employed
No                62.20     37.80
Yes               62.23     37.77
```

- The picture shows percentage loan status distribution, loan status by education and loan status by self employment.

LOGISTIC REGRESSION MODEL

```
Numeric columns: ['no_of_dependents', 'income_annum', 'loan_amount', 'loan_term', 'cibil_score', 'commercial_assets_value', 'luxury_assets_value', 'ban
k_asset_value']
Categorical columns: ['education', 'self_employed']

Target distribution:
 loan_status_binary
1    0.622
0    0.378
Name: proportion, dtype: float64

CV Accuracy: 0.9180 ± 0.0162
CV ROC AUC: 0.9664 ± 0.0082

===== Test Performance =====
Accuracy: 0.9239
Precision: 0.9551
Recall: 0.9209
F1 Score: 0.9377
ROC AUC: 0.9734
```



ROC Curve (AUC = 0.973)
Interpretation:

ROC Curve (Receiver Operating Characteristic) measures the model's ability to distinguish between two classes (Approved vs. Rejected).

X-axis: False Positive Rate shows how often the model incorrectly predicts "Approved" when it should be "Rejected".

Y-axis: True Positive Rate shows how often it correctly predicts "Approved".

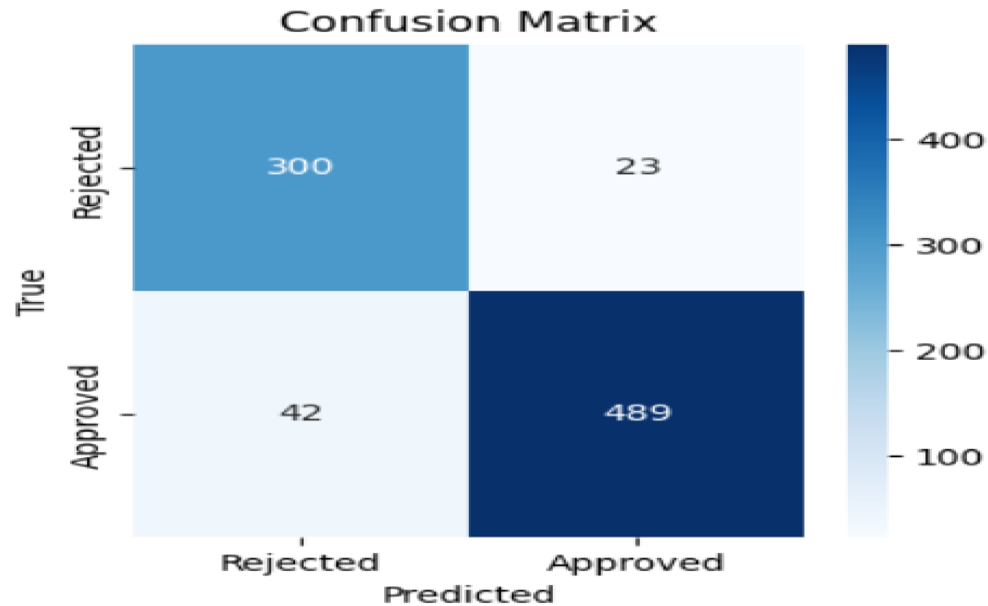
Key insight:

The curve hugs the top-left corner, which means the model performs extremely well at separating classes.

AUC (Area Under Curve) = 0.973, or 97.3% accuracy in discrimination, it indicates the model almost perfectly differentiates approved vs. rejected cases.

AUC STANDARD:
AUC = 0.5 → random guessing

AUC = 1.0 → perfect model



Confusion Matrix

Interpretation:

True Positives (489): Approved applications correctly predicted as approved.

True Negatives (300): Rejected applications correctly predicted as rejected.

False Positives (23): Rejected applications incorrectly predicted as approved.

False Negatives (42): Approved applications incorrectly predicted as rejected.

Classification Report:

	precision	recall	f1-score	support
0	0.8772	0.9288	0.9023	323
1	0.9551	0.9209	0.9377	531
accuracy			0.9239	854
macro avg	0.9161	0.9248	0.9200	854
weighted avg	0.9256	0.9239	0.9243	854

Metrics:

Accuracy: (300 + 489) / (300 + 23 + 42 + 489) = 0.93 → 93%

Precision (Approved): 489 / (489 + 23) = 95.5%

Recall (Approved): 489 / (489 + 42) = 92.1%

F1-score: ≈ 93.8%

```
===== Top 10 Most Influential Features =====
               feature  coefficient  abs_coef
0           cibil_score     4.228211  4.228211
1          income_annum    -1.480854  1.480854
2           loan_amount     1.238355  1.238355
3             loan_term    -0.783198  0.783198
4    education_Graduate     0.390096  0.390096
5     self_employed_Yes     0.336227  0.336227
6      self_employed_No     0.319772  0.319772
7  education_Not Graduate   0.265903  0.265903
8      bank_asset_value     0.174677  0.174677
9   luxury_assets_value     0.128979  0.128979
```

- Above are features that influences loan rejection or approval.

CONCLUSION

- Overall Summary

  The model is highly reliable (AUC = 0.973, Accuracy ≈ 93%).

  It performs slightly better at rejecting false approvals (low false positive rate) than catching every approval (few false negatives).

  In business terms: it's safer - rarely approves something that should have been rejected, though it misses a few valid approvals.

THANK YOU