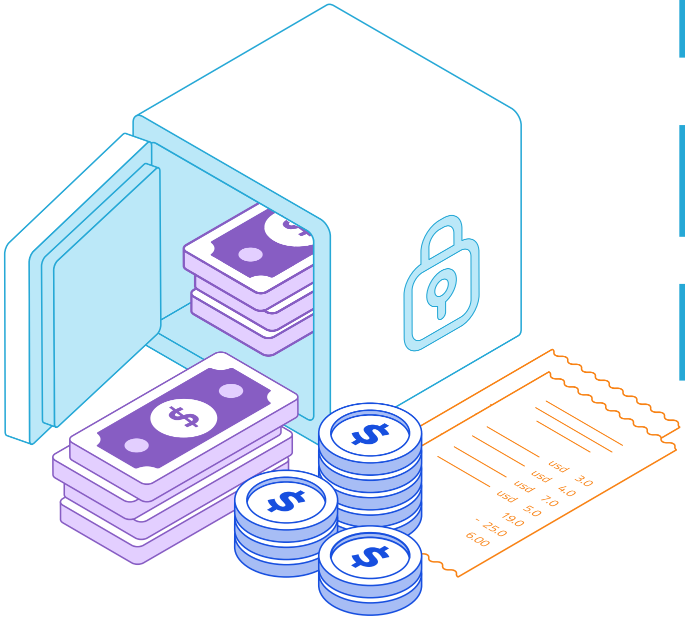


Loan Approval Prediction

Project 4

Group 4
Jenipher Flores | Andrew Hawthorne
Ryan Blais | Elizabeth Lawal | Fidel Carrillo

Table of Contents

Project Overview

01

Objective

Using loan approval data, we will develop a supervised learning model to predict whether or not a loan application will be approved based on past approvals/rejections.

Our Dataset

Loan-Approval-Prediction-Dataset (Kaggle)

"The loan approval dataset is a collection of financial records and associated information used to determine the eligibility of individuals or organizations for obtaining loans from a lending institution. It includes various factors such as cibil score, income, employment status, loan term, loan amount, assets value, and loan status. This dataset is commonly used in machine learning and data analysis to develop models and algorithms that predict the likelihood of loan approval based on the given features."

Exploratory Analysis

02

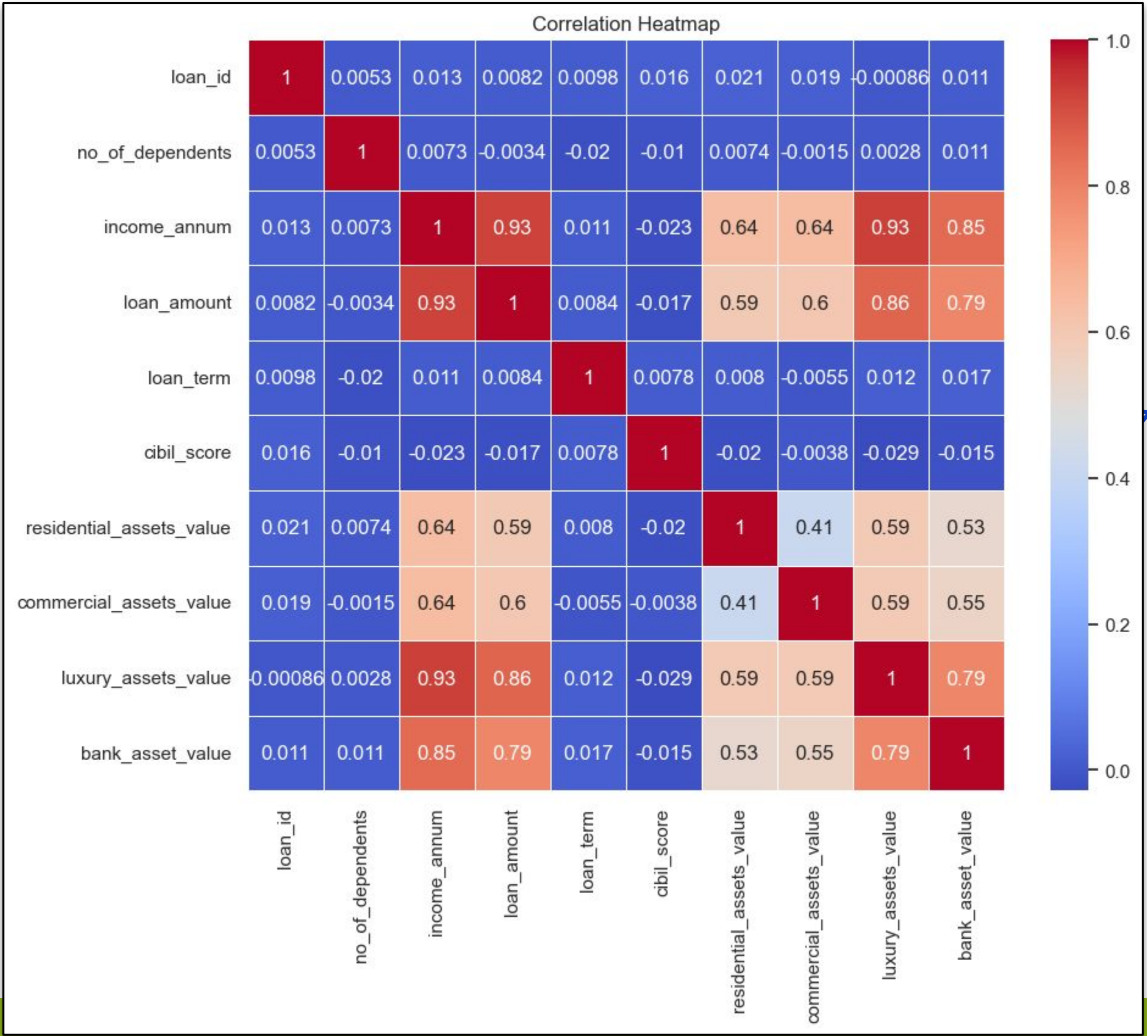
What does the data look like?

	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
0	1	2	Graduate	No	9600000	29900000	12	778	2400000	17600000	22700000	8000000	Approved
1	2	0	Not Graduate	Yes	4100000	12200000	8	417	2700000	2200000	8800000	3300000	Rejected
2	3	3	Graduate	No	9100000	29700000	20	506	7100000	4500000	33300000	12800000	Rejected
3	4	3	Graduate	No	8200000	30700000	8	467	18200000	3300000	23300000	7900000	Rejected
4	5	5	Not Graduate	Yes	9800000	24200000	20	382	12400000	8200000	29400000	5000000	Rejected

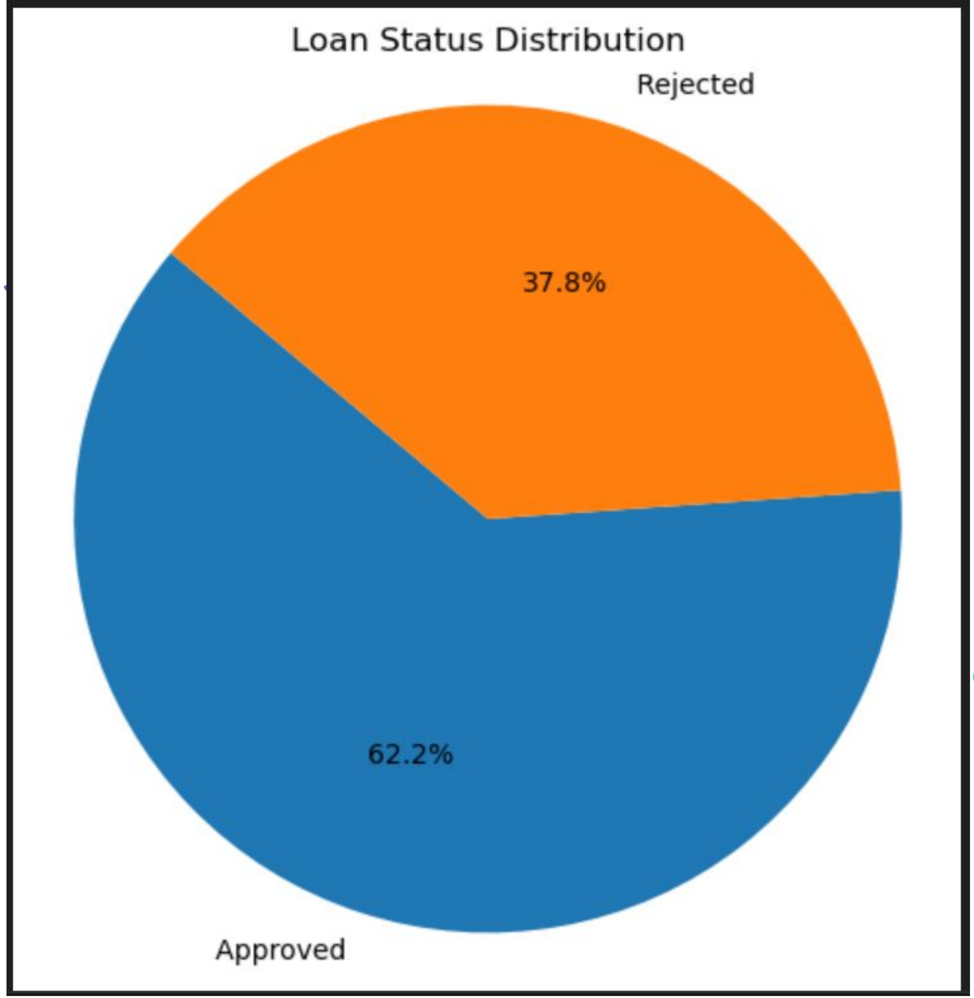
```
data.describe()
```

	loan_id	no_of_dependents	income_annum	loan_amount	loan_term	cibil_score	residential_assets_value	commercial_assets_value	luxury_assets_value	bank_asset_value
count	4269.000000	4269.000000	4.269000e+03	4.269000e+03	4269.000000	4269.000000	4.269000e+03	4.269000e+03	4.269000e+03	4.269000e+03
mean	2135.000000	2.498712	5.059124e+06	1.513345e+07	10.900445	599.936051	7.472617e+06	4.973155e+06	1.512631e+07	4.976692e+06
std	1232.498479	1.695910	2.806840e+06	9.043363e+06	5.709187	172.430401	6.503637e+06	4.388966e+06	9.103754e+06	3.250185e+06
min	1.000000	0.000000	2.000000e+05	3.000000e+05	2.000000	300.000000	-1.000000e+05	0.000000e+00	3.000000e+05	0.000000e+00
25%	1068.000000	1.000000	2.700000e+06	7.700000e+06	6.000000	453.000000	2.200000e+06	1.300000e+06	7.500000e+06	2.300000e+06
50%	2135.000000	3.000000	5.100000e+06	1.450000e+07	10.000000	600.000000	5.600000e+06	3.700000e+06	1.460000e+07	4.600000e+06
75%	3202.000000	4.000000	7.500000e+06	2.150000e+07	16.000000	748.000000	1.130000e+07	7.600000e+06	2.170000e+07	7.100000e+06
max	4269.000000	5.000000	9.900000e+06	3.950000e+07	20.000000	900.000000	2.910000e+07	1.940000e+07	3.920000e+07	1.470000e+07

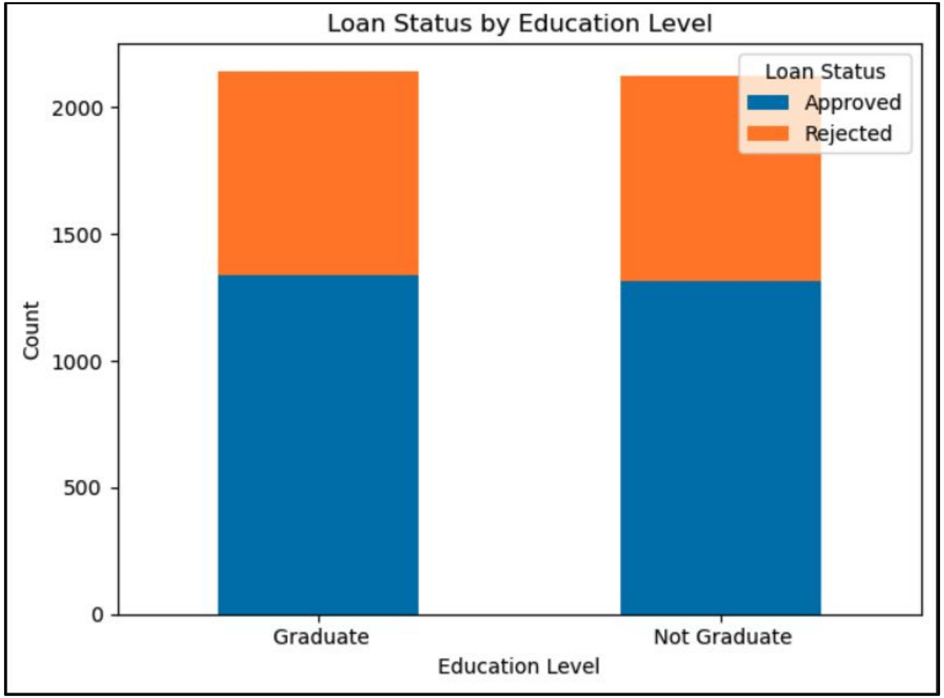
Correlation Matrix

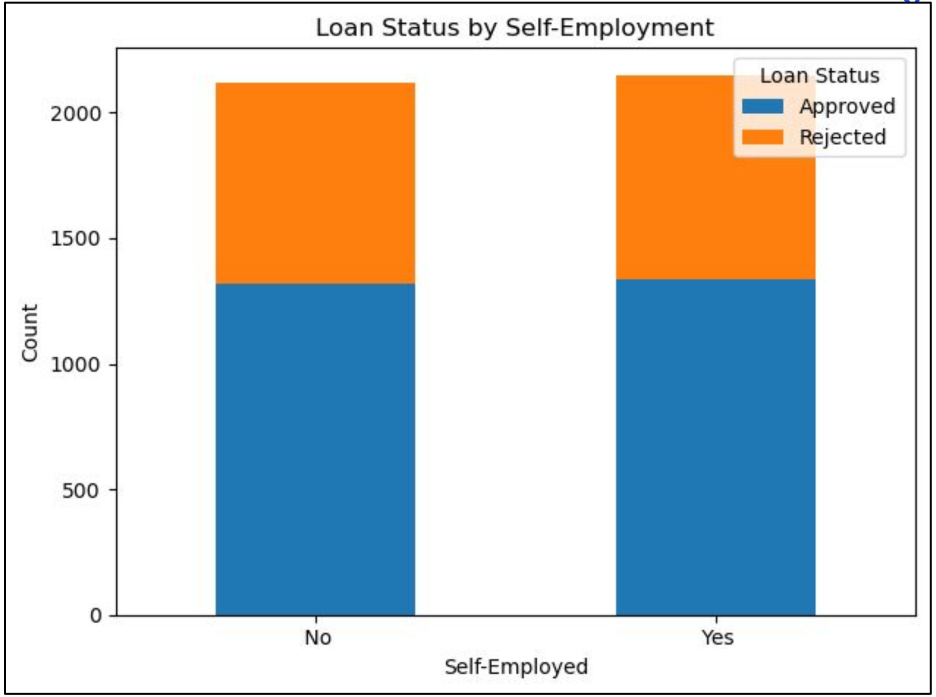


Loan Distribution

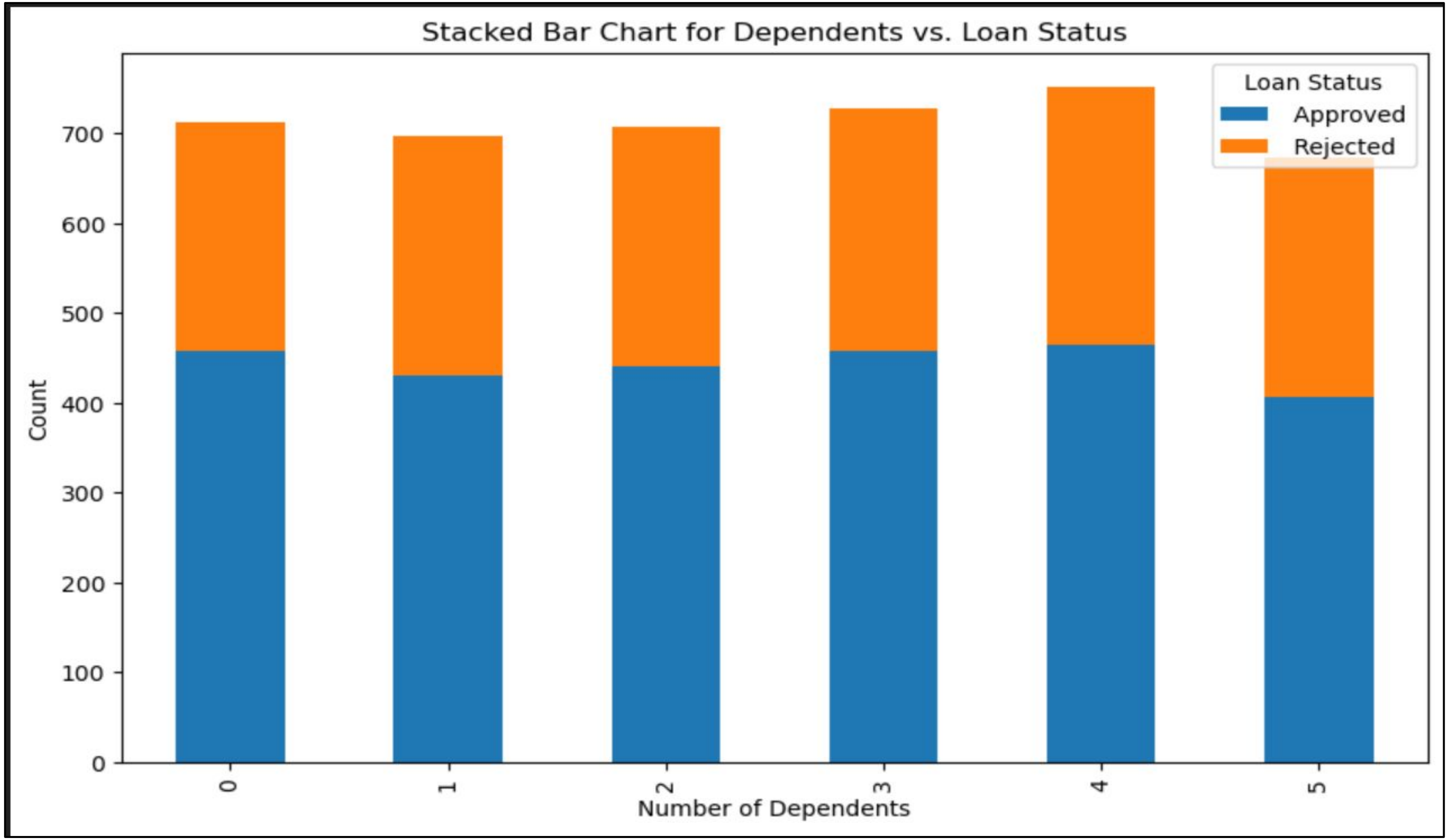


Education & Self-Employment

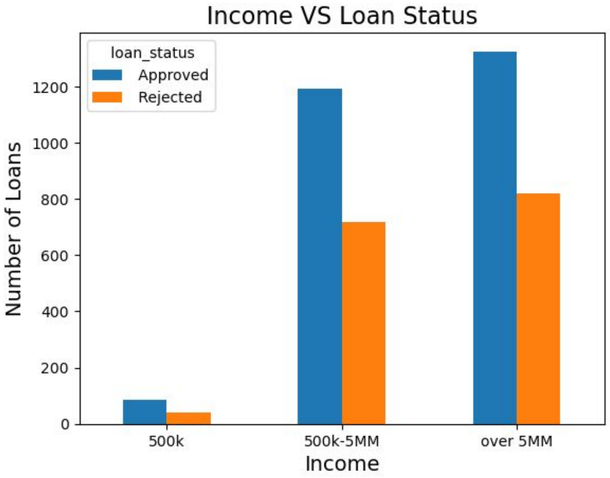


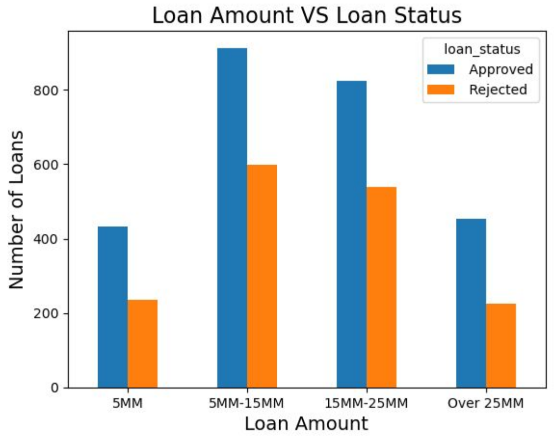


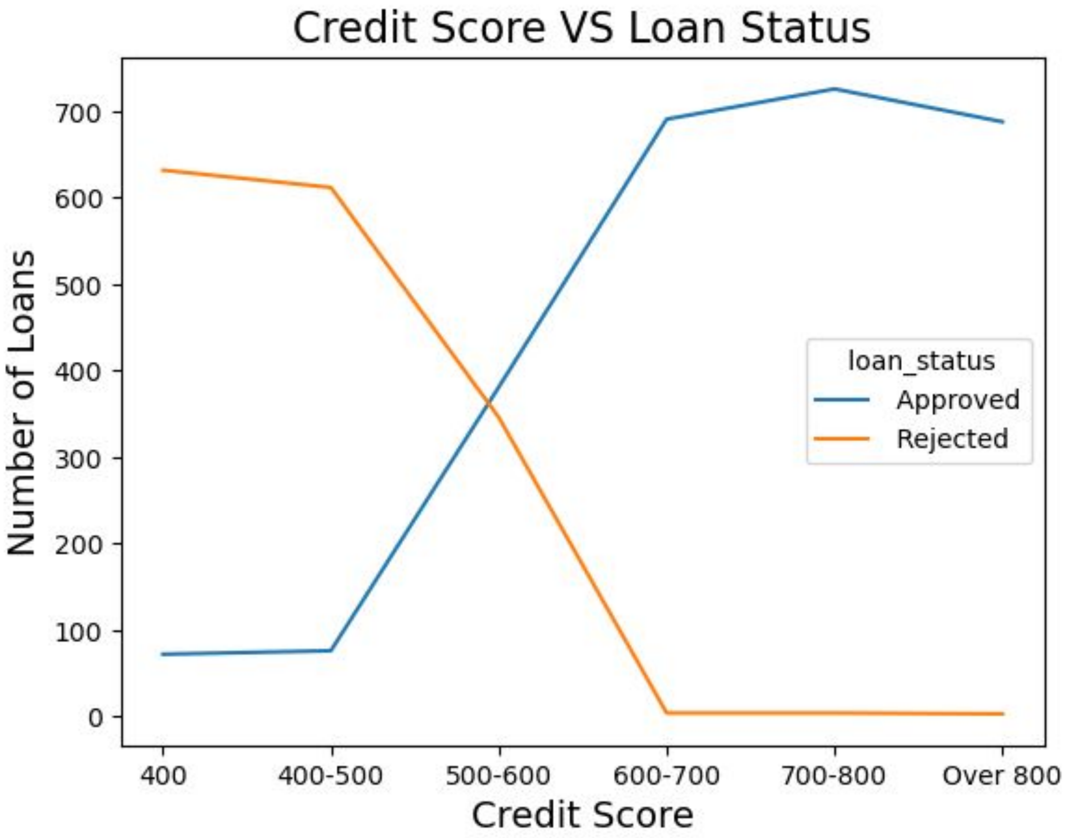
Number of Dependants by Loan Status



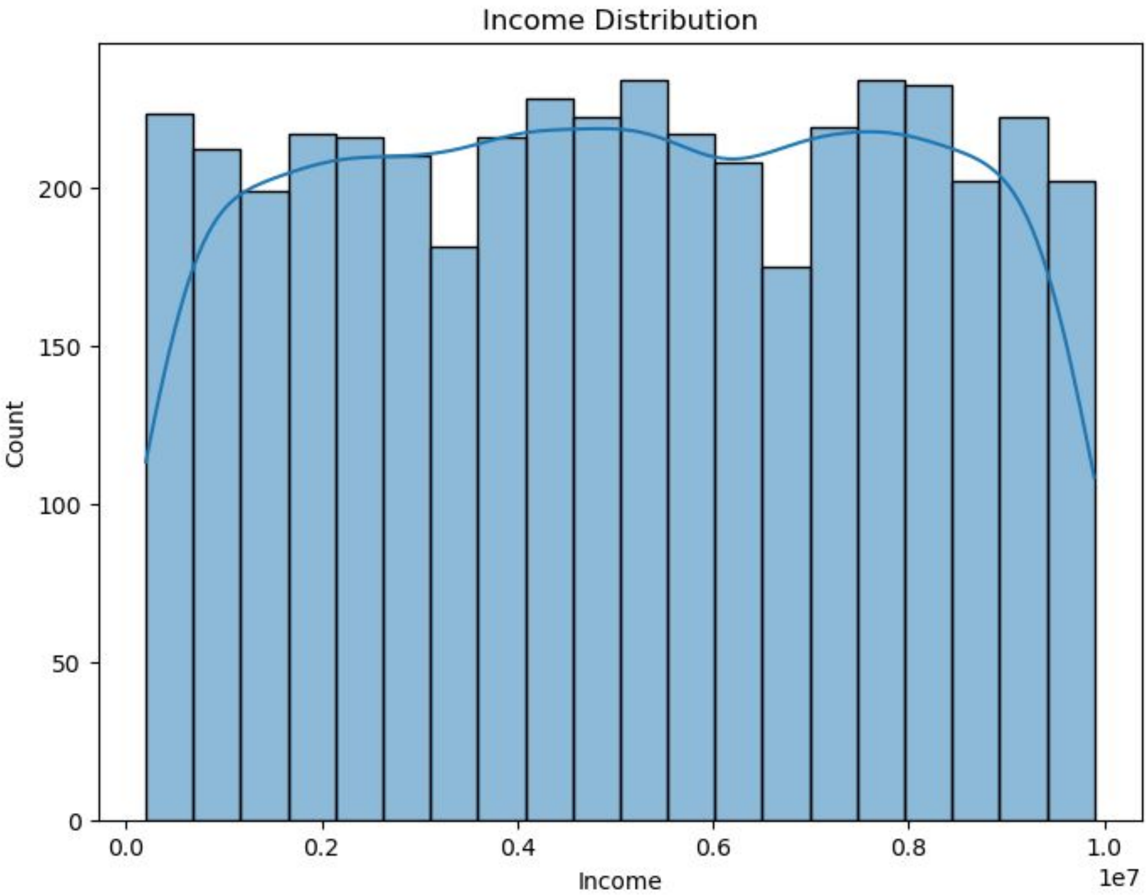
Additional Analysis by Loan Status



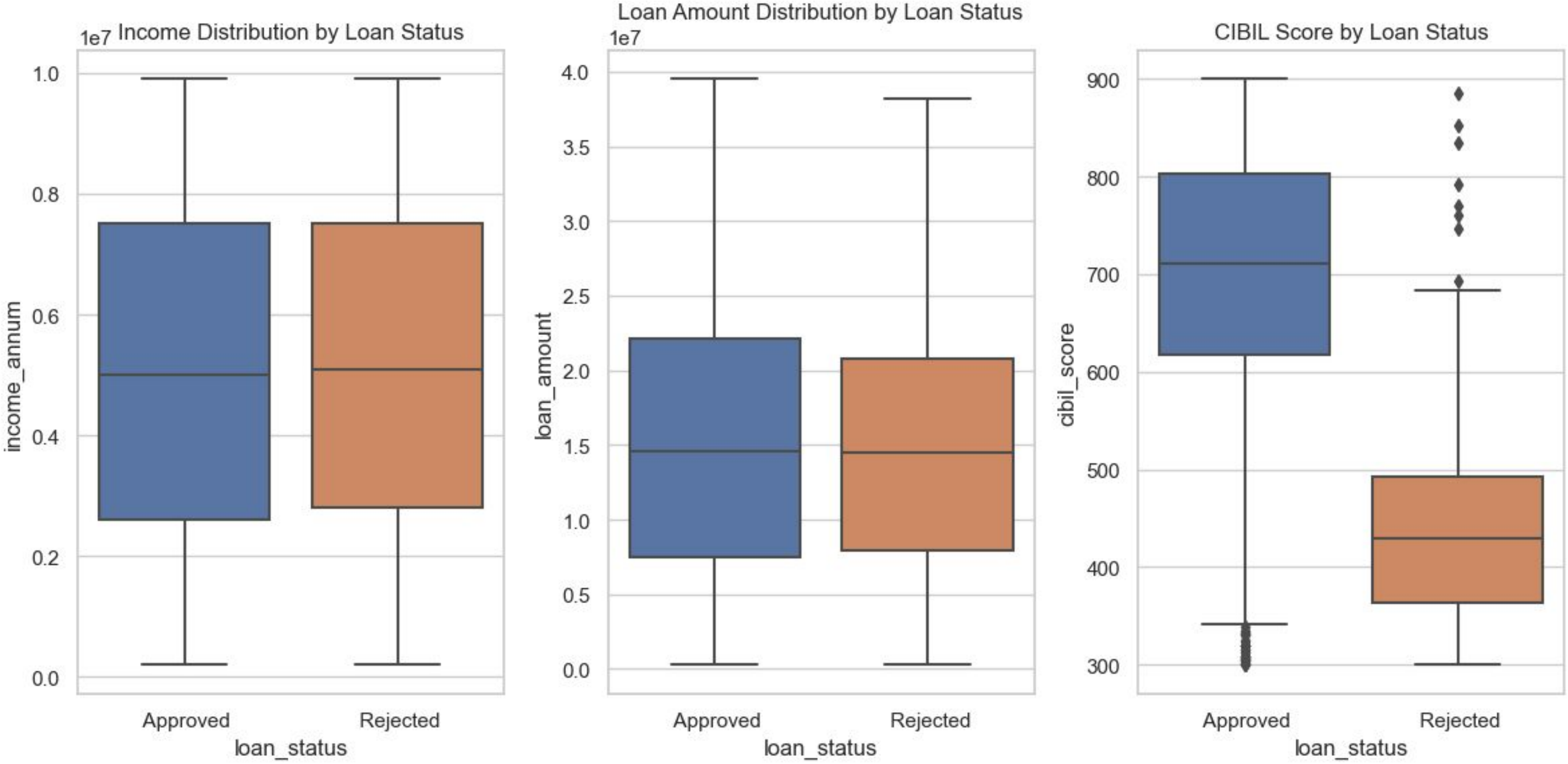




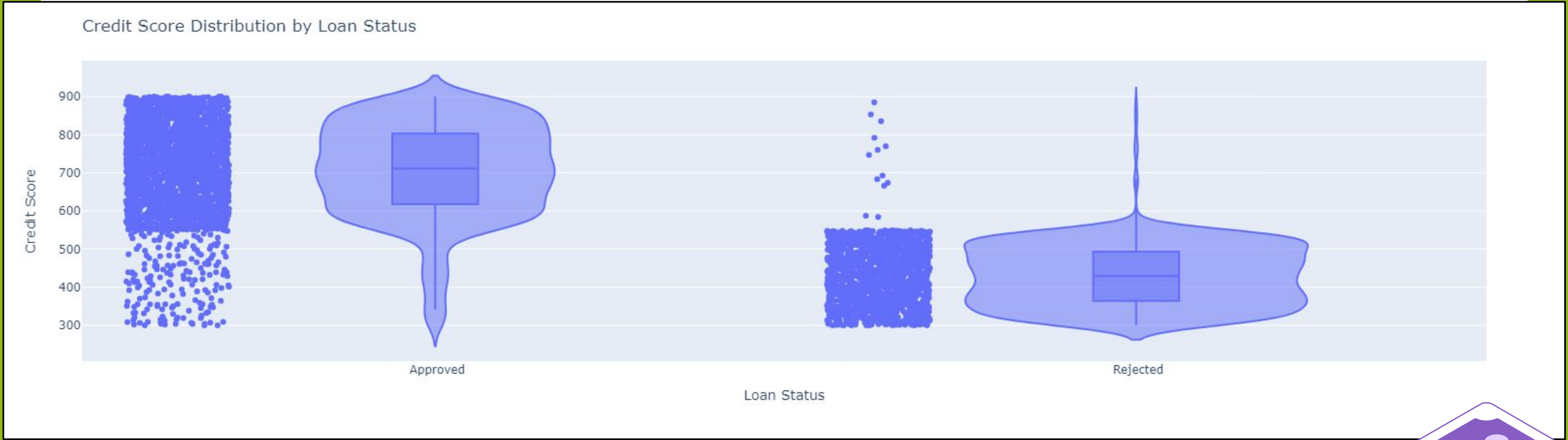
Income Distribution

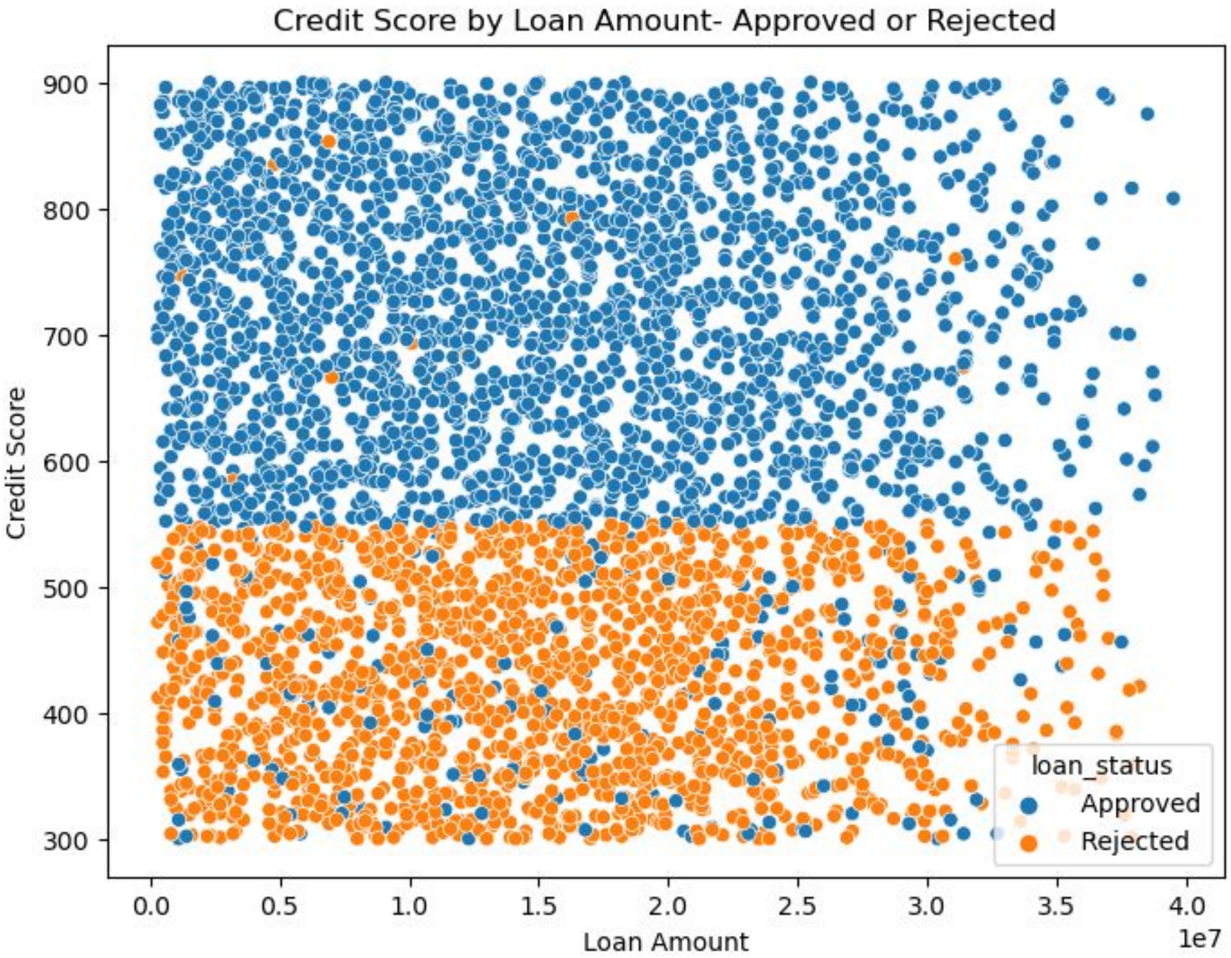


Boxplot Analysis by Loan Status

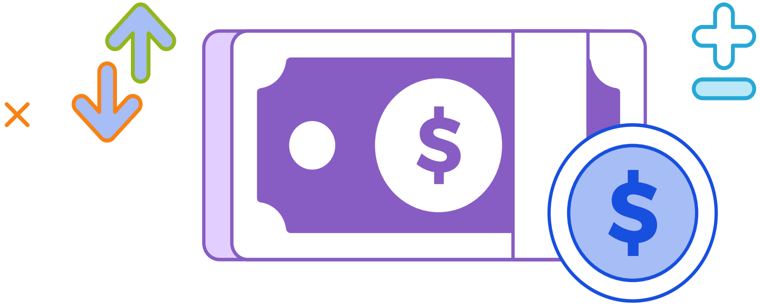


Credit Score by Loan Status




Credit Score by Loan Amount- Approved or Rejected
Credit Score
900
800
700
600
500
400
300
Loan Amount
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
1e7
loan_status
Approved
Rejected

Supervised Learning Model



03

Models Attempted

- Random Forest (x7)
- KN Neighbors (x4)
- Decision Tree (x1)
- Logistic Regression (x4)

Best Model - Logistic Regression

- Stripped leading/trailing whitespaces in column names
- Dropped ‘loan_id’ column as it is not beneficial
- Target (y) = ‘loan_status’ (Approved or Rejected)
- Features (X) = remaining columns
- Used pd.get_dummies to encode categorical variables (‘education’ and ‘self_employed’)
- Split data into training and testing sets
- Scaled the data using StandardScaler
- Created & trained the model and made predictions
- Optimization attempts were made by adjusting the features included, but the highest score was with keeping all of the features

Features Snapshots

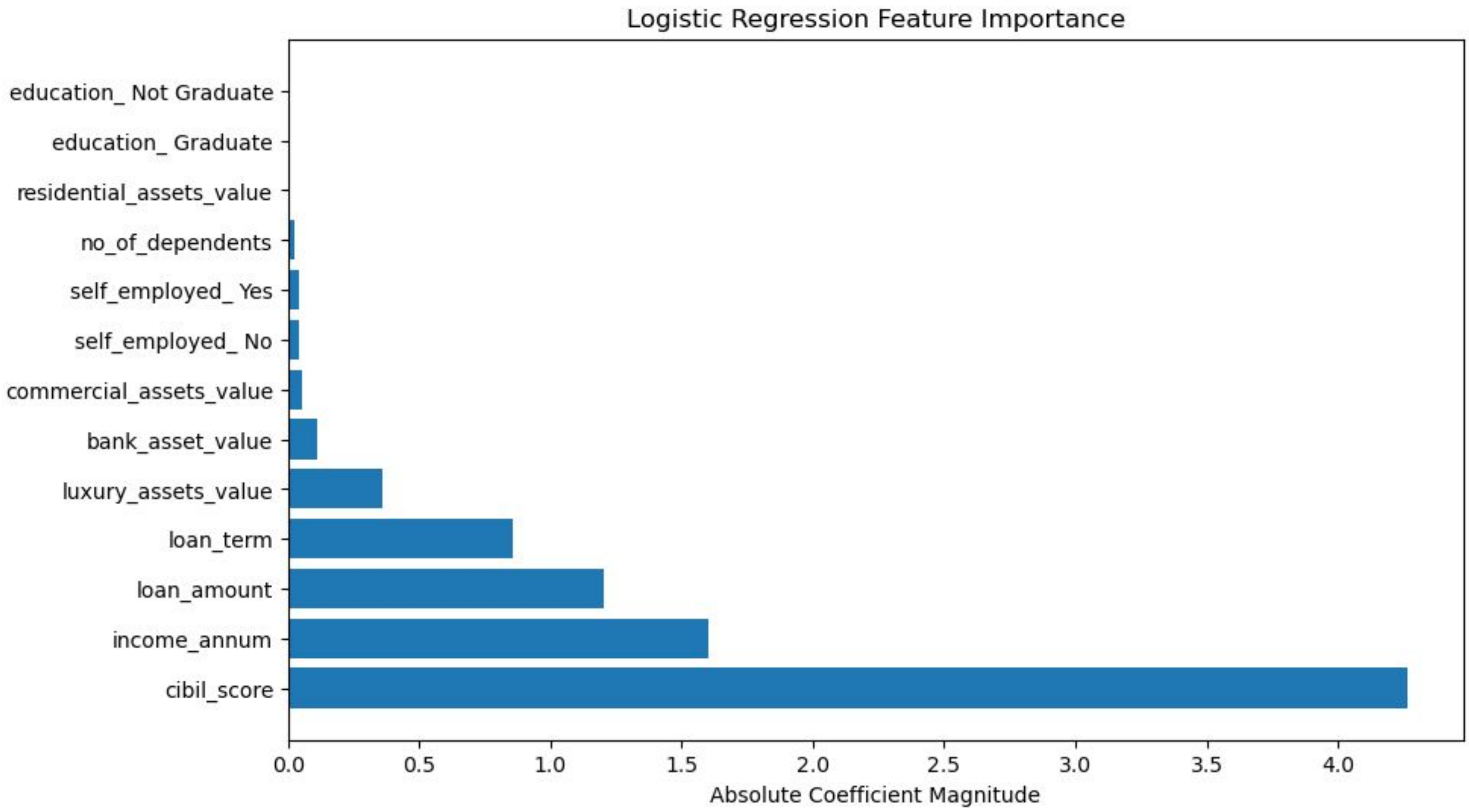
```
1  # Preview the features data
2  X.head()
```

	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
0	2	Graduate	No	9600000	29900000	12	778	2400000	17600000	22700000	8000000
1	0	Not Graduate	Yes	4100000	12200000	8	417	2700000	2200000	8800000	3300000
2	3	Graduate	No	9100000	29700000	20	506	7100000	4500000	33300000	12800000
3	3	Graduate	No	8200000	30700000	8	467	18200000	3300000	23300000	7900000
4	5	Not Graduate	Yes	9800000	24200000	20	382	12400000	8200000	29400000	5000000

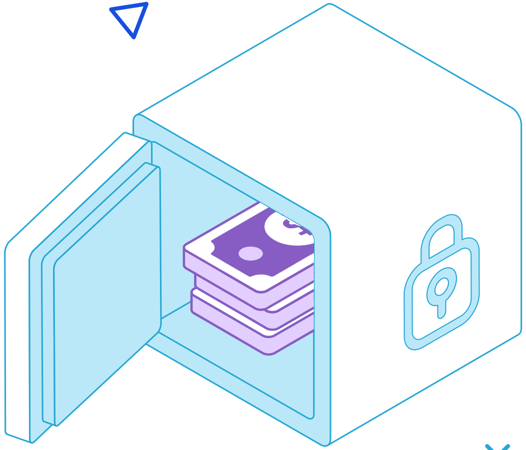
```
1  # Review the features data
2  X.head()
```

Pyth

	no_of_dependents	income_annum	loan_amount	loan_term	cibil_score	residential_assets_value	commercial_assets_value	luxury_assets_value	bank_asset_value	education_ Graduate	education_ Not Graduate	self_employed_ No	self_employed_ Yes
0	2	9600000	29900000	12	778	2400000	17600000	22700000	8000000	1	0	1	0
1	0	4100000	12200000	8	417	2700000	2200000	8800000	3300000	0	1	0	1
2	3	9100000	29700000	20	506	7100000	4500000	33300000	12800000	1	0	1	0
3	3	8200000	30700000	8	467	18200000	3300000	23300000	7900000	1	0	1	0
4	5	9800000	24200000	20	382	12400000	8200000	29400000	5000000	0	1	0	1


Logistic Regression Feature Importance
education_ Not Graduate
education_ Graduate
residential_assets_value
no_of_dependents
self_employed_ Yes
self_employed_ No
commercial_assets_value
bank_asset_value
luxury_assets_value
loan_term
loan_amount
income_annum
cibil_score
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
Absolute Coefficient Magnitude

Model Results

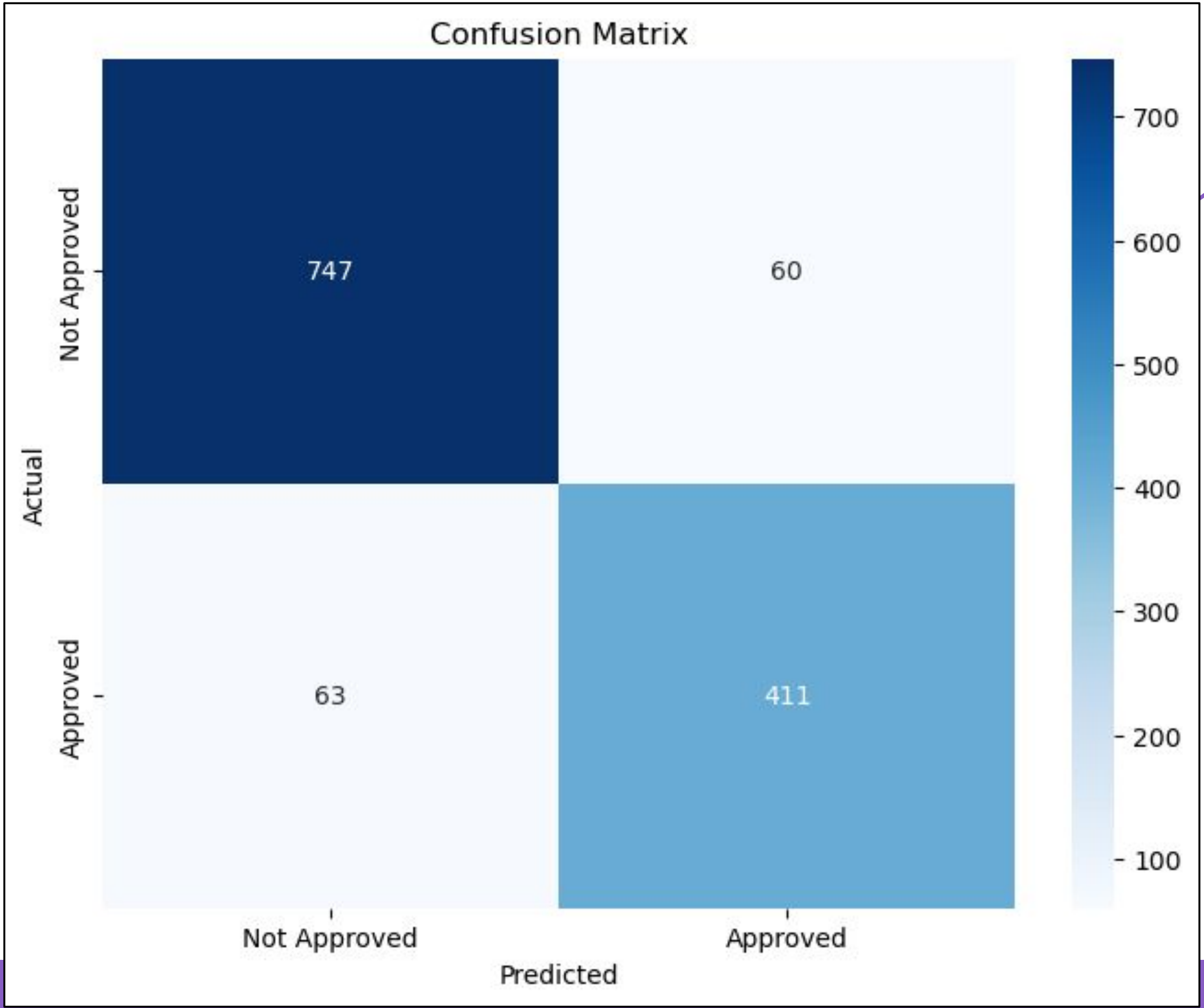


04

Final accuracy score: 90%

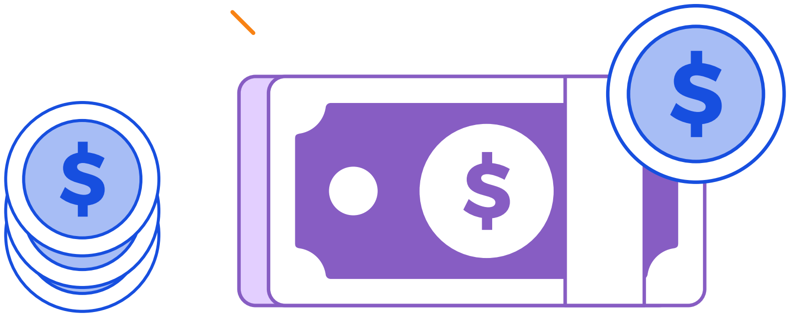
```
Classification Report:
              precision    recall  f1-score   support

    Approved       0.93      0.92      0.92       810
    Rejected       0.87      0.87      0.87       471

    accuracy                           0.90      1281
   macro avg       0.90      0.90      0.90      1281
weighted avg       0.90      0.90      0.90      1281
```

Confusion Matrix



Takeaway & Next Steps

05



Takeaways:

Key feature: credit score

Data Limitations:

- Loan types are not specified
- Additional factors, such as foreclosures or bankruptcies, that were not included
- Debt-to-income ratio
- Interest rates

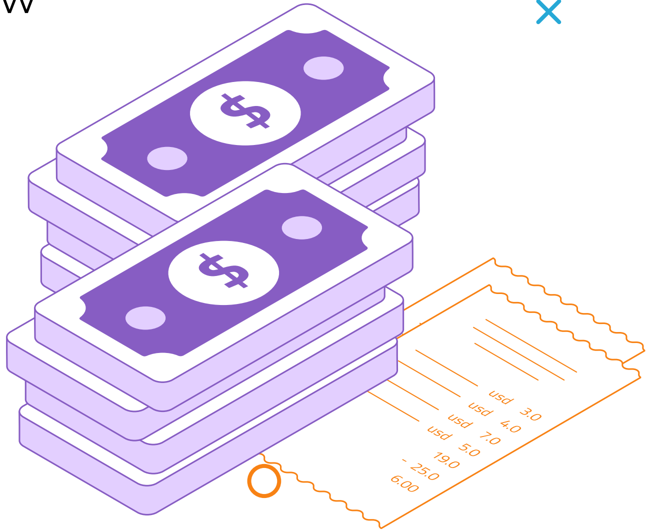
Other considerations:

- Demographics

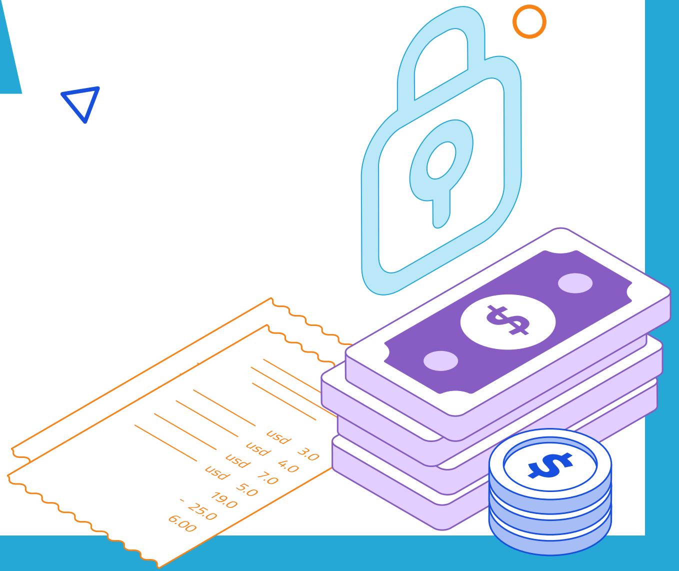
Next Steps

With more time, we would...

- Further identify factors to make model more efficient with fewer dependent features
- Test product for users to be able to quickly know how likely their loan request is to be approved



Q&A



Thank You!

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon** and infographics & images by **Freepik**