Standard Bank

Exploratory Data Analysis for Home Loan Application

Author: Aboubakr Aakaou

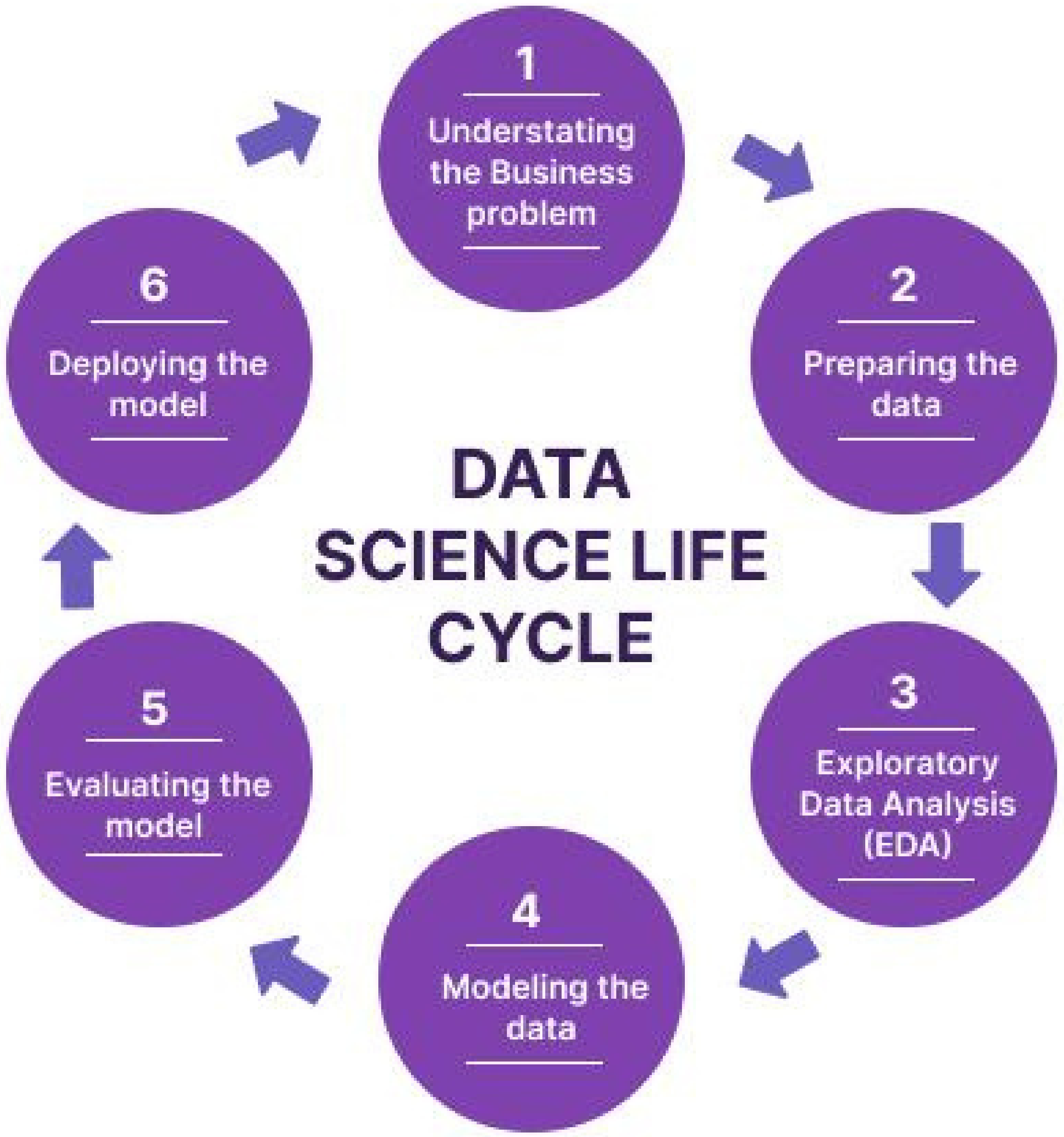
Agenda

- Data Science lifecycle
- Project Overview
- Overview of the Data
- Analysis
- Modeling
- Evaluation
- Deployment
- Conclusion

Data Science lifecycle

Unlocking Insights: The Data Science Lifecycle

The data science lifecycle comprises six stages: Business Understanding, Data Understanding, Data Preparation, Modeling, Evaluation, and Deployment, guiding the process from problem definition to actionable insights.

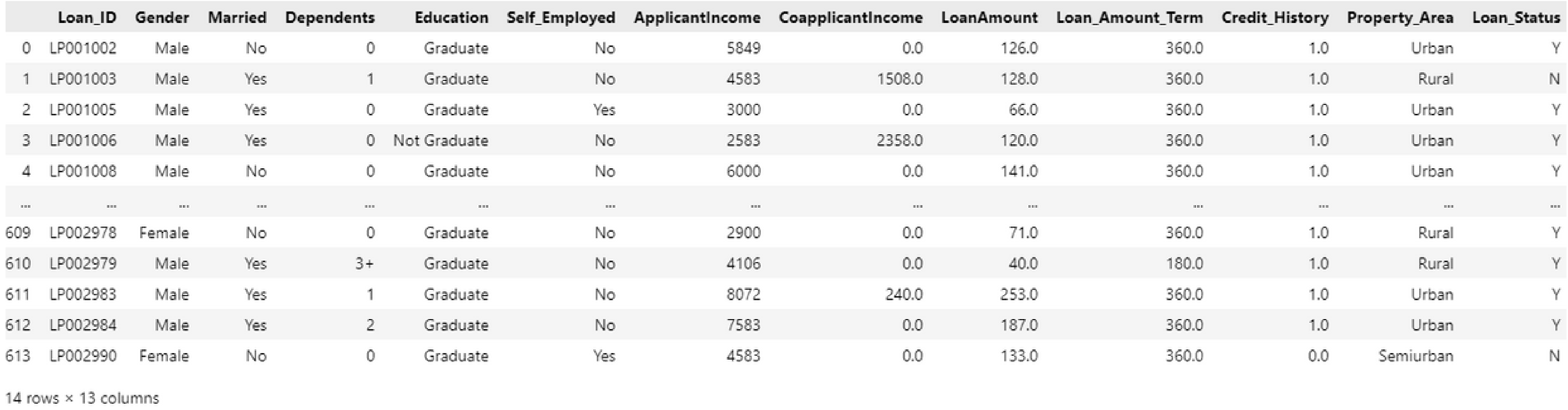


Project Overview

1. Standard Bank's digital transformation initiative for home loan applications.
2. **Objective**: Predict loan default risk and provide instant responses.
3. Adhering to the data science lifecycle (CRISP-DM).
4. **Key analysis objectives:** data overview, data quality, loan status, dependents, income comparison, and credit history impact.
5. A blend of automation and traditional methods for enhanced efficiency.
6. Improving customer service and experience.

Overview of the Data

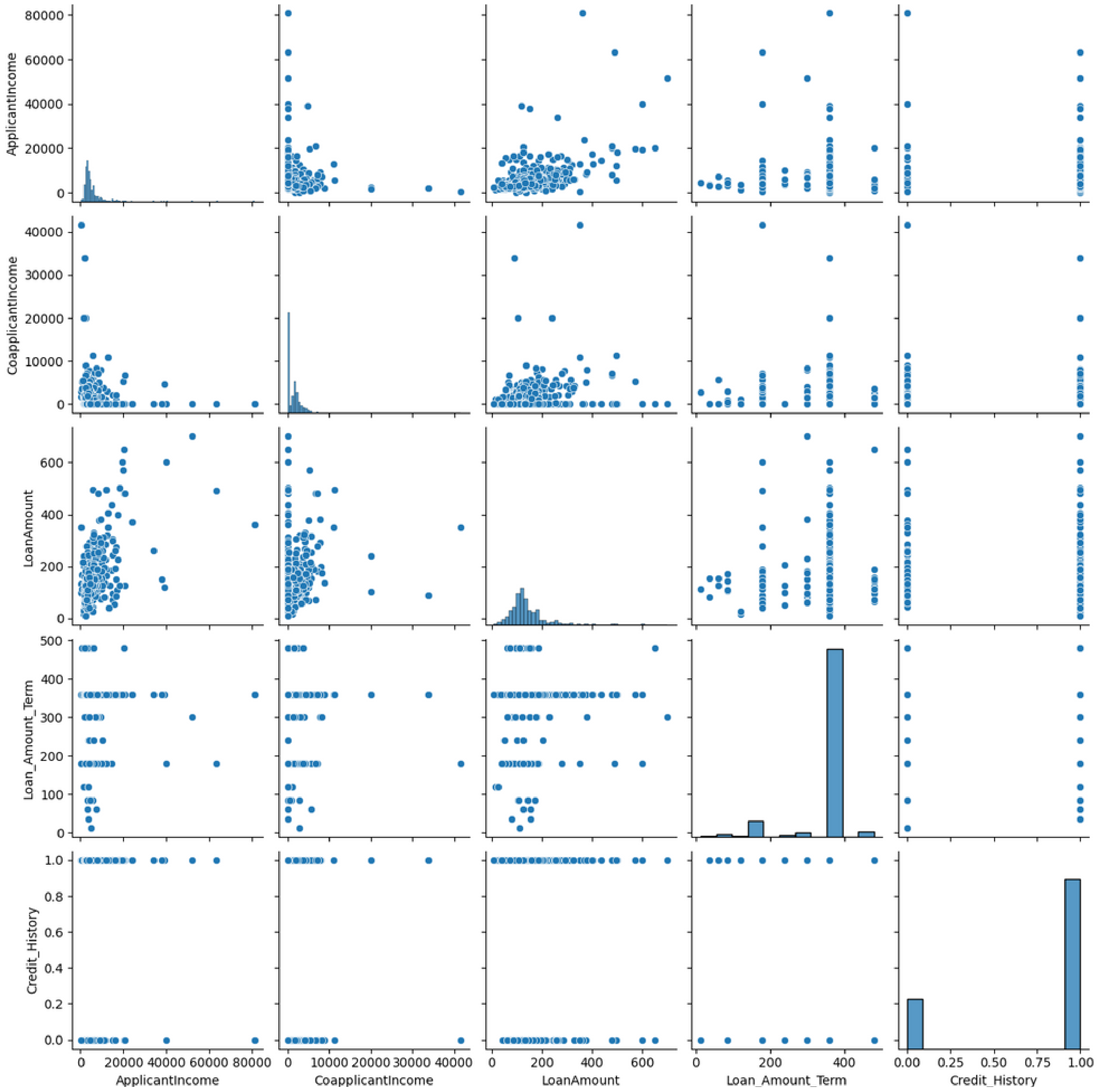
This pair plot is a graphical representation that shows pairwise relationships between variables in a dataset, typically through scatterplots. It helps visualize correlations, distributions, and patterns between multiple variables in the data.



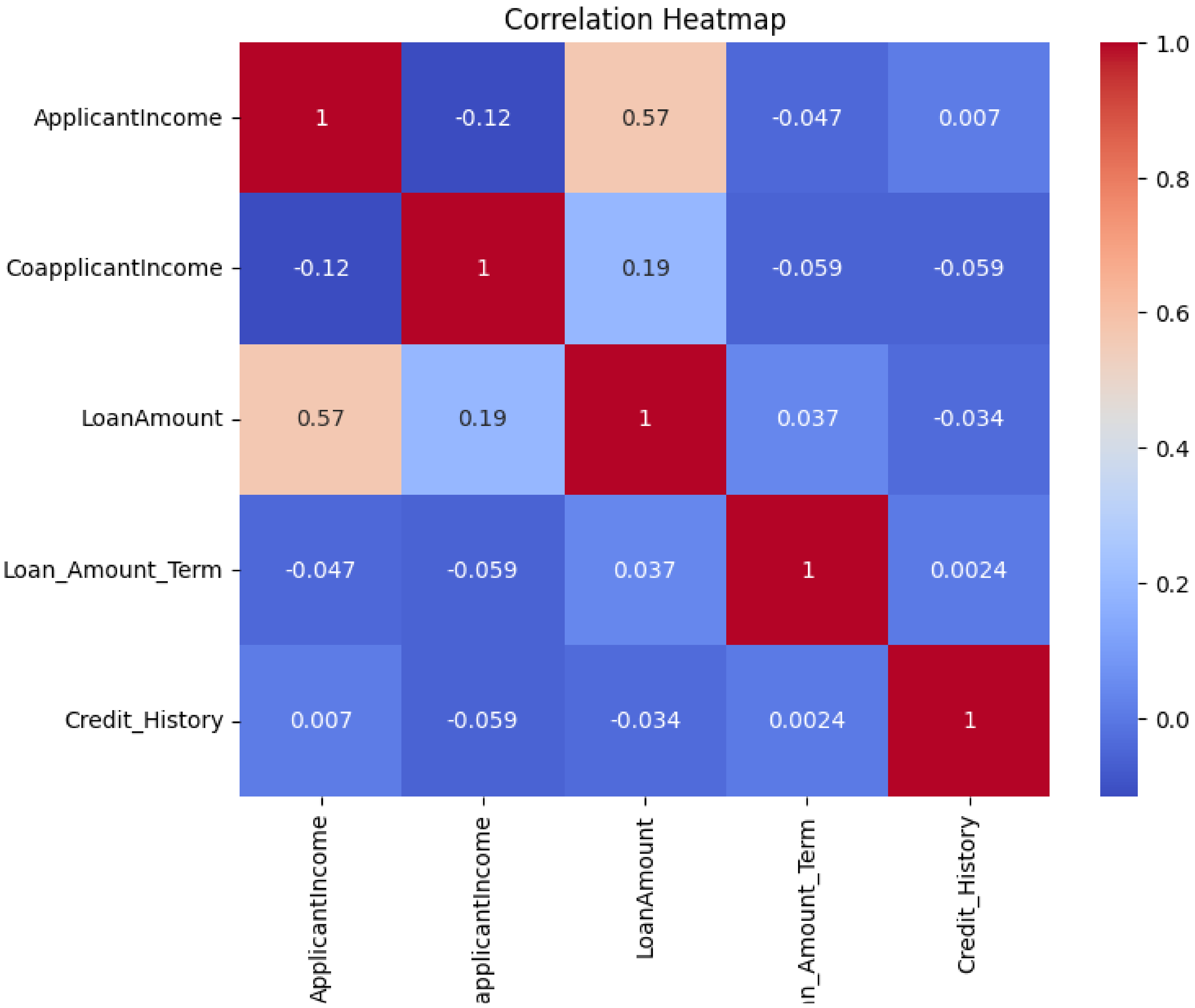
	Loan_ID	Gender	Married	Dependents	Education	Self_Employed	ApplicantIncome	CoapplicantIncome	LoanAmount	Loan_Amount_Term	Credit_History	Property_Area	Loan_Status
0	LP001002	Male	No	0	Graduate	No	5849	0.0	126.0	360.0	1.0	Urban	Y
1	LP001003	Male	Yes	1	Graduate	No	4583	1508.0	128.0	360.0	1.0	Rural	N
2	LP001005	Male	Yes	0	Graduate	Yes	3000	0.0	66.0	360.0	1.0	Urban	Y
3	LP001006	Male	Yes	0	Not Graduate	No	2583	2358.0	120.0	360.0	1.0	Urban	Y
4	LP001008	Male	No	0	Graduate	No	6000	0.0	141.0	360.0	1.0	Urban	Y
...	...	...	...	...	...	...	...	...	...	...	...	...	...
609	LP002978	Female	No	0	Graduate	No	2900	0.0	71.0	360.0	1.0	Rural	Y
610	LP002979	Male	Yes	3+	Graduate	No	4106	0.0	40.0	180.0	1.0	Rural	Y
611	LP002983	Male	Yes	1	Graduate	No	8072	240.0	253.0	360.0	1.0	Urban	Y
612	LP002984	Male	Yes	2	Graduate	No	7583	0.0	187.0	360.0	1.0	Urban	Y
613	LP002990	Female	No	0	Graduate	Yes	4583	0.0	133.0	360.0	0.0	Semiurban	N

14 rows × 13 columns

Data set after preprocessing

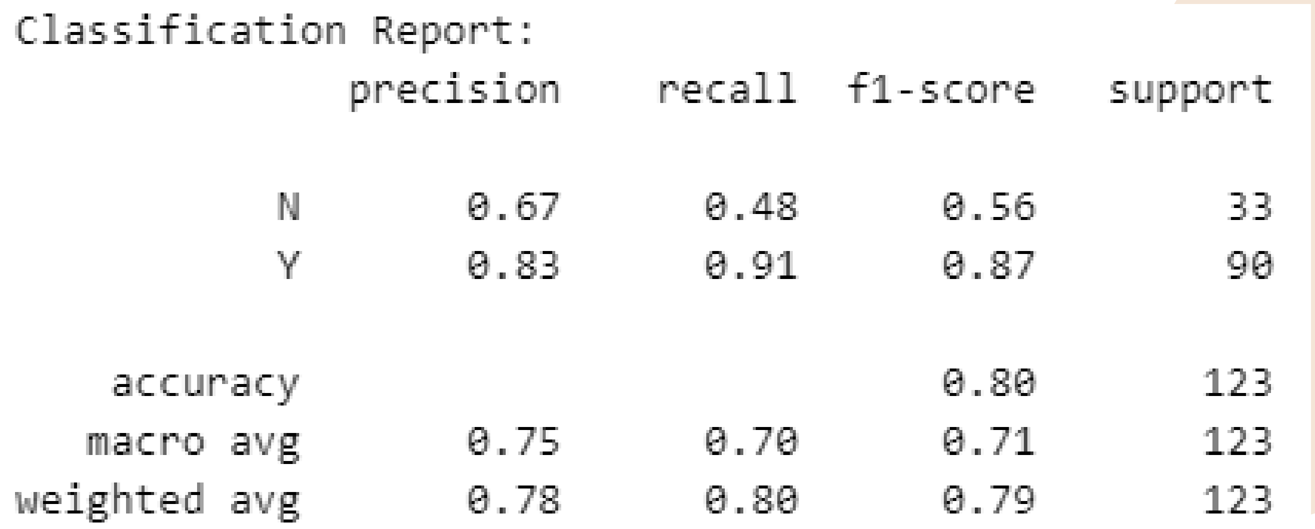


Analysis

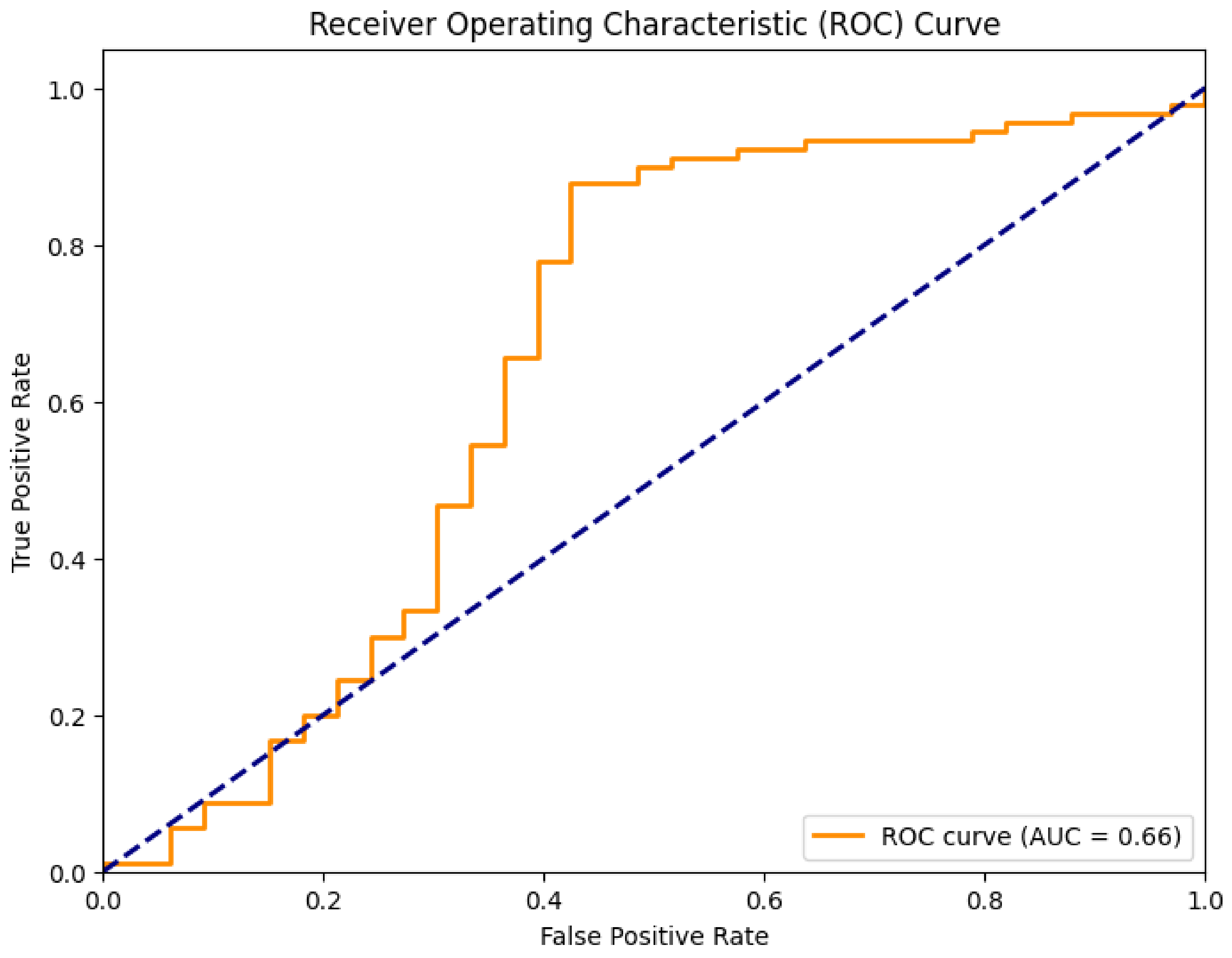

Correlation Heatmap
ApplicantIncome
CoapplicantIncome
LoanAmount
Loan_Amount_Term
Credit_History
1
-0.12
0.57
-0.047
0.007
-0.12
1
0.19
-0.059
-0.059
0.57
0.19
1
0.037
-0.034
-0.047
-0.059
0.037
1
0.0024
0.007
-0.059
-0.034
0.0024
1
ApplicantIncome
applicantIncome
LoanAmount
ın_Amount_Term
Credit_History
1.0
0.8
0.6
0.4
0.2
0.0

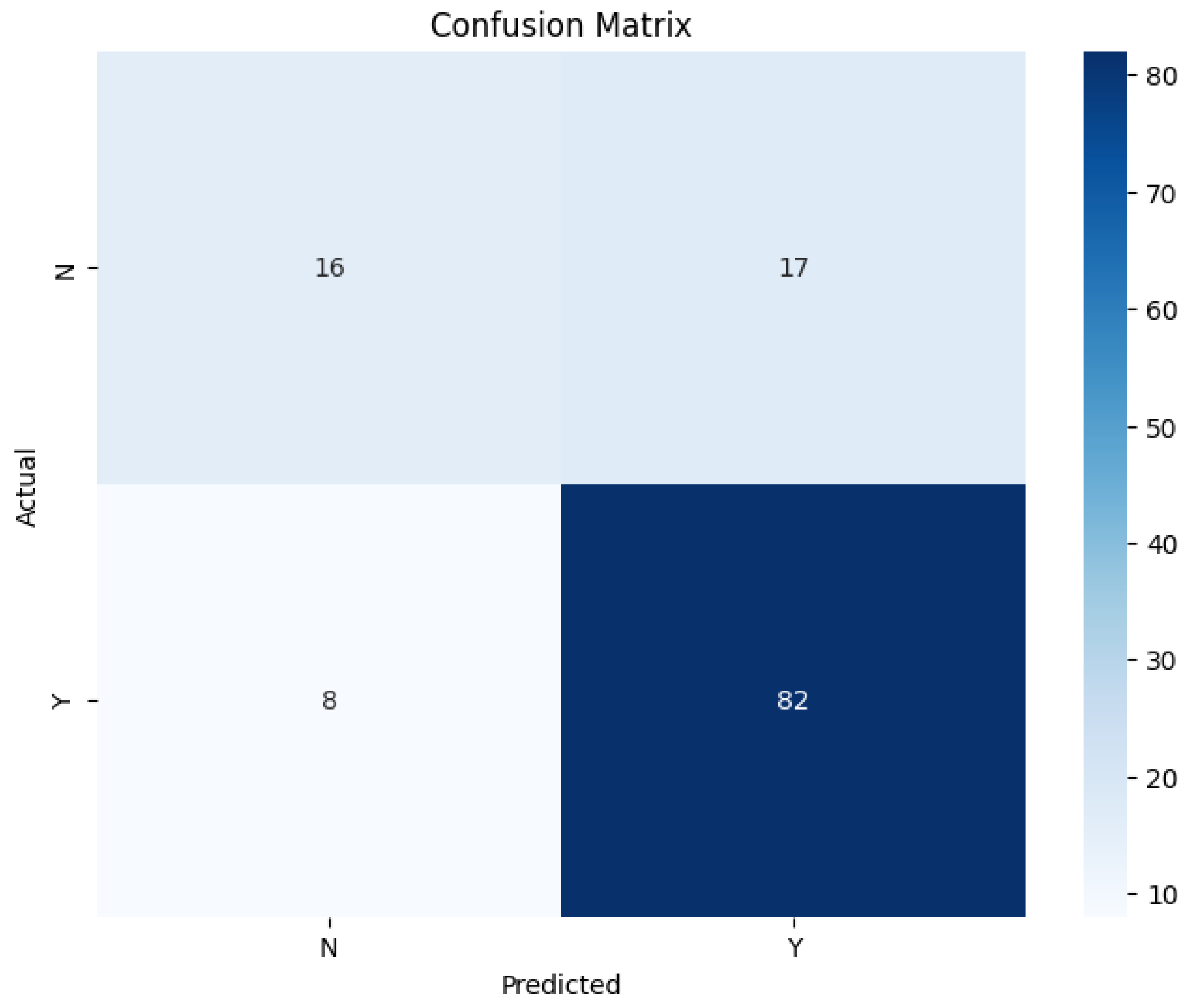
Modeling

Logistic Regression



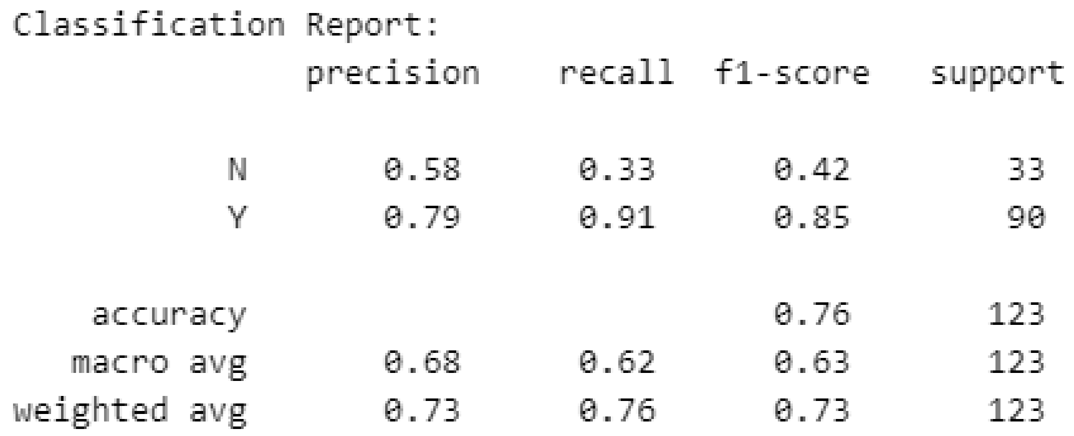
```
Classification Report:
              precision    recall  f1-score   support

           N       0.67      0.48      0.56        33
           Y       0.83      0.91      0.87        90

    accuracy                           0.80       123
   macro avg       0.75      0.70      0.71       123
weighted avg       0.78      0.80      0.79       123
```



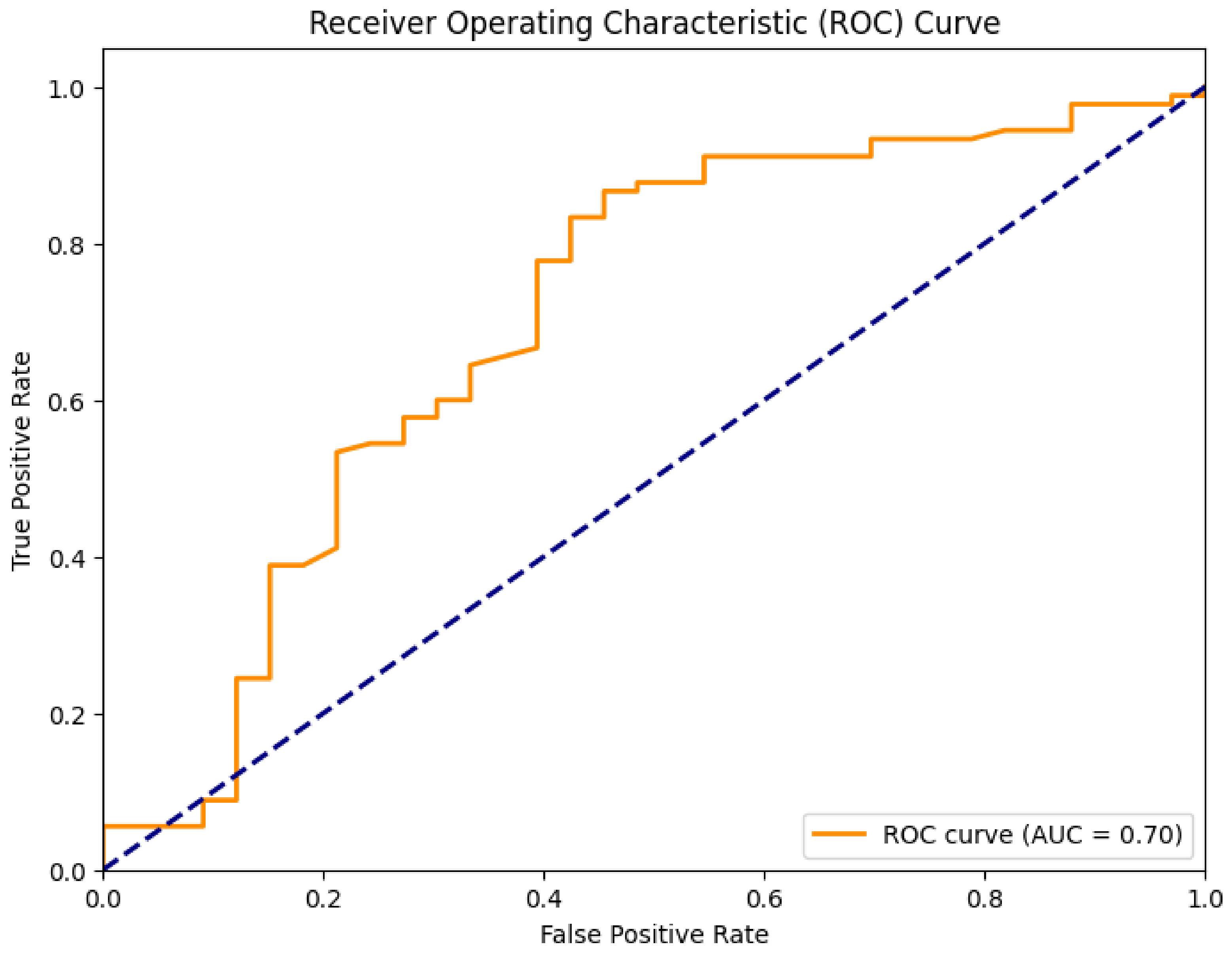


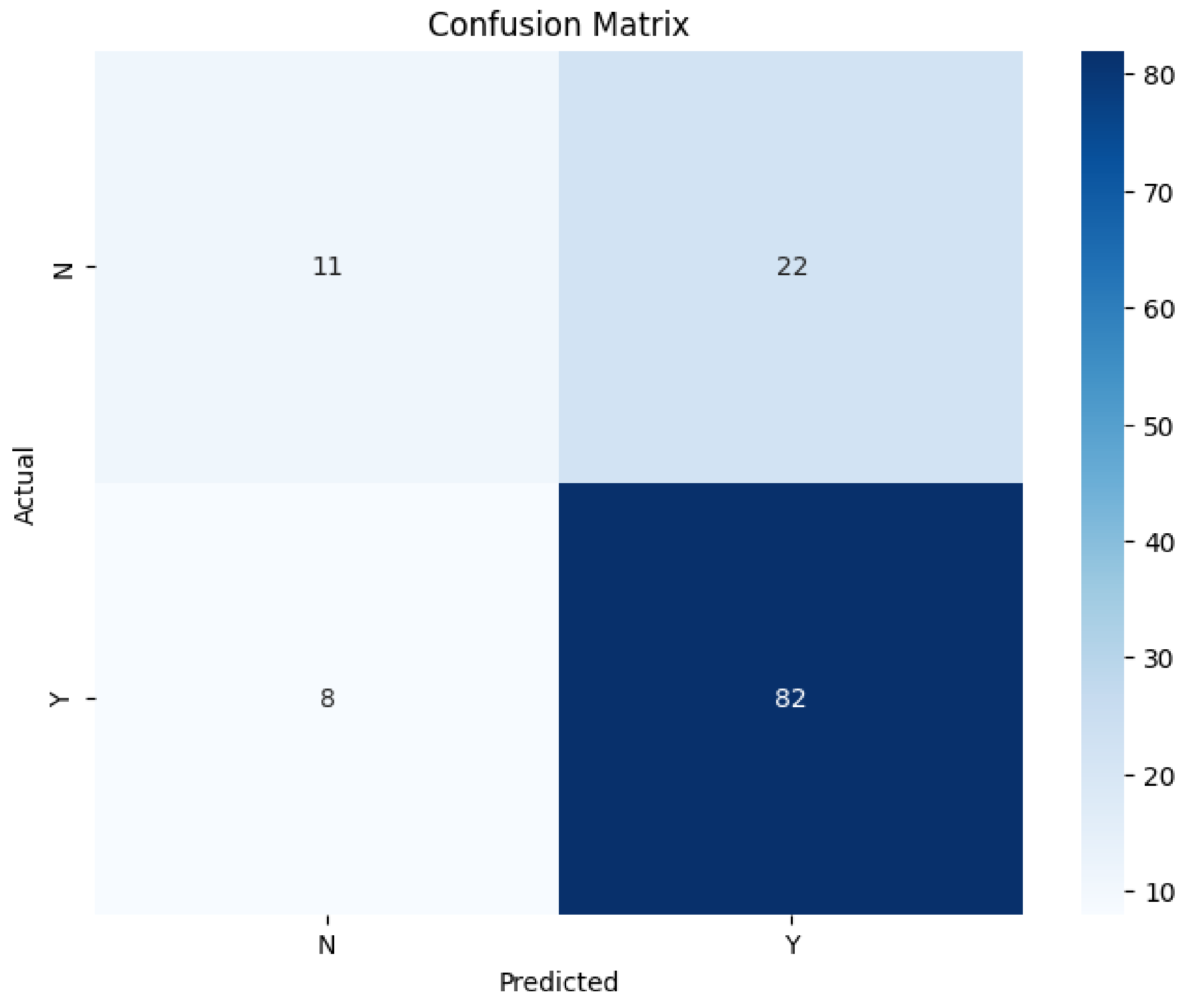
Modeling

Gradient boosting



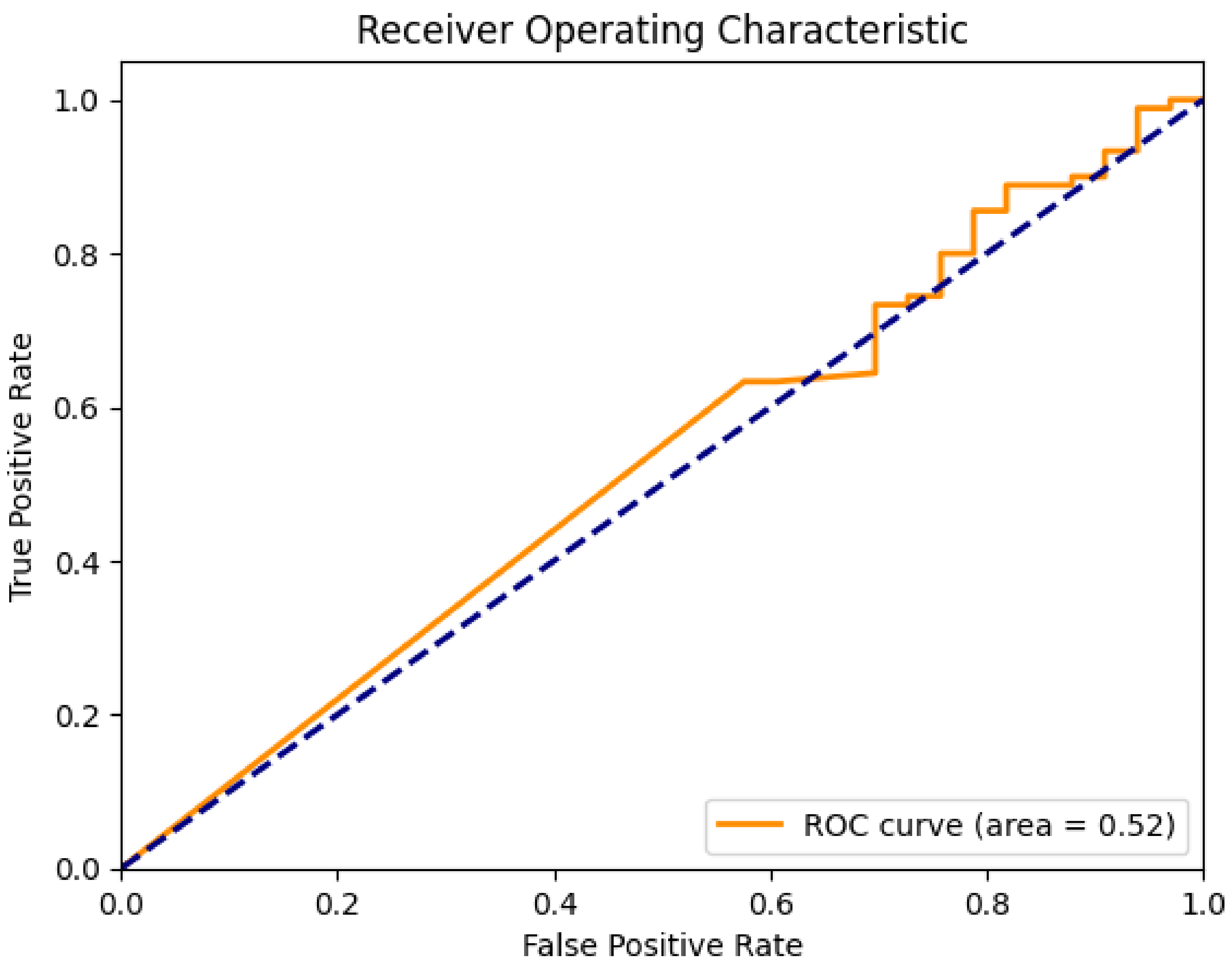
```
Classification Report:
              precision    recall  f1-score   support

           N       0.58      0.33      0.42        33
           Y       0.79      0.91      0.85        90

    accuracy                           0.76       123
   macro avg       0.68      0.62      0.63       123
weighted avg       0.73      0.76      0.73       123
```

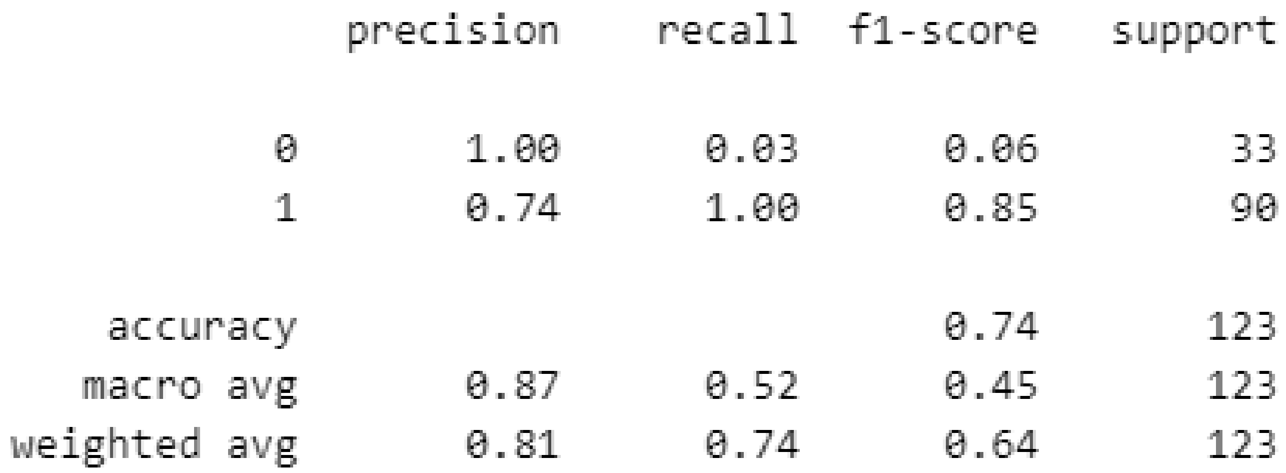




Modeling

CNN model





```
              precision    recall  f1-score   support

           0       1.00      0.03      0.06        33
           1       0.74      1.00      0.85        90

    accuracy                           0.74       123
   macro avg       0.87      0.52      0.45       123
weighted avg       0.81      0.74      0.64       123
```

Evaluation

Models	Logistic Regression	Gradient boosting	CNN
Accuracy	80%	76%	74%

Deployment

In the deployment phase, we put our models into action within the real-world context of Standard Bank's home loan application process. This ensures that our predictive solutions are actively contributing to quicker, more informed decisions.

Conclusion

Our journey through the data science lifecycle has brought us valuable insights and promising results. Logistic Regression, with an 80% accuracy, stands out as a robust choice, closely followed by Gradient Boosting at 76% and the CNN model at 74%. This project underscores the potential of data-driven decision-making, and it's a significant step towards enhancing the efficiency and customer experience in home loan applications at Standard Bank.

Thank you for your attention