Master's Project

LOAN APPROVAL PREDICTION

- UP2205708

MSc. Data Analytics

Contents

ABSTRACT

- Loan approval is a crucial process for banking firms.
- Recovery of loans significantly impacts a bank’s financial statements.
- Predicting an applicant’s ability to repay the loan is a common challenge.
- Machine learning (ML) models are effective in predicting outcomes for large datasets.
- The project aims to streamline the loan approval process, reduce the risk of defaults, and improve decision-making for financial institutions.

AIM and OBJECTIVE

- **Loan Approval Prediction:** Develop machine learning models with more accuracy predict the outcomes of loan applications, helping banks make more informed decisions.
- **Applicant Behaviour:** Build models to differentiate between defaulters and non-defaulters, offering insights into the risk profile of applicants.
- **Association Rules:** Generate association rules to provide actionable insights that assist banks and loan firms in refining their loan approval processes and minimising risk.
- The primary objective is to go Conduct a comprehensive literature review of current tools and machine learning approaches used in loan approval prediction.
- Planing methodolgy to complete the study of this project
- Execution: Utilise the available dataset for exploratory data analysis (EDA), followed by data preparation, model development, testing, and performance evaluation.
- Summarise findings, discuss limitations, and suggest directions for future research and improvements

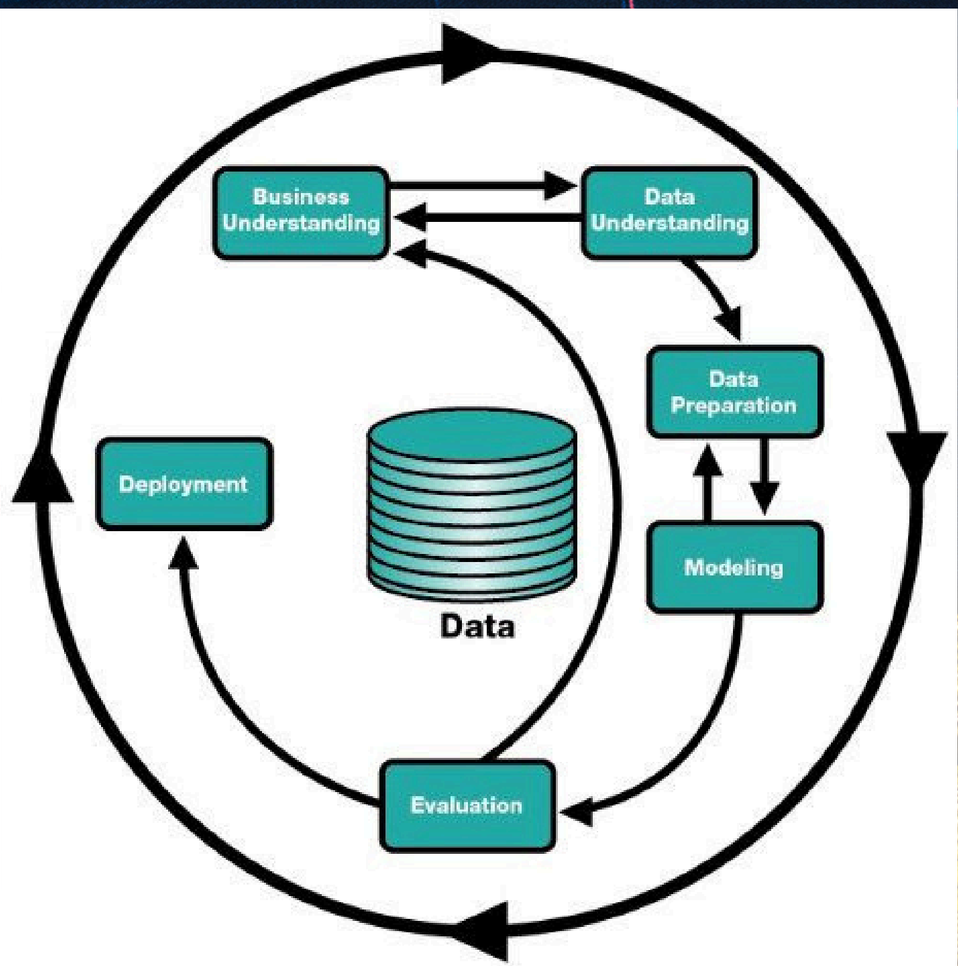
LITERATURE REVIEW

- The loan approval methods began in the mid 20th century using traditional manual procedures

1960s–1970s	1980s–1990s	2000s–2010s	2010s–Present
Discriminant Analysis	Logistic Regression	Neural Networks	Machine Learning Big Data Analytics
Simple Linear Models	Probability Estimates	Non-linear Relationships	Complex Pattern Recognition Real-time Analysis

METHODOLOGY

- **CRISP-DM Framework:**
  - Business Understanding: Defines objectives and scope.
  - Data Understanding: Collect and analyse data.
  - Data Preparation: Clean and prepare data for modeling.
  - Modeling: Apply machine learning models (logistic regression, decision trees, random forests,etc).
  - Evaluation: Assess model performance using metrics like accuracy, precision, recall, and AUC.
  - Deployment: Suggest to implement the model in a real-world setting.



DATASET

- **Source**: **Kaggle**, contributed by Shivam Kapoor (2021).
- **Current_app Dataset** (166.13 MB): Contains information on existing loan applications, including payment difficulties. (**Size**: 307,511 rows and 122 columns.)
- **Previous_app Dataset** (404.97 MB): Includes details on previous loan applications with statuses like Approved, Cancelled, Refused, or Unused. (**Size**: 1,607,214 rows and 37 columns.)
- **Data Privacy:** The datasets are anonymised to ensure data privacy without the need for individual consent.
- **Currency:** All financial data is in INR for the purpose of this analysis.
- **Data Partitioning: 60%** for **training**, **20%** for **testing**, and **20%** for **validation** to fit models, evaluate performance, and tune parameters, ensuring robust model development.

```
[ ] filtered_previous_app.head()
```

	SK_ID_PREV	SK_ID_CURR	NAME_CONTRACT_TYPE	AMT_ANNUITY	AMT_APPLICATION	AMT_CREDIT	AMT_DOWN_PAYMENT	AMT_GOODS_PRICE	WEEKDAY_APPR_PROCESS_START
0	1038818	100002	Consumer loans	9251.775	179055.0	179055.0	0.0	179055.0	SATURDAY
1	2636178	100003	Consumer loans	64567.665	337500.0	348637.5	0.0	337500.0	SUNDAY
2	1564014	100004	Consumer loans	5357.250	24282.0	20106.0	4860.0	24282.0	FRIDAY
3	2827850	100006	Revolving loans	NaN	0.0	0.0	NaN	NaN	THURSDAY
4	2730157	100007	Cash loans	13010.985	225000.0	284400.0	NaN	225000.0	FRIDAY

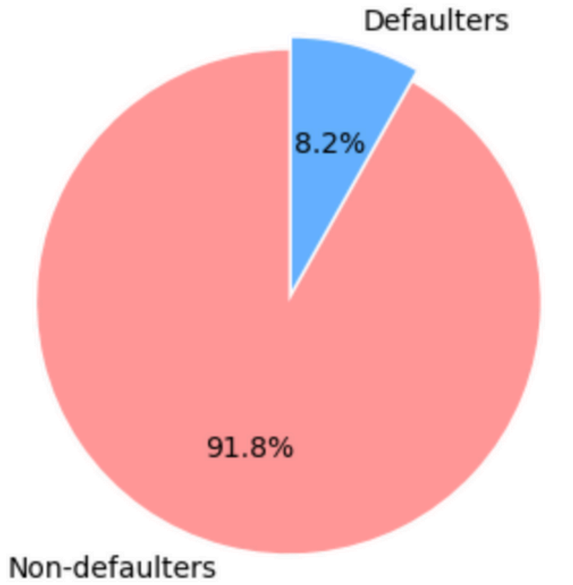
```
[ ] filtered_current_app.head()
```

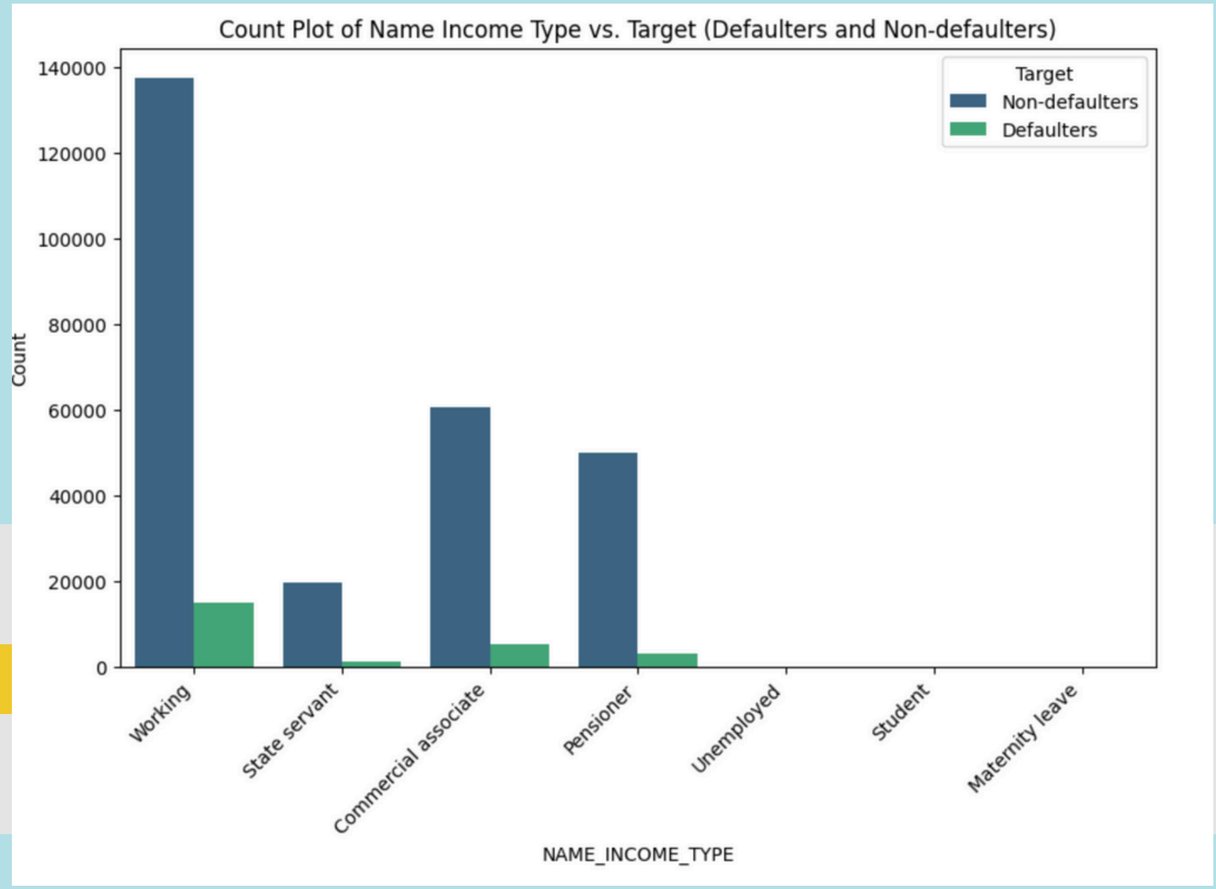
	SK_ID_CURR	TARGET	NAME_CONTRACT_TYPE	CODE_GENDER	CNT_CHILDREN	AMT_INCOME_TOTAL	AMT_CREDIT	AMT_ANNUITY	AMT_GOODS_PRICE	NAME_TYPE_SUITE
0	100002	1	Cash loans	M	0	202500.0	406597.5	24700.5	351000.0	Unaccompanied
1	100003	0	Cash loans	F	0	270000.0	1293502.5	35698.5	1129500.0	Family
2	100004	0	Revolving loans	M	0	67500.0	135000.0	6750.0	135000.0	Unaccompanied
3	100006	0	Cash loans	F	0	135000.0	312682.5	29686.5	297000.0	Unaccompanied
4	100007	0	Cash loans	M	0	121500.0	513000.0	21865.5	513000.0	Unaccompanied

DATA VISUALISATION : Loan Applicant Behaviour

- **Distribution of applicants for ‘Defaulters’ Vs ‘Non-Defaulters**
- **Distribution of Loan Defaulters and Non-Defaulters by Income type**

```
TARGET
0    267212
1     23845
Name: count, dtype: int64
```





Classification: Loan Applicant Behaviour(Analysing Defaulters or Non-Defaulters)

This classification model, developed from the data, aims to predict whether an applicant is a defaulter (class 1) or a non-defaulter (class 0). It will provide banks and loan firms with a behaviour analysis of loan applicants.

Algorithms that have been selected to perform classification on TARGET feature.

1. Logistic Regression Classification
2. Decision Tree Classification
3. Gradient Boosting Classification

CLASSIFICATION: Loan Applicant Behaviour

```
Logistic Regression - Training Set Evaluation
Confusion Matrix :
[[126329  33998]
 [ 40561 119766]]
Classification Report :
              precision    recall  f1-score   support

           0       0.76      0.79      0.77    160327
           1       0.78      0.75      0.76    160327

    accuracy                           0.77    320654
   macro avg       0.77      0.77      0.77    320654
weighted avg       0.77      0.77      0.77    320654

Accuracy : 0.7674783411402946
AUC : 0.8393667657804622
```

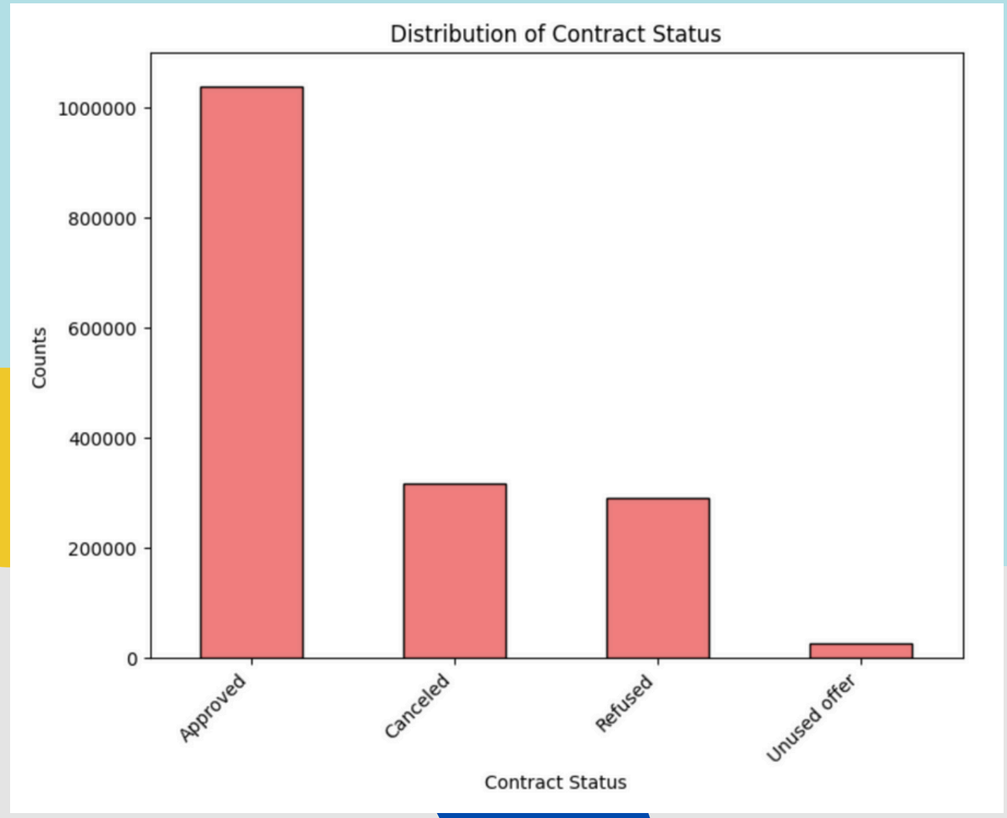
```
Decision Tree - Training Set Evaluation
Confusion Matrix :
[[148243  12084]
 [ 32021 128306]]
Classification Report :
              precision    recall  f1-score   support

           0       0.82      0.92      0.87    160327
           1       0.91      0.80      0.85    160327

    accuracy                           0.86    320654
   macro avg       0.87      0.86      0.86    320654
weighted avg       0.87      0.86      0.86    320654

Accuracy : 0.8624529867084146
AUC : 0.9300415457321152
```

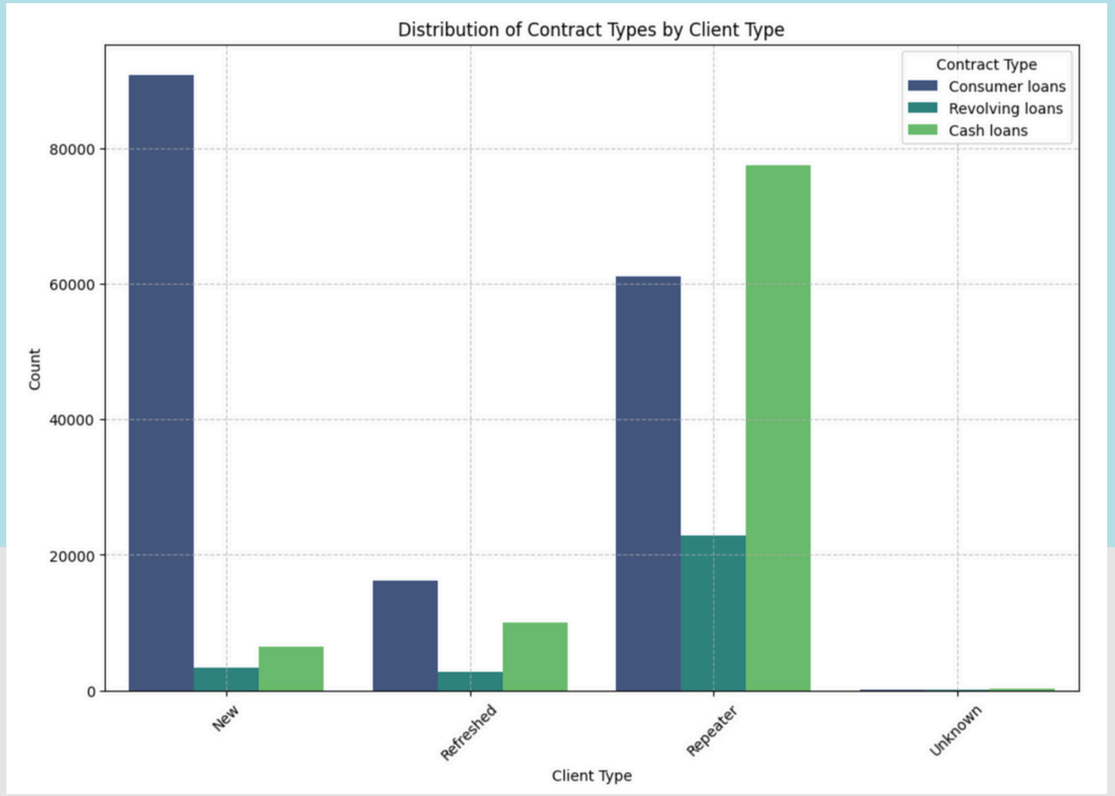
```
Gradient Boosting - Training Set Evaluation
Confusion Matrix :
[[153283   7044]
 [ 38825 121502]]
Classification Report :
              precision    recall  f1-score   support

           0       0.80      0.96      0.87    160327
           1       0.95      0.76      0.84    160327

    accuracy                           0.86    320654
   macro avg       0.87      0.86      0.86    320654
weighted avg       0.87      0.86      0.86    320654

Accuracy : 0.8569517299020127
AUC : 0.9168320421358388
```

Model	Accuracy (train Set)	Cross-Validated Accuracy	Accuracy (Test Set)
Logistic Regression	0.76	0.59	0.74
Decision Tree	0.86	0.65	0.84
Gradient Boosting	**0.85**	**0.91**	**0.88**

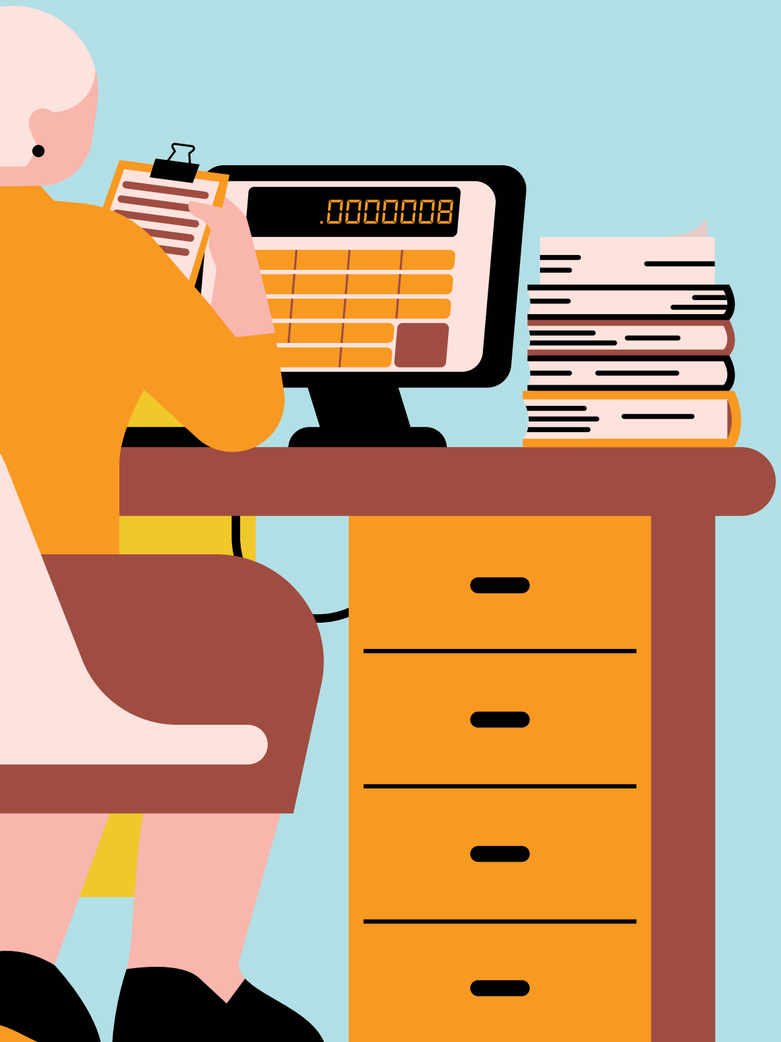
Data Visualisation : Loan Approval Prediction

- Count plot of ‘Contract status’
- Distribution of Contract Types by Client Type





CLASSIFACATION: Loan Approval Prediction(Analysing Loan applications is Approved, Refused, Cancelled or Unused Offer)



This classification model aims to predict the results of loan applications based on various applicant and loan characteristics. It will provide banks and loan firms with the outcomes of loan applications.

Algorithms that have been selected to perform classification on NAME_CONTRACT_STATUS feature.

1. Logistic Regression Classification
2. Decision Tree Classification
3. Random Forest Classification

CLASSIFICATION: Loan Approval Prediction

```
Logistic Regression
Confusion Matrix (Test Set):
[[30593  8260  3815    72]
 [  804  2286  3312   320]
 [    2  1159  6661     4]
 [   29     1     0   894]]
Mean Absolute Error: 0.37995602281316565
Mean Squared Error: 0.5325362468219611
Root Mean Squared Error: 0.7297508114568706
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       0.97      0.72      0.82     42740
           2       0.20      0.34      0.25      6722
           3       0.48      0.85      0.62      7826
           4       0.69      0.97      0.81       924

    accuracy                           0.69     58212
   macro avg       0.59      0.72      0.62     58212
weighted avg       0.81      0.69      0.73     58212

Accuracy (Test Set): 0.6945990517419088
AUC (Test Set): 0.8843750988623812
```

```
Random Forest
Confusion Matrix (Test Set):
[[41625   678   416    21]
 [   16  4685  1690   331]
 [    1  6329  1495     1]
 [    0    19     0   905]]
Mean Absolute Error: 0.17712842712842713
Mean Squared Error: 0.2056448842163128
Root Mean Squared Error: 0.4534808531970372
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       1.00      0.97      0.99     42740
           2       0.40      0.70      0.51      6722
           3       0.42      0.19      0.26      7826
           4       0.72      0.98      0.83       924

    accuracy                           0.84     58212
   macro avg       0.63      0.71      0.65     58212
weighted avg       0.85      0.84      0.83     58212

Accuracy (Test Set): 0.8367690510547653
AUC (Test Set): 0.971203878094806
```

```
Decision Tree
Confusion Matrix (Test Set):
[[41649   688   384    19]
 [  119  4643  1631   329]
 [    2  6337  1486     1]
 [    4    22     2   896]]
Mean Absolute Error: 0.17730021301449872
Mean Squared Error: 0.20499209784924072
Root Mean Squared Error: 0.45276053035709807
Classification Report (Test Set):
              precision    recall  f1-score   support

           1       1.00      0.97      0.99     42740
           2       0.40      0.69      0.50      6722
           3       0.42      0.19      0.26      7826
           4       0.72      0.97      0.83       924

    accuracy                           0.84     58212
   macro avg       0.63      0.71      0.64     58212
weighted avg       0.85      0.84      0.83     58212

Accuracy (Test Set): 0.8361506218649076
AUC (Test Set): 0.964563315068163
```

Model	Accuracy (train Set)	Cross-Validated Accuracy	Accuracy (Test Set)
Logistic Regression	0.695	0.77	0.69
Decision Tree	0.836	0.94	0.83
Random Forest	**0.837**	**0.95**	**0.83**

Association Rules:

Row - 35354	Row - 35354 - Items : frozenset({'Consumer loans', 'high', 'POS'}) => frozenset({'Approved'})
	Row - 35354 - Antecedent :Consumer loans, high, POS
	Row - 35354 - Consequent :Approved
	Row - 35354 - Support :0.168355953
	Row - 35354 - Confidence :0.944123314
	Row - 35354 - Lift :1.288567834

Row - 29105	Row - 29105 - Items :frozenset({'Not Specified', 'Cash loans', 'Credit and cash offices'}) => frozenset({'Repeater'})
	Row - 29105 - Antecedent :Not Specified, Cash loans, Credit and cash offices
	Row - 29105 - Consequent :Repeater
	Row - 29105 - Support :0.206150146
	Row - 29105 - Confidence :0.842696629
	Row - 29105 - Lift :1.496442062

Row - 34451	Row - 34451 - Items : frozenset({'POS household with interest', 'Consumer electronics'}) => frozenset({'Cash through the bank', 'Unknown'})
	Row - 34451 - Antecedent :POS household with interest, Consumer electronics
	Row - 34451 - Consequent :Cash through the bank, Unknown
	Row - 34451 - Support :0.176601958
	Row - 34451 - Confidence :0.874893617
	Row - 34451 - Lift :1.307846878

Row - 5565	Row - 5565 - Items : frozenset({'Repeater', 'Credit and cash offices', 'Accepted After Initial Rejection'}) => frozenset({'Cash loans', 'Cash'})
	Row - 5565 - Antecedent :Repeater, Credit and cash offices, Accepted After Initial Rejection
	Row - 5565 - Consequent :Cash loans, Cash
	Row - 5565 - Support :0.167840577
	Row - 5565 - Confidence :0.803453947
	Row - 5565 - Lift :2.478487243

Thank You.

Reference

- kapoorshivam. (2021, February 20). Credit Analysis using EDA. Kaggle.com; Kaggle. https://www.kaggle.com/code/kapoorshivam/credit-analysis-using-eda/input
- Kaggle. (2022). *Kaggle: Your Home for Data Science*. Kaggle.com. https://www.kaggle.com/