Title: Credit Risk Assessment

CSE 343: Machine Learning Mid-Project Presentation

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation

The lending landscape has undergone a significant transformation with the advent of platforms like LendingClub, the world's largest peer-to-peer lending platform.

In response to a surge in loan applications, the importance of precise credit risk assessment has never been more pronounced. This report presents a comprehensive case study focused on LendingClub, a pioneering peer-to-peer lending company based in the United States.

Our primary objective is to uncover the intricacies of risk analytics within the banking and financial services sector, particularly within the context of urban customers.

At the heart of this challenge lies the decision-making process upon receiving a loan application. Striking a balance between two types of risks is imperative:

1. Potential Loss of Business: Denying a loan to a credible applicant directly translates to a loss of business for the company.

2. Risk of Default: Conversely, approving a loan for an applicant likely to default could lead to financial repercussions.

The provided dataset encompasses historical data of past loan applicants and their respective repayment behaviors.

Our aim is to unearth discernible patterns that act as indicators of default.

These insights could play a pivotal role in making informed decisions, such as adjusting loan amounts, assigning higher interest rates to riskier applicants, or even in some cases, denying loans.

Through a thorough examination of the data using exploratory data analysis (EDA) and leveraging machine learning techniques, we aspire to construct a robust framework for identifying high-risk loan applicants.

Ultimately, our goal is to empower lending institutions with the knowledge and tools to minimize credit loss and optimize their lending portfolios.

Literature Review

In the realm of credit risk assessment, significant strides have been made, with researchers exploring various facets of lending risk prediction. This section provides an overview of seminal works that have paved the way for understanding and addressing challenges in this domain.

1. Consumer Credit Risk Assessment - Jonathan N. Crook, David B. Edelman, and Lyn C. Thomas delve into the realm of classification algorithms and profit scoring, shedding light on their application in the prediction of lending risks [1]

Methods -

This study focuses on consumer credit risk assessment, employing logistic regression and various classifiers. While support vector machines offer high accuracy, practical application is impeded by data quality issues. Training on accepted applicants introduces challenges in setting cut-off values. Profit scoring and the Basel 2 accord significantly influence research directions.

Conclusion -

This review delves into contemporary research in consumer credit risk assessment, emphasizing logistic regression's prevalence. It underscores credit scoring's evolution and the Basel 2 Accord's impact on global banking regulations. The study sets the stage for refining credit risk assessment models.

2. P2P Loan Acceptance & Default Prediction with AI Kenneth Kennedy’s thesis provides valuable insights into the challenges faced in the development of credit scorecards, with a particular focus on issues like class imbalance and low-default portfolios [2]

Methods -

This thesis conducts a comprehensive analysis of credit scoring, a critical aspect of lending decisions for financial institutions. It addresses challenges in developing accurate credit scorecards, focusing on imbalanced datasets. The study evaluates various supervised classification techniques, highlighting the intricacies of class imbalance. Additionally, it addresses the low-default portfolio problem by comparing semi-supervised classification methods with logistic regression.

Conclusion -

This research underscores the paramount importance of credit scoring in lending decisions for financial institutions and the broader economy. It highlights four key contributions, including insights into handling imbalanced datasets and addressing the low-default portfolio problem. The study emphasizes the need for precision in dataset implementation and introduces artificial data as a resourceful solution when real-world data is unavailable.

3. Deep Learning Credit Risk Modeling - J. D. Turiel and T. Aste employ cutting-edge deep learning techniques to forecast loan acceptance and default risk in peer-to-peer lending credit risk models [3].

Methods -

This research employs logistic regression (LR), support vector machine algorithms, and deep neural networks (DNNs) to replicate lending decisions and predict default probabilities. A two-phase model is proposed, where the first phase predicts loan rejection, and the second predicts default risk for approved loans. LR performs exceptionally in the first phase, achieving a test set recall macro score of 77.4%.

Conclusion -

This study underscores the critical role of credit scoring in lending decisions for financial institutions and the broader economy. It introduces a novel two-phase model for loan assessment, demonstrating the effectiveness of LR and DNNs. The research not only advances credit risk modeling but also presents a pathway to making credit access fairer and more accessible, particularly for small businesses.

Dataset Description

Utilized the LendingClub loan data from Kaggle spanning 2007 to 2018.

The original dataset had 2 million rows and 151 features.

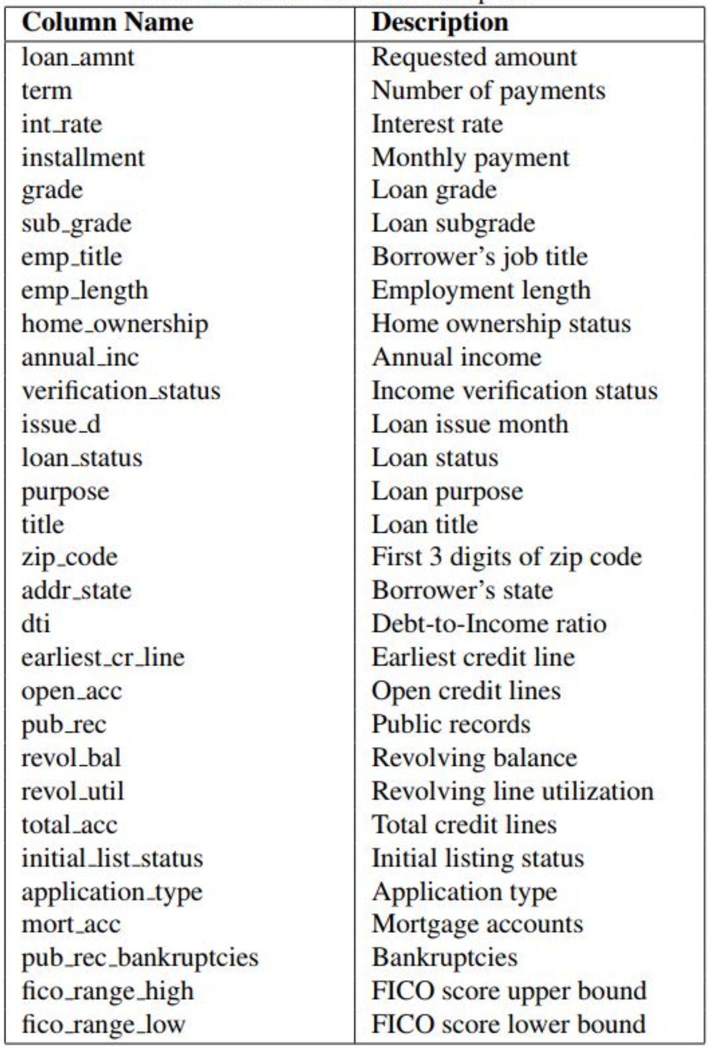
Carefully selected relevant features based on the LC dictionary.

- Excluded features with substantial missing values for precision.
- Ensured each chosen feature had a clear definition and relevance.

Performed In -depth EDA to gain valuable insights from the data

Attributes

Table 1. Selected Columns Description



Column Name	Description
loan_amnt	Requested amount
term	Number of payments
int_rate	Interest rate
installment	Monthly payment
grade	Loan grade
sub_grade	Loan subgrade
emp_title	Borrower's job title
emp_length	Employment length
home_ownership	Home ownership status
annual_inc	Annual income
verification_status	Income verification status
issue_d	Loan issue month
loan_status	Loan status
purpose	Loan purpose
title	Loan title
zip_code	First 3 digits of zip code
addr_state	Borrower's state
dti	Debt-to-Income ratio
earliest_cr_line	Earliest credit line
open_acc	Open credit lines
pub_rec	Public records
revol_bal	Revolving balance
revol_util	Revolving line utilization
total_acc	Total credit lines
initial_list_status	Initial listing status
application_type	Application type
mort_acc	Mortgage accounts
pub_rec_bankruptcies	Bankruptcies
fico_range_high	FICO score upper bound
fico_range_low	FICO score lower bound

A target variable (loan status) is defined to include only ’Fully Paid’ and ’Charged Off’ statuses.

Focused on loans with conclusive outcomes, excluding ongoing loans and credit policy exceptions.

Feature Selection & Feature Engineering:

Removed certain features, such as issue_d, to prevent data leakage.

Focused on features available at the time of loan origination to ensure accurate assessments.

Constructed an additional feature by computing the average FICO score utilizing the 'fico_range_low' and 'fico_range_high' attributes, enhancing the model's predictive capability.

Data Encoding

One-Hot Encoding - Applied one-hot encoding to specific categorical features, including 'verification_status,' 'purpose,' 'initial_list_status,' 'application_type,' 'home_ownership,' and 'addr_state,' creating dummy variables for each category.

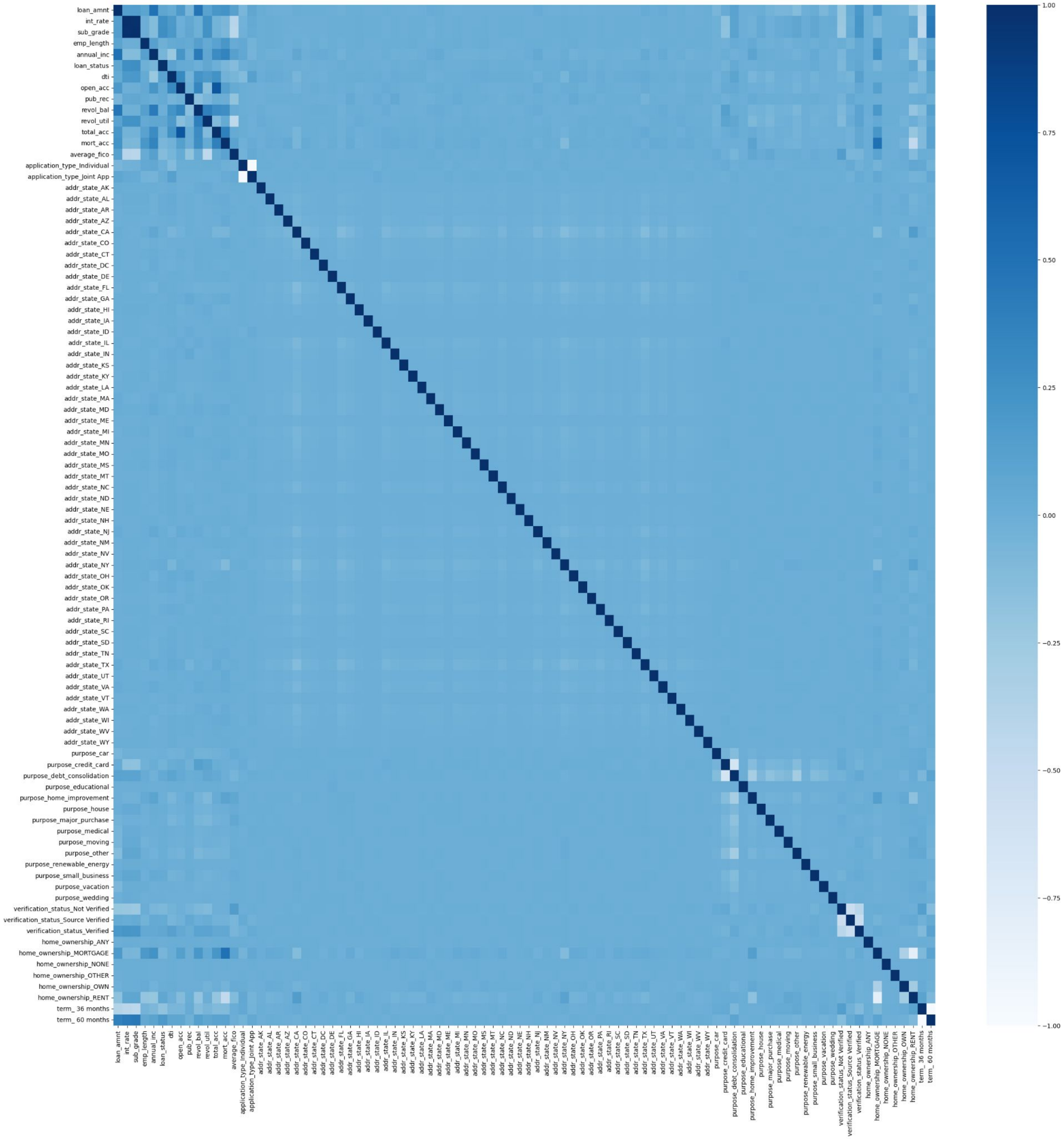
Label Encoding - Utilized label encoding with numerical conversion for categorical variables such as 'emp_length' and 'sub_grade,' ensuring compatibility with the model's requirements.

Data Encoding

Conversion to Numerical Data - Cleaned and converted selected columns to numeric format, ensuring their suitability for further analysis and modeling.

Data Cleaning - Eliminated rows with missing values amounting to less than 5% of total data in the newly processed features, ensuring a cleaner and more complete dataset for analysis.

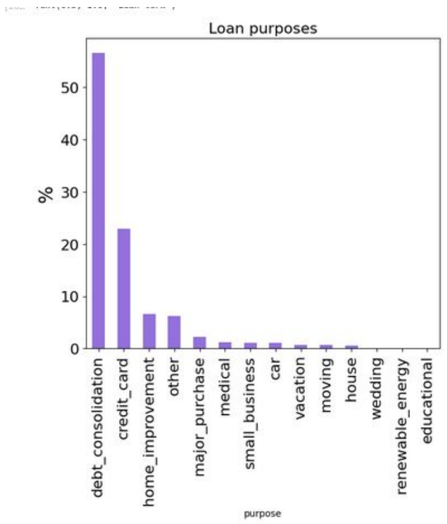
Visualization

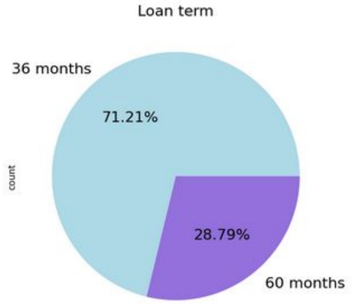


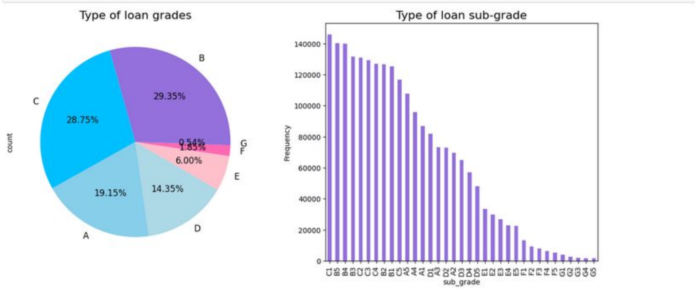
Heatmap to verify the correlation between columns.

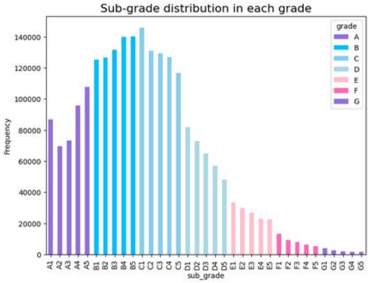
Majority of the columns are not highly correlated.

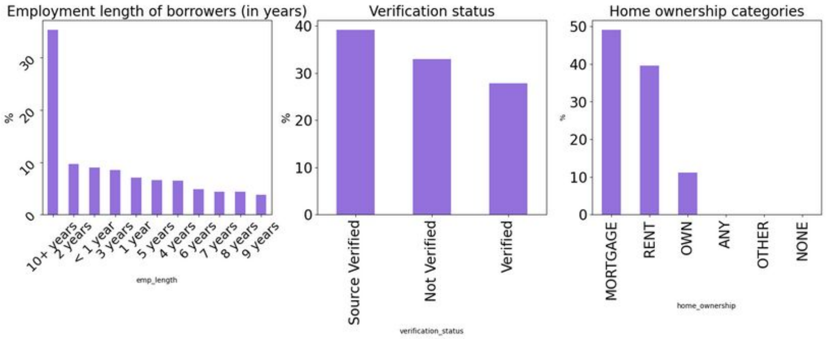
Visualization (EDA)

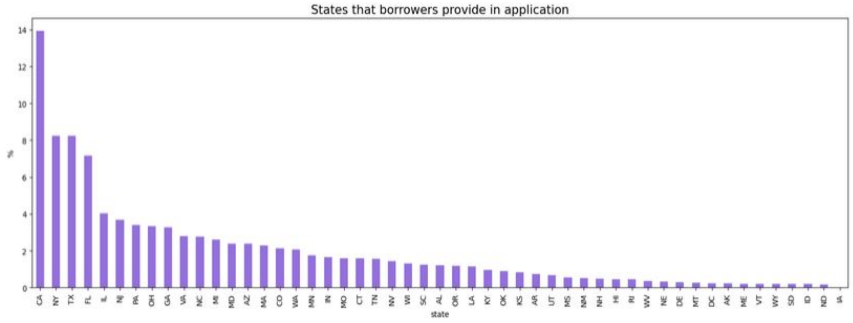




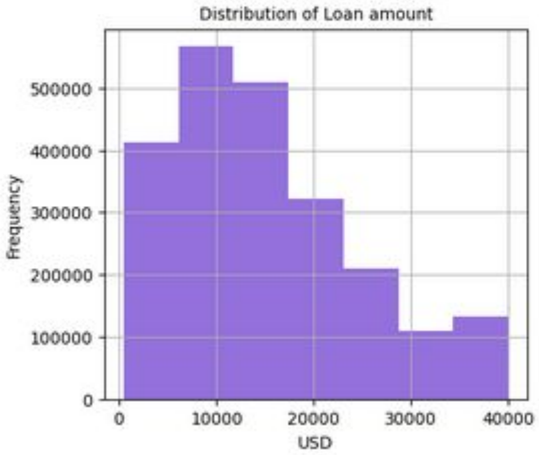


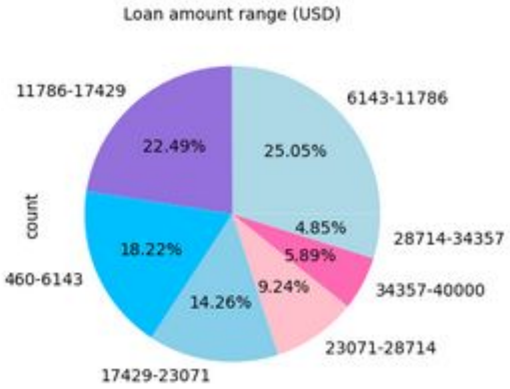


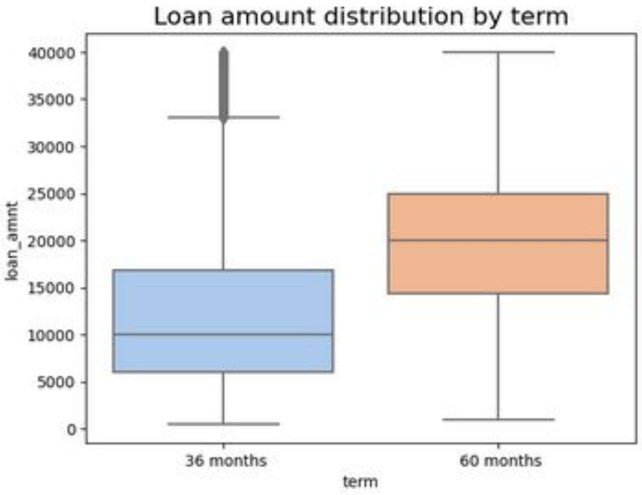


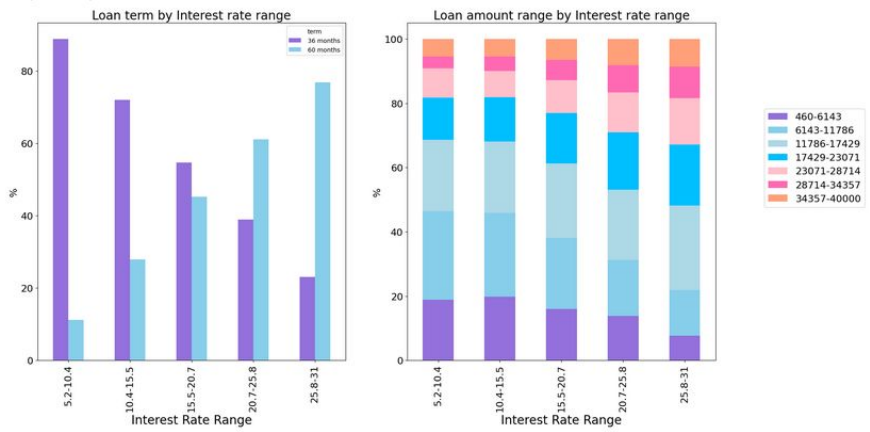


Visualization (EDA)







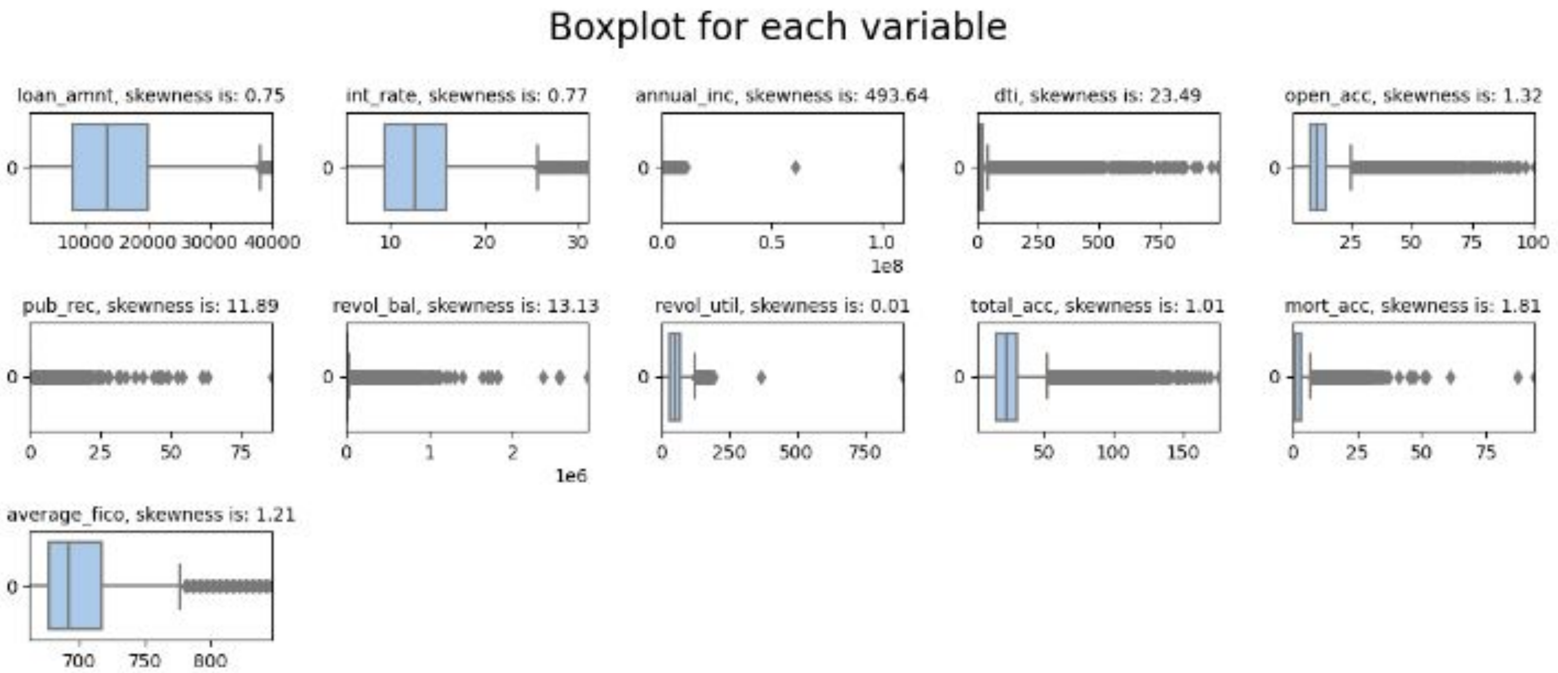


EDA Inferences

1. **Loan Amount Trends:**
   - Varied loan amounts from $460 to $40,000 USD.
   - Common range: $6,143-$11,786, indicating a prevalent choice.
   - Skewed distribution, fewer loans above $17,429.
2. **Interest Rate Insights:**
   - 41.4% with interest rates between 10.4% and 15.5%.
   - Right-skewed loan amount data suggests rates of 5.2% to 15.5%.
3. **FICO Scores and Applicants:**
   - 71% with good FICO scores; 675 is most frequent.
   - Rarity in applicants with FICO scores above 800.
4. **Loan Purpose and Term Preferences:**
   - Majority for debt consolidation and credit cards.
   - 71.2% prefer 36-month terms over 60 months.
5. **Loan Grades and Risk Assessment:**
   - Primary grades: A (19%), B (29%), C (28%).
   - Higher-risk categories (E, F, G) have fewer loans.
6. **Applicant Characteristics:**
   - 10% unemployed or less than a year of work; 35% employed for over ten years.
   - 50% have a mortgage; 40% pay rent.
7. **Loan Status and Types:**
   - 48% fully paid; 40% current; 30% charged off.
   - Majority (94%) are individual applications; 6% joint applications.

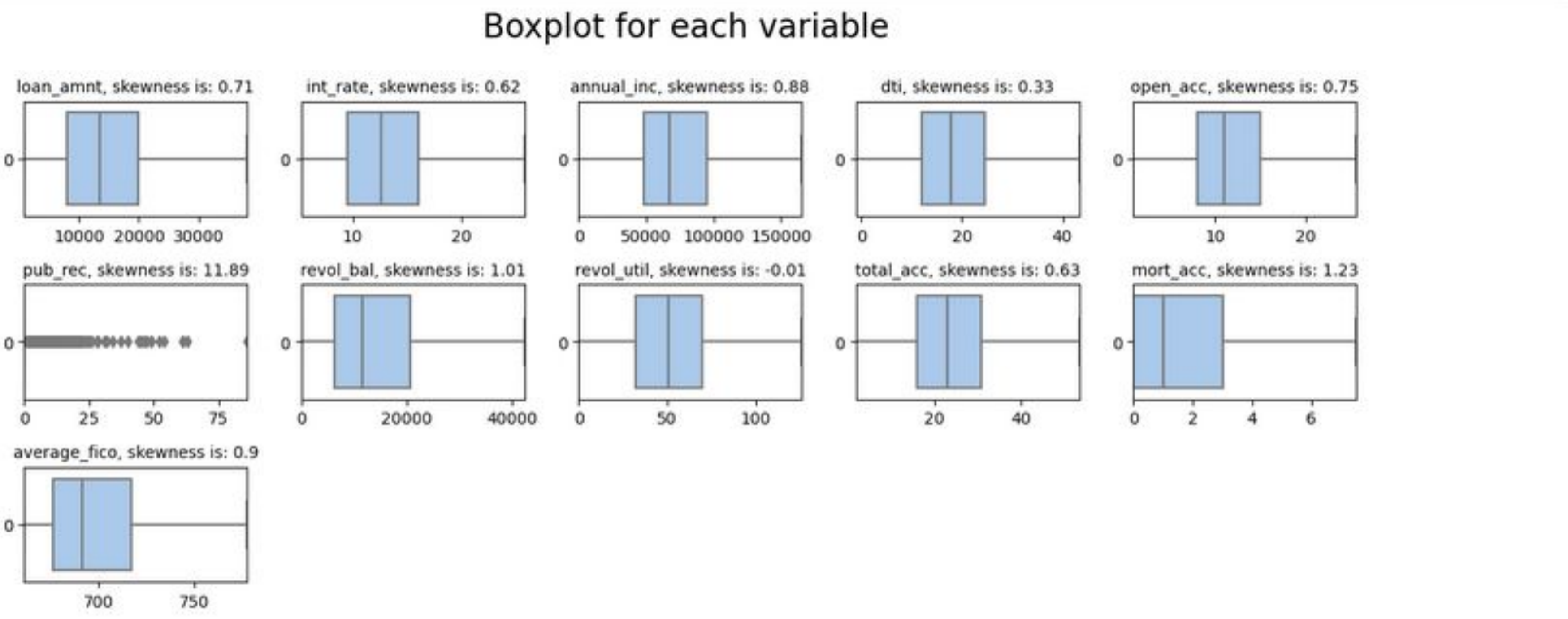
Visualization

Box plots to check for potential outliers.



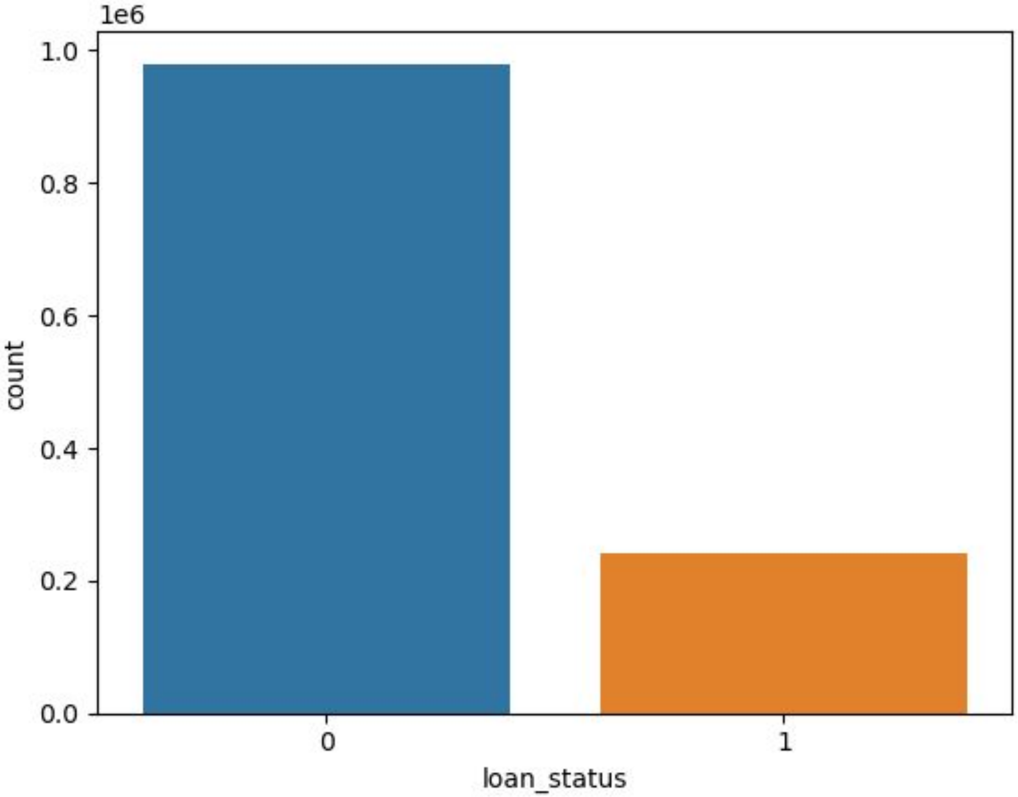
Visualization

Box plots of features after outlier capping process



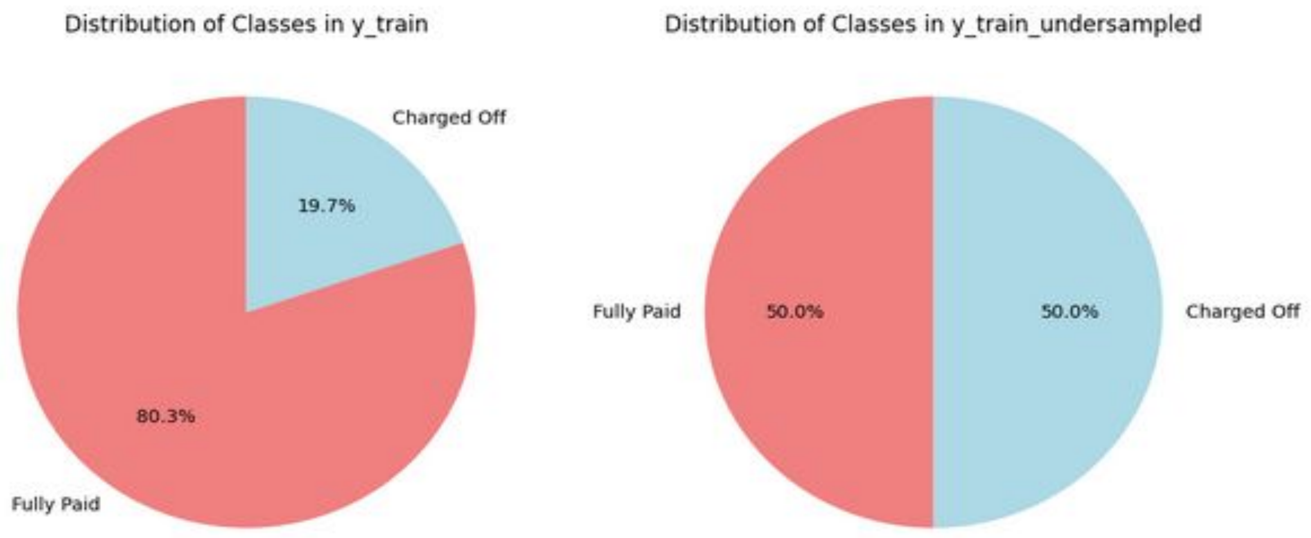
Visualization

Observing the imbalance in dataset.



Visualization

Mitigated class imbalance through random undersampling of the data.



Methodology

Logistic Regression - We used logistic regression model for binary classification using L2 regularization and log loss with inverse of regularization constant = 1. This model is evaluated using accuracy, confusion matrix, and an ROC curve.

Random Forest - We used random forest of decision trees with gini criterion. This model is evaluated using ROC-AUC score.

Naive Bayes - We also used Naive Bayes model for our prediction. This model is evaluated based on accuracy, a confusion matrix and a classification report.

Methodology

K-nearest neighbor: A K-Nearest Neighbors Classifier is trained after data imputation and standardization. This model is evaluated based on ROC-AUC score.

Support Vector Machine (SVM): This model is evaluated using accuracy, a confusion matrix, and a ROC curve.

Methodology

XG-Boost: This model is evaluated using accuracy, a confusion matrix, a training curve vs epochs, and an ROC curve.

ANN: This model is also evaluated using accuracy, a confusion matrix, a training curve vs epochs, and an ROC curve.

Note -

All the models are based on under-sampled data.

Analysis

Logistic Regression -

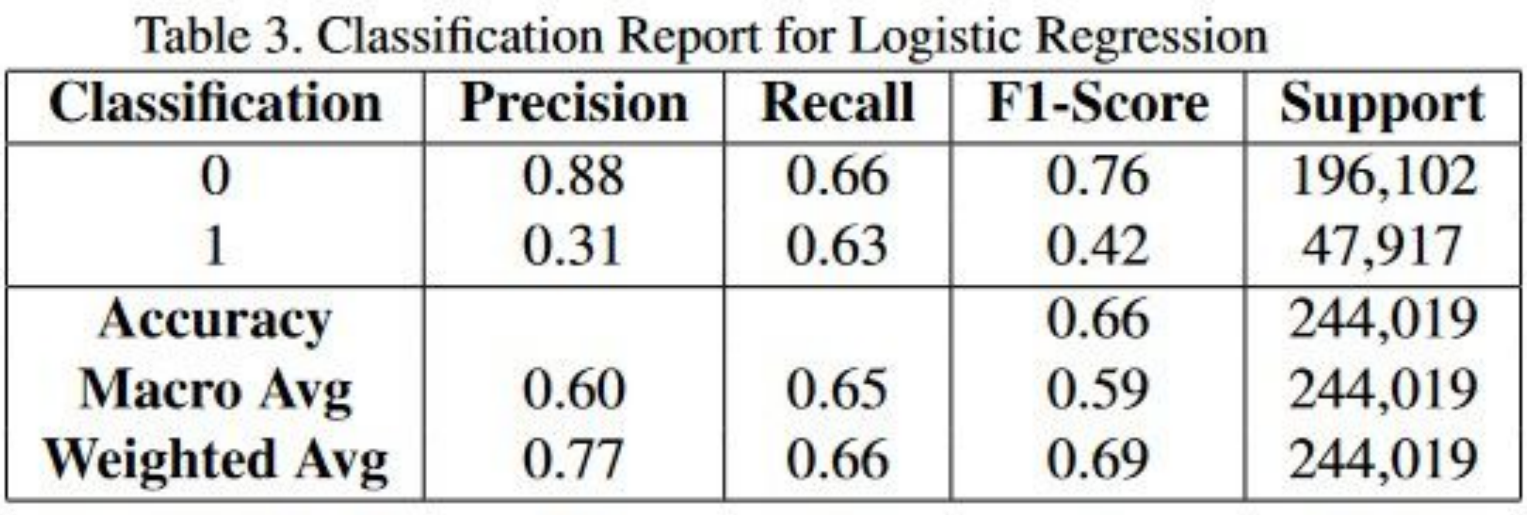
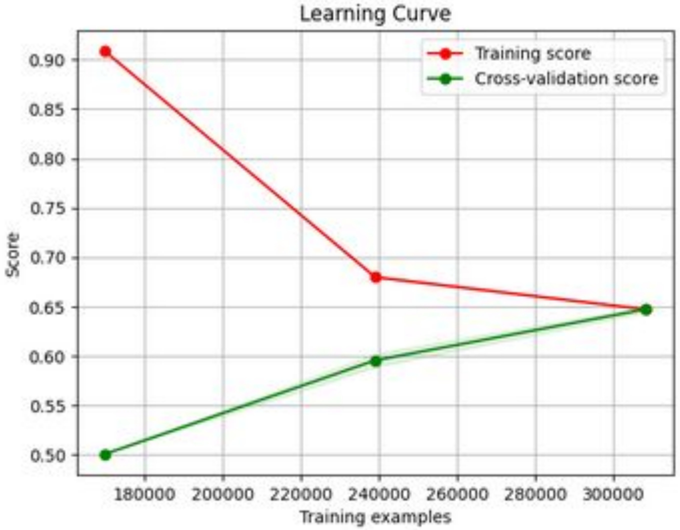


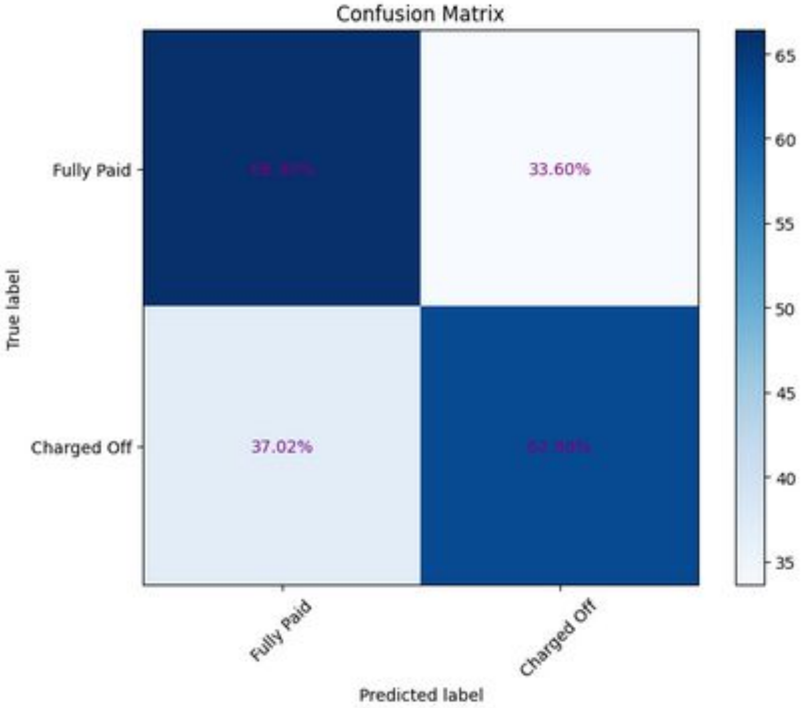
Table 3. Classification Report for Logistic Regression

Classification	Precision	Recall	F1-Score	Support
0	0.88	0.66	0.76	196,102
1	0.31	0.63	0.42	47,917
Accuracy			0.66	244,019
Macro Avg	0.60	0.65	0.59	244,019
Weighted Avg	0.77	0.66	0.69	244,019

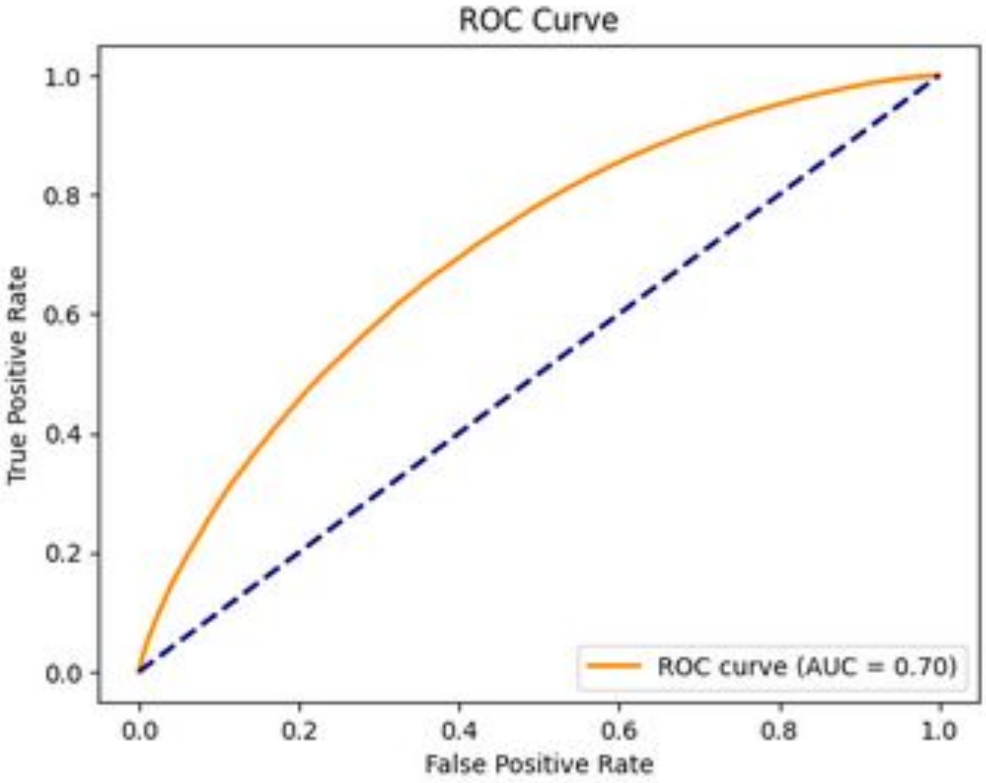
Analysis

Logistic Regression -





Logistic Regression -



Random Forest -

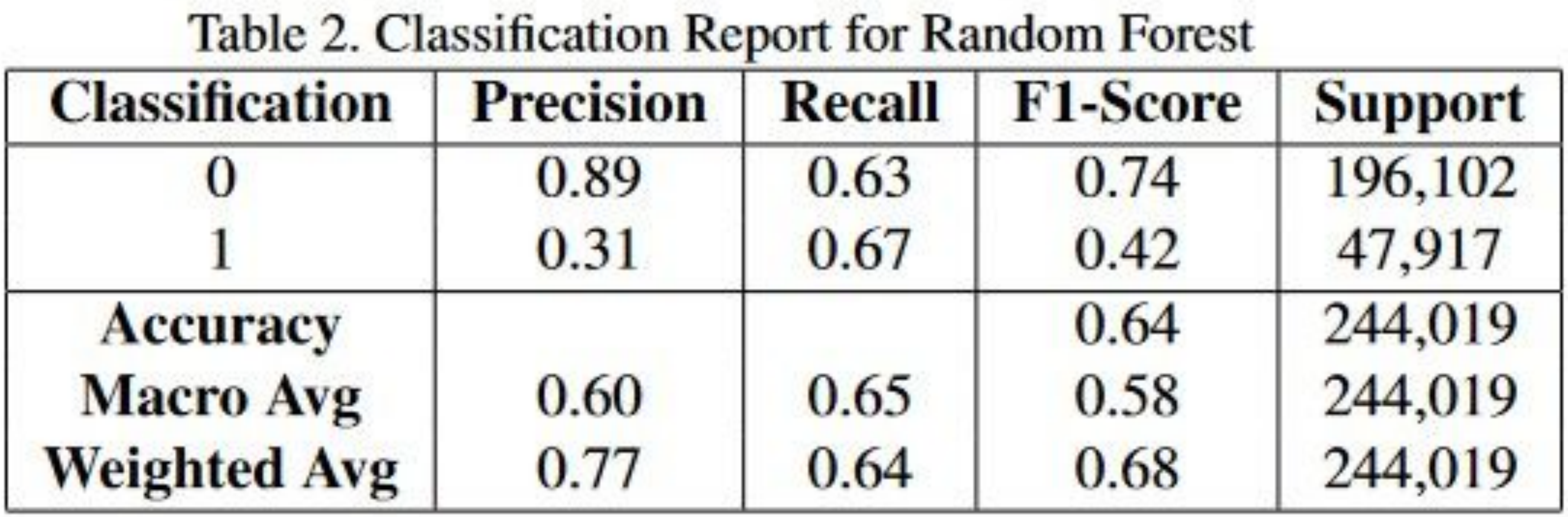
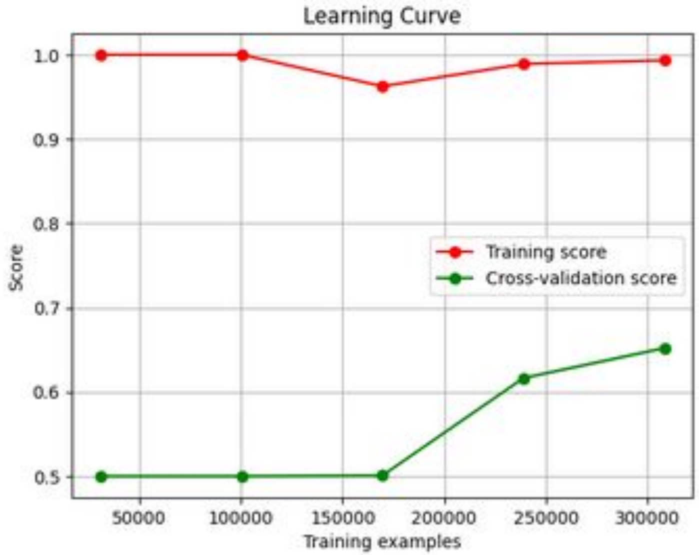


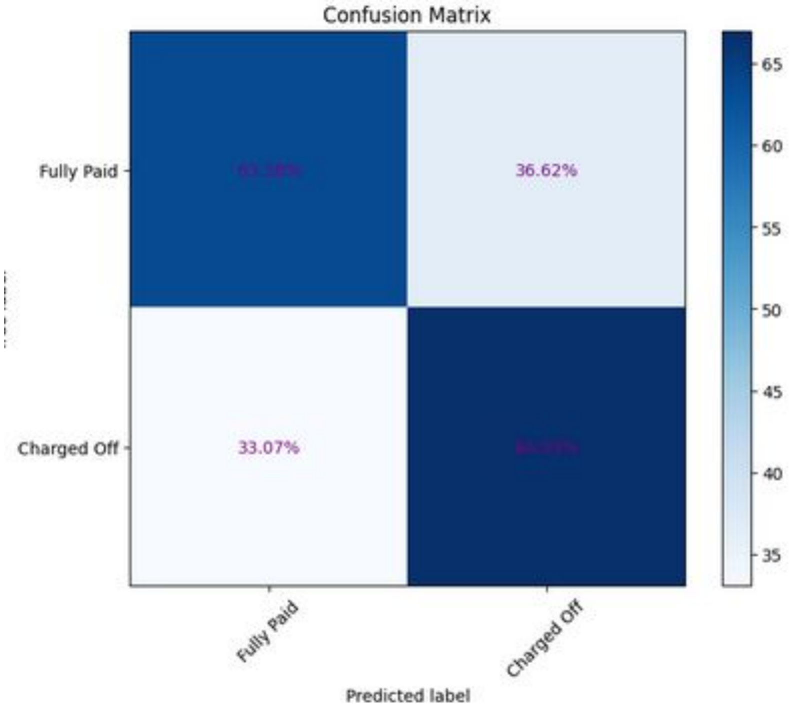
Table 2. Classification Report for Random Forest

Classification	Precision	Recall	F1-Score	Support
0	0.89	0.63	0.74	196,102
1	0.31	0.67	0.42	47,917
Accuracy			0.64	244,019
Macro Avg	0.60	0.65	0.58	244,019
Weighted Avg	0.77	0.64	0.68	244,019

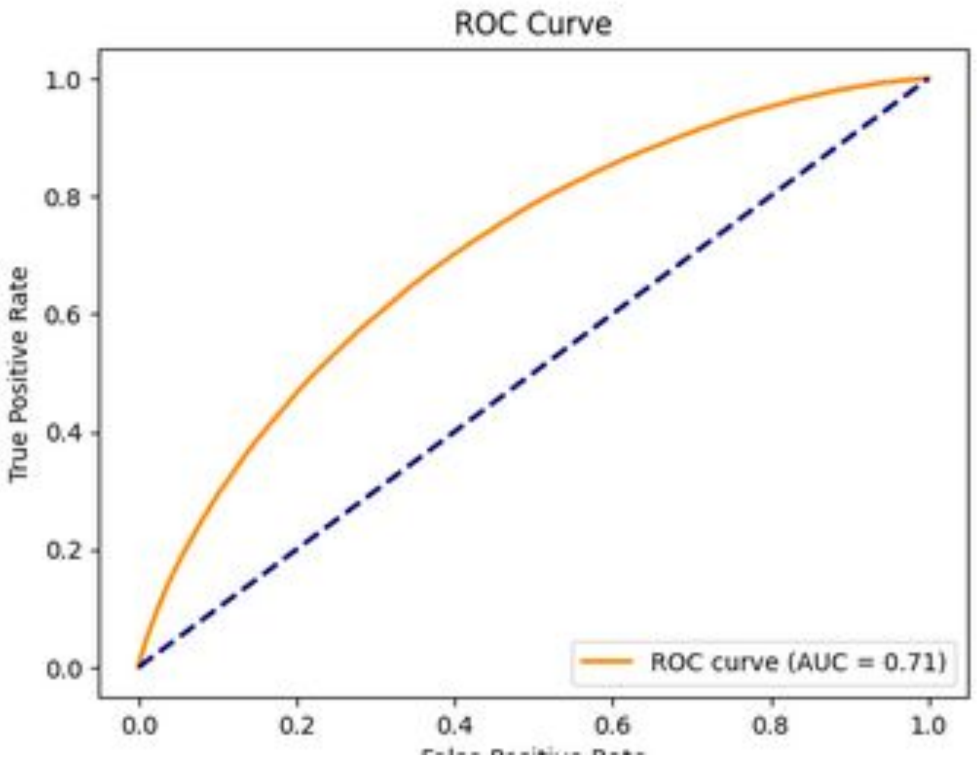
Analysis

Random Forest-





Random Forest -



K - Nearest Neighbors -

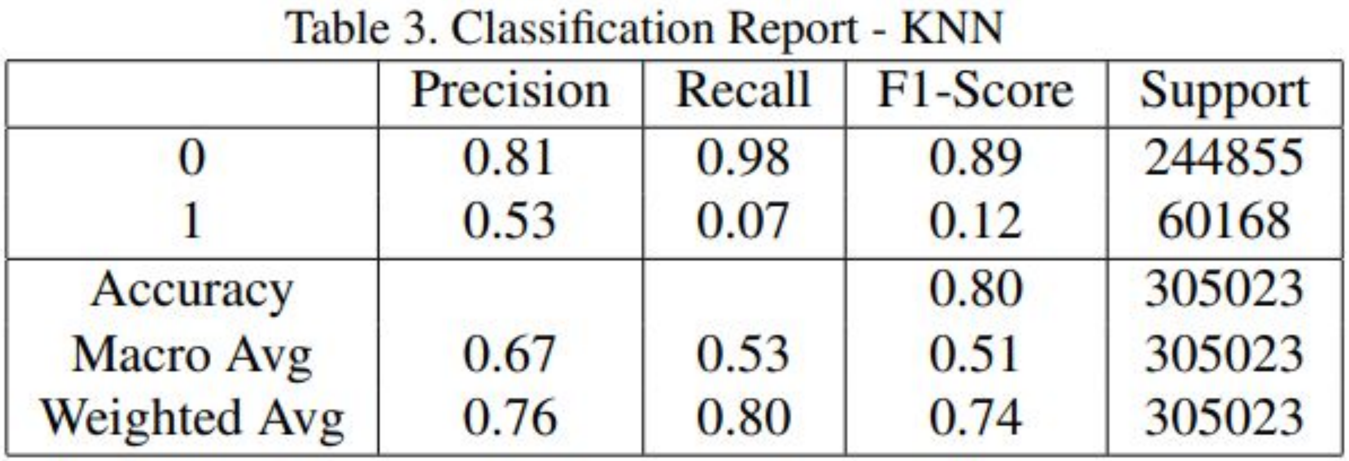
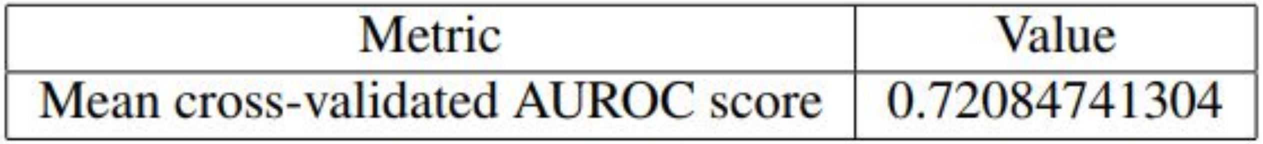


Table 3. Classification Report - KNN

	Precision	Recall	F1-Score	Support
0	0.81	0.98	0.89	244855
1	0.53	0.07	0.12	60168
Accuracy			0.80	305023
Macro Avg	0.67	0.53	0.51	305023
Weighted Avg	0.76	0.80	0.74	305023



Metric	Value
Mean cross-validated AUROC score	0.72084741304

XG Boost -

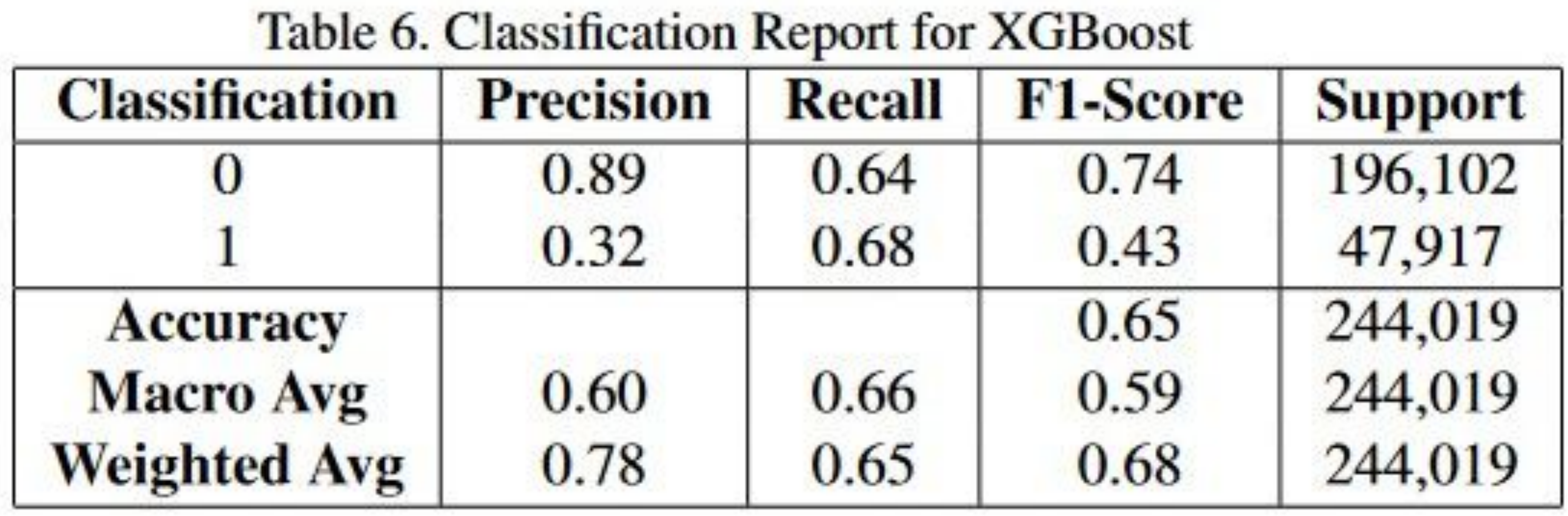
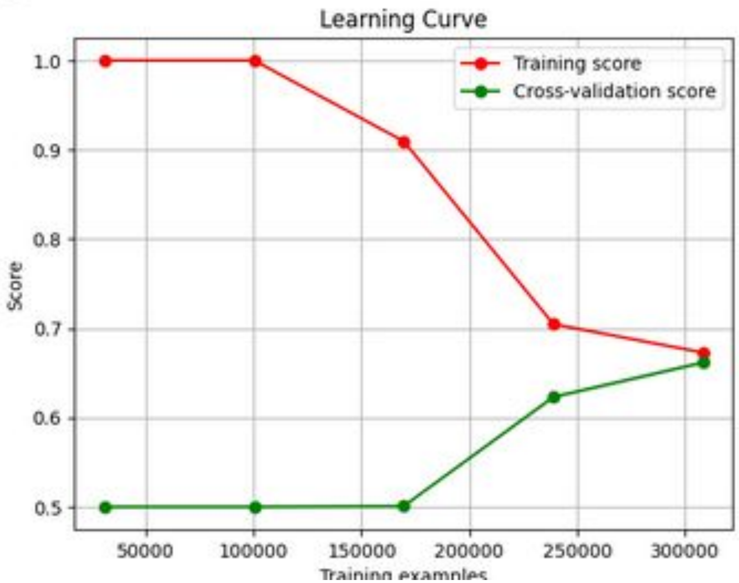
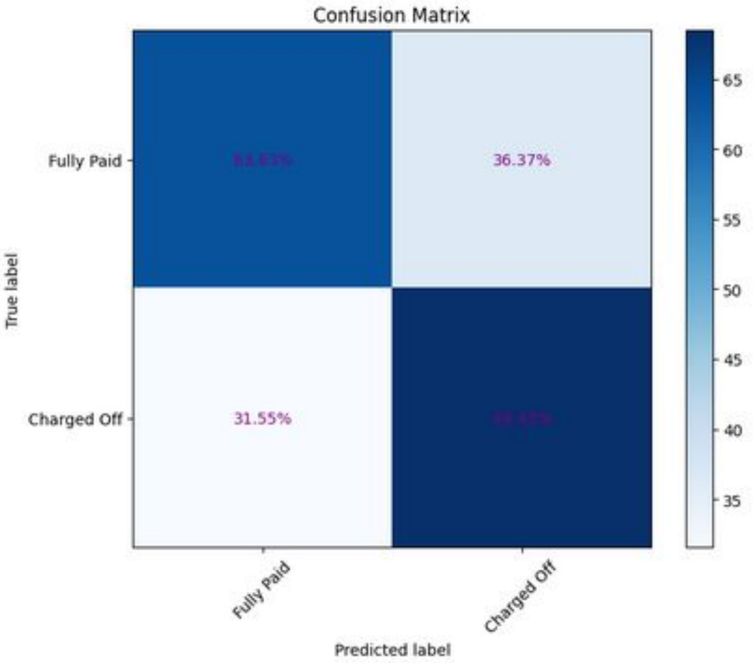


Table 6. Classification Report for XGBoost

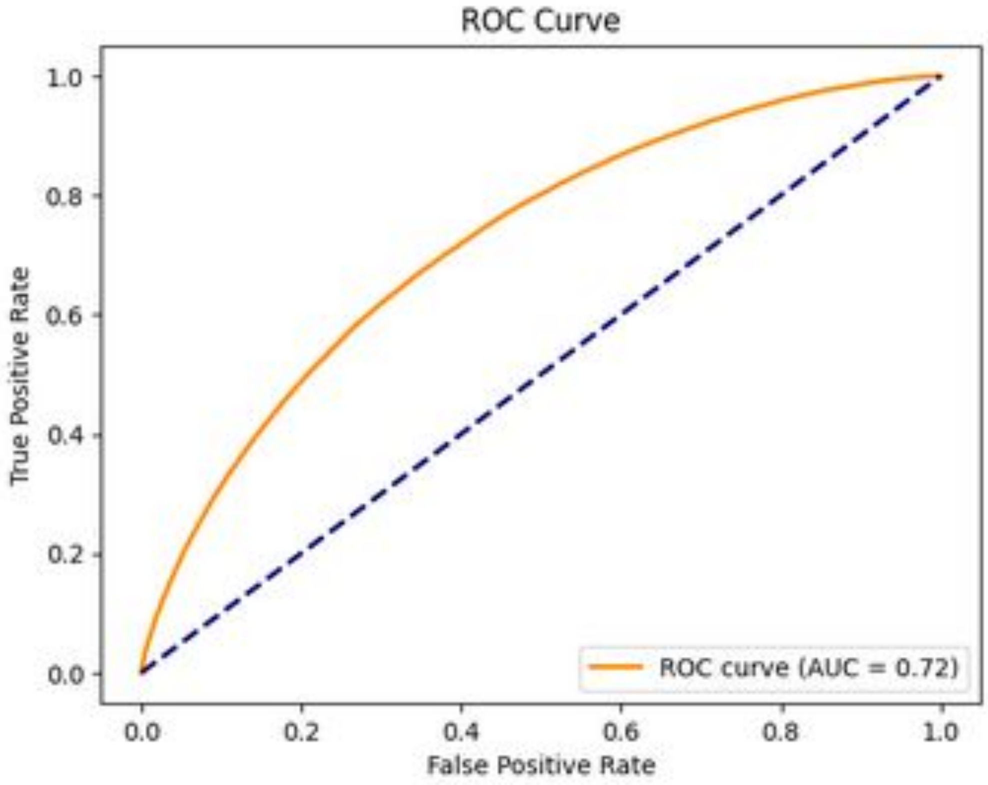
Classification	Precision	Recall	F1-Score	Support
0	0.89	0.64	0.74	196,102
1	0.32	0.68	0.43	47,917
Accuracy			0.65	244,019
Macro Avg	0.60	0.66	0.59	244,019
Weighted Avg	0.78	0.65	0.68	244,019

XG Boost -


Learning Curve
Training score
Cross-validation score
Score
1.0
0.9
0.8
0.7
0.6
0.5
50000
100000
150000
200000
250000
300000
Training examples

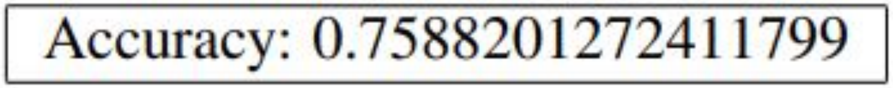

Confusion Matrix
True label
Fully Paid
Charged Off
63.63%
36.37%
31.55%
68.45%
Predicted label
Fully Paid
Charged Off
65
60
55
50
45
40
35

XG Boost -



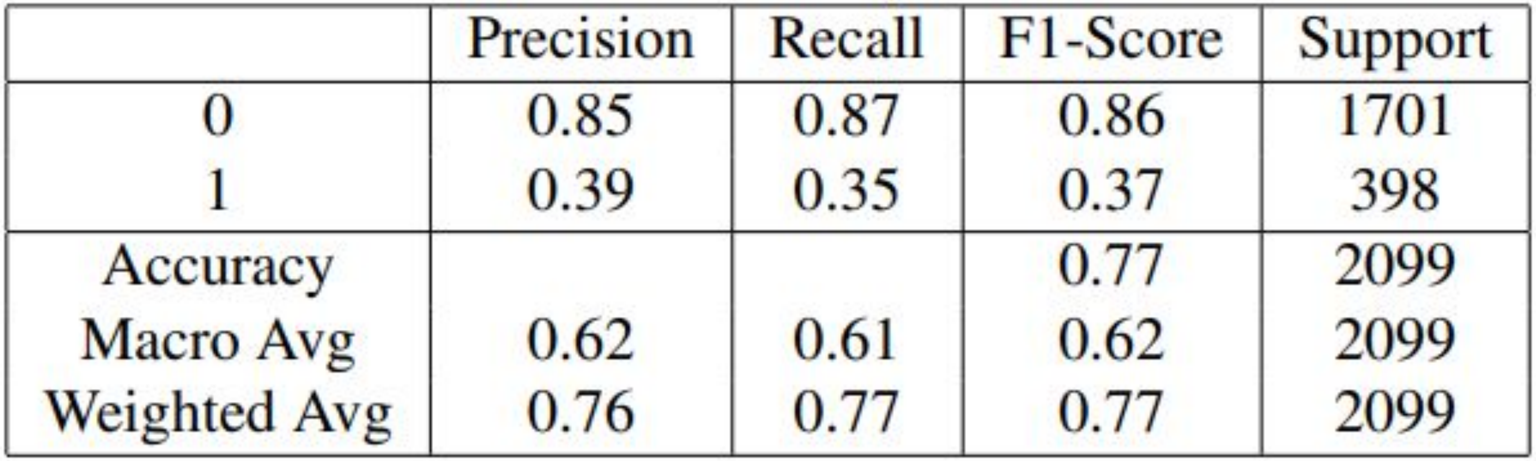
```
Accuracy for XGBoost: 64.50%
```

SVM -



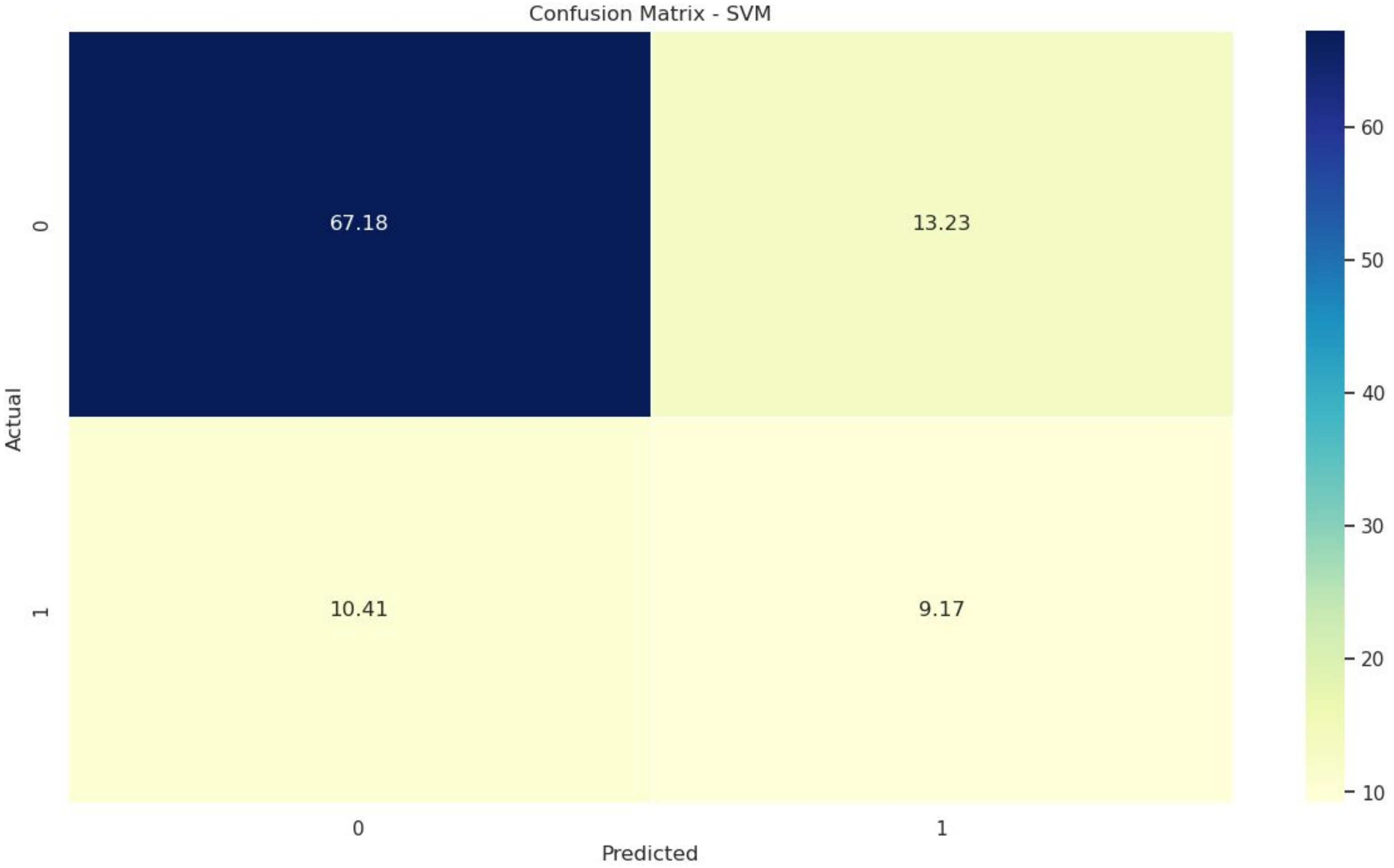
```
Accuracy: 0.7588201272411799
```

Table 4. Classification Report - SVM

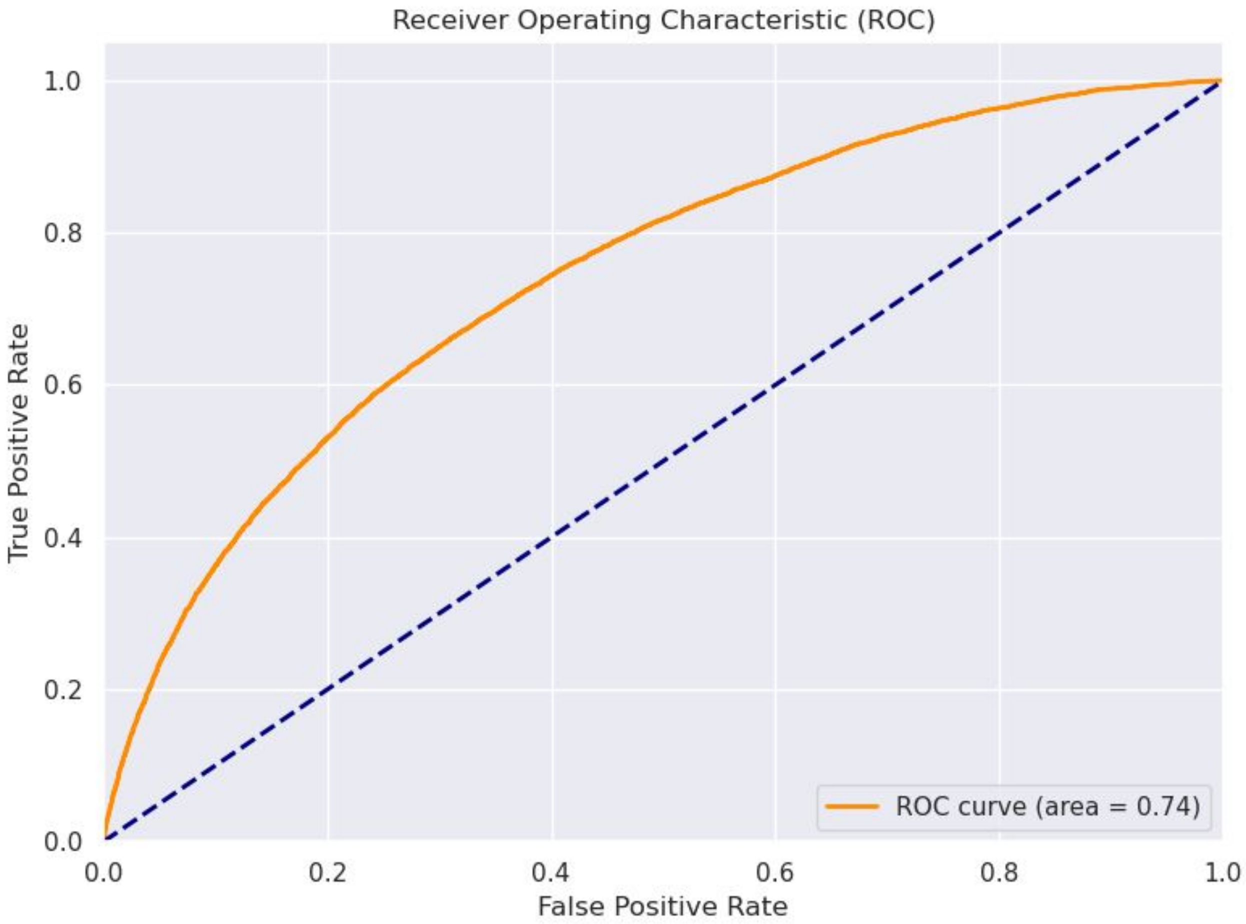


	Precision	Recall	F1-Score	Support
0	0.85	0.87	0.86	1701
1	0.39	0.35	0.37	398
Accuracy			0.77	2099
Macro Avg	0.62	0.61	0.62	2099
Weighted Avg	0.76	0.77	0.77	2099

SVM -


Confusion Matrix - SVM
67.18
13.23
10.41
9.17
Actual
0
1
Predicted
0
1
60
50
40
30
20
10

SVM -


Receiver Operating Characteristic (ROC)
True Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0
ROC curve (area = 0.74)

Naive Bayes -

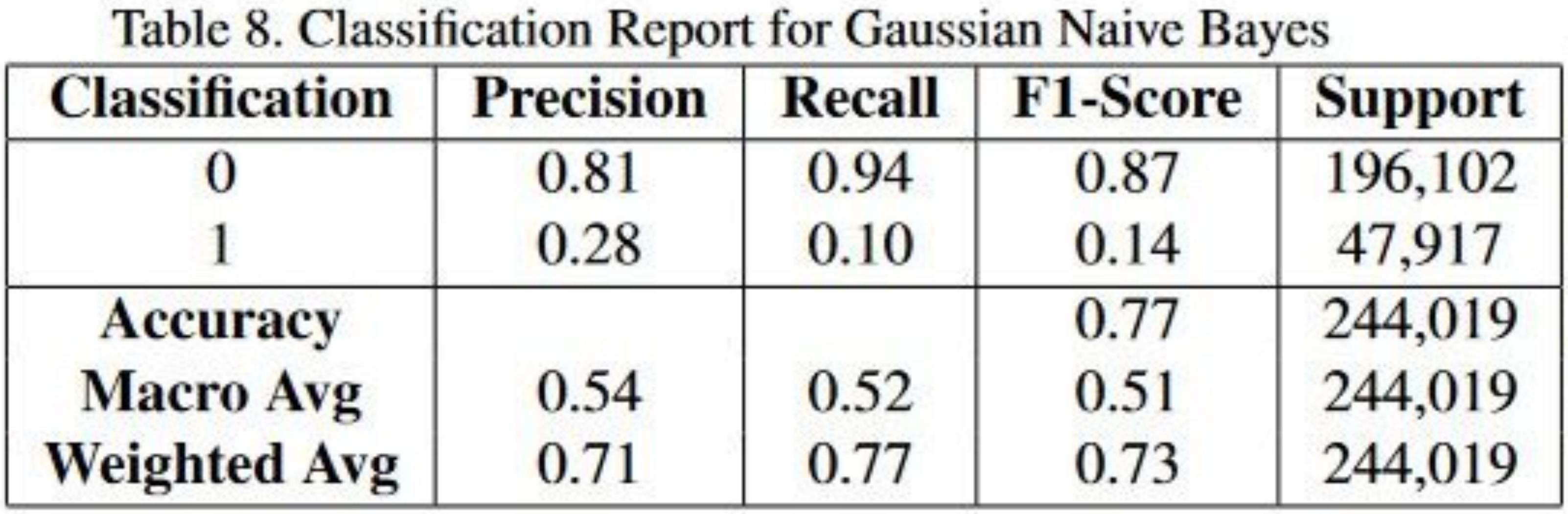
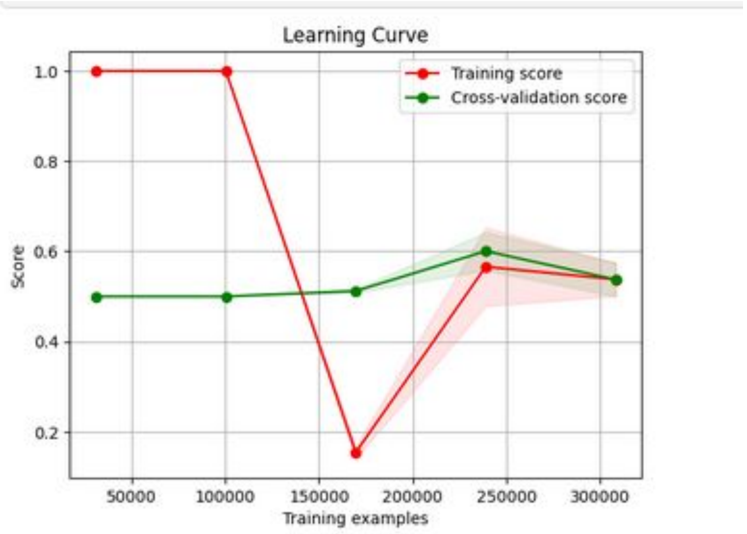
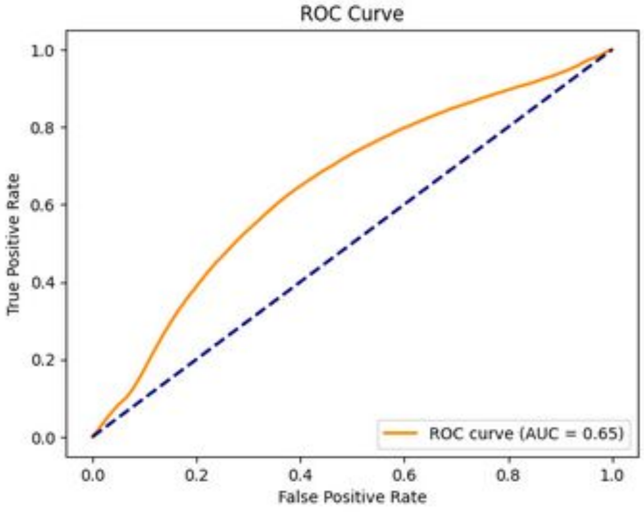


Table 8. Classification Report for Gaussian Naive Bayes

Classification	Precision	Recall	F1-Score	Support
0	0.81	0.94	0.87	196,102
1	0.28	0.10	0.14	47,917
Accuracy			0.77	244,019
Macro Avg	0.54	0.52	0.51	244,019
Weighted Avg	0.71	0.77	0.73	244,019

Naive Bayes -


Learning Curve
Training score
Cross-validation score
Score
1.0
0.8
0.6
0.4
0.2
50000
100000
150000
200000
250000
300000
Training examples


ROC Curve
True Positive Rate
1.0
0.8
0.6
0.4
0.2
0.0
ROC curve (AUC = 0.65)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate

ANN -

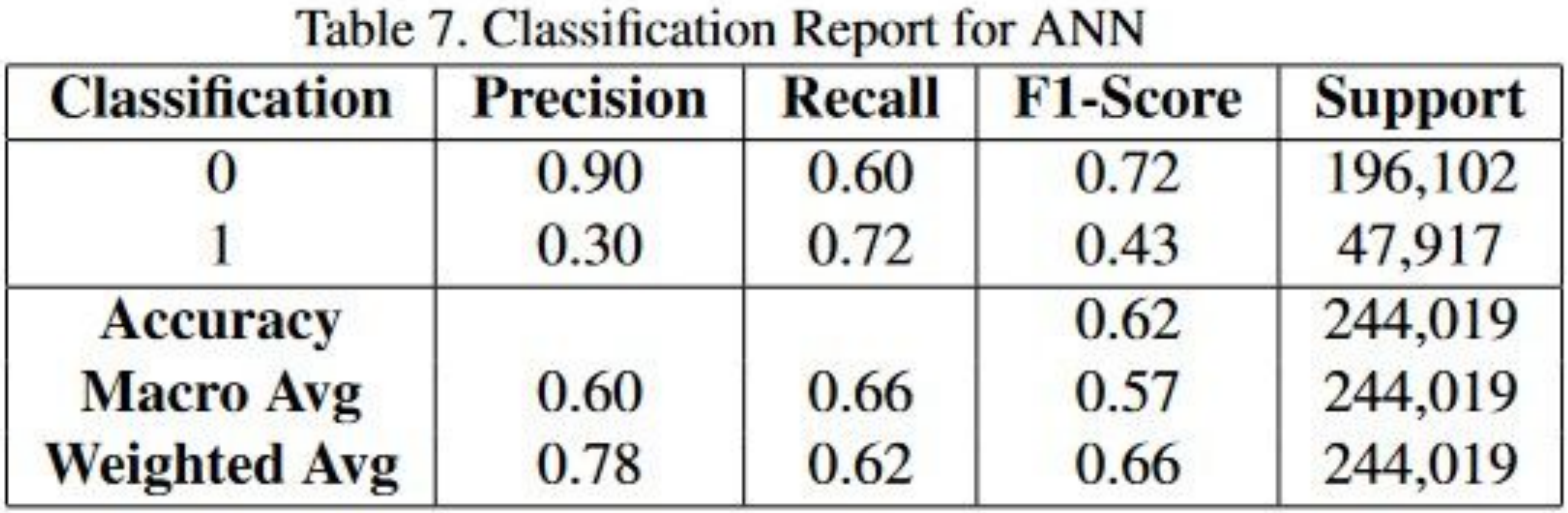
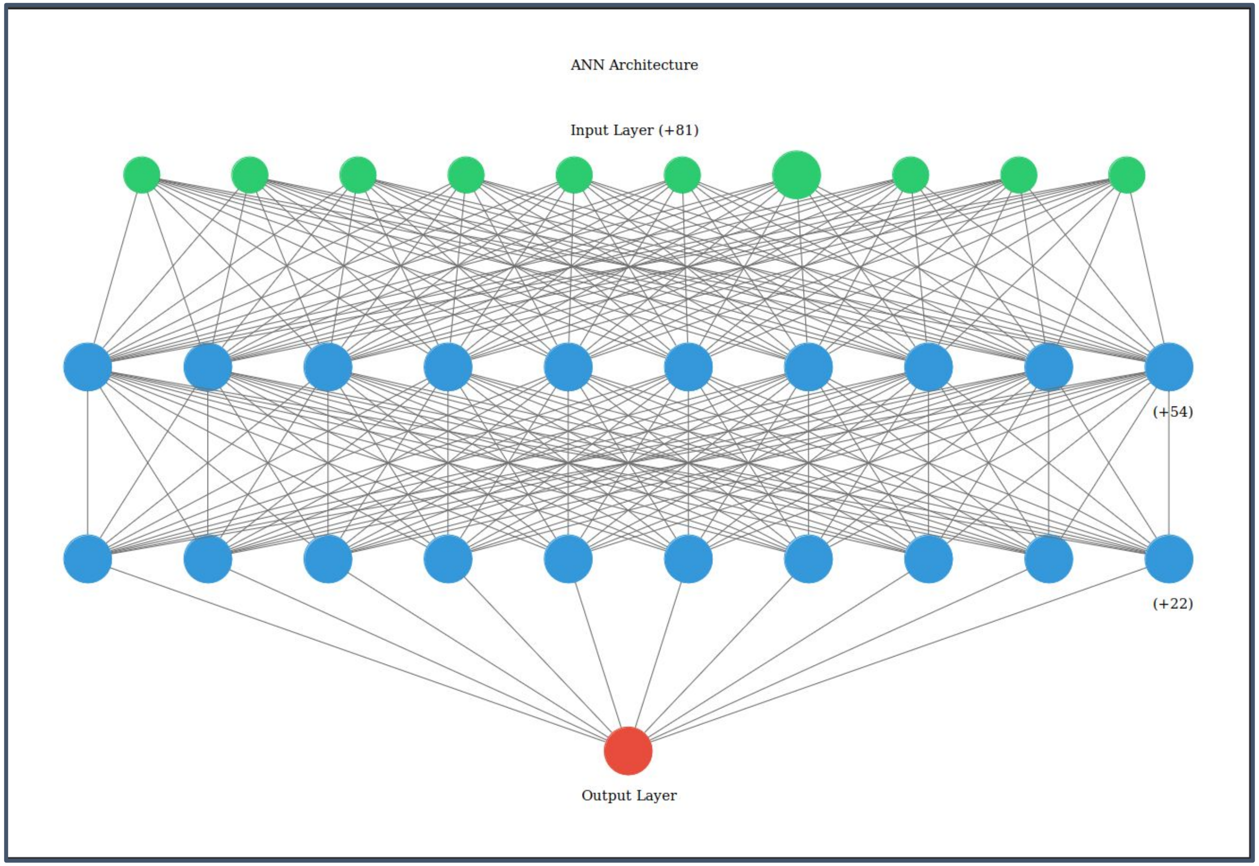


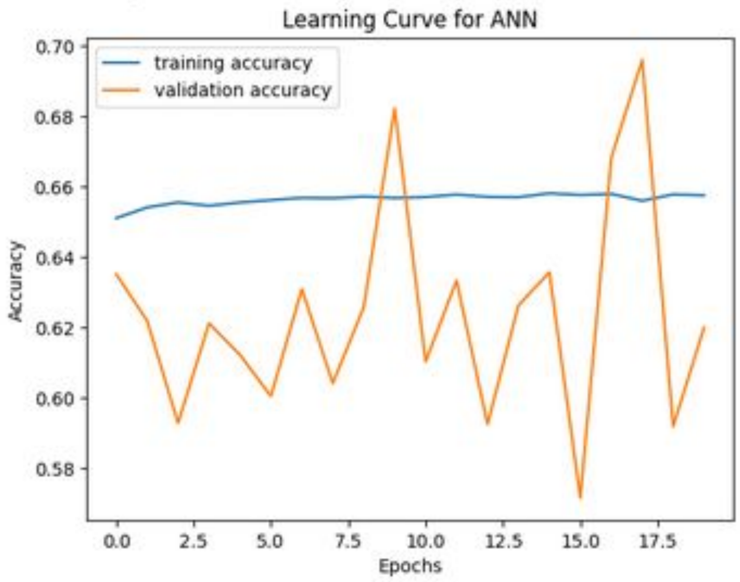
Table 7. Classification Report for ANN

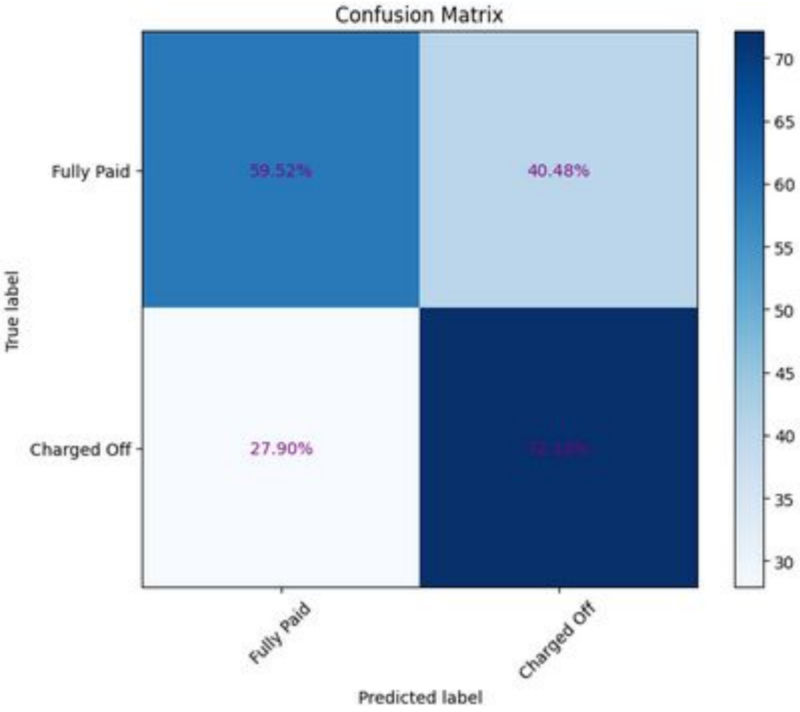
Classification	Precision	Recall	F1-Score	Support
0	0.90	0.60	0.72	196,102
1	0.30	0.72	0.43	47,917
Accuracy			0.62	244,019
Macro Avg	0.60	0.66	0.57	244,019
Weighted Avg	0.78	0.62	0.66	244,019

ANN -



ANN -


Learning Curve for ANN
training accuracy
validation accuracy
Accuracy
0.70
0.68
0.66
0.64
0.62
0.60
0.58
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
Epochs


Confusion Matrix
True label
Fully Paid
Charged Off
59.52%
40.48%
27.90%
72.10%
Predicted label
Fully Paid
Charged Off
70
65
60
55
50
45
40
35
30

Inference

Logistic Regression model was the most effective because of it’s **higher precision**, thus, **lowering actual defaults.**

Lowering defaults is one of the main goals for this prediction.

XG Boost is next to **Logistic Regression** in terms of performance.

Inference

Precision -

Use - Focuses on the accuracy of positive predictions. It's the proportion of true positives out of all positive predictions.

Considerations - Useful when false positives are costly (e.g., approving a risky loan).

Recall (Sensitivity, True Positive Rate) -

Use - Focuses on capturing as many actual positives as possible. It's the proportion of true positives out of all actual positives.

Considerations - Important when false negatives are costly (e.g., missing a genuine loan).

Inference

F1-Score -

Use - Harmonic mean of precision and recall. Balances both precision and recall.

Considerations - Suitable when you want a balance between false positives and false negatives.

Based on Average F1-Scores too, **Logistic Regression** has comparatively better performance, with the weighted average F1-Score of **69%**.

Mid-Project Timeline

Here’s the proposed weekly timeline for the project -

- Week 1 - Literature review on credit risk assessment.
- Week 2 - Collect and preprocess credit data.
- Week 3 - Implement and train ML models.
- Week 4 - Experiment with deep learning methods.
- Week 5 - Evaluate model performance metrics.
- Week 6 - Explore alternative data integration.
- Week 7 - Fine-tune models and hyperparameters.
- Week 8 - Write and finalize project report.
- Week 9 - Prepare and practice project presentation.
- Week 10 - Submit project report, give presentation.

Final–Project Timeline

- Week 5 - Experiment with deep learning methods.
- Week 6 - Explore alternative data integration.
- Week 7 - Fine-tune models and hyperparameters.
- Week 8 - Write and finalize project report.
- Week 9 - Submit project report, give presentation

Timeline

We’ve been able to follow the timeline, as proposed earlier, including experimenting with some Deep Learning Techniques, such as ANN.

We’ve also included models such as XG-Boost, in our final presentation as well, in addition to all the other previous models.

Contributions (Group 52)

Nikhil Suri (2021268) - Implementation of ML models, Mid-Project Report, Slides, Inferences, Final Project Report.

Maanas Gaur (2021537) - Data Pre-Processing & Datacleaning, Slides, ANN & XG-Boost, re-training models after undersampling.

Adish Jain (2021227) - Debugging ML training code, Slides, Verification of Results & Inferences, ANN.