ML Mid Project Presentation

Machine Learning / CSE-343

INDRAPRASTHA INSTITUTE *of*
INFORMATION TECHNOLOGY
DELHI

Motivation & Problem Statement

Problem Statement

Predicting Loan Defaulting using Machine Learning Techniques

Motivation

- The rapid growth of India's banking sector has increased loan applicants, amplifying the challenge of accurately predicting loan defaults.
- Recent data from the Reserve Bank of India (RBI) reveals loan write-offs by banks surged to Rs 209,144 crore in the fiscal year ending March 2023, which is a 19.53 % increase from FY22.
- Through harnessing the potential of machine learning (ML), we aim to formulate an advanced predictive model to bolster the precision of loan default prediction.
- This innovation holds the potential to equip Indian banks with proactive measures against potential defaults, mitigating financial losses and fortifying the lending landscape.

Literature Review

In recent years, the application of machine learning algorithms to predict loan defaults has gained significant attention due to their potential to enhance predictive accuracy and risk assessment.

Some relevant studies and literature on loan default prediction:

- “A study on predicting loan default based on the random forest algorithm” by Lin Zhu et al. (2019).
- “Loan Default Prediction Using Machine Learning Techniques” by E. Praynlin et al. (2023).

ELSEVIER

Procedia Computer Science

Volume 162, 2019, Pages 503-513

A study on predicting loan default based on the random forest algorithm

Lin Zhu [a], Dafeng Qiu [a], Daji Ergu [a] ✉, Cai Ying [a], Kuiyi Liu [b]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.procs.2019.12.017

Get rights and content

IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661,p-ISSN: 2278-8727, Volume 25, Issue 5, Ser. 1 (Sept. – October. 2023), PP 15-18
www.iosrjournals.org

Loan Default Prediction Using Machine Learning Techniques

Srinivasulu M[1] , Ambika B A[2], Sukanya M [3], Ambika T[4], Amrutha D[5].
[1]*assistant Professor, Department Of Master Of Computer Application, Ubdtce,Davangere*
[2,3,4,5] *Student, Department Of Mca, Ubdtce, Davangere*

Abstract –
Loan business is one of the major income sourcesfor bank. Loan default problem is a major issue for loan business. Loans, specifically whether borrowers repay the loan or default on it, have a significant impact on a bank's profitability. By anticipating loan defaulters, the bank is ableto reduce its non-performing assets. Three primary predictiveanalytics techniques—I Data Collection, II Data Cleaning, and III Performance Assessment—are used to research the prediction of loan defaulters. Experimental investigations reveal that when it comes to loan forecasting, the KNN modelperforms better than the Decision tree model.

Key Words: *Machine learning, Loan prediction, Banking, Decision tree, KNN.*

Dataset: Lending Club Loan Dataset

- LendingClub is a US peer-to-peer **lending company**, in California. It contains a **complete record of all the loans** from 2007 to 2015.
- This dataset offered a heterogeneous view of the loan data.
- It included several critical attributes like **loan amount**, **interest rates**, **employee title**, **employment years**, **loan purpose**, and **credit history**, among others.
- **Huge Dataset**: An extensive **2,260,668 rows x 145 columns**.
- A **subset of these 145 features** was obtained from various **pre-processing techniques.**

Dataset: Lending Club Loan Dataset

Feature	Description
loan_amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
funded_amnt	The total amount committed to that loan at that point in time.
funded_amnt_inv	The total amount committed by investors for that loan at that point in time.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
int_rate	Interest Rate on the loan
installment	The monthly payment owed by the borrower if the loan originates.
grade	Lending Club assigned loan grade
sub_grade	Lending Club assigned loan subgrade
emp_length	Employment length in years. Possible values are between 0 and 10 where 0 means less than one year and 10 means ten or more years.
annual_income	The self-reported annual income provided by the borrower during registration.
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
delinq_2y	The number of 30+ days past-due incidences of delinquency in the borrower's credit file for the past 2 years
open_acc	The number of open credit lines in the borrower's credit file.
pub_rec	Number of derogatory public records
revol_bal	Total credit revolving balance
revol_util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
total_acc	The total number of credit lines currently in the borrower's credit file
recoveries	post charge off gross recovery
acc_now_delinq	The number of accounts on which the borrower is now delinquent.
tot_cur_bal	Total current balance of all accounts
total_bal_il	Total current balance of all installment accounts
all_util	Balance to credit limit on all trades
total_cu_tl	Number of finance trades
avg_cur_bal	Average current balance of all accounts
bc_util	Ratio of total current balance to high credit/credit limit for all bankcard accounts.
delinq_amnt	The past-due amount owed for the accounts on which the borrower is now delinquent.
num_sats	Number of satisfactory accounts
pub_rec_bankruptcies	Number of public record bankruptcies
hardship_flag	Flags whether or not the borrower is on a hardship plan
State	US State to which the borrower belongs

Feature Details

Data Visualization

Dataset Description:-

- Dimensions: **2,260,668 rows x 145 columns**
- 2 primary datatypes: float (representing numerical values) & object (representing categorical values).
- A statistical analysis of all variables by calculating mean, median, mode, value counts, and percentile values across all variables helped us develop a foundational understanding of the dataset.

	loan_amnt	funded_amnt	funded_amnt_inv	int_rate	installment	annual_inc	dti	delinq_2yrs	inq_last_6mths	open_acc	...
count	2070597.00	2070597.00	2070597.00	2070597.00	2070597.00	2.070597e+06	2070597.00	2070597.00	2070597.00	2070597.00	...
mean	15236.81	15231.42	15212.77	13.08	450.60	7.991642e+04	18.58	0.31	0.58	11.72	...
std	9213.27	9211.57	9215.93	4.82	267.71	1.159295e+05	11.41	0.88	0.89	5.66	...
min	500.00	500.00	0.00	5.31	4.93	0.000000e+00	-1.00	0.00	0.00	0.00	...
25%	8000.00	8000.00	8000.00	9.49	256.23	4.800000e+04	11.89	0.00	0.00	8.00	...
50%	13000.00	13000.00	13000.00	12.62	383.72	6.700000e+04	17.76	0.00	0.00	11.00	...
75%	20000.00	20000.00	20000.00	15.99	599.50	9.500000e+04	24.33	0.00	1.00	15.00	...
max	40000.00	40000.00	40000.00	30.99	1719.83	1.100000e+08	999.00	58.00	32.00	101.00	...

Data Visualization

- Target Variable: loan_status

```
df['loan_status'].value_counts()
```

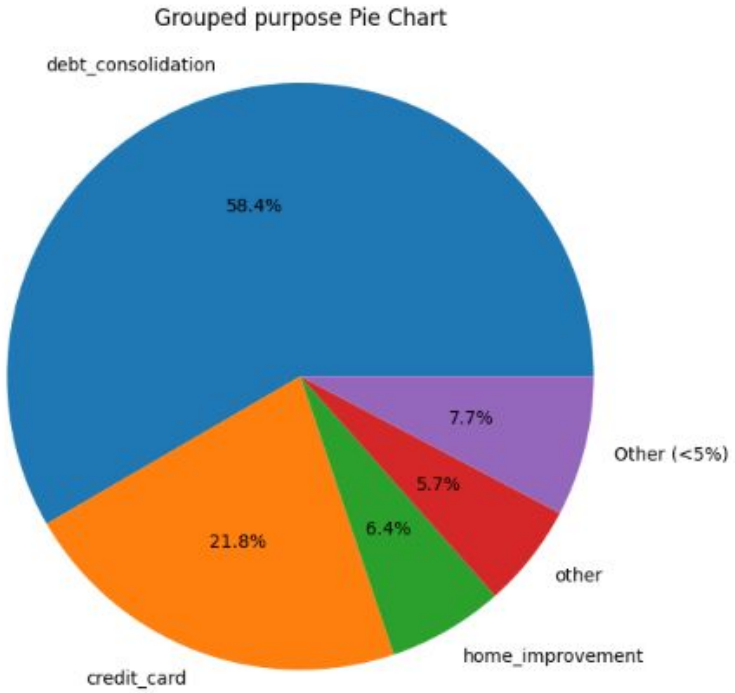
```
Fully Paid                                           971118
Current                                              832137
Charged Off                                          233914
Late (31-120 days)                                    19436
In Grace Period                                        8191
Late (16-30 days)                                      3214
Does not meet the credit policy. Status:Fully Paid     1862
Does not meet the credit policy. Status:Charged Off     697
Default                                                  28
Name: loan_status, dtype: int64
```

- Given our objective of classifying loans as either ’good’ or ’bad,’ we scrutinized the distribution of loan statuses. Rows with a ’current’ loan status were excluded from consideration due to their inconclusive nature.

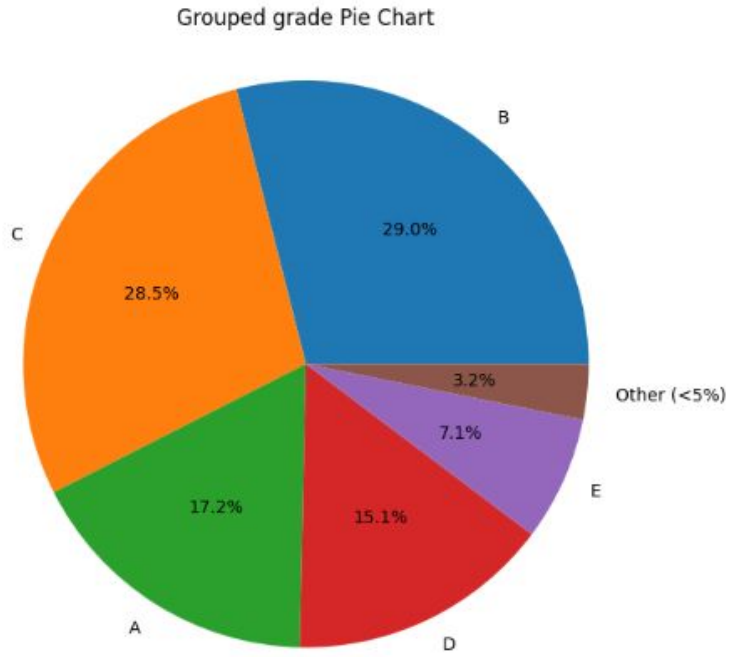
Data Visualization

Categorical Feature Insights

- To gain insights into categorical features, we employed pie charts to visualize their distributions.

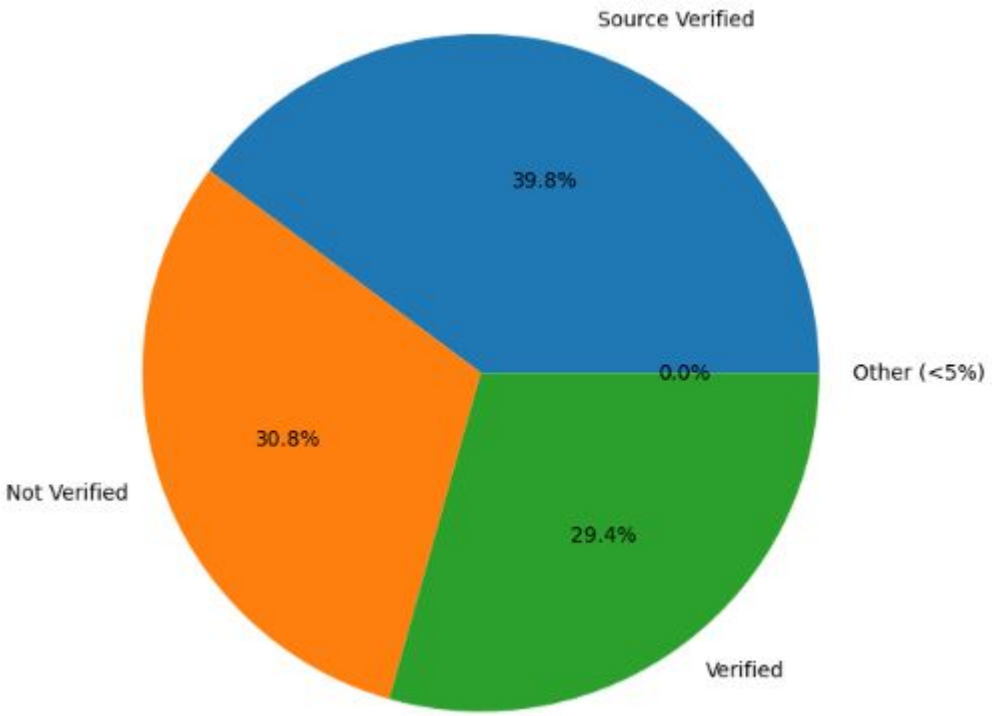


Distribution of Loan Purpose

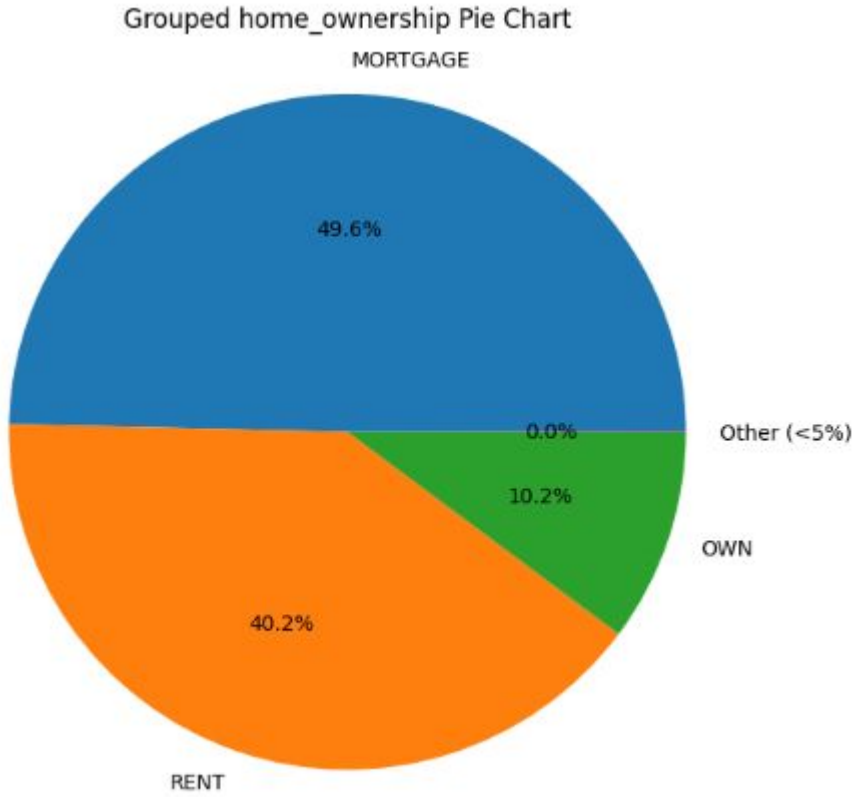


Distribution of Employee Grade

Data Visualization



Distributions of Verification

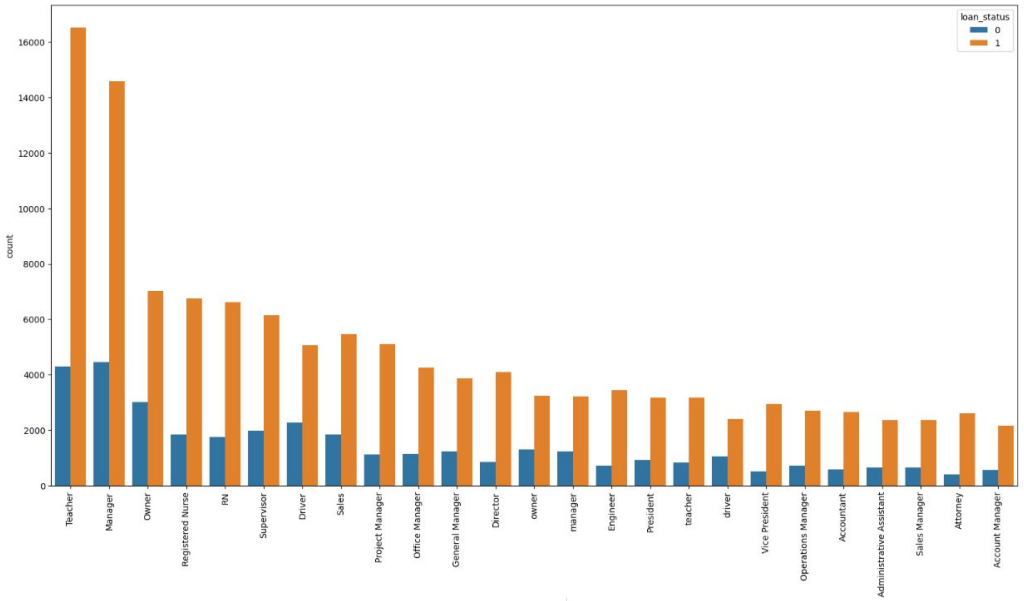


Distribution of Home Ownership Status

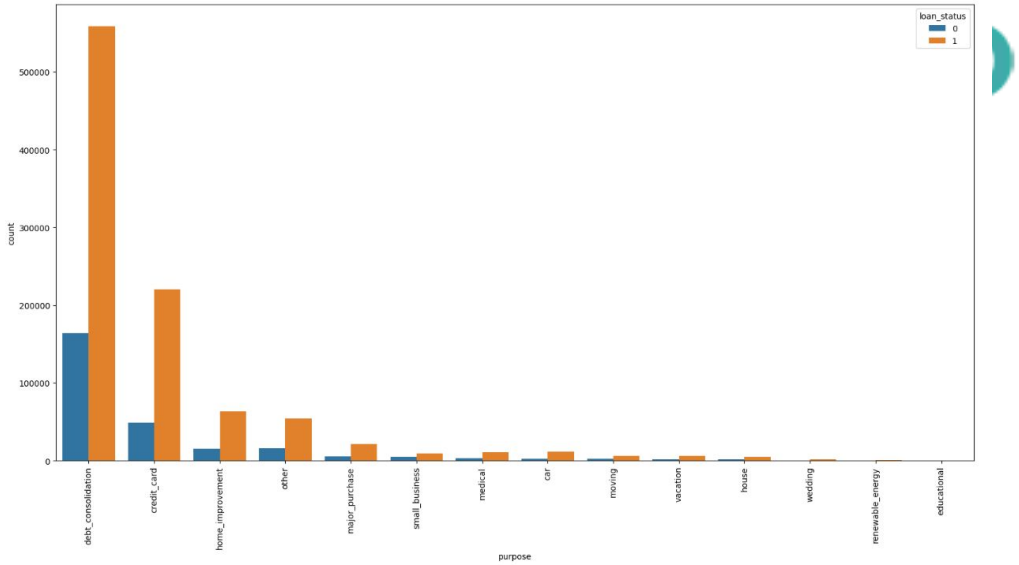
Data Visualization

Feature and Target Variable Relationship

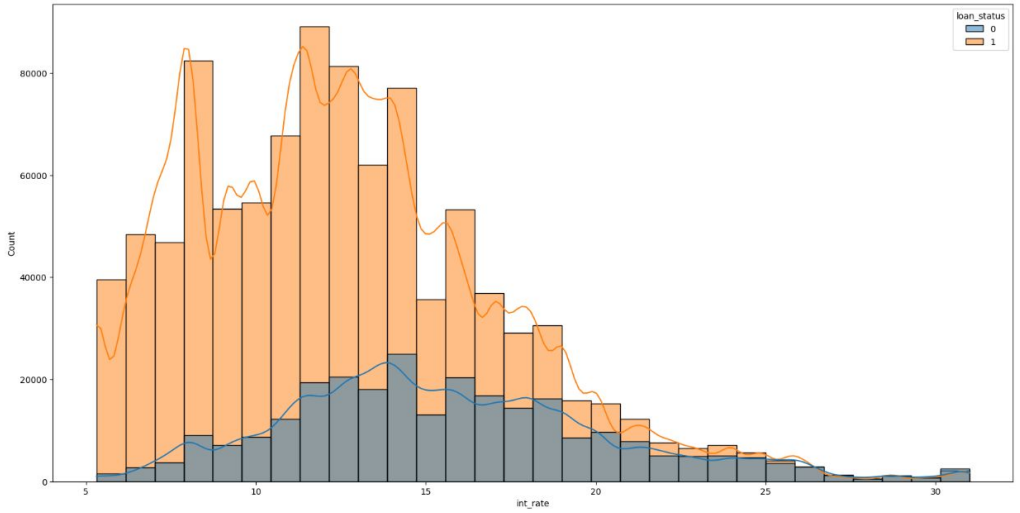
- Helped in identifying features that could potentially enhance our modeling efforts.
- Some common observations were:-
  - Loan amounts predominantly fell within the range of $5,000 to $25,000.
  - Loans with lower interest rates showed higher repayment rates, while those with interest rates exceeding 20% experienced a higher default rate.
  - Common loan purposes encompassed debt consolidation, credit card refinancing, and home improvement.
- This exploratory analysis laid the foundation for our modeling efforts, offering invaluable insights into the dataset's key features influencing loan outcomes.



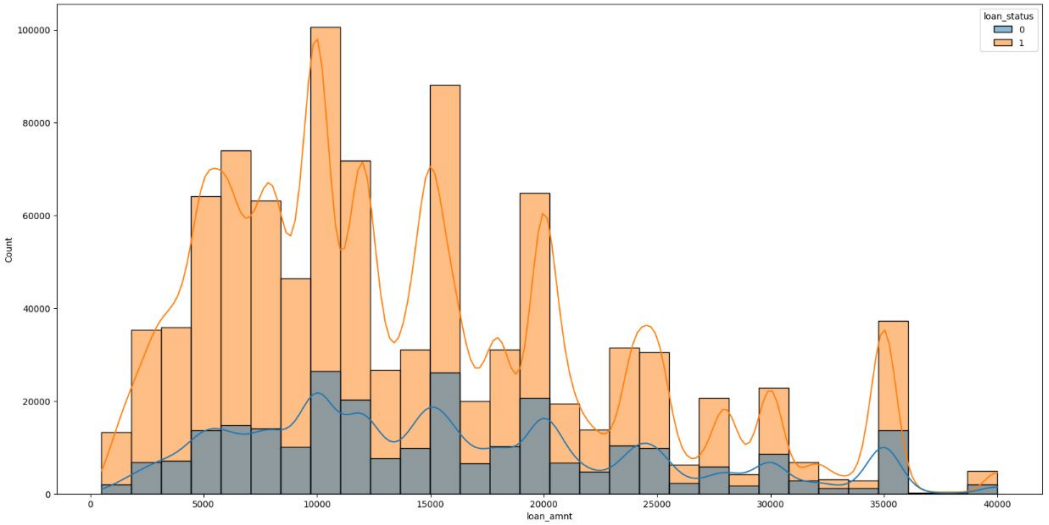
Employee Title



Loan Purpose



Interest Rates



Loan Amount

Data Preprocessing

Handling Missing Data

Identification & Removal

- All samples tested for missing target variables.
- Dropped columns having more than 50 percent missing values.
- Overall, missing values formed 9.87 percent of total values.
- All samples found missing one or more categorical data were dropped.

Imputation

- Common imputation strategies are replacing by mean, median, mode. To ensure if mean would be a good estimate, the distribution of data was studied via qq plots.
- ‘median’ was chosen as the imputation strategy.
- Total missing values post removal and imputation - 0

Data Preprocessing

Data Cleaning

Duplicate Removal

- Checked numerical columns for duplication. None found.
- Checked categorical columns for duplication. Columns ‘title’ and ‘purpose’ had the same categorical data.
- Data under ‘purpose’ was better presented. Hence, ‘title’ was dropped.

Data Relevance

- The target variable ‘loan status’ contained several categorical values.
- As per the problem statement, only two classes of data were relevant - ‘Charged Off’ and ‘Fully Paid’.
- ‘Charged Off’ was represented by 1 and ‘Fully Paid’ by 0. Samples related to all other classes were dropped.

Data Preprocessing

Data Transformation

Encoding categorical variables

- Categorical variables were converted into numerical representations, using techniques like -
  - one-hot encoding - term, grade, emp length, etc.
  - label encoding - : home ownership, verification status, purpose, etc.
  - manual mapping - debt settlement flag, pymnt plan, hardship flag, etc.

Dropped Features

- zip code
- title
- emp title
- others like ’total rec late fee ’, ’open il 12m’, ’open acc 6m’, ’next pymnt d’ etc.

Methodology

After the preprocessing steps, we did a 70-30 train-test split of the data. We standardized the remaining non-encoded columns. We calculated the accuracy, F1 score, precision, and recall after training the different models.

Different ML models used are:

- Logistic Regression
- Decision Trees
- Random Forest
- XGBoost

FEATURE IMPORTANCE IN DECISION TREES

Feature Name	Importance
recoveries	0.9659
sub_grade	0.0113
all_util	0.0068
annual_inc	0.0007
loan_amnt	0.0003
dti	0.0016

Results & Analysis – Logistic Regression

L1 REGULARIZATION **(Acc = 92.6%)**

Precision	Recall	F1-Score	ROC-AUC
99.888	63.339	77.522	89.782

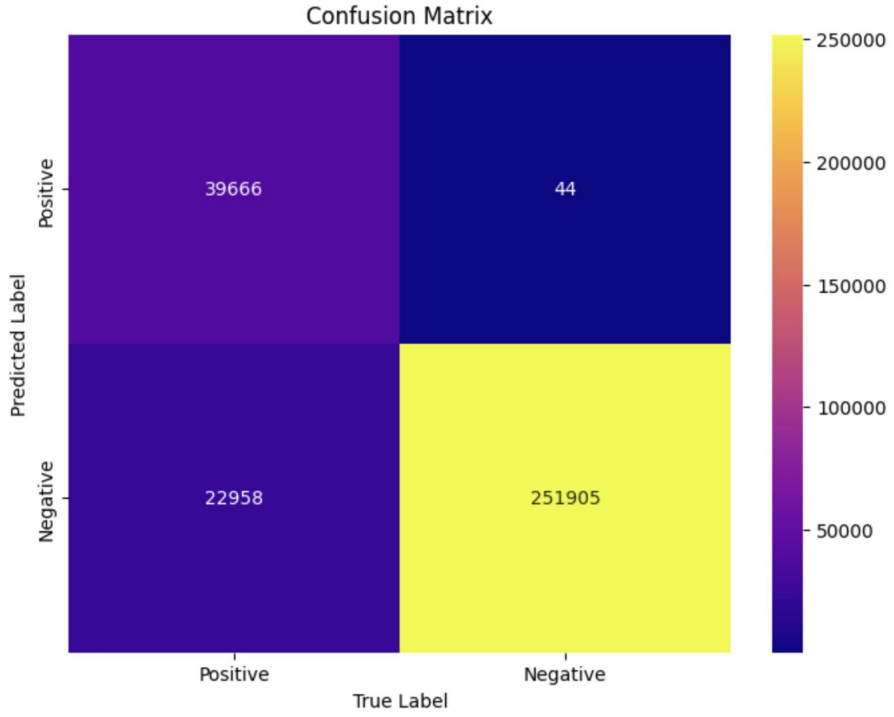


Figure 8. Logistic Regression with L1 Reg.

L2 REGULARIZATION **(Acc = 93.04%)**

Precision	Recall	F1-Score	ROC-AUC
99.877	65.128	78.843	90.031

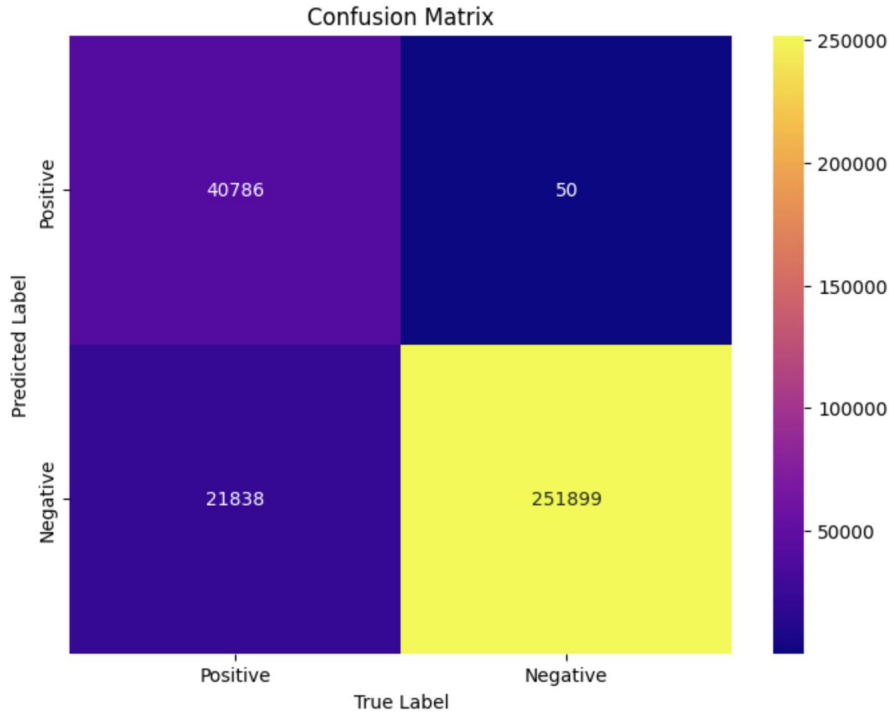


Figure 7. Logistic Regression with L2 Reg.

Results & Analysis

DECISION TREE (Acc = 93.8%)

Precision	Recall	F1-Score	ROC-AUC
99.981	68.254	81.126	91.667

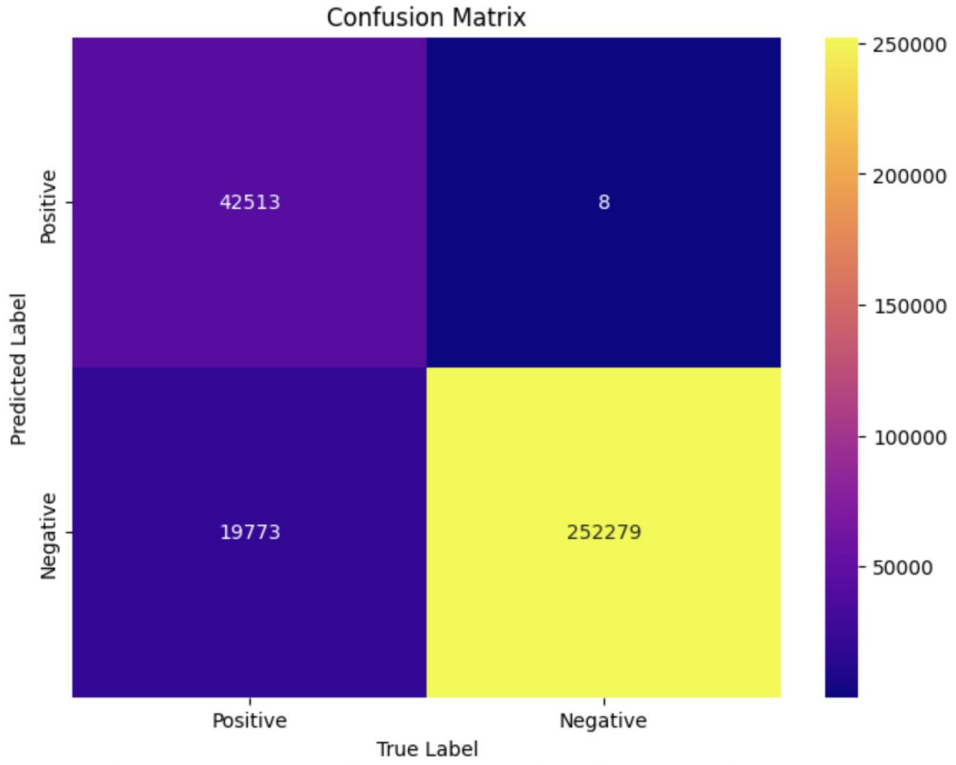


Figure 9. Confusion Matrix for Decision Tree

RANDOM FOREST (Acc = 93.7%)

Precision	Recall	F1-Score	ROC-AUC
100	68.246	81.126	91.128

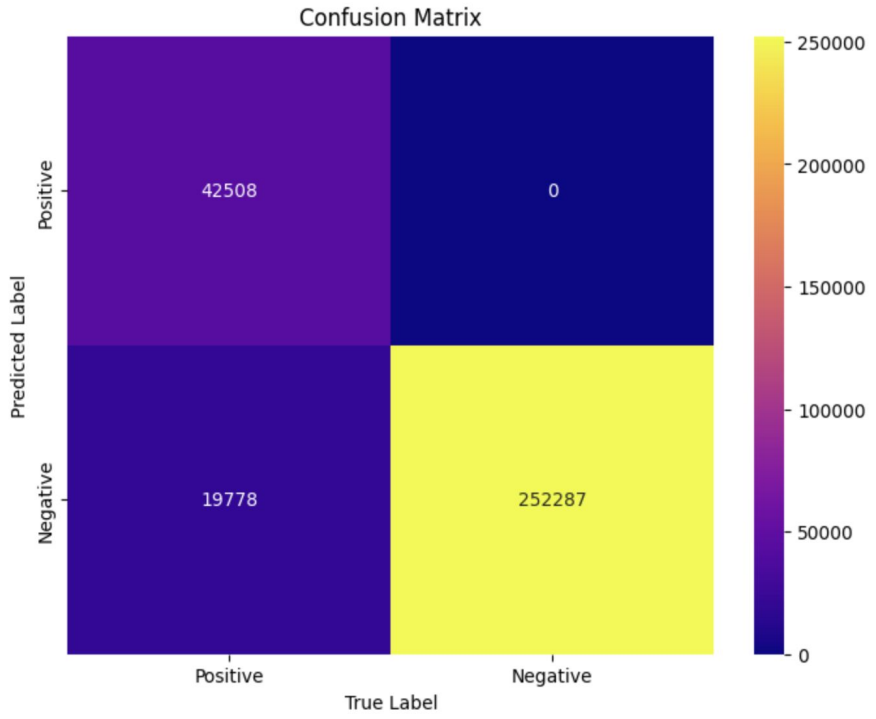


Figure 10. Confusion Matrix for Random Forest

Results & Analysis

XGBOOST (Acc = 93.6%)

Precision	Recall	F1-Score	ROC-AUC
99.772	68.198	81.017	92.461

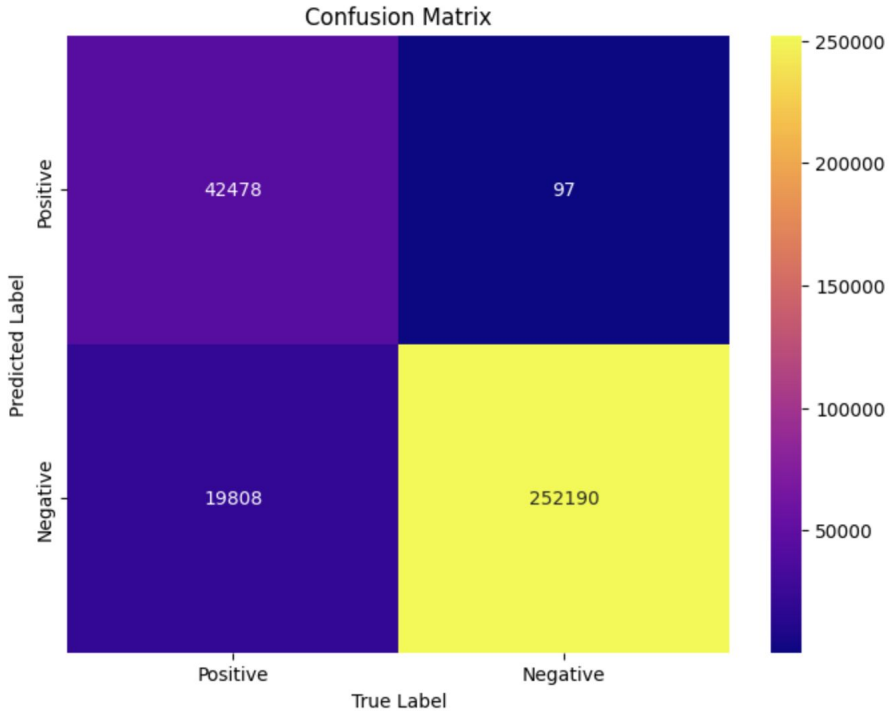


Figure 11. Confusion Matrix for XGboost

FEATURE IMPORTANCE

XGBOOST

Feature Name	Importance
recoveries	0.9335
grade	0.0085
sub_grade	0.00091
all_util	0.0014
term	0.0055
encoded_w	0.0010
encoded_WV	0.0008
encoded_MS	0.0007

RANDOM FOREST

Feature Name	Importance
recoveries	0.8060
grade	0.0515
sub_grade	0.0422
all_util	0.0034
avg_cur_bal	0.0047
loan_amnt	0.0029
dti	0.0066

Conclusion

- **Best Precision:** Random Forest achieved the highest precision, making it the safest choice when minimizing false positives is crucial, ensuring minimal unnecessary loan rejections.

- **Best Recall:** Decision Tree and XGBoost had the highest recall, indicating their effectiveness in capturing most default cases. They are suitable when identifying potential loan defaults accurately is a priority.

- **Best Overall Performance:** XGBoost demonstrated the highest ROC-AUC score, showcasing the best overall discrimination ability. It is the top choice for achieving a balanced and robust performance in loan default prediction.

Future Timeline

- We will try other boosting methods like AdaBoost and LightGBM Boost and try hyperparameter tuning for these and XGboost.
- Next, we will evaluate the performance of Support Vector Classification and ANNs on this task.
- We will also try incorporating various fairness measures such as demographic parity, equalized odds and equalized opportunity in the classification process and test models on the same.

Quick RECAP

- **Problem Statement :** Predicting Loan Defaulting using Machine Learning Techniques
- **Dataset :** Lending Club Loan Dataset
- **Models Used(till now) :**
  - Logistic Regression
  - Decision Trees
  - Random Forest
  - XGBoost
- **Results :**
  - Best Precision - Random Forest
  - Best Recall - Decision Trees & XGBoost
  - Best Overall Performance - XGBoost

Results & Analysis – Gaussian Naive Bayes & SVM

Gaussian Naive Bayes
(Acc = 54.48%)

Precision	Recall	F1-Score
99.704	65.631	79.157

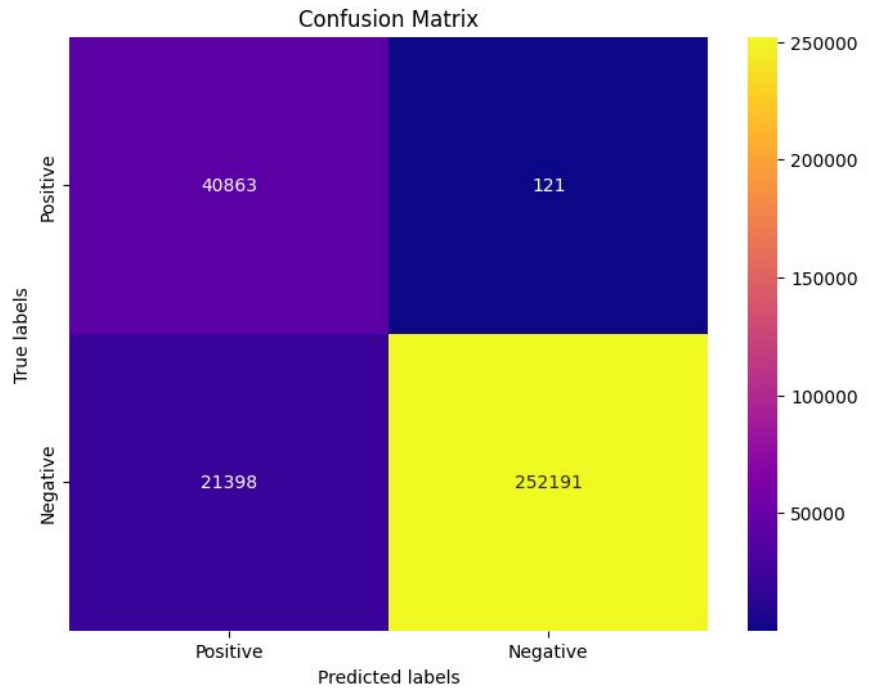


Fig: Confusion matrix for Gaussian Naive Bayes

SVM **(Acc = 54.22%)**

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

           0       0.92      0.99      0.95    252312
           1       0.93      0.66      0.77     62261
```

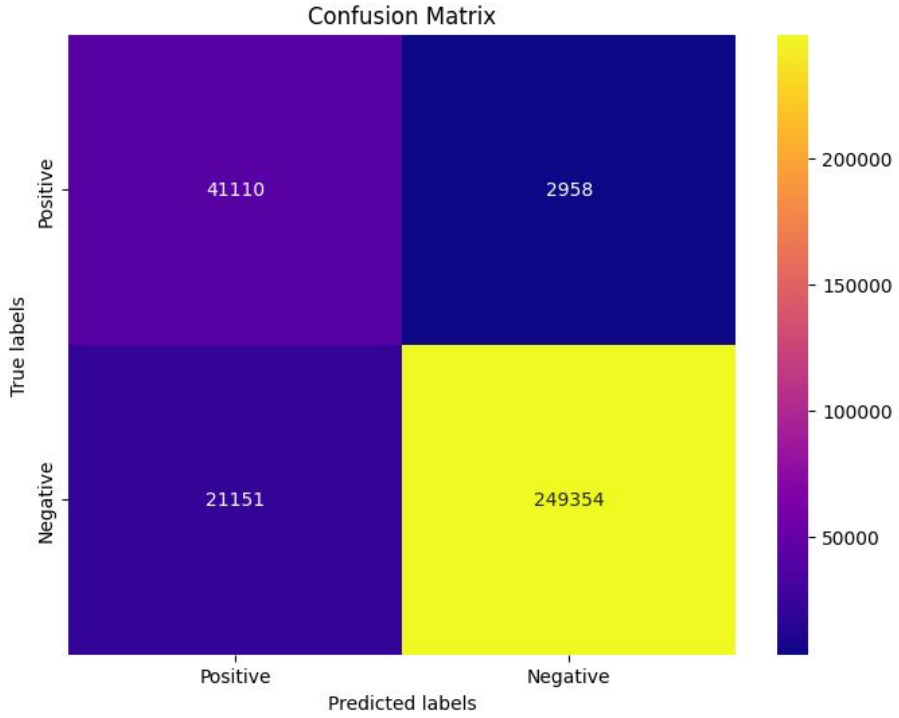


Fig: Confusion matrix for SVM

Results & Analysis – MLP

MLP (Acc = 93.07%)

Precision	Recall	F1-Score	ROC-AUC
99.48	65.33	78.87	90.45

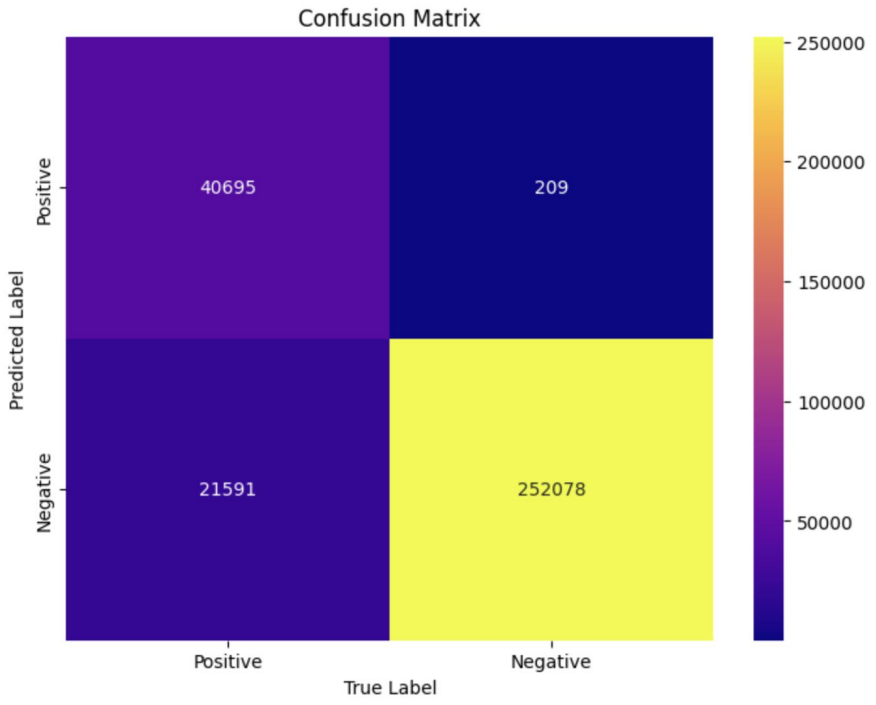


Fig: Confusion matrix for MLP

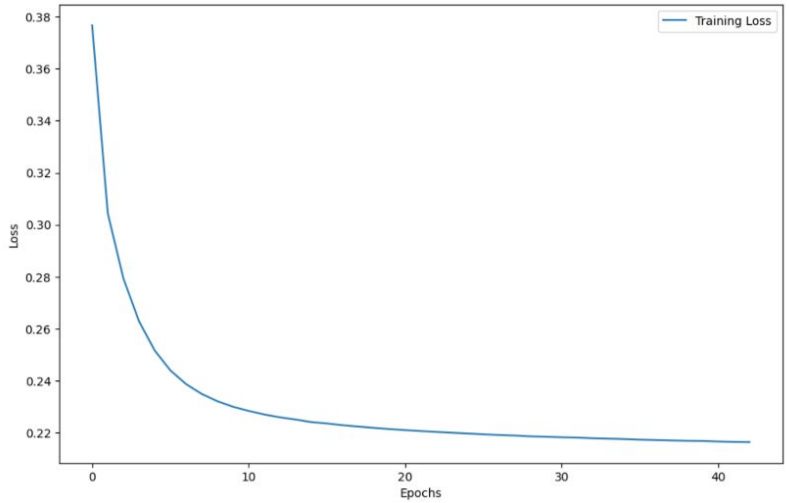


Fig: Training Loss v/s Epochs

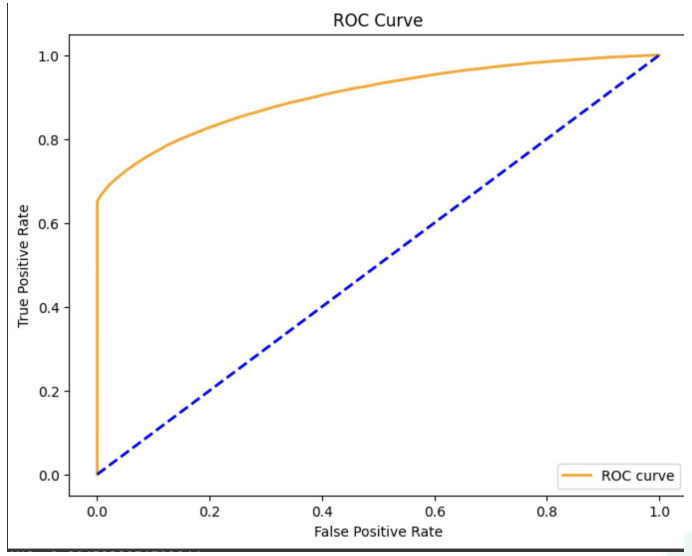


Fig: ROC Curve

Results & Analysis – AdaBoost & LightGBM

AdaBoost (Acc = 93.71%)

Precision	Recall	F1-Score
99.997	68.235	81.118

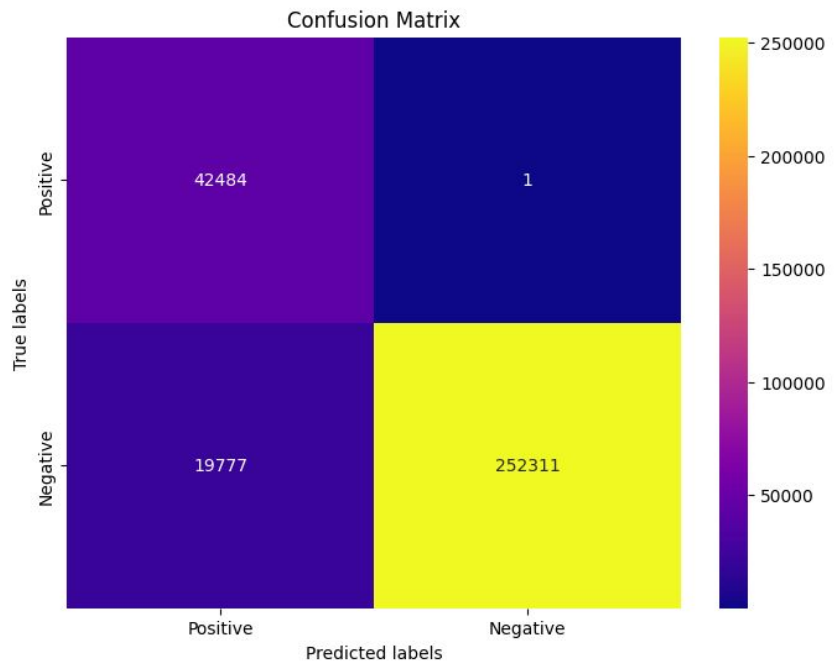


Fig: Confusion matrix for AdaBoost

LightGBM (Acc = 93.72%)

Precision	Recall	F1-Score	ROC-AUC
99.943	68.302	81.147	92.601

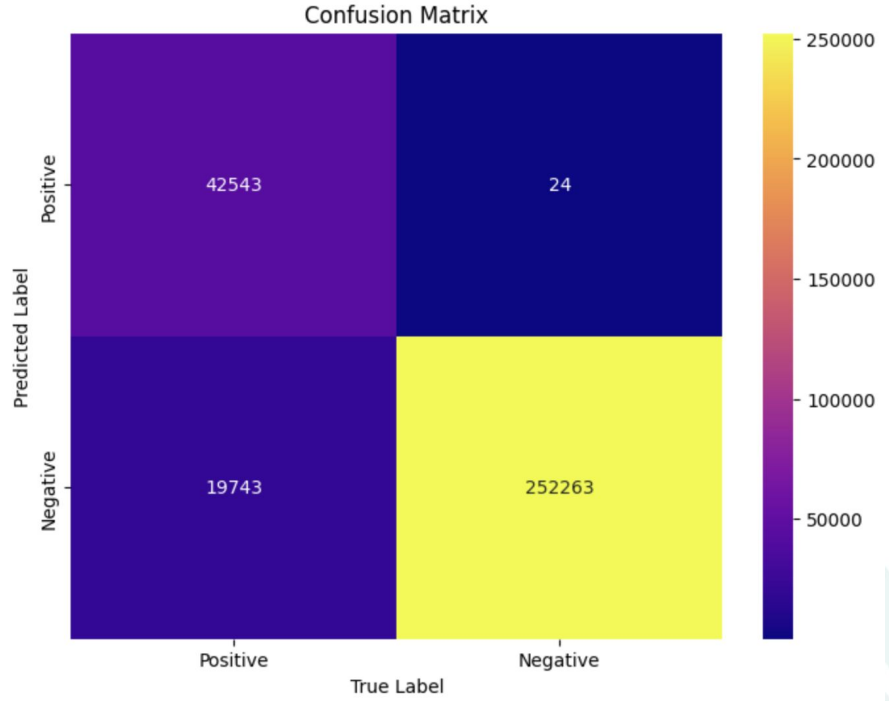


Fig: Confusion matrix for LightGBM

Fairness Evaluation

- Fairness in ML refers to treating everyone justly without bias, avoiding discrimination based on sensitive features.
- Fairness is crucial for responsible AI to prevent biased models from creating discriminatory outcomes & worsening social inequalities.
- Thus, evaluating a model goes beyond measuring its accuracy and assessing potential biases in predictions becomes essential.

Fairness Evaluation (Methodology)

- We introduced a bias variable for each dataset row, randomly assigned based on a binary probability distribution, but it's not used in training.
- For defaulted loans (class 1), bias is set with a 65% chance of being 0; for non-defaulted loans (class 0), it's assigned with a 35% chance of being 0.
- After model training, we compute metrics, like Demographic Parity, Equalized Odds etc. focusing on parity between privileged (bias_variable = 1) and non-privileged groups.
- Fairness is checked by aiming for metric ratios close to 1.0 for different groups. Significantly above or below 1.0 may indicate bias, prompting a closer examination.

Fairness Evaluation (Results)

- **Demographic Parity Difference (0.1463):** This value indicates the difference in the rate of positive predictions between the two groups (0 and 1 in your bias_variable). Our results suggests a moderate disparity, with one group receiving positive outcomes at a rate approximately 14.63% higher than the other. This can imply a potential bias in how the model treats different groups.
- **Equalized Odds Difference (0.0182)**: This metric measures the difference in both false positive rates (FPR) and true positive rates (TPR) between groups. A lower value, like 0.0182, indicates that both TPR and FPR are relatively similar across groups, suggesting that the model is more fair in terms of both false alarms and correctly identified positive cases.
- **True Positive Rate (TPR)**: Group 0 has a TPR of 0.6981, and Group 1 has a TPR of 0.6799. These values are relatively close, indicating that both groups have a similar likelihood of being correctly identified as positive by the model.
- **False Positive Rate (FPR)**: Group 0 has an FPR of 0.0088, while Group 1 has a significantly lower FPR of 0.0001. This suggests that Group 0 is more likely to experience false alarms compared to Group 1.

Contributions

- **Ashhar Aziz :** Feature Selection, Logistic Regression, XGboost, Fairness Evaluation, Report
- **Lakshya Goel :** Exploratory Data Analysis, Data Visualization, Data Preprocessing, SVM, Naive Bayes, Report
- **Sanmay Sood :** Decision Trees, Random Forest, XGboost, MLP, LightGBM, Report
- **Srimant Mohanty :** Exploratory Data Analysis, Data Visualization, Data Preprocessing, Fairness Evaluation, Report