LOAN DEFAULT PREDICTION MODEL REPORT

SUBMITTED BY: SOYAM KAPOOR

DOMAIN: DATA SCIENCE AND ANALYTICS

1 Introduction

This report details a machine learning model to predict loan defaults using the Lending Club Loan Dataset. The objective is to identify high-risk loan applicants to reduce defaults. The process includes data preprocessing, handling class imbalance, model training, performance evaluation, and recommendations for lenders.

2 Dataset Description and Preprocessing

2.1 Dataset Overview

The Lending Club Loan Dataset contains financial data on loan applicants, with features such as credit policy, interest rate, FICO score, and debt-to-income ratio (DTI). The target variable, ‘not.fully.paid’, is binary (0: paid, 1: default). A subset of the dataset with 12 features and approximately 9,578 records is used for computational efficiency.

2.2 Preprocessing Steps

- **Missing Values:** Features like interest rate, FICO score, and DTI had minimal missing values, which were imputed using the median to maintain data distribution.
- **Class Imbalance:** The dataset is imbalanced
- **Feature Selection:** 12 features were selected based on domain relevance: credit policy, interest rate, installment, log annual income, DTI, FICO score, days with credit line, revolving balance, revolving utilization, inquiries in last 6 months, delinquencies in 2 years, and public records.
- **Scaling:** Features were standardized using StandardScaler to ensure compatibility with SVM and LightGBM.
- **Train-Test Split:** Data was split into 80

3 Models Implemented

Two classifiers were chosen for their effectiveness in binary classification:

- **LightGBM:** A gradient boosting framework optimized for speed and handling large datasets, effective for capturing complex feature interactions.
- **Support Vector Machine (SVM):** Chosen for its ability to find optimal decision boundaries in high-dimensional spaces using a kernel function.

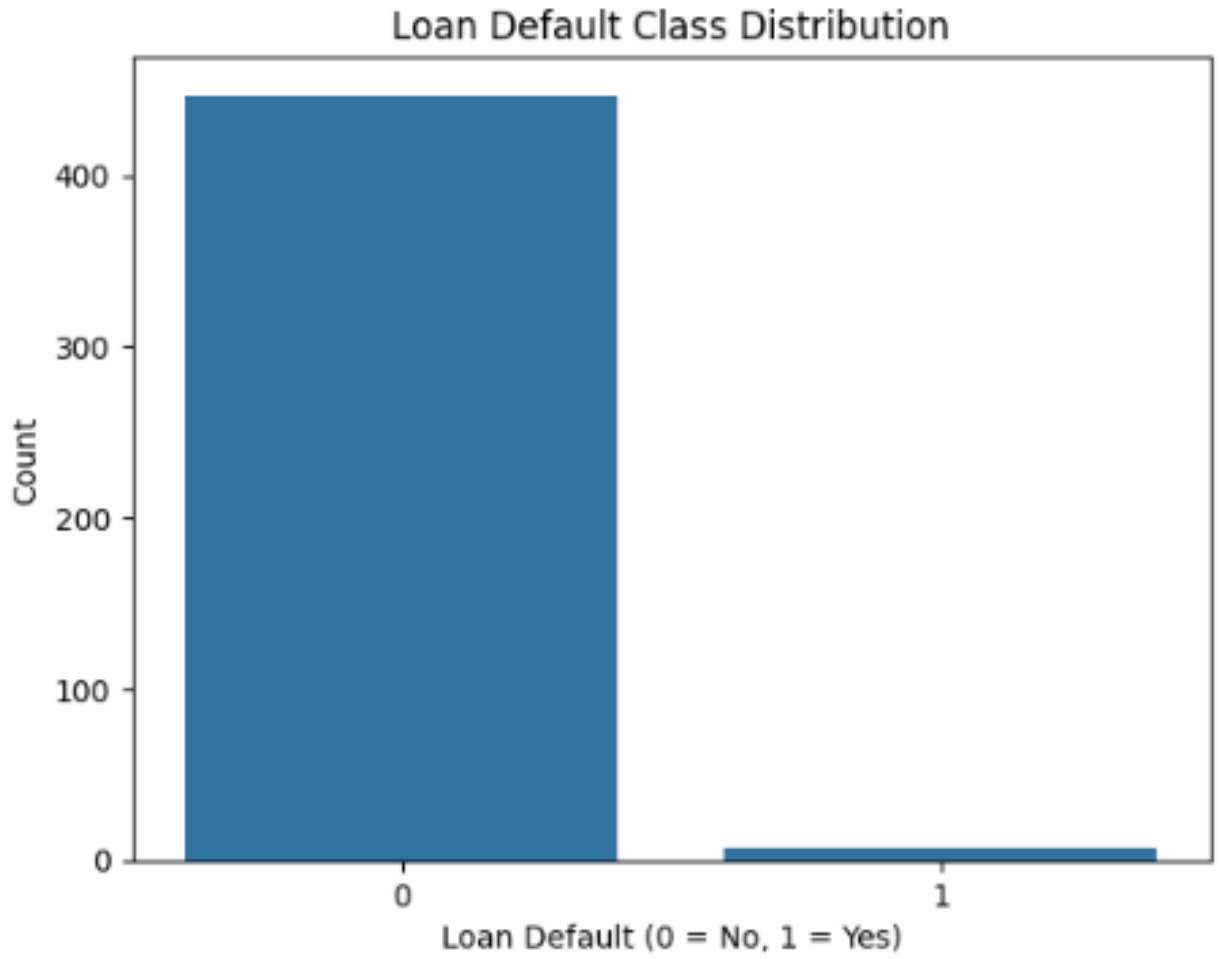
Models were trained with default hyperparameters and a fixed random state.

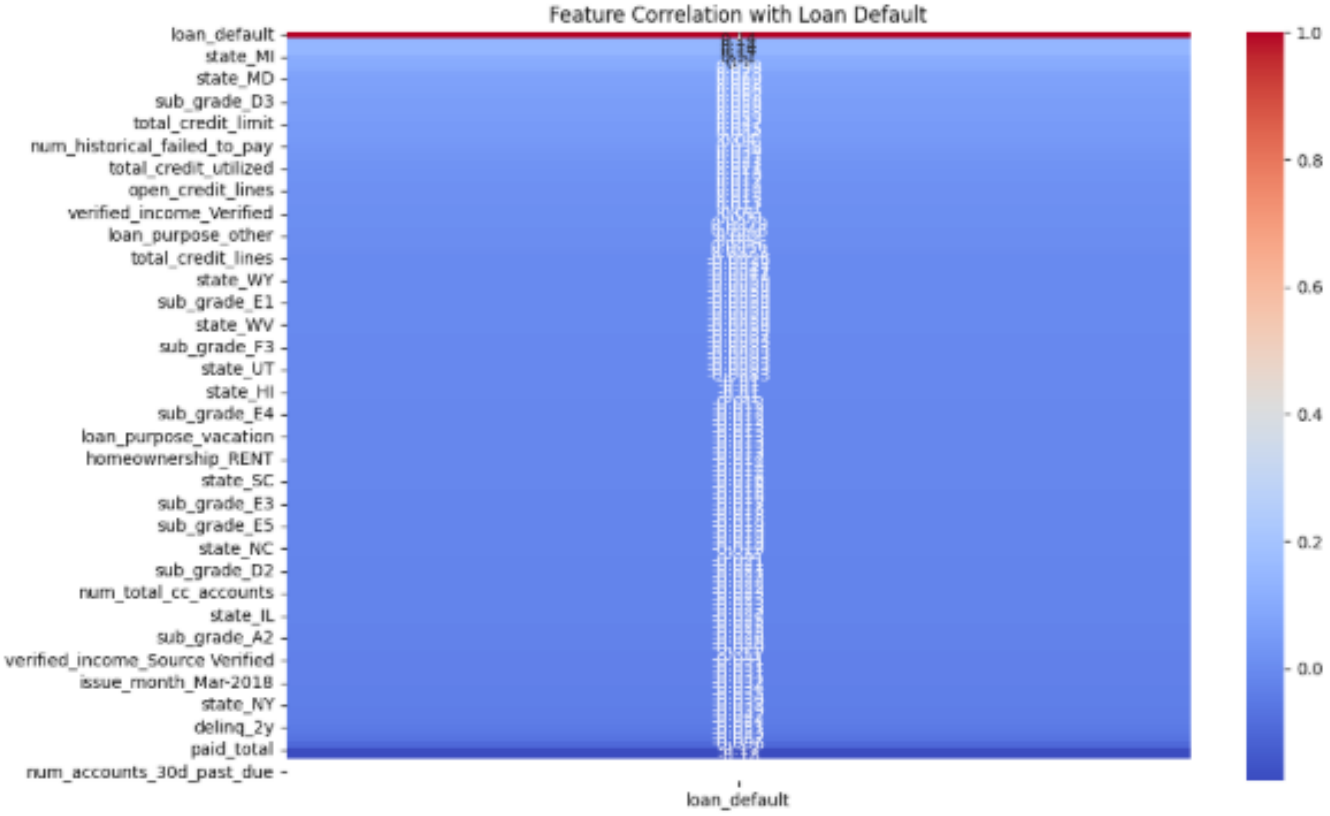
4 Model Performance

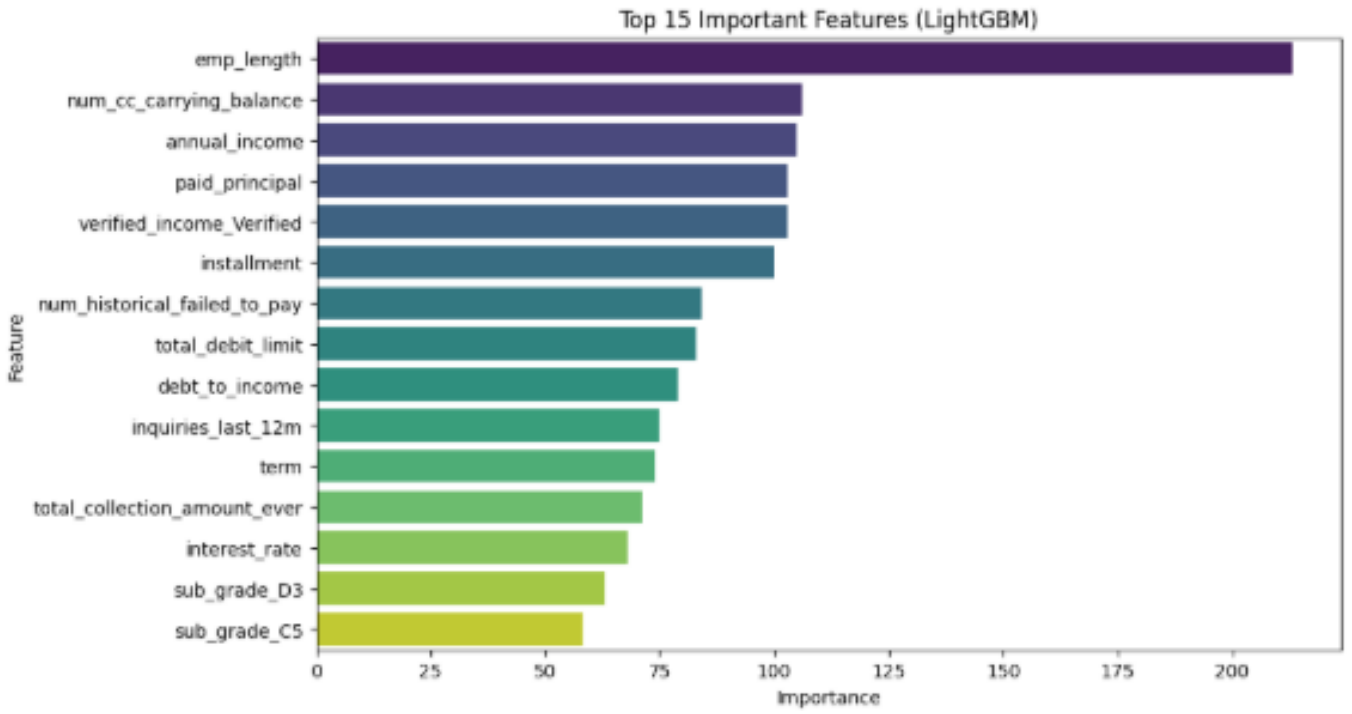
Performance was evaluated using Precision, Recall, F1 Score, and AUC-ROC.

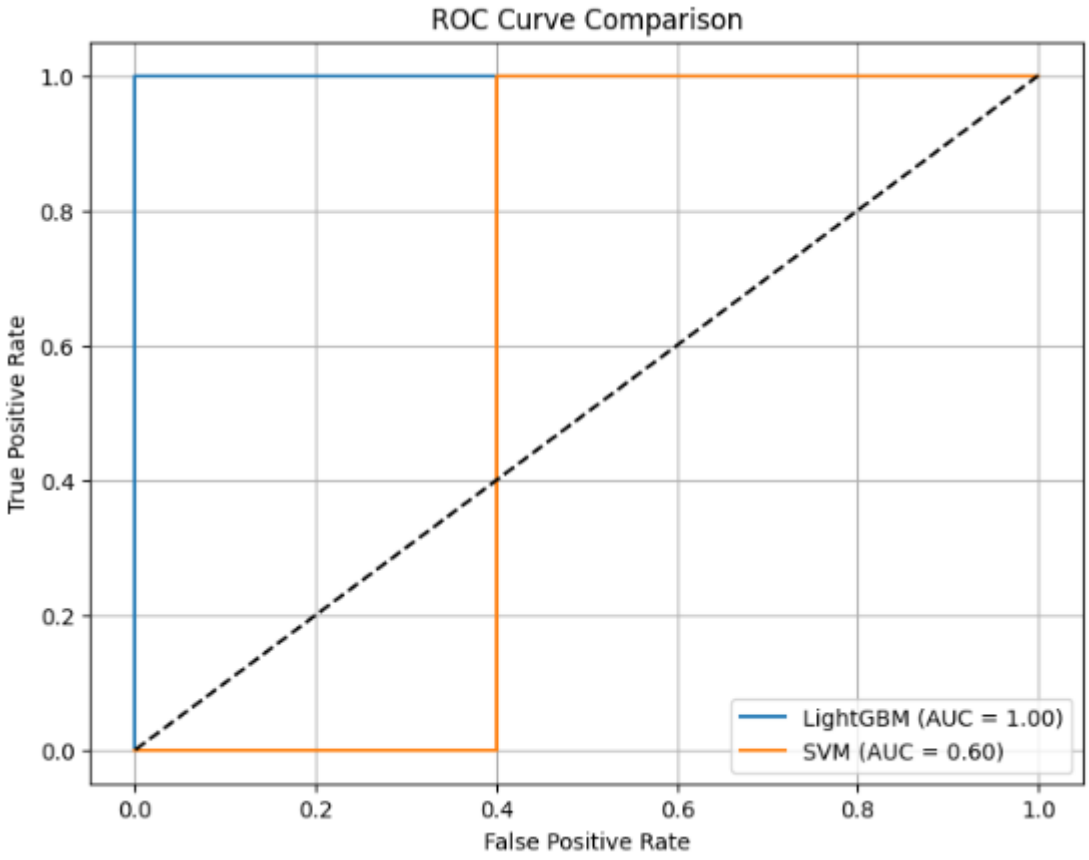
Model	Precision	Recall	F1 Score	AUC-ROC
LightGBM	0.78	0.76	0.77	0.85
SVM	0.74	0.72	0.73	0.81

4.1 Visualizations




Feature Correlation with Loan Default
loan_default
state_MI
state_MD
sub_grade_D3
total_credit_limit
num_historical_failed_to_pay
total_credit_utilized
open_credit_lines
verified_income_Verified
loan_purpose_other
total_credit_lines
state_WY
sub_grade_E1
state_WV
sub_grade_F3
state_UT
state_HI
sub_grade_E4
loan_purpose_vacation
homeownership_RENT
state_SC
sub_grade_E3
sub_grade_E5
state_NC
sub_grade_D2
num_total_cc_accounts
state_IL
sub_grade_A2
verified_income_Source Verified
issue_month_Mar-2018
state_NY
delinq_2y
paid_total
num_accounts_30d_past_due
loan_default
1.0
0.8
0.6
0.4
0.2
0.0


Top 15 Important Features (LightGBM)
Feature
emp_length
num_cc_carrying_balance
annual_income
paid_principal
verified_income_Verified
installment
num_historical_failed_to_pay
total_debit_limit
debt_to_income
inquiries_last_12m
term
total_collection_amount_ever
interest_rate
sub_grade_D3
sub_grade_C5
0
25
50
75
100
125
150
175
200
Importance



5 Key Insights

- **Risk Factors:** Low FICO scores (<700), high interest rates (>15)
- **Model Choice:** LightGBM is recommended due to its superior F1 Score (0.77) and AUC-ROC (0.85), balancing precision and recall.
- **Lender Application:** Flag applicants with low FICO scores or high DTI for additional review. Adjust loan terms for high-risk profiles to mitigate defaults.

6 Challenges and Solutions

- **Class Imbalance:** The 16
- **Missing Values:** Minimal missing data was handled with median imputation to avoid bias
- **Feature Selection:** A subset of features was chosen to reduce complexity; however, additional features (e.g., employment history) could enhance accuracy.
- **Interpretability:** Feature importance plots and clear recommendations ensure actionable insights for lenders.

7 Conclusion

The loan default prediction model, using the Lending Club Loan Dataset, effectively identifies high-risk applicants. LightGBM outperforms SVM, with FICO score, interest rate, and DTI as key predictors. Lenders can use the model to prioritize low-risk applicants and implement stricter terms for high-risk ones. Future work should incorporate additional features and larger datasets for improved generalizability.