LENDING CLUB

Supervised Learning for
Credit Risk Assessment

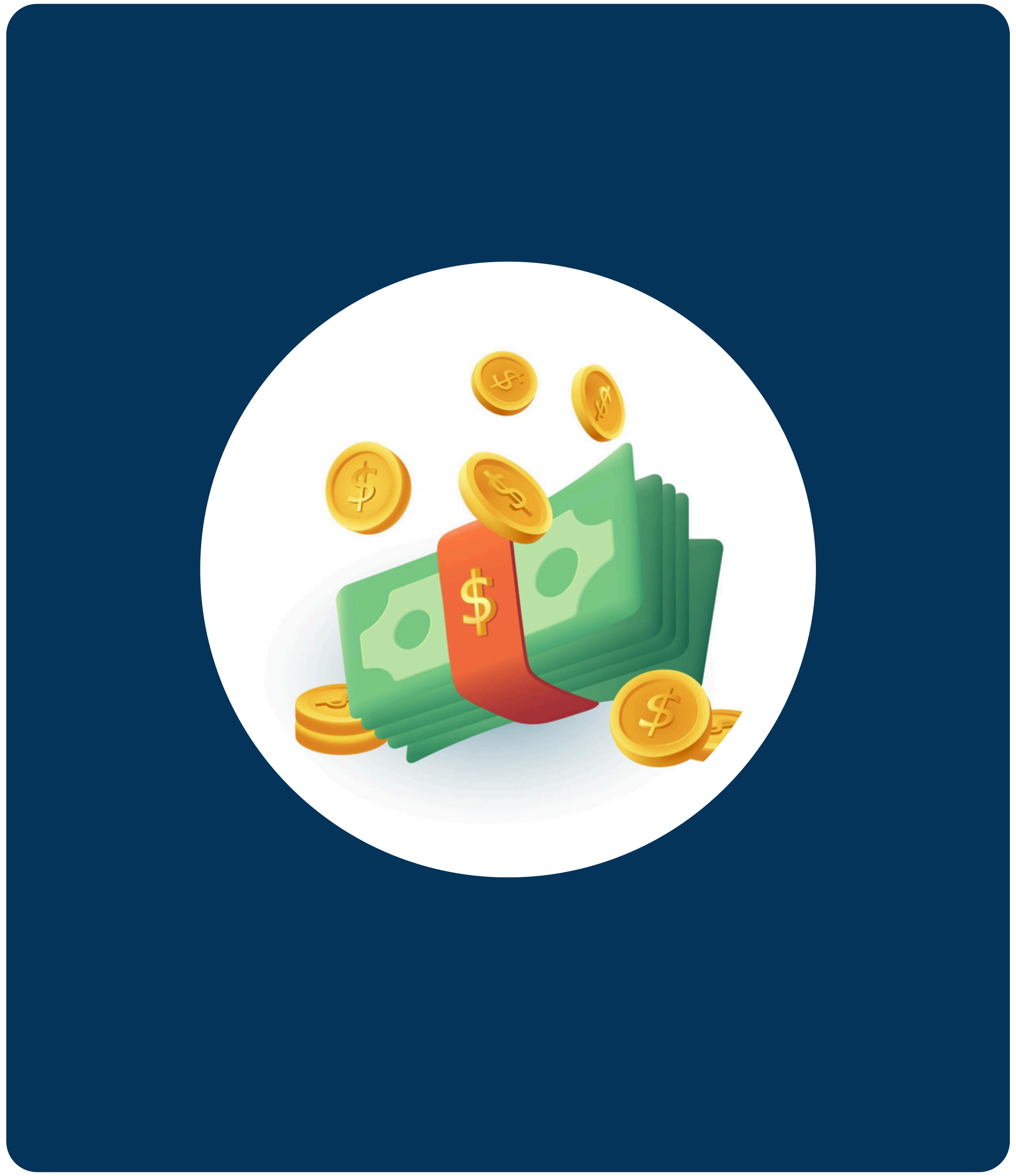
Raissa Angnged
ESADE MiBA

Introduction & Business Context

- Lending Club is a lending platform that **assesses credit risk** to optimize loan approvals and profitability

- The goal of this project is to apply **supervised learning techniques to predict loan defaults**, minimizing financial losses while maximizing returns

- **Machine learning enhances decision-making** by balancing risk and reward in lending strategies



Overview of ML Lifecycle

1

Define Goals

To leverage machine learning for **smarter loan approvals**, minimizing defaults while maximizing profitability

2

Prepare Data

1. Explore the data
2. Clean the data
3. Analyze relevant features
4. Data transformations

3

Create Model

1. Define the target variable
2. Train the models
3. Analyze the performance metrics

4

Interpret Model

1. Evaluate model using most relevant metric
2. Compare the confusion matrices
3. Interpret results

5

Implement Model

1. Analyze results' business impact
2. Apply to "open loans"
3. Assess strategies

Data Cleaning

Dropped the following columns:
1. **Joint Application**: Different borrower profile and will lead to data imbalance

2. **Post-Outcome Variables**: Risk of data leakage

3. **Hardship & Settlement Features**: Future information bias

4. **Metadata**: Non-predictive

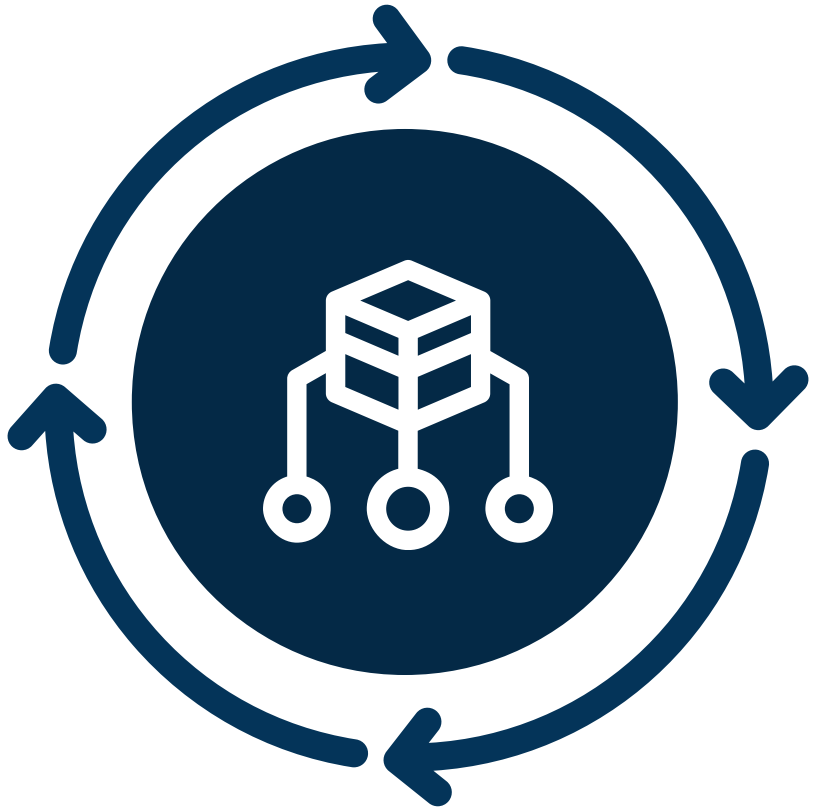
```python
columns_to_drop = [
    #Metadata and ID features
    'Unnamed: 0', 'id', 'zip_code', 'title', 'addr_state',

    #Joint application features
    'annual_inc_joint', 'dti_joint', 'verification_status_joint', 'revol_bal_joint',
    'sec_app_fico_range_low', 'sec_app_fico_range_high', 'sec_app_earliest_cr_line',
    'sec_app_inq_last_6mths', 'sec_app_mort_acc', 'sec_app_open_acc', 'sec_app_revol_util',
    'sec_app_open_act_il', 'sec_app_num_rev_accts', 'sec_app_chargeoff_within_12_mths',
    'sec_app_collections_12_mths_ex_med',

    #Post-outcome features
    'recoveries', 'collection_recovery_fee', 'last_pymnt_amnt', 'last_pymnt_d',
    'total_rec_prncp', 'total_rec_int', 'total_pymnt', 'total_pymnt_inv',
    'last_fico_range_low', 'last_fico_range_high',
    'out_prncp', 'out_prncp_inv', 'total_rec_late_fee',
    'next_pymnt_d', 'months_since_last_pymnt_d',

    #Hardship and settlement features
    'hardship_flag', 'hardship_type', 'hardship_reason', 'hardship_status',
    'hardship_amount', 'hardship_start_date', 'hardship_end_date', 'payment_plan_start_date',
    'hardship_length', 'hardship_dpd', 'hardship_loan_status',
    'debt_settlement_flag', 'debt_settlement_flag_date',
    'settlement_status', 'settlement_date', 'settlement_amount',
    'settlement_term', 'settlement_percentage'
]
```

Data Cleaning



Loading the Dataset

Efficiently processed the dataset through a **combined approach of using chunks and samples**

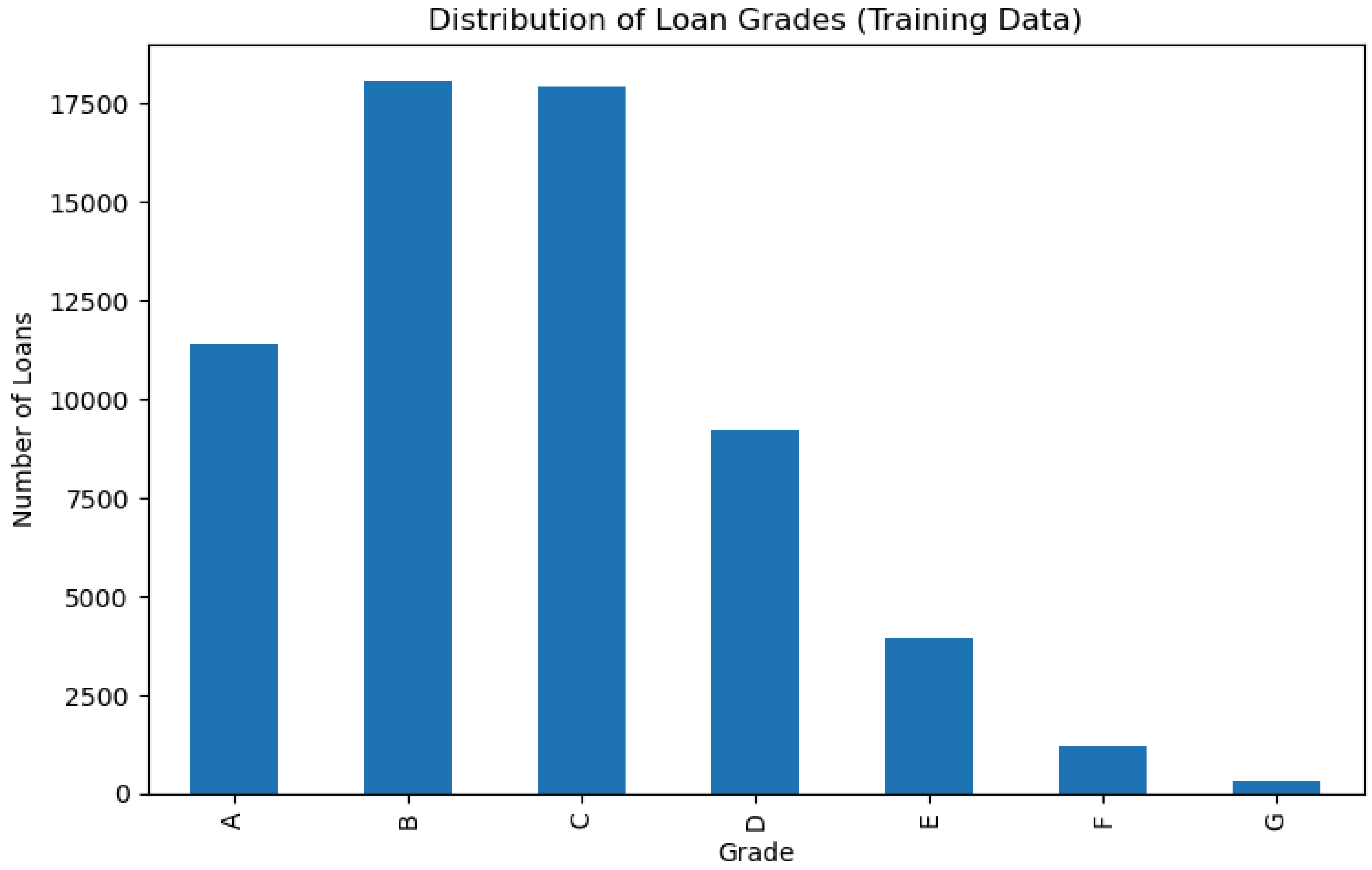
Handling Categorical Features

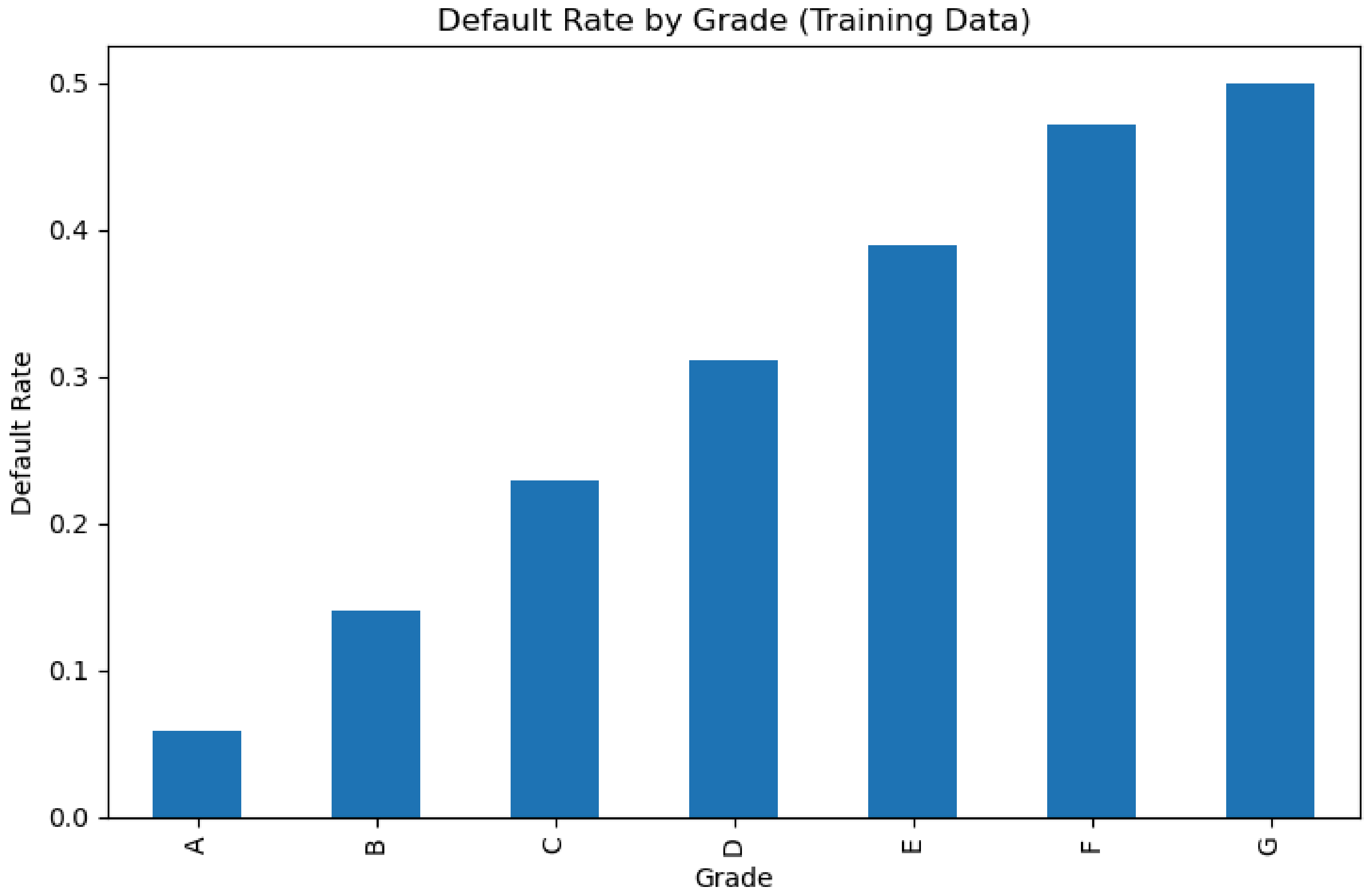
Removed features that are **statistically insignificant** (Chi-squared test) **and too unique** (variance test)

Handling Numerical Features

Removed **highly correlated variables** (correlation matrix) and the **least predictive features** (feature importance)

Data Exploration

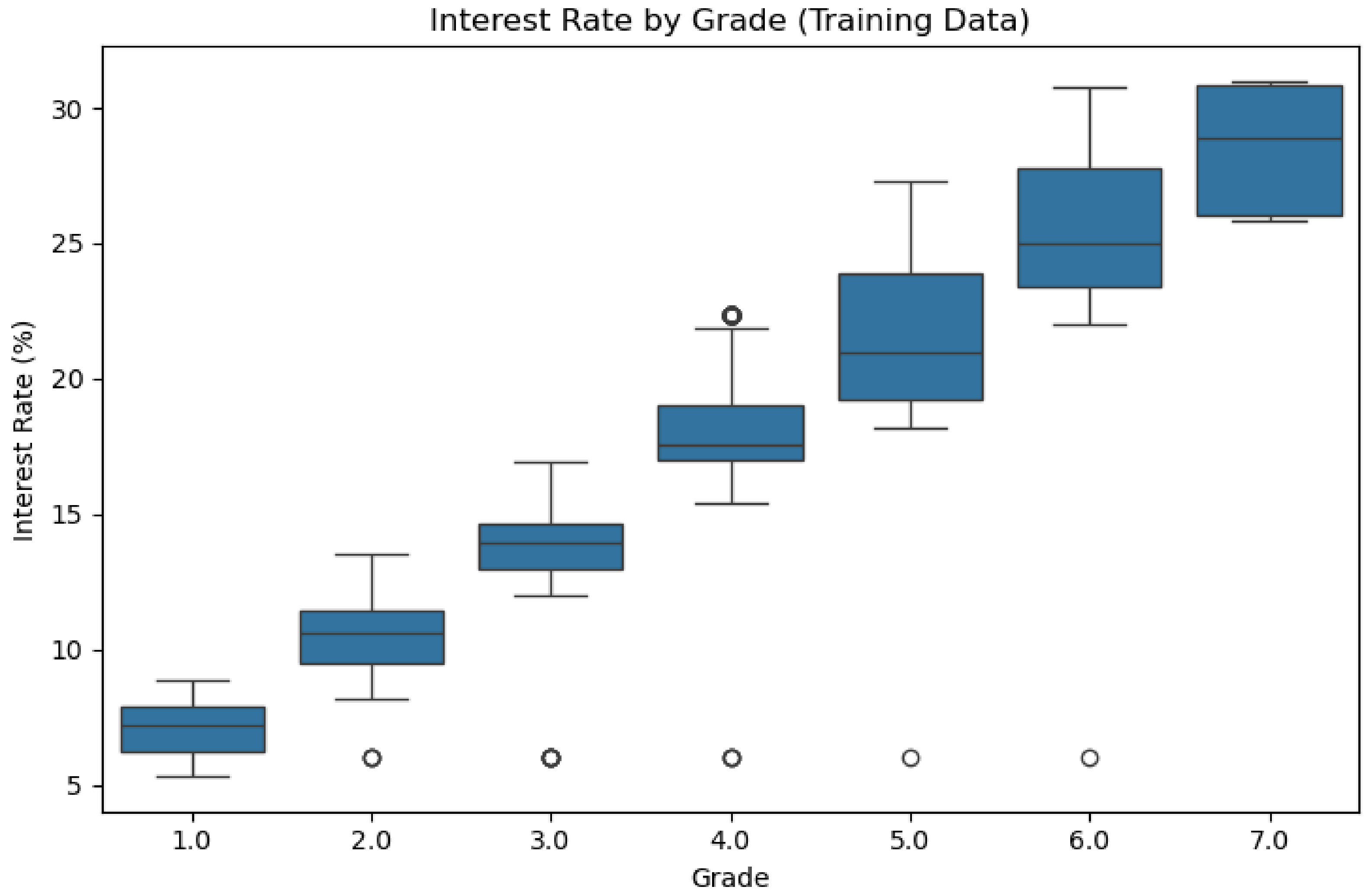


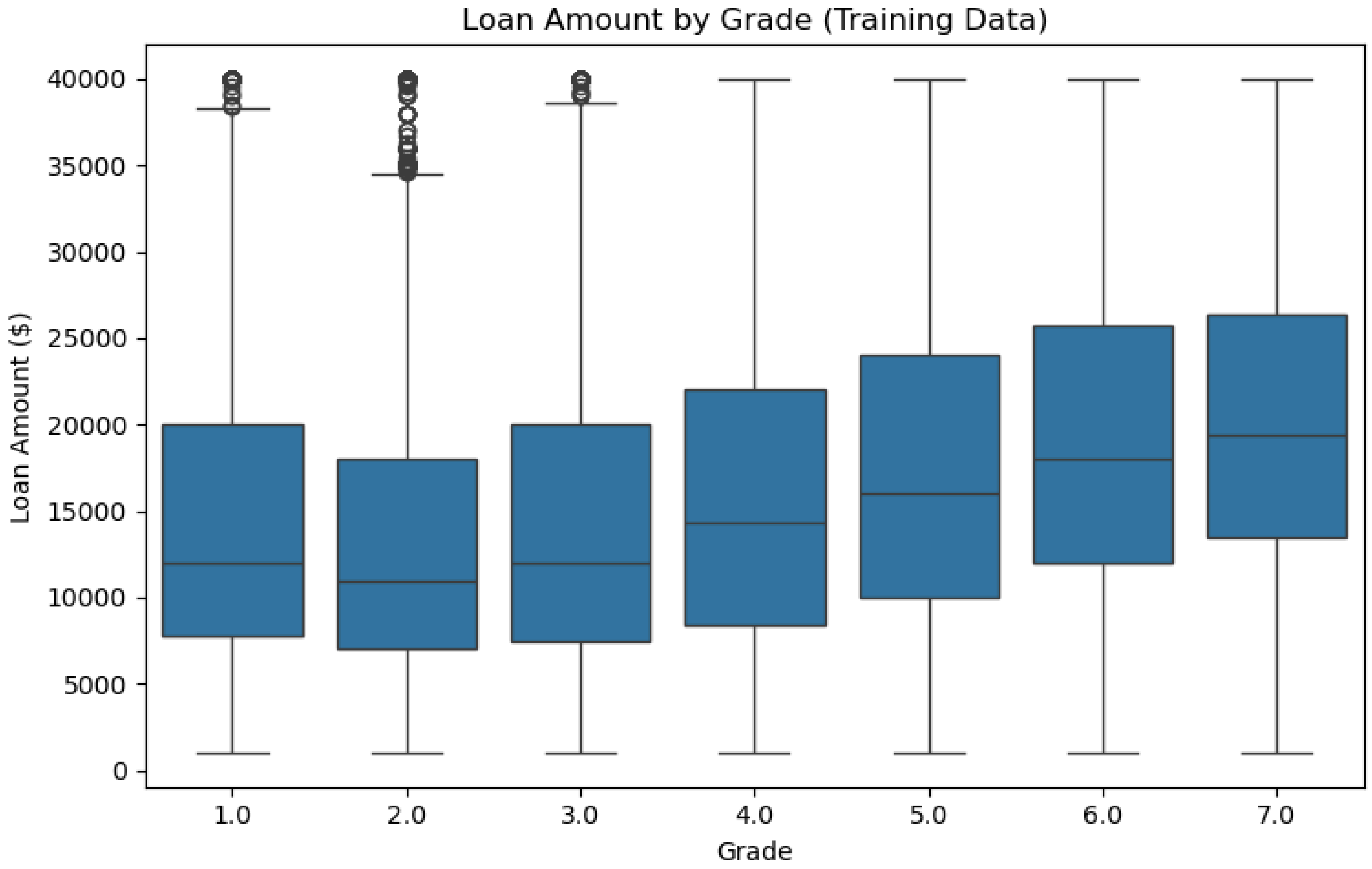


Key Insights:

- Majority of loans are B & C grades, indicating **mid-tier borrowers dominate the dataset**
- Fewer loans in high-risk (E, F, G) and low-risk (A) categories
- Default rates increase sharply as loan grade worsens
- Grade A has the lowest default rate, while G has the highest - confirms that **loan grade is a strong predictor of default risk**

Data Exploration





Key Insights:

- Higher loan grades (A, B) have lower interest rates while lower grades (F, G) have higher interest rates, reflecting borrower risk
- **Strong positive trend** between grade and interest rate
- Loan amounts are relatively stable across grades
- Some outliers in lower-grade loans suggest **riskier borrowers may take larger loans**

Model Creation and Interpretation

1. Define the target variable

Created a new variable, **loan_outcome**, which is based on the loan_status feature - explored using SMOTEENN to balance dataset

- Loan outcome = 0 if loan was fully paid
- Loan outcome = 1 if loan was defaulted or charged off

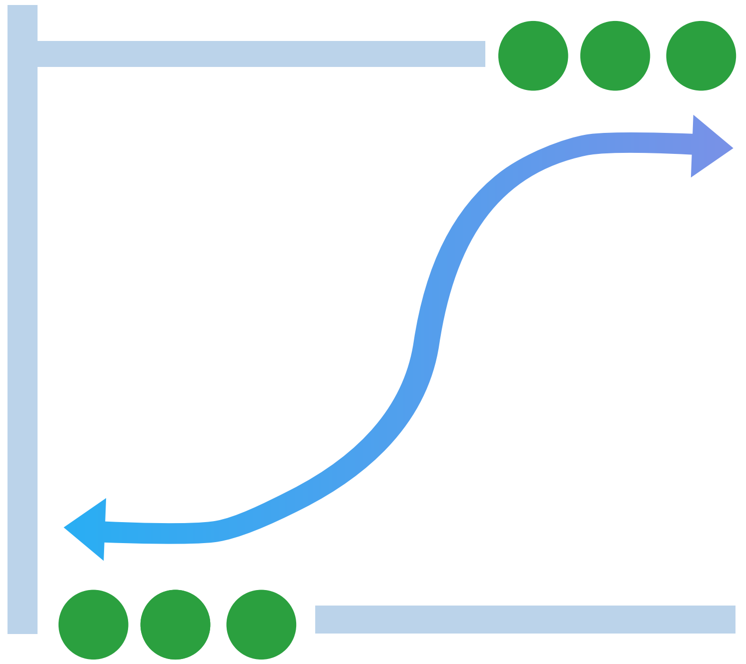
2. Train multiple classification models

There were **7 classification models trained**, as they each vary in complexity, efficiency, and approach in handling different data complexities - each with **unique strengths in predicting binary outcomes like loan classification**

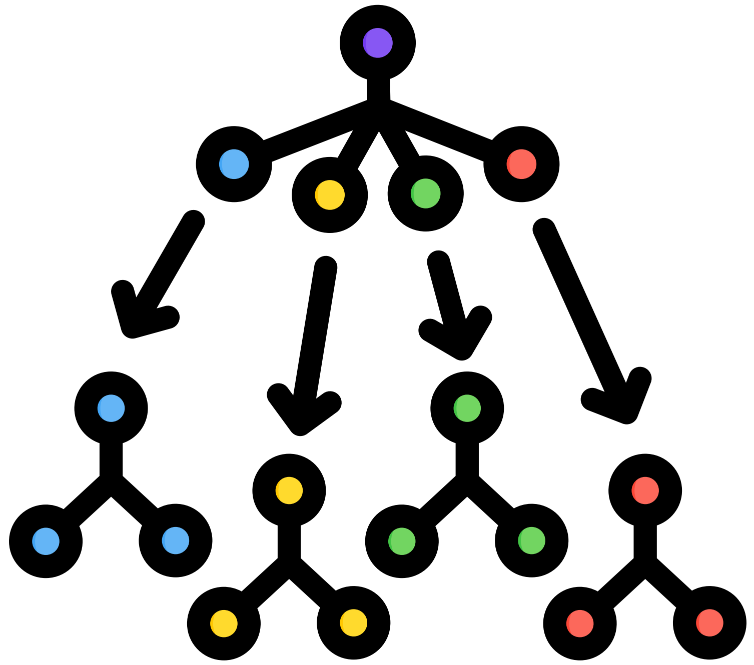
3. Interpret models' performance metrics

Analyzed the models based on these **five metrics: accuracy, precision, recall, F1, and ROC AUC**. Together, they provide a holistic view of the model's performance, particularly in terms of handling imbalances between good and bad loan predictions.

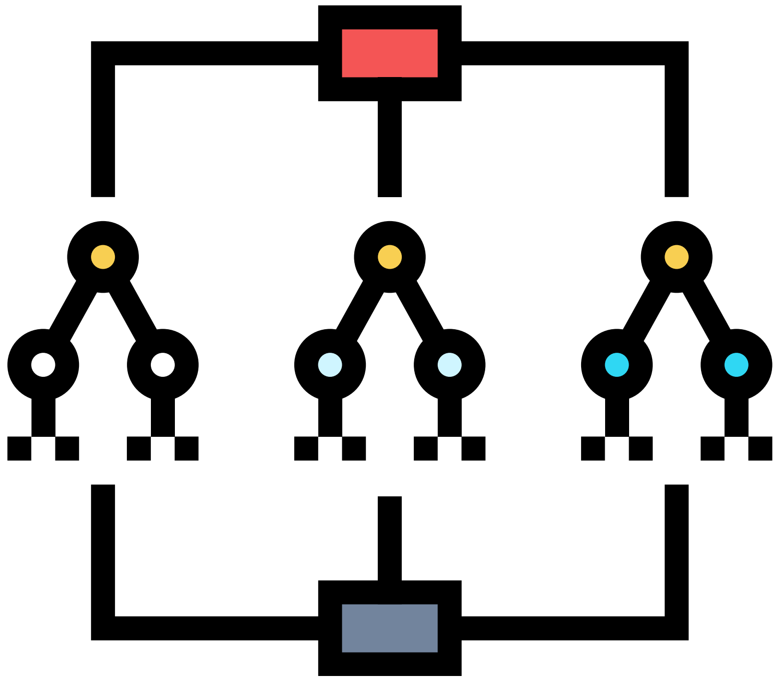
Seven Classification Models Used



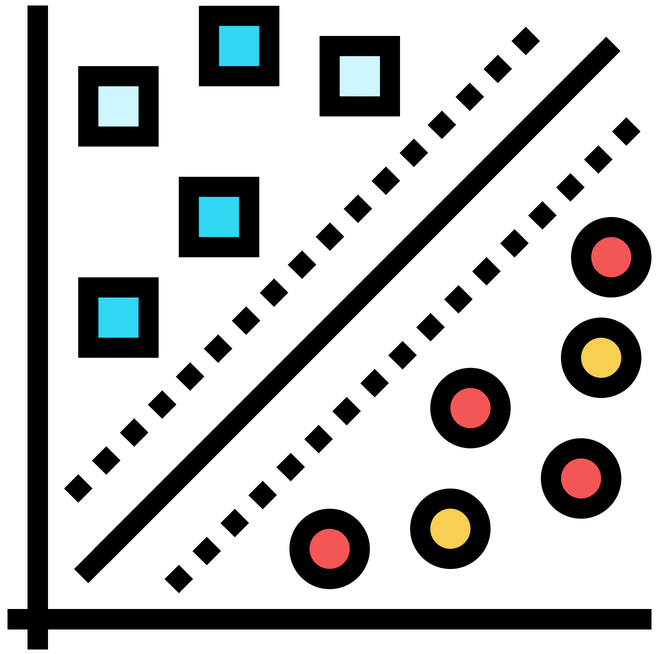
1. Logistic Regression



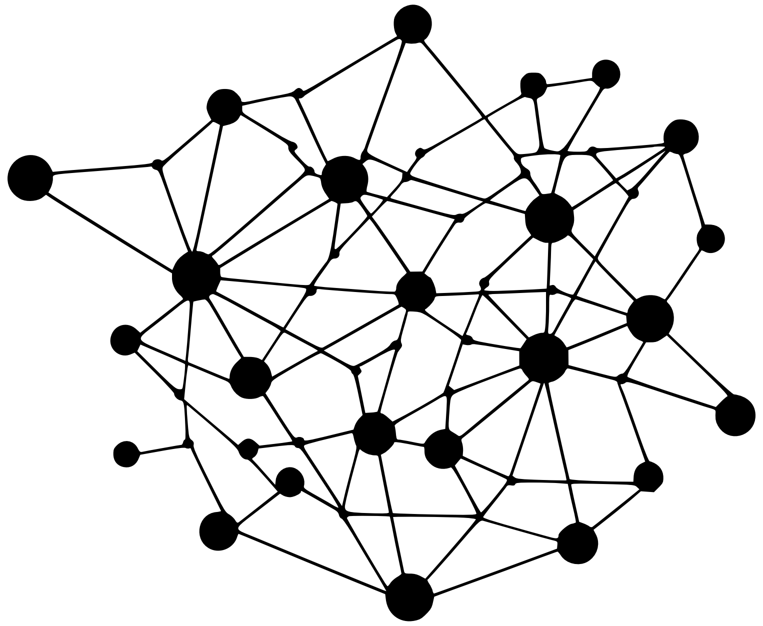
2. Random Decision Forests



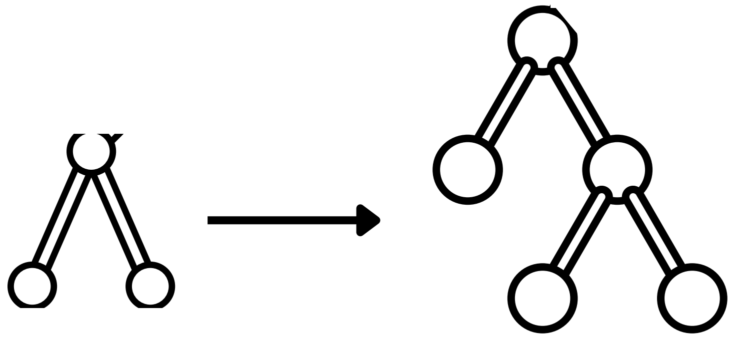
3. XGBoost



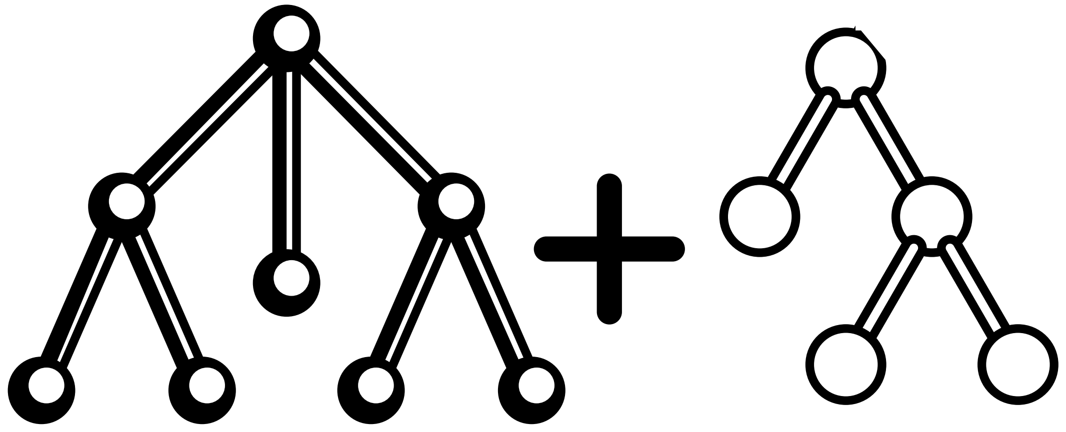
4. Support Vector Machines



5. Neural Networks



6. LightBGM



7. CatBoost

Analyzing Model Performance

01 Most Relevant Metrics

- Main focus should be on **Recall, Precision, F1-score, and ROC AUC**
  - Given the business cost of False Negatives (granting risky loans) and False Positives (rejecting good borrowers)

02 Balanced Trade-off: LightGBM & CatBoost

- Good recall, which means a number of defaulters were correctly identified
  - Important for risk management.
- These models would help minimize bad loans but may also reject some good borrowers

03 Best Precision: Random Forest

- Due to the good precision score, this model avoids rejecting too many good borrowers
- However, its lower Recall means it may miss many actual defaulters.

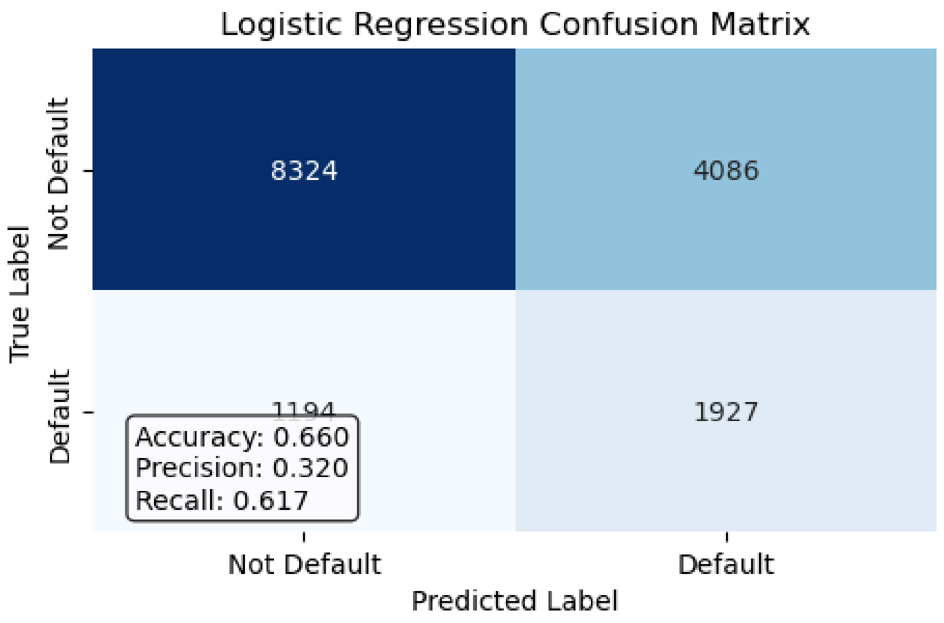
=== Model Performance Comparison ===

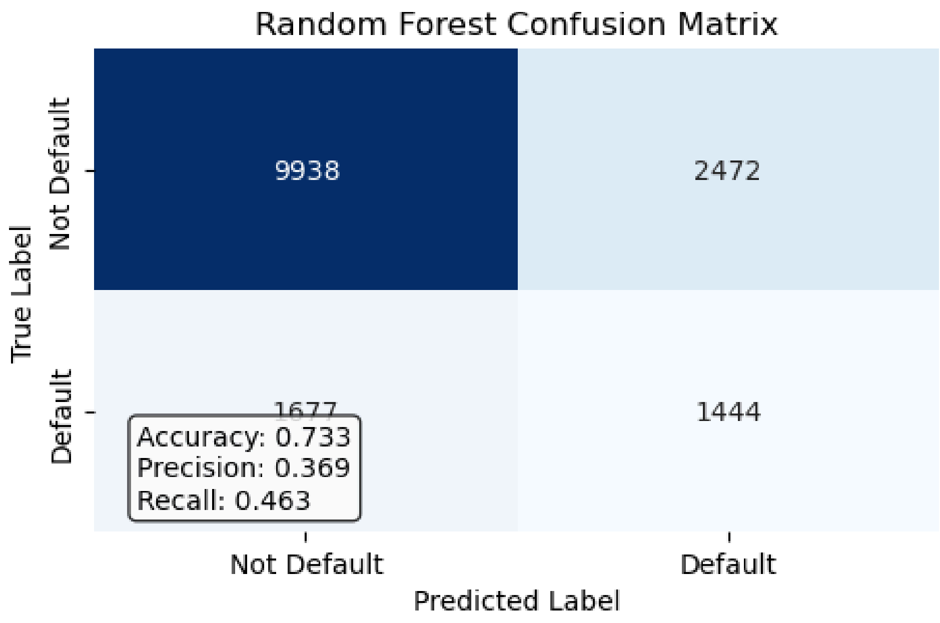
Model	Accuracy	Precision	Recall	F1	ROC AUC
Logistic Regression	0.6600	0.3205	0.6174	0.4219	0.6986
Random Forest	0.7329	0.3687	0.4627	0.4104	0.6994
XGBoost	0.6490	0.3153	0.6370	0.4218	0.7045
SVM	0.5245	0.2427	0.6443	0.3526	0.5973
Neural Network	0.7991	1.0000	0.0003	0.0006	0.5000
LightGBM	0.6485	0.3177	0.6527	0.4274	0.7076
CatBoost	0.6391	0.3123	0.6623	0.4245	0.7052

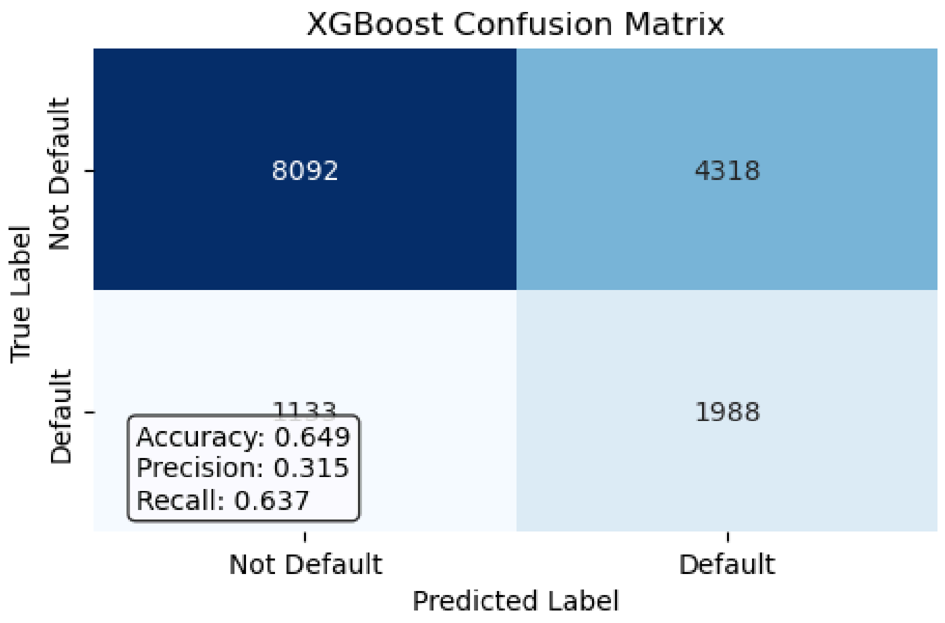
04 Neural Network Anomaly

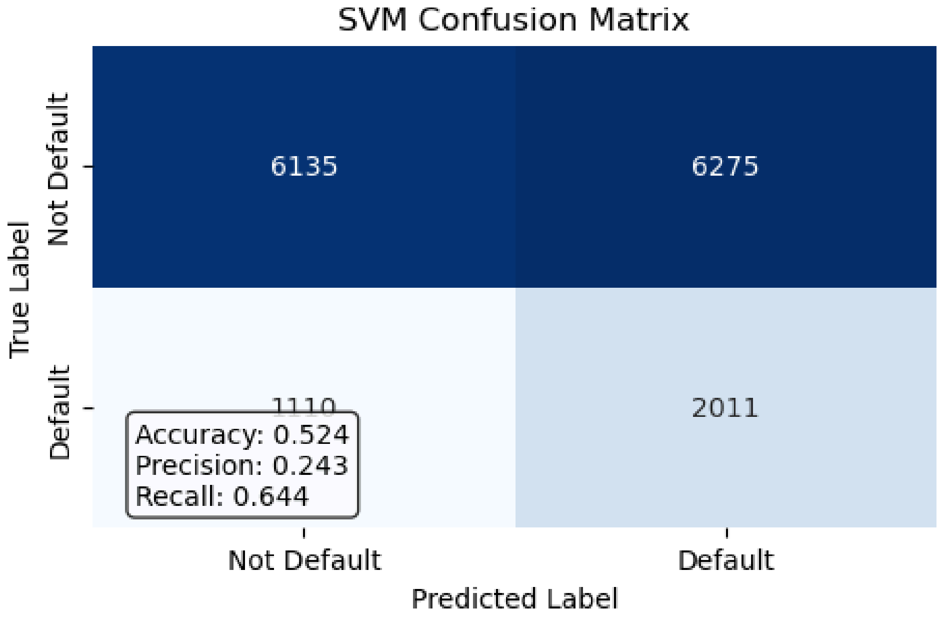
- Model is severely biased towards predicting non-defaults, classifying almost all cases as "fully paid" and ignoring defaulters
- Likely misleading due to extreme class imbalance

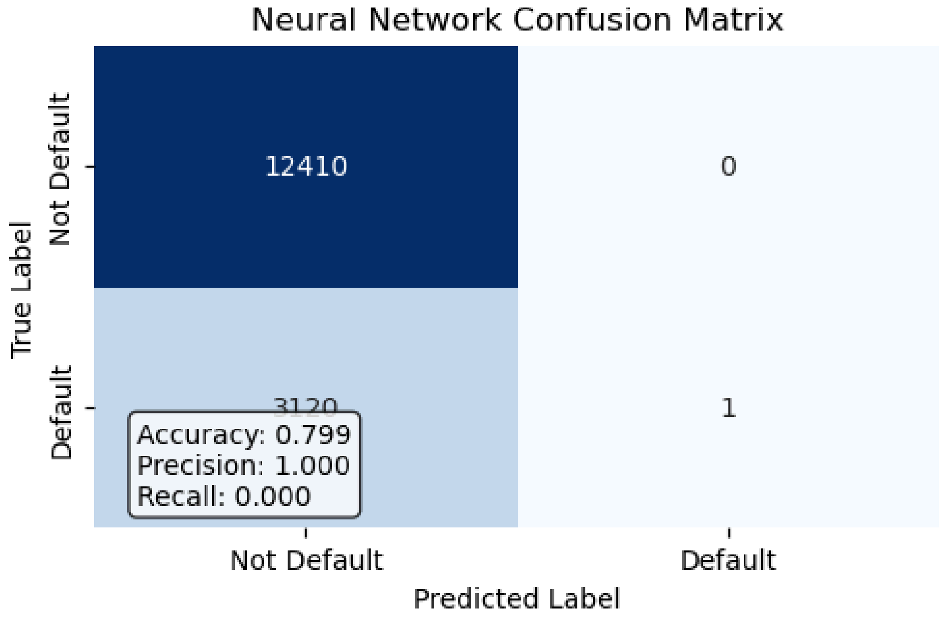
Confusion Matrices

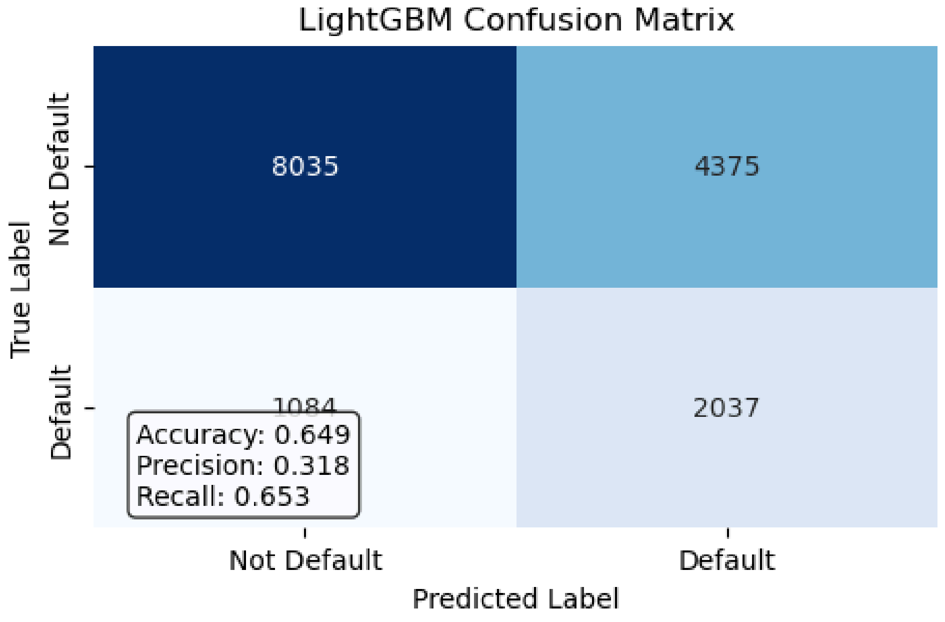


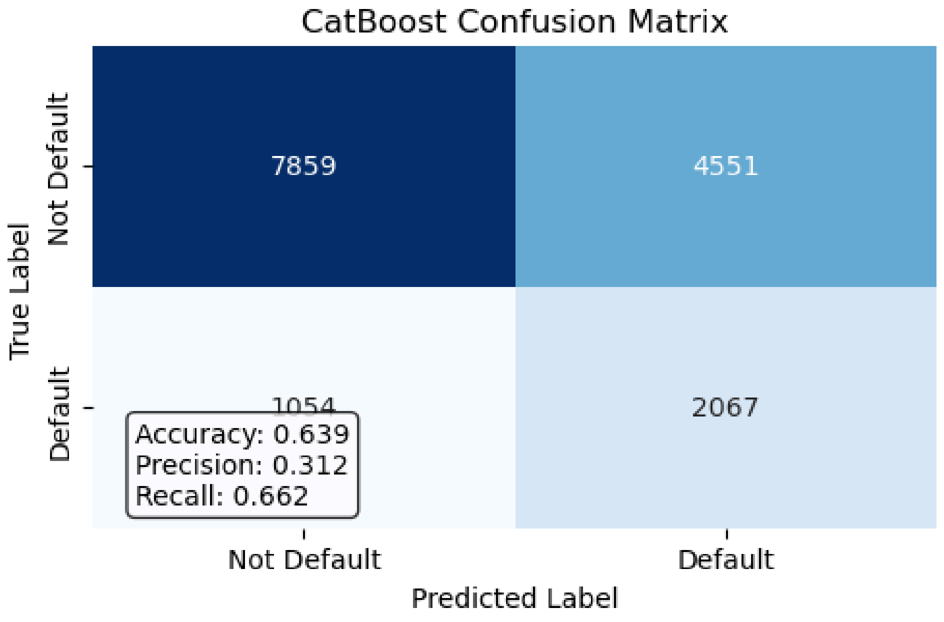












Key Insights

- High False Positives vs. False Negatives Trade-off:
  - Random Forest has the lowest false positive rate (19.9%), meaning fewer fully paid loans are mistakenly flagged as defaults.
  - SVM has the highest false positive rate (50.6%), meaning it wrongly classifies many good loans as defaults
- If you want to minimize falsely rejecting good borrowers, Random Forest is safest. If you care about catching more defaulters but at a risk, XGBoost or Logistic Regression might work better

Investment Strategies: Insights and Implications

1. Conservative Lending Strategy (Avoid Rejecting Good Loans)

a. Best Model: Random Forest (FPR: 19.9%)

i. Lowest false positive rate, meaning it rejects fewer good borrowers

b. Audience: Best for low-risk lending, prioritizing stable borrowers and minimizing lost revenue from falsely rejected loans

c. Trade-off: It misses some defaulters (lower recall) - a few bad loans may still slip through

2. Aggressive Risk-Averse Strategy (Capture Every Risky Loan)

a. Best Models: CatBoost (Recall: 66.2%), LightGBM (Recall: 65.3%)

i. Detect the most defaulters, reducing the risk of bad loans in the portfolio

b. Audience: Best for high-risk lenders, such as those offering subprime loans or credit cards

c. Trade-off: More false positives, meaning some good borrowers are wrongly denied

3. Balanced Strategy (Trade-Off Between False Positives and False Negatives

a. Best Models: XGBoost (Precision: 31.5%, Recall: 63.7%)

i. Achieves a reasonable balance between catching defaulters and not over-rejecting good loans.

b. Implication: Useful for mainstream banks and lenders who want both risk mitigation and loan approvals.

c. Trade-off: Still makes a fair number of false predictions, but it’s not extreme in either direction.

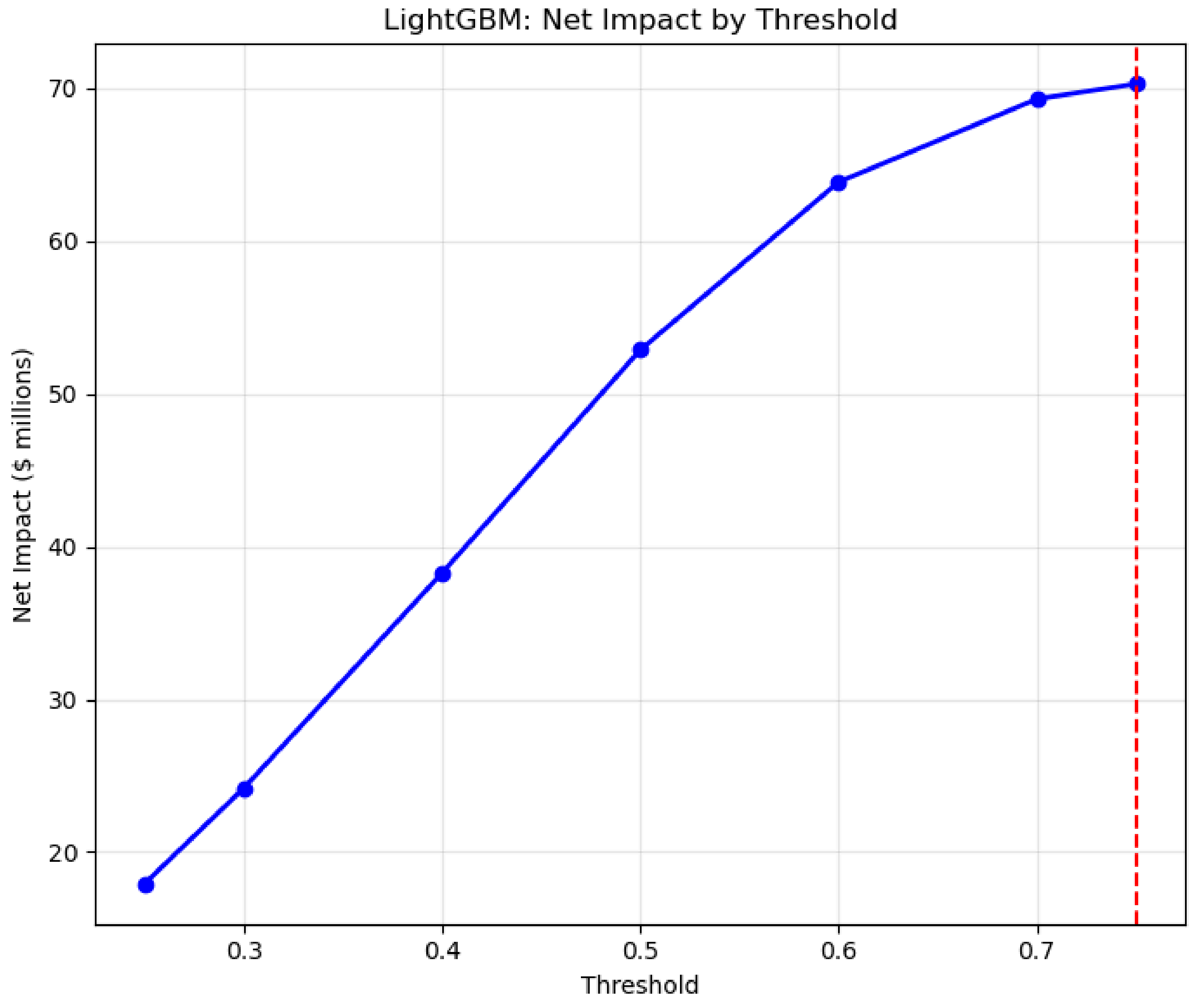
Business Impact of Best Performing Model: LightGBM

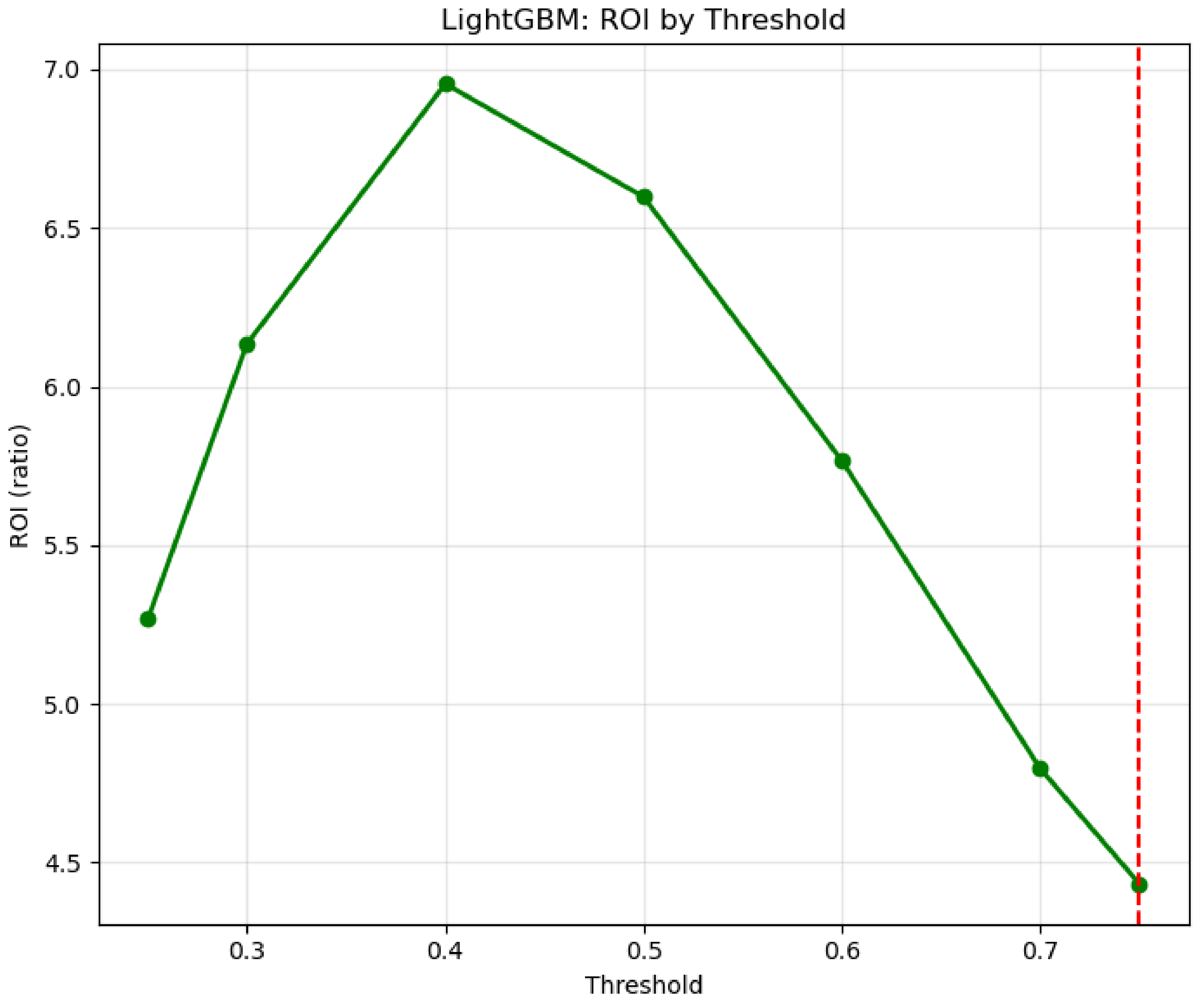
=== LightGBM ANALYSIS ===

Optimal threshold: 0.75

Business impact by threshold:

Threshold	TN	FP	FN	TP	Net Impact	ROI
0.25	2556	9854	166	2955	$17,923,800.00	5.27x
0.30	3470	8940	270	2851	$24,169,000.00	6.13x
0.40	5619	6791	585	2536	$38,253,700.00	6.95x
0.50	8035	4375	1084	2037	$52,857,000.00	6.60x
0.60	10108	2302	1694	1427	$63,841,400.00	5.77x
0.70	11575	835	2400	721	$69,283,000.00	4.80x
0.75	12045	365	2715	406	$70,271,500.00	4.43x

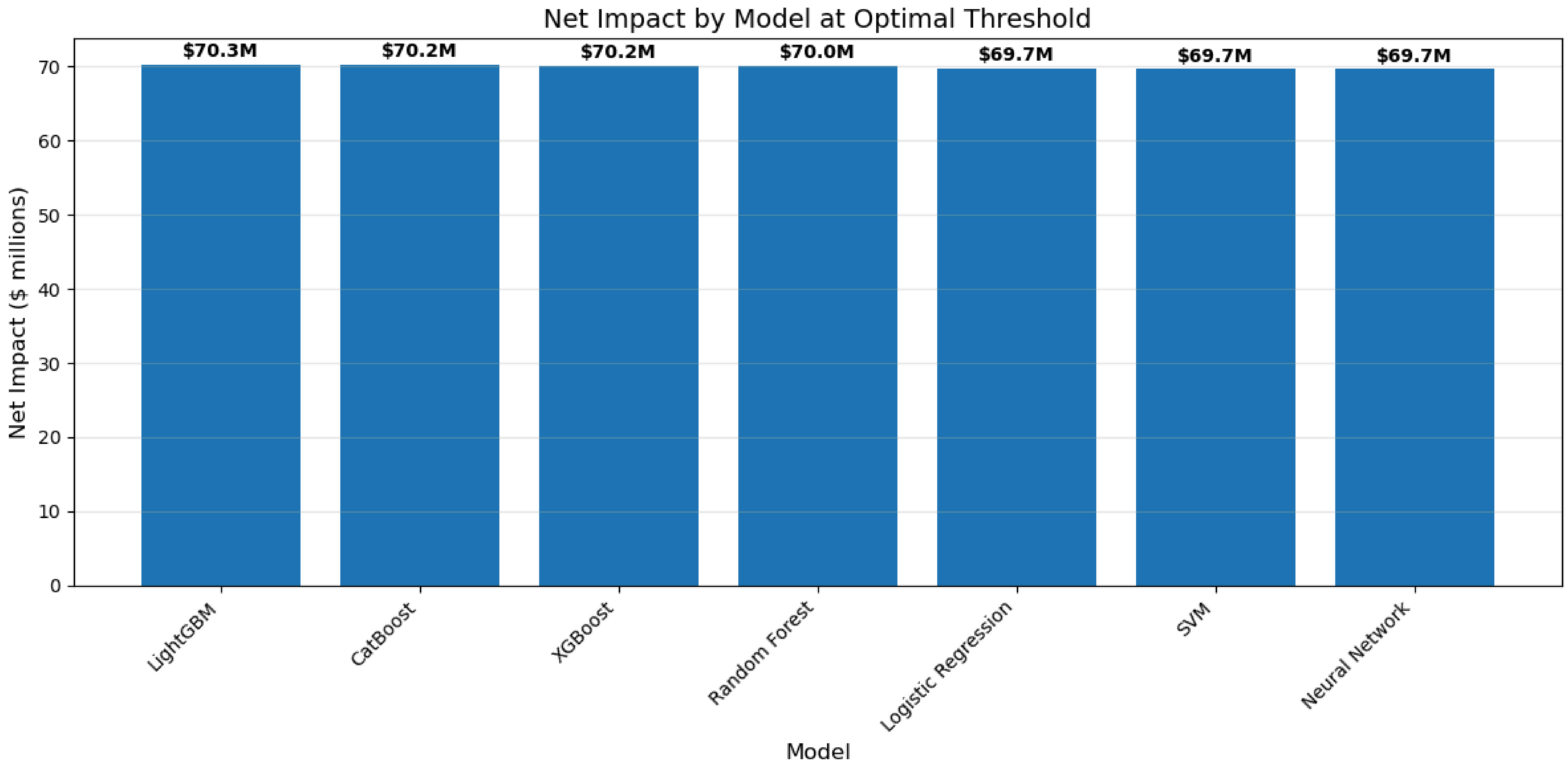




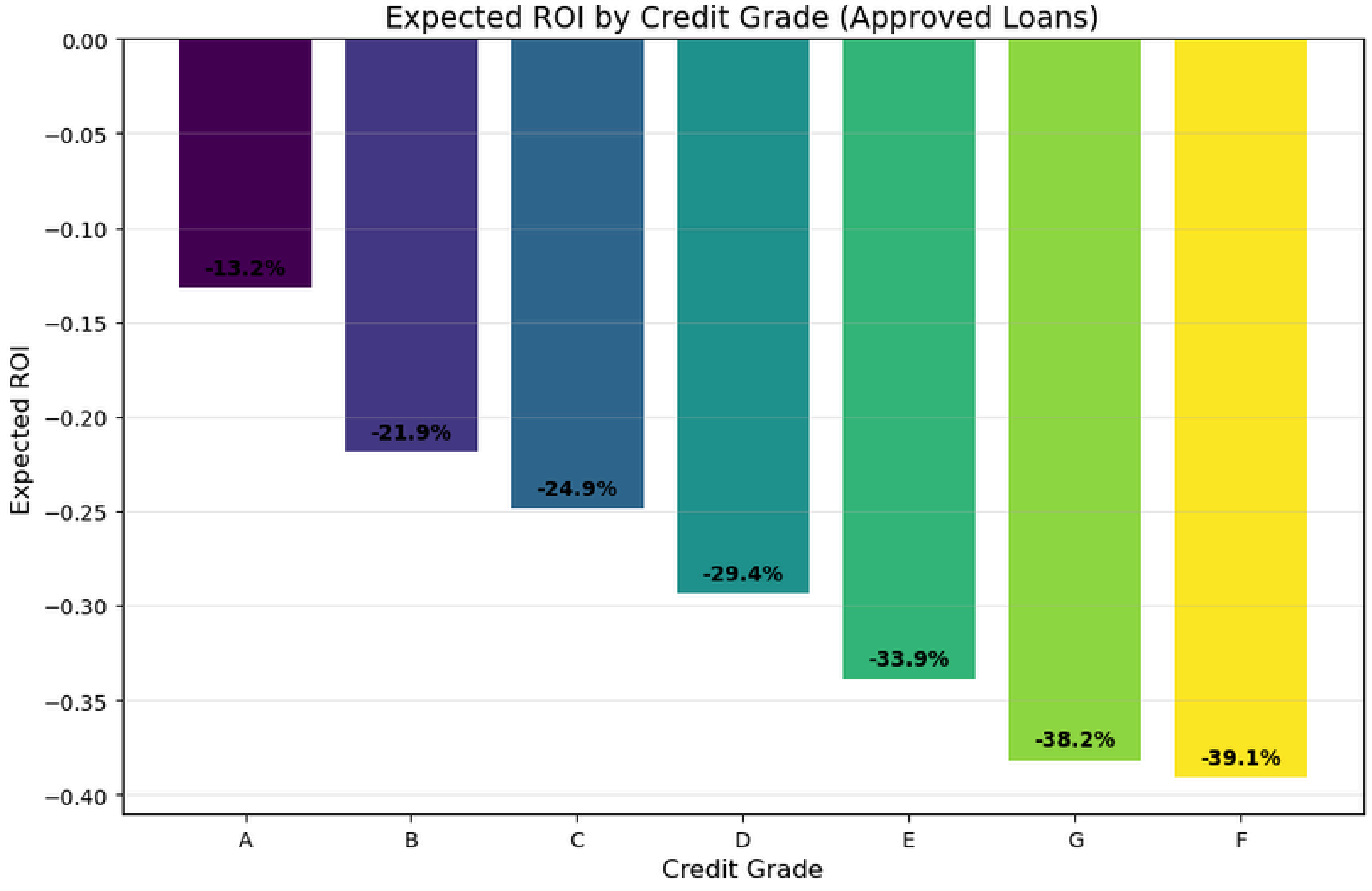
Optimal Threshold Comparison of All Models

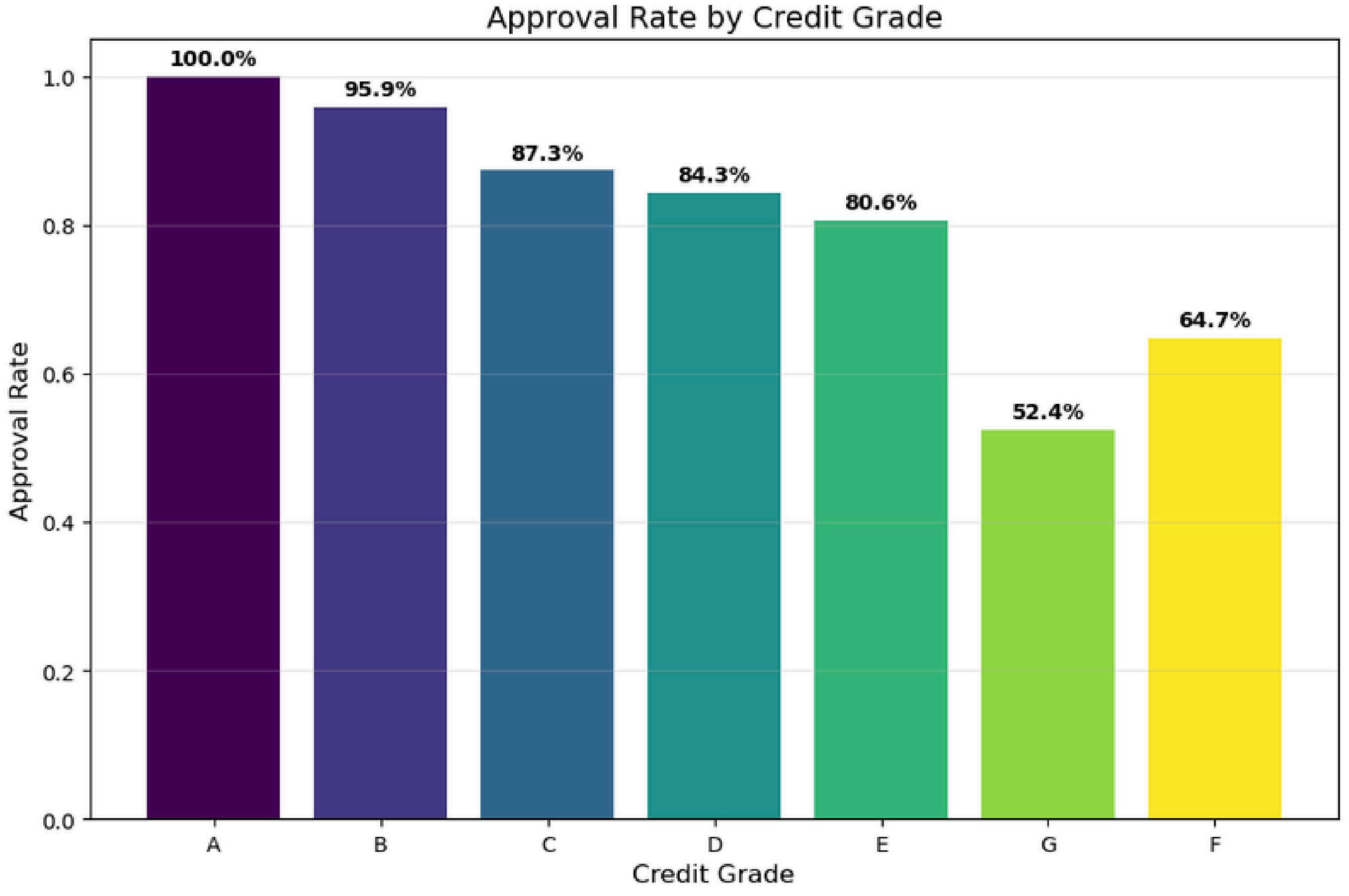
=== OPTIMAL THRESHOLD COMPARISON ===

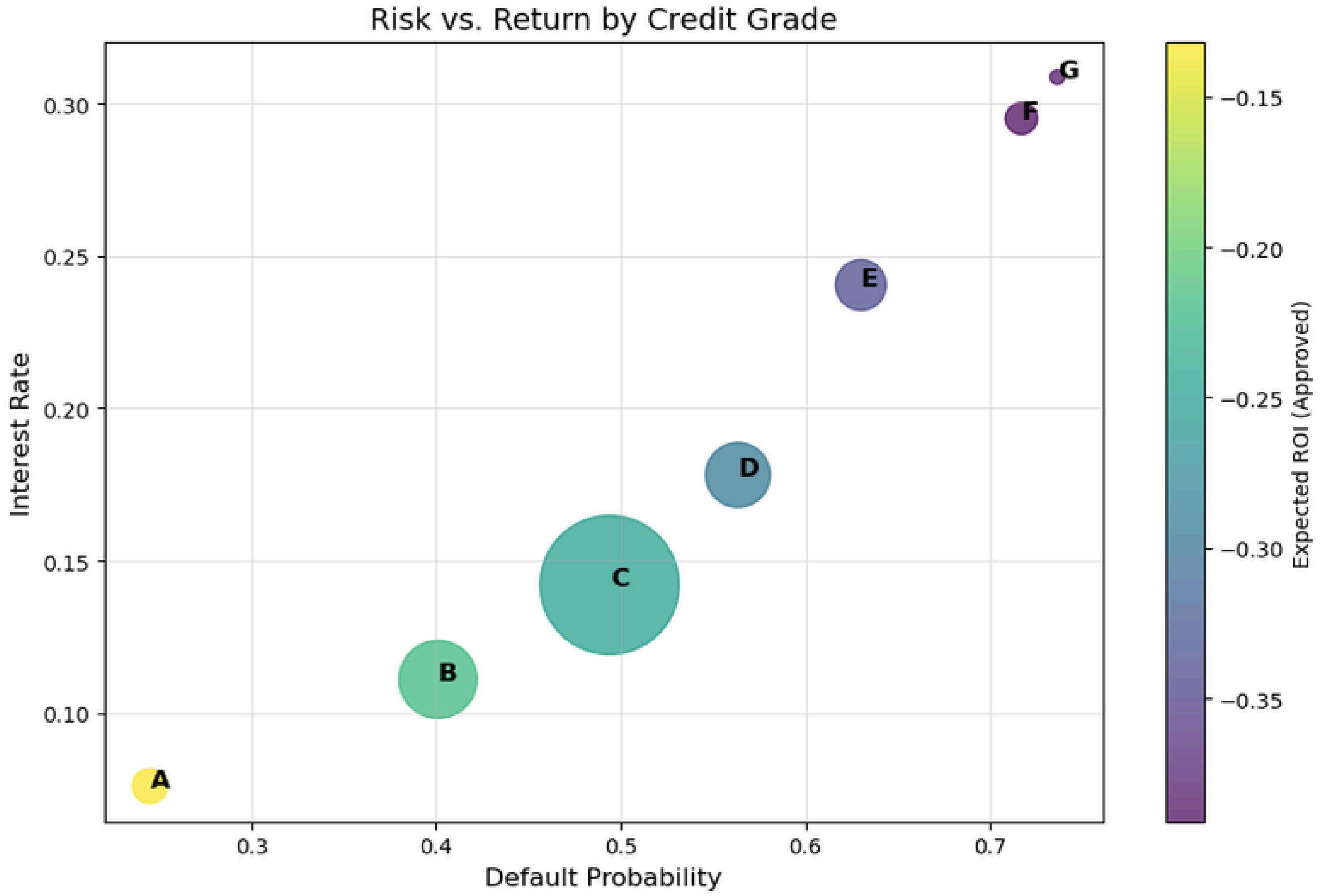
Model	Optimal Threshold	Net Impact	ROI	TN	FP	FN	TP
LightGBM	0.75	$70,271,500.00	4.43x	12045.0	365.0	2715.0	406.0
CatBoost	0.75	$70,227,900.00	4.38x	12083.0	327.0	2755.0	366.0
XGBoost	0.75	$70,150,200.00	4.43x	12024.0	386.0	2710.0	411.0
Random Forest	0.70	$70,048,500.00	4.28x	12145.0	265.0	2833.0	288.0
Logistic Regression	0.75	$69,693,600.00	4.49x	11887.0	523.0	2638.0	483.0
SVM	0.40	$69,676,000.00	3.98x	12410.0	0.0	3120.0	1.0
Neural Network	0.25	$69,676,000.00	3.98x	12410.0	0.0	3120.0	1.0

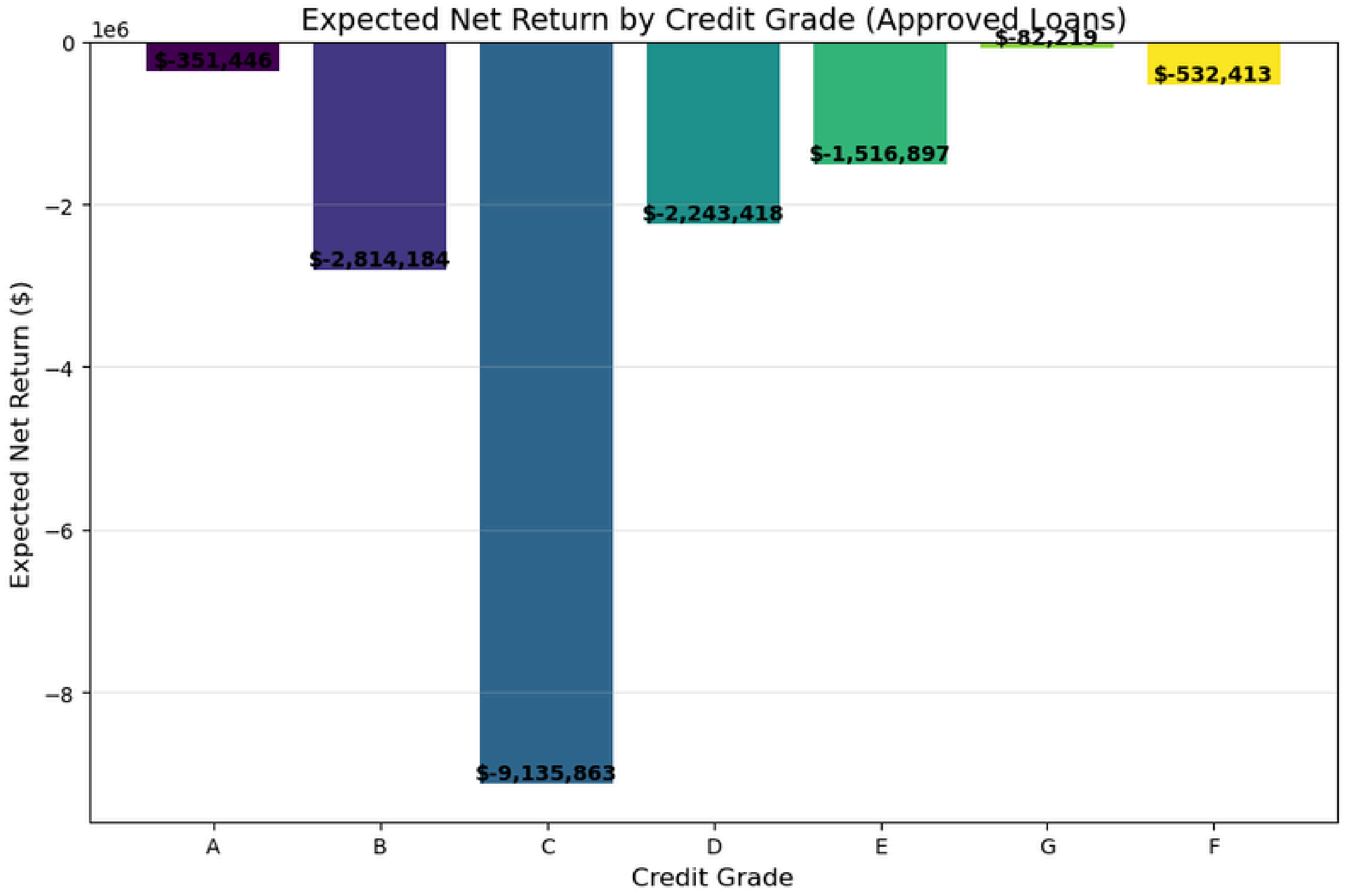


Open Loans Prediction (Based on Closed Loans Model)









Open Loans Prediction Strategy

1. Investment Strategies:

1. High-Grade Strategy (A & B)
   - Pros: High approval rates, low default probabilities.
   - Cons: Low interest rates, negative expected returns.
   - Assessment: Low risk but results in losses due to inadequate interest rates compared to defaults.
2. Mid-Grade Strategy (C & D)
   - Pros: Higher interest rates than A & B.
   - Cons: Higher default probabilities, significant negative net return (especially for C).
   - Assessment: Riskier than A & B but still unprofitable due to high default rates outweighing interest gains.
3. High-Risk, High-Return Strategy (E, F, G)
   - Pros: Highest interest rates, lower approval rates controlling exposure.
   - Cons: Very high default probabilities, potential for extreme losses.
   - Assessment: Only E and G show marginally positive net returns, making them selectively viable.

Optimal Strategy to Maximize Returns & Manage Risk:

1. Avoid lending to C & D entirely due to poor risk-return balance.
2. Selectively invest in E & G, focusing on loans with moderate risk indicators within these grades.
3. Minimize exposure to A & B since they are safe but generate losses.
4. Implement risk-adjusted approval criteria to optimize the mix of loans.