Smart Loan Advisor For Better Loan Recommendations

A Machine Learning Project by Syed Asad Ali Kazmi

Project Description

- Developed and assessed 8 distinct models, both with and without SMOTE, for loan approval prediction.
- Utilized logistic regression, random forest classifier, and a hybrid model with voting.
- Objective: Automate loan approval decisions via binary classification.
- Employed a synthesized dataset featuring customer attributes and loan application details.
- Models predict whether loans are approved (1) or not (0).
- Emphasis on data preprocessing, feature selection, model training, evaluation, and visualization.
- Gained insights into practical machine learning applications and predictive model creation.

DATA QUALITY ASSESSMENT

- Examined key statistics and data distribution.
- Removed duplicate entries, addressed missing values, lowercase and standardized variable formats.
- Validated responses against source documents, corrected inaccuracies, and resolved outliers.
- Checked for logical consistency in data columns, ensuring data coherence.

DATA OVERVIEW

The dataset contain information about loan applicants. A brief summary of the columns:

1. `person_age`: The age of the loan applicant.

2. `person_income`: The income of the loan applicant.

3. `person_home_ownership`: The type of home ownership for the applicant (e.g., RENT, OWN, MORTGAGE).

4. `person_emp_length`: The length of employment for the applicant.

5. `loan_intent`: The purpose or intent of the loan (e.g., PERSONAL, EDUCATION, VENTURE).

6. `loan_grade`: The loan grade assigned to the applicant (e.g., A, B, C).

DATA OVERVIEW

7. `loan_amnt`: The requested loan amount.

8. `loan_int_rate`: The interest rate for the loan.

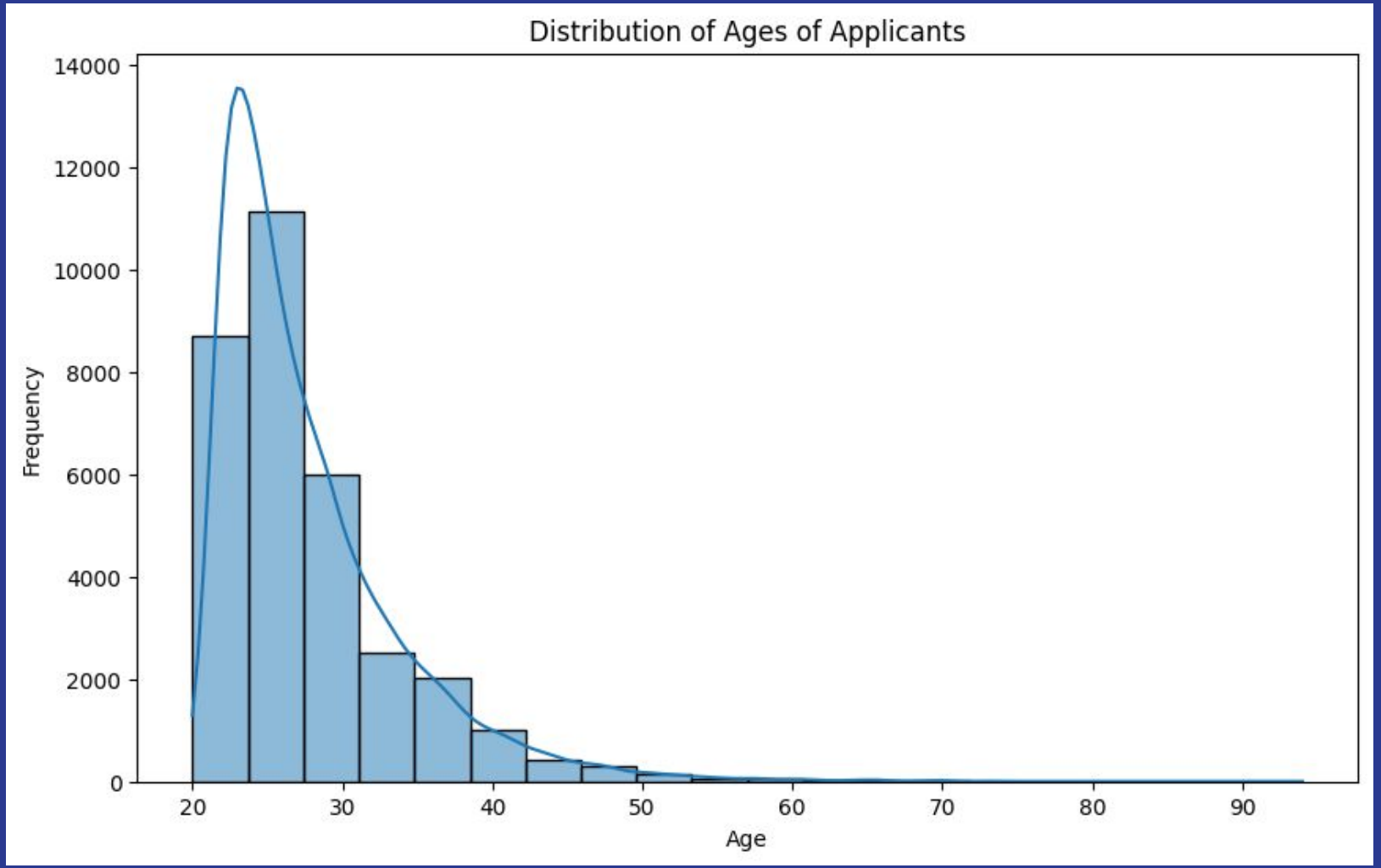
9. `loan_status`: The loan approval status (0 for not approved, 1 for approved).

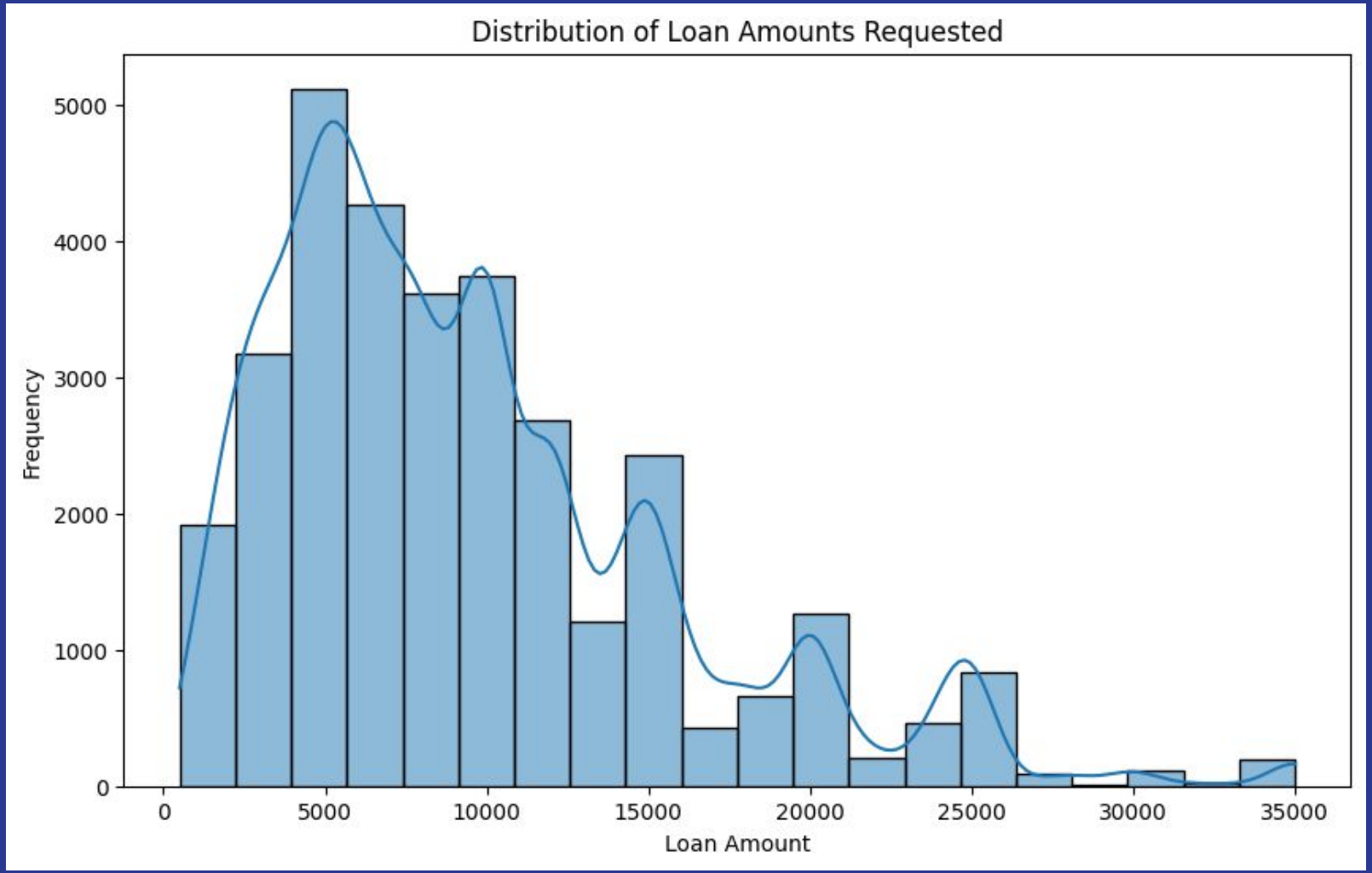
10. `loan_percent_income`: The loan amount as a percentage of the applicant's income.

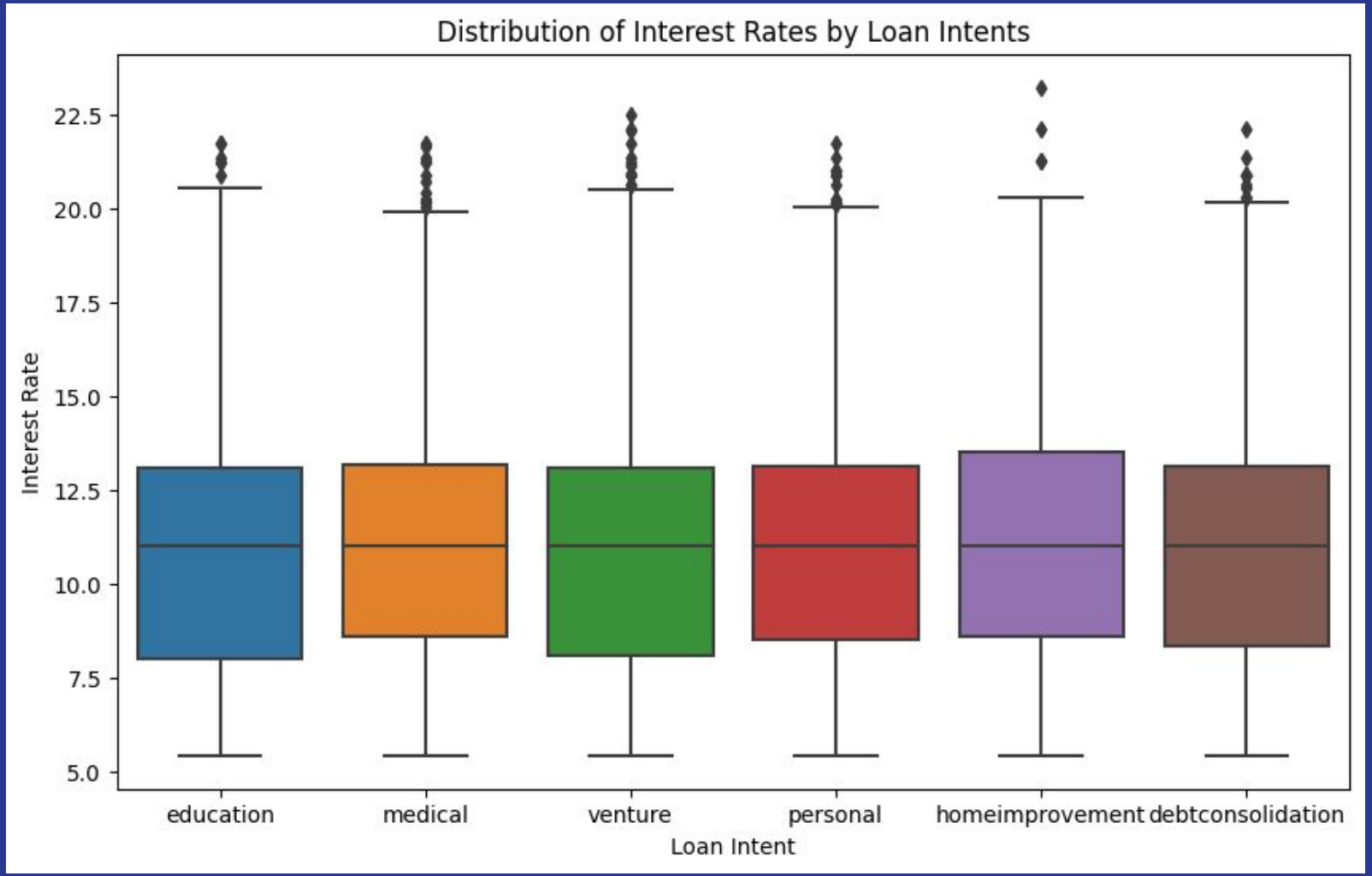
11. `cb_person_default_on_file`: Whether the applicant has a default on file (Y or N).

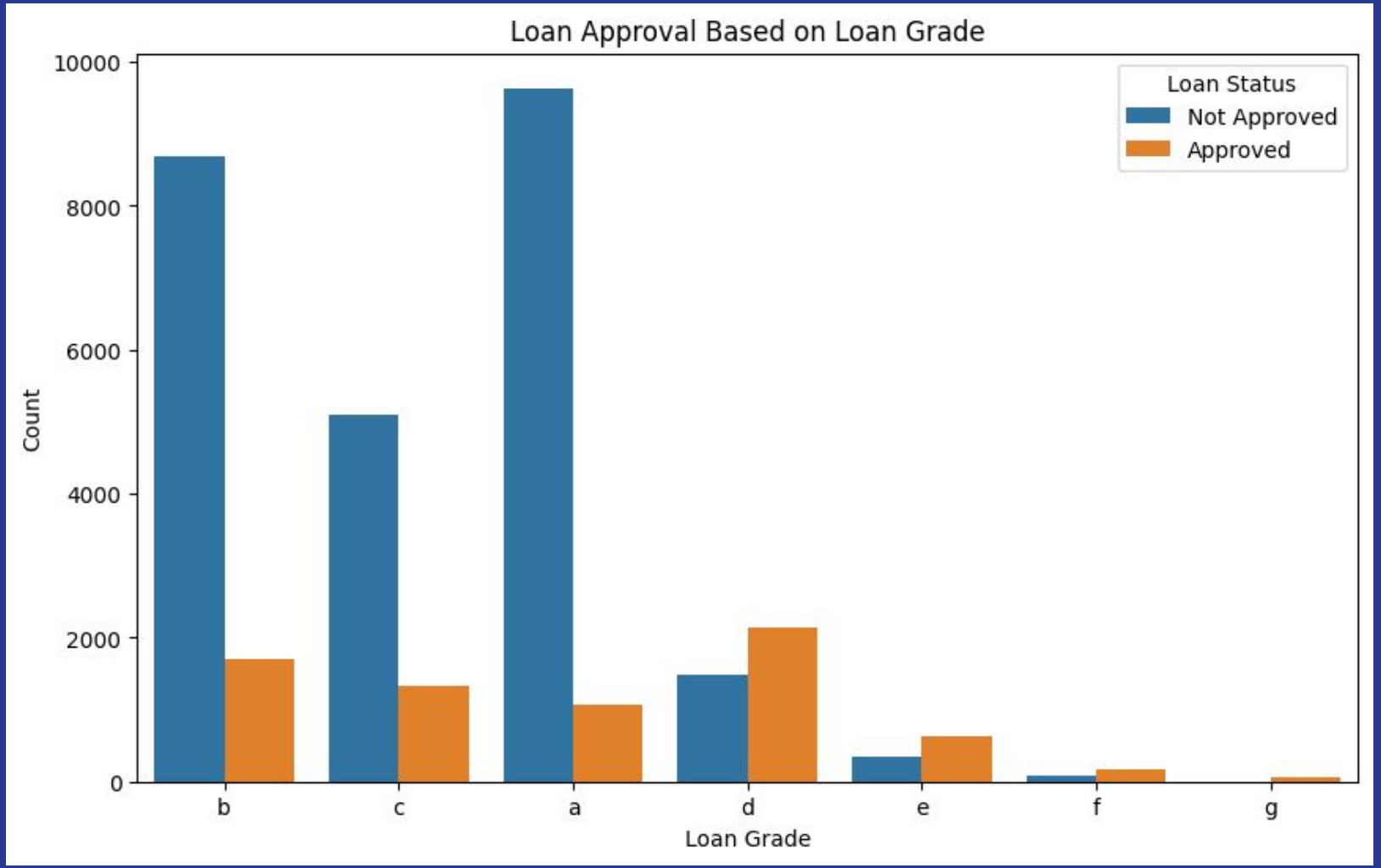
12. `cb_person_cred_hist_length`: The length of the applicant's credit history.

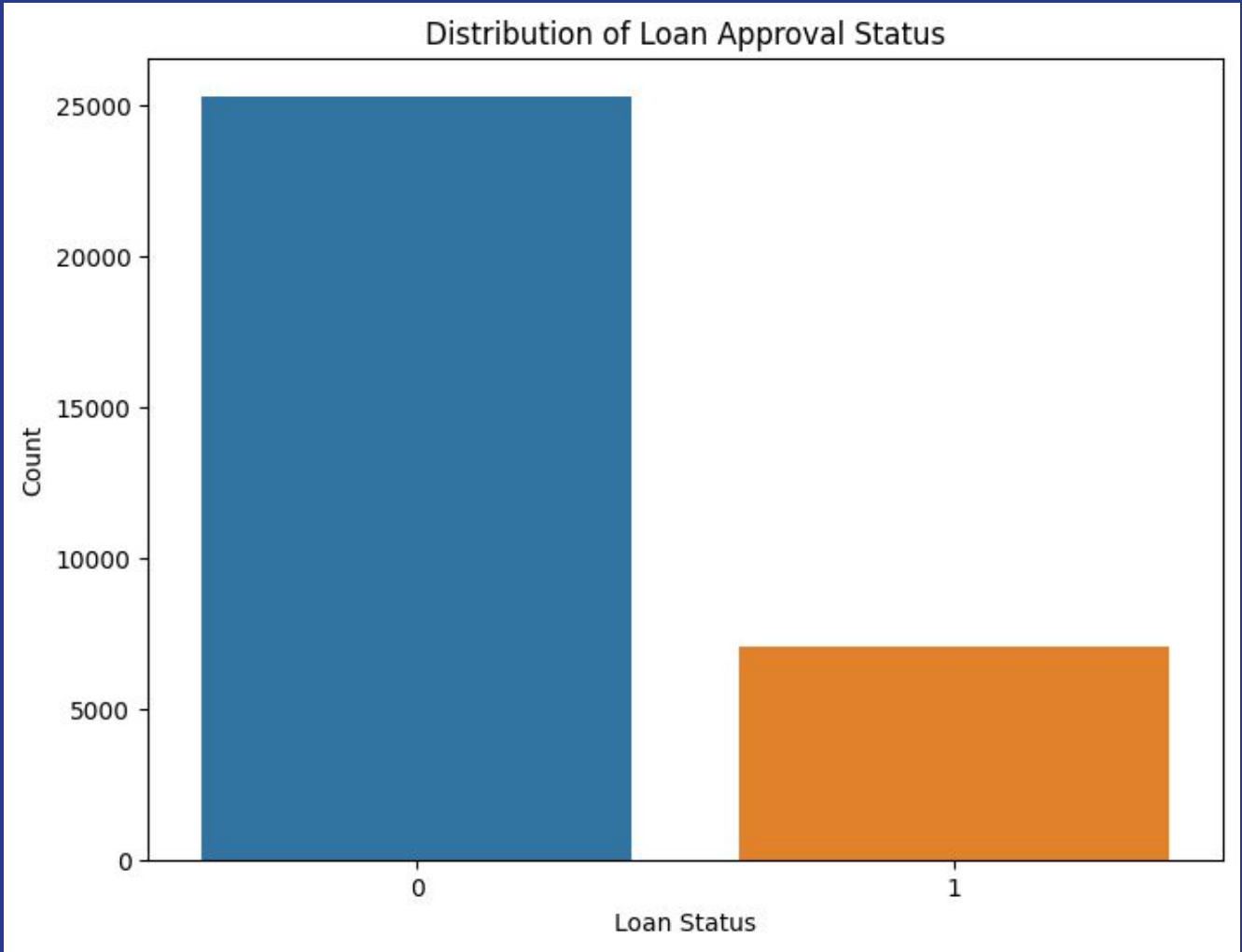
EXPLORATORY DATA ANALYSIS

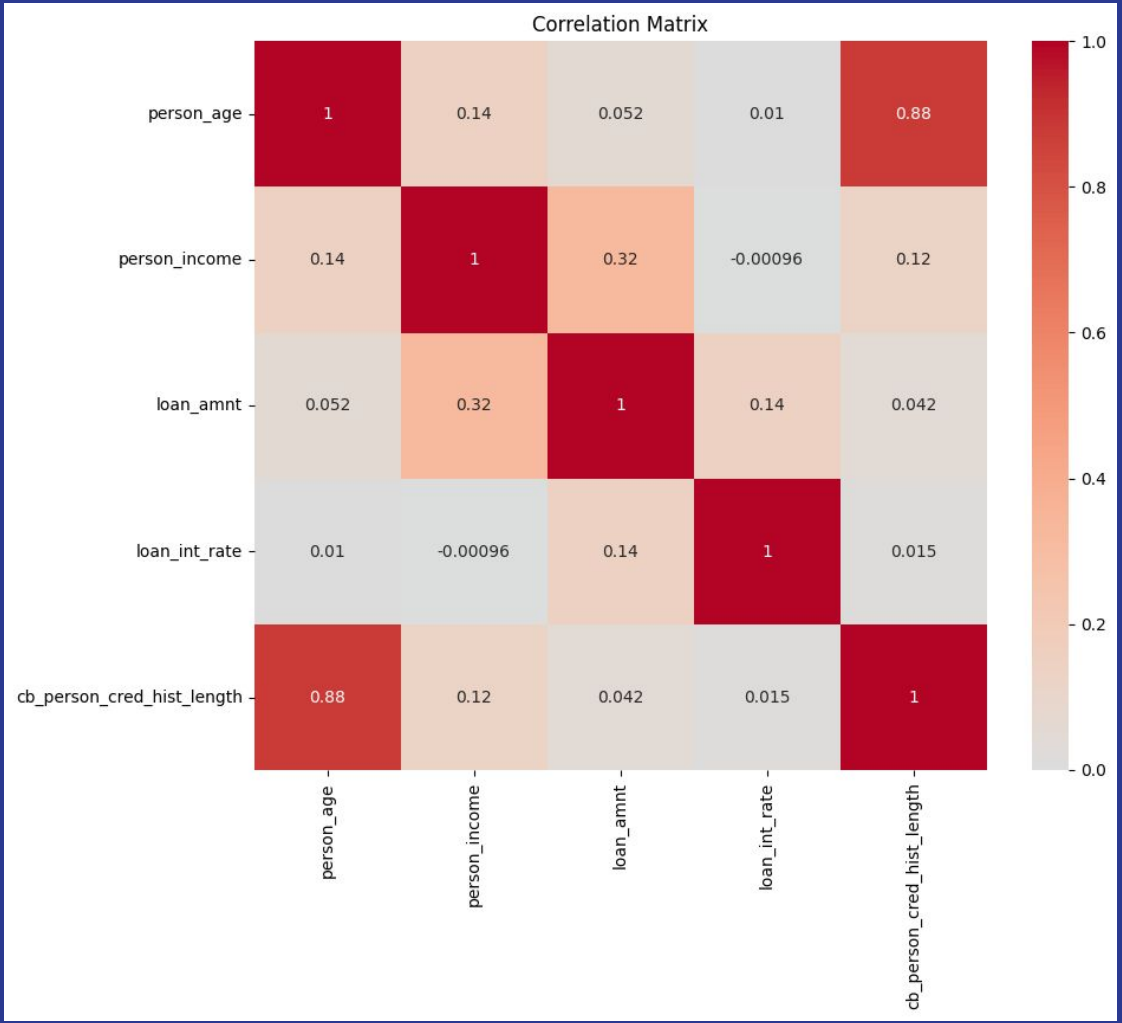

Distribution of Ages of Applicants
Frequency
Age
0
2000
4000
6000
8000
10000
12000
14000
20
30
40
50
60
70
80
90

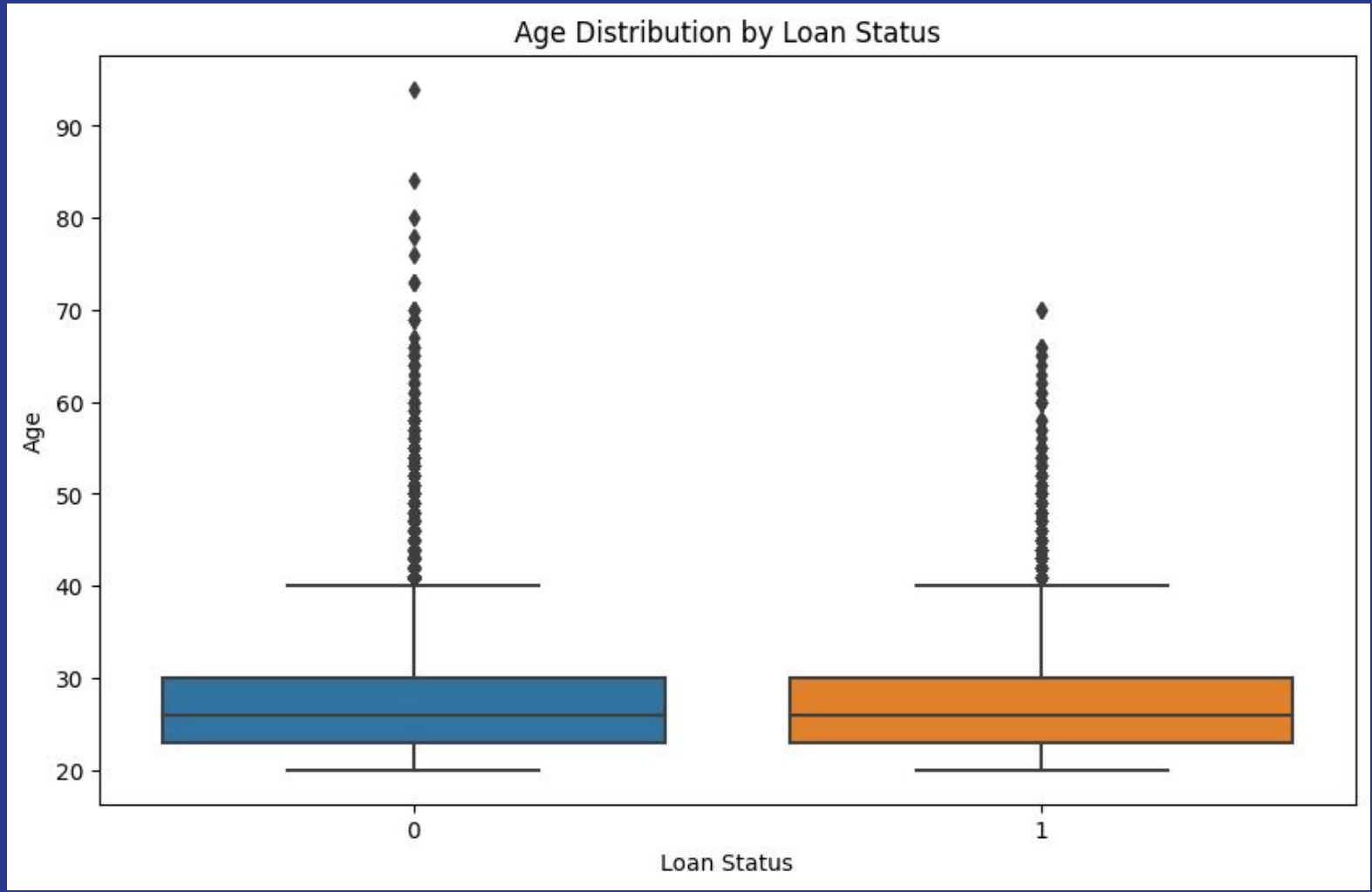

Distribution of Loan Amounts Requested
Frequency
Loan Amount
0
1000
2000
3000
4000
5000
0
5000
10000
15000
20000
25000
30000
35000

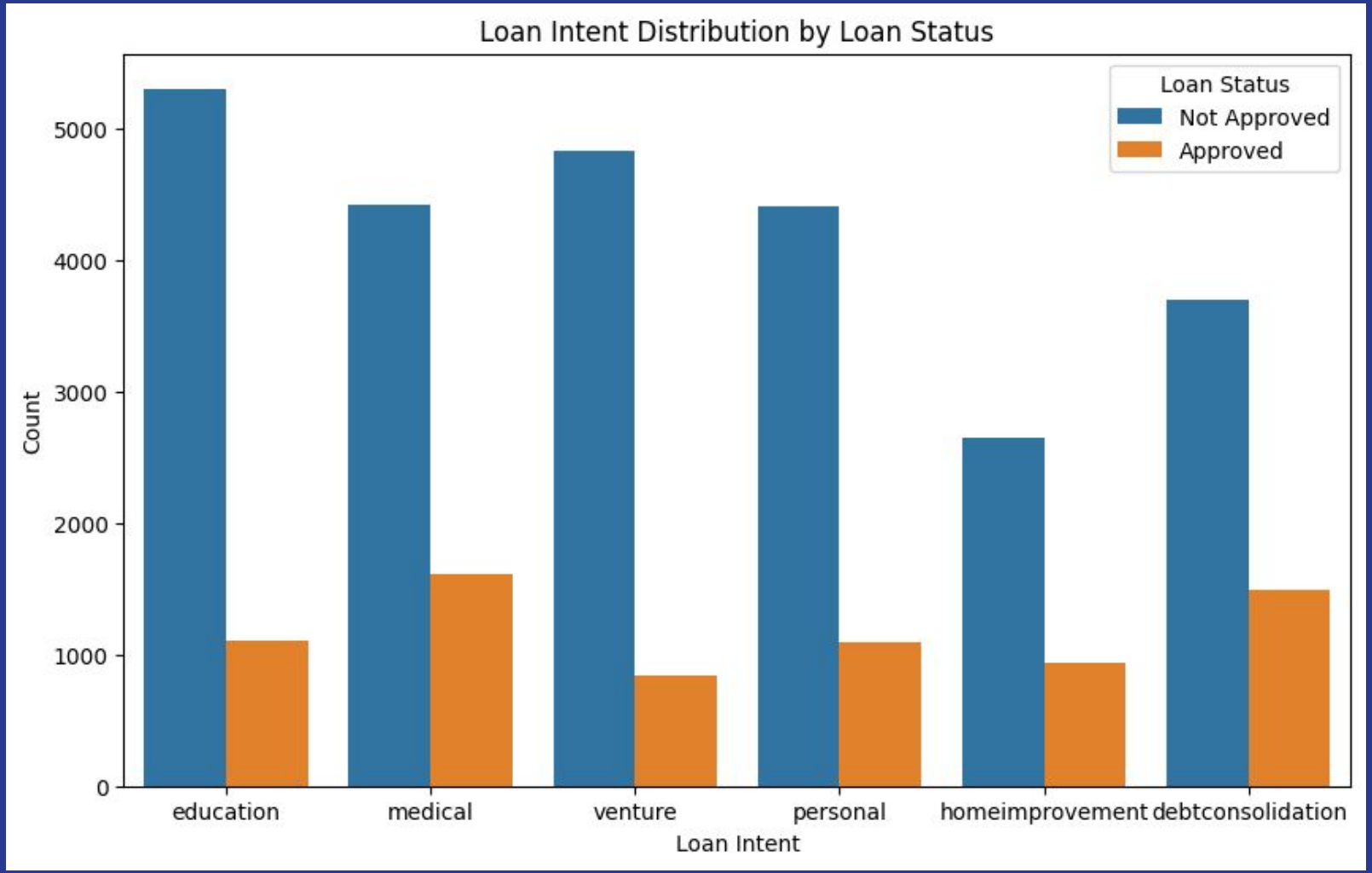

Distribution of Interest Rates by Loan Intents
Interest Rate
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
education
medical
venture
personal
homeimprovement
debtconsolidation
Loan Intent

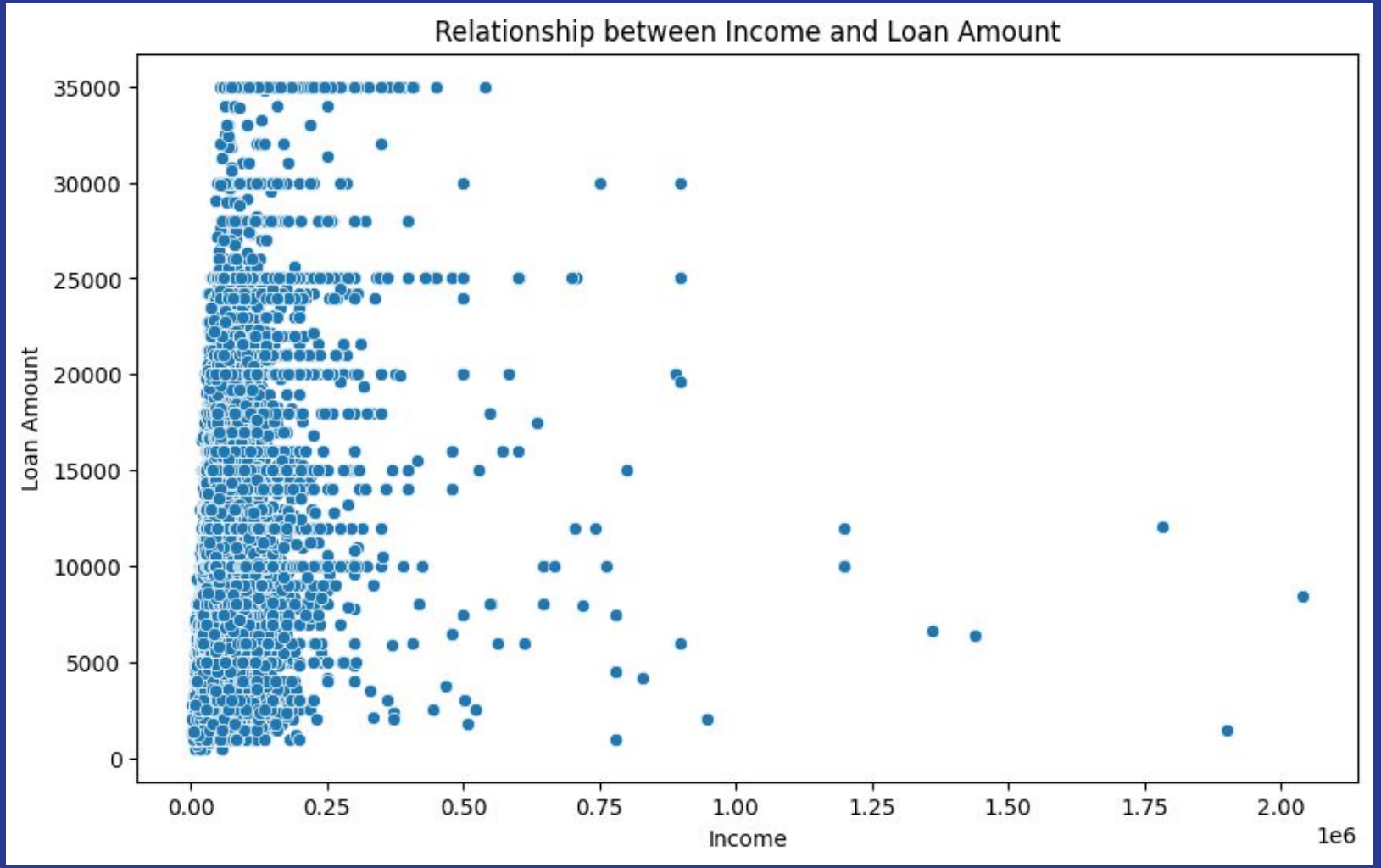

Loan Approval Based on Loan Grade
Loan Status
Not Approved
Approved
Count
10000
8000
6000
4000
2000
0
b
c
a
d
e
f
g
Loan Grade

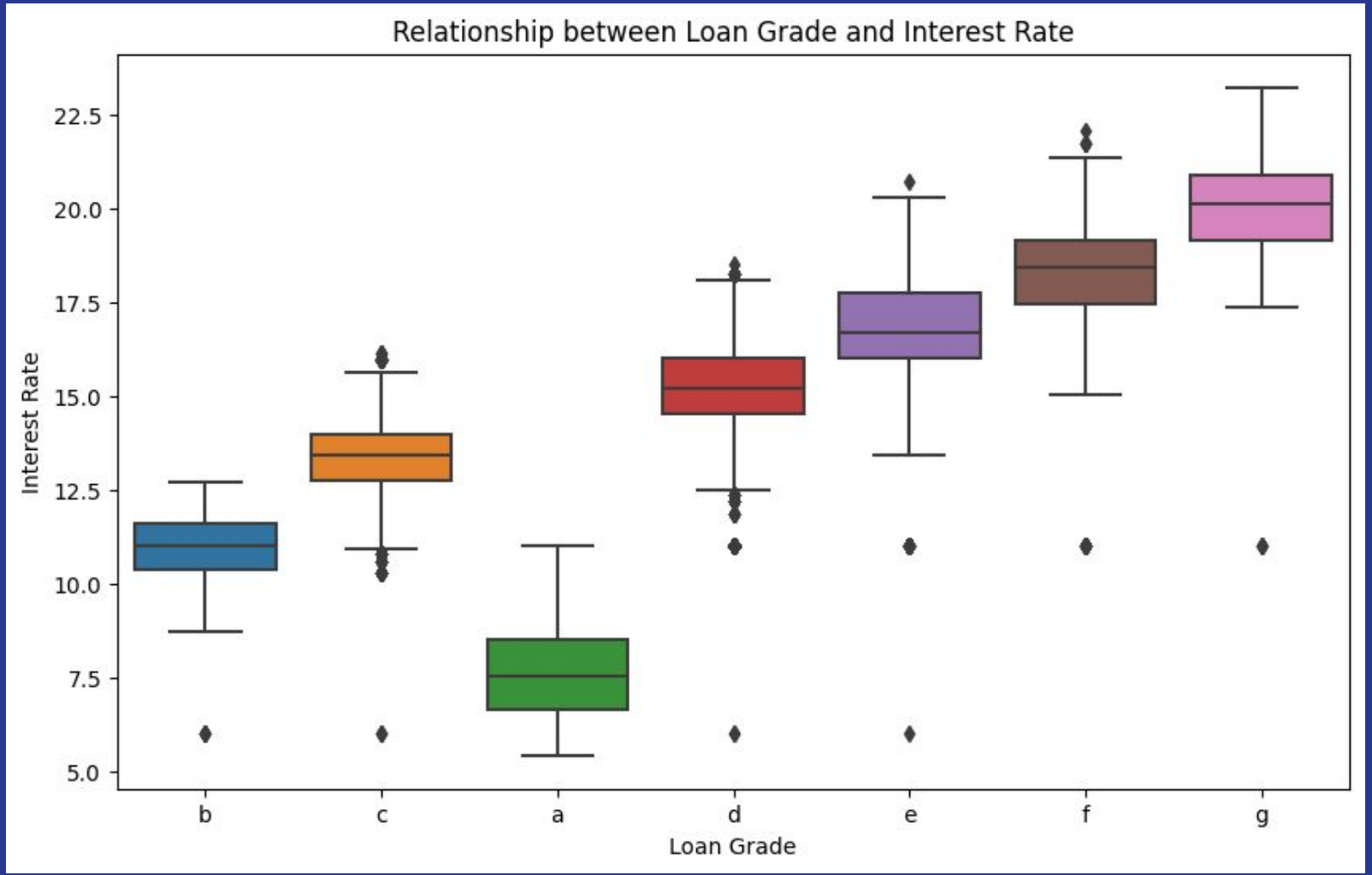

Distribution of Loan Approval Status
25000
20000
15000
10000
5000
0
Count
0
1
Loan Status

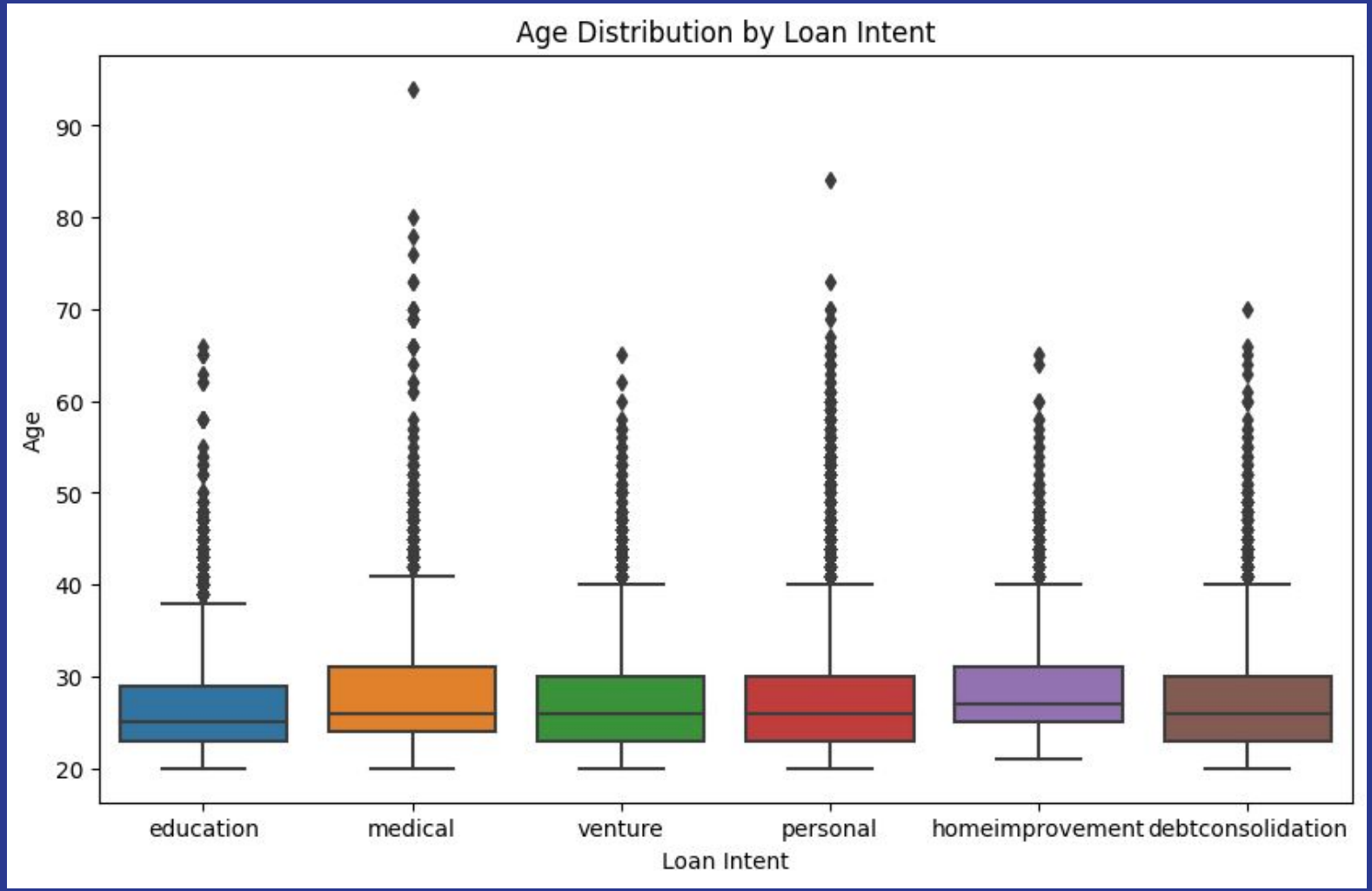

Correlation Matrix
person_age
person_income
loan_amnt
loan_int_rate
cb_person_cred_hist_length
1
0.14
0.052
0.01
0.88
0.14
1
0.32
-0.00096
0.12
0.052
0.32
1
0.14
0.042
0.01
-0.00096
0.14
1
0.015
0.88
0.12
0.042
0.015
1
1.0
0.8
0.6
0.4
0.2
0.0

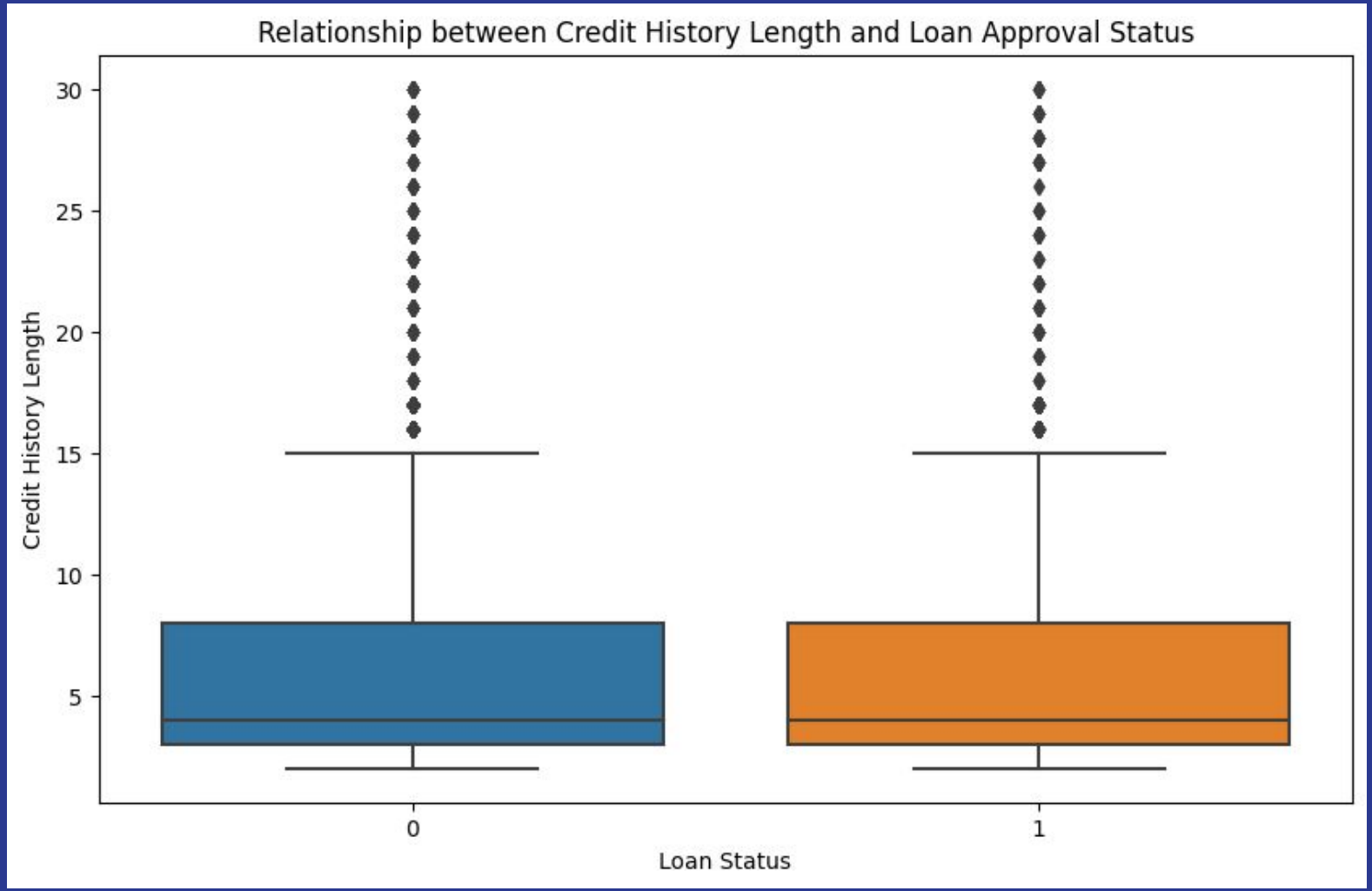

Age Distribution by Loan Status
Age
90
80
70
60
50
40
30
20
0
1
Loan Status


Loan Intent Distribution by Loan Status
Loan Status
Not Approved
Approved
Count
0
1000
2000
3000
4000
5000
education
medical
venture
personal
homeimprovement
debtconsolidation
Loan Intent


Relationship between Income and Loan Amount
Loan Amount
35000
30000
25000
20000
15000
10000
5000
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
2.00
Income
1e6


Relationship between Loan Grade and Interest Rate
Interest Rate
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
b
c
a
d
e
f
g
Loan Grade


Age Distribution by Loan Intent
Age
90
80
70
60
50
40
30
20
education
medical
venture
personal
homeimprovement
debtconsolidation
Loan Intent


Relationship between Credit History Length and Loan Approval Status
Credit History Length
30
25
20
15
10
5
0
1
Loan Status

Non-Parametric Test:

Kruskal-Wallis Test
p-value: 0.0099

Mann-Whitney U Test
p-value: (approximately zero)

There are differences in interest rates among different loan intents.

The ages of individuals with approved loans and not approved loans are significantly different.

Hypothesis Testing

T-Test p-value for Income between Defaulted and Not Defaulted: 0.6501136224274147.

- The income levels of these two groups are not significantly different.

ANOVA p-value for Loan Amount by Loan Intent: (approximately zero)

- The average loan amounts are not the same across all loan intents.

Point Biserial Correlation: -0.016440588272711207

- It indicates a very weak negative correlation between credit history length and the likelihood of loan approval (loan_status). The negative sign indicates that as credit history length increases, the likelihood of loan approval slightly decreases. However, the correlation coefficient is close to zero, suggesting that the relationship is negligible.

ML Models Used

Model 1: Logistic Regression with Scalar

Model 2: Logistic Regression with SMOTE

Model 3: Logistic Regression with SMOTE & Scalar

Model 4: Logistic Regression with SMOTE & Scalar & Grid Search

Model 5: Random Forest Classifier

Model 6: Random Forest Classifier with SMOTE & Grid Search

Model 7: Hybrid Model (RFC+XGBC+SVC)

Model 8: Hybrid Model (RFC+XGBC+SVC) with Scalar & SMOTE

Evaluation Metrics

Model	Accuracy	ROC AUC	Precision (Class 0)	Recall (Class 0)	F1 Score (Class 0)	Precision (Class 1)	Recall (Class 1)	F1 Score (Class 1)	Remarks
1	0.82	0.64	0.83	0.96	0.89	0.7	0.31	0.43	
2	0.7	0.71	0.91	0.68	0.78	0.39	0.75	0.52	
3	0.72	0.72	0.91	0.72	0.8	0.42	0.73	0.53	
4	0.72	0.72	0.91	0.72	0.8	0.42	0.73	0.53	
5	0.88	0.78	0.89	0.95	0.92	0.78	0.6	0.68	1st
6	0.85	0.8	0.92	0.88	0.9	0.64	0.72	0.68	3rd
7	0.86	0.73	0.87	0.97	0.92	0.8	0.49	0.61	4th
8	0.86	0.81	0.92	0.9	0.91	0.66	0.72	0.69	2nd

Choosing Best ML Model

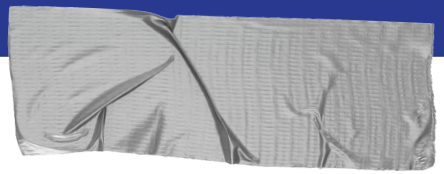
Model 5 (Random Forest Classifier): This model seems to have the highest accuracy and ROC AUC score. It also shows a good balance between precision, recall, and F1 score for both classes. It's performing quite well on various metrics.

Model 8 (Hybrid Model with Scaling and SMOTE):

This model also has competitive performance with high ROC AUC, precision, and recall values. It's a strong contender.

Model 6 (Ensemble Model with SMOTE and Grid Search):

This model has good performance across most metrics as well.



Thank You!