Loan Approval Prediction – Checkpoint 1

Group: Predictive Minds

Team Members:

- **Suraj Ravindra Rao (ASU ID: 1233990435)**
- **Om Gaurav Patel (ASU ID: 1233379110)**
- **Darren Alaistair Ferrao (ASU ID: 1237189124)**
- **Mohammad Asgar Khan Dalwai (ASU ID: 1237671684)**

Background & Problem

Loan approval prediction is essential for financial institutions.

Problem: Binary classification → Approved (0) vs Default (1).

Importance:

Risk Reduction

Reduces loan defaults and financial risk.

Process Automation

Automates decisions for faster processing.

Fair & Scalable

Ensures fairness and scalability.

Dataset Overview

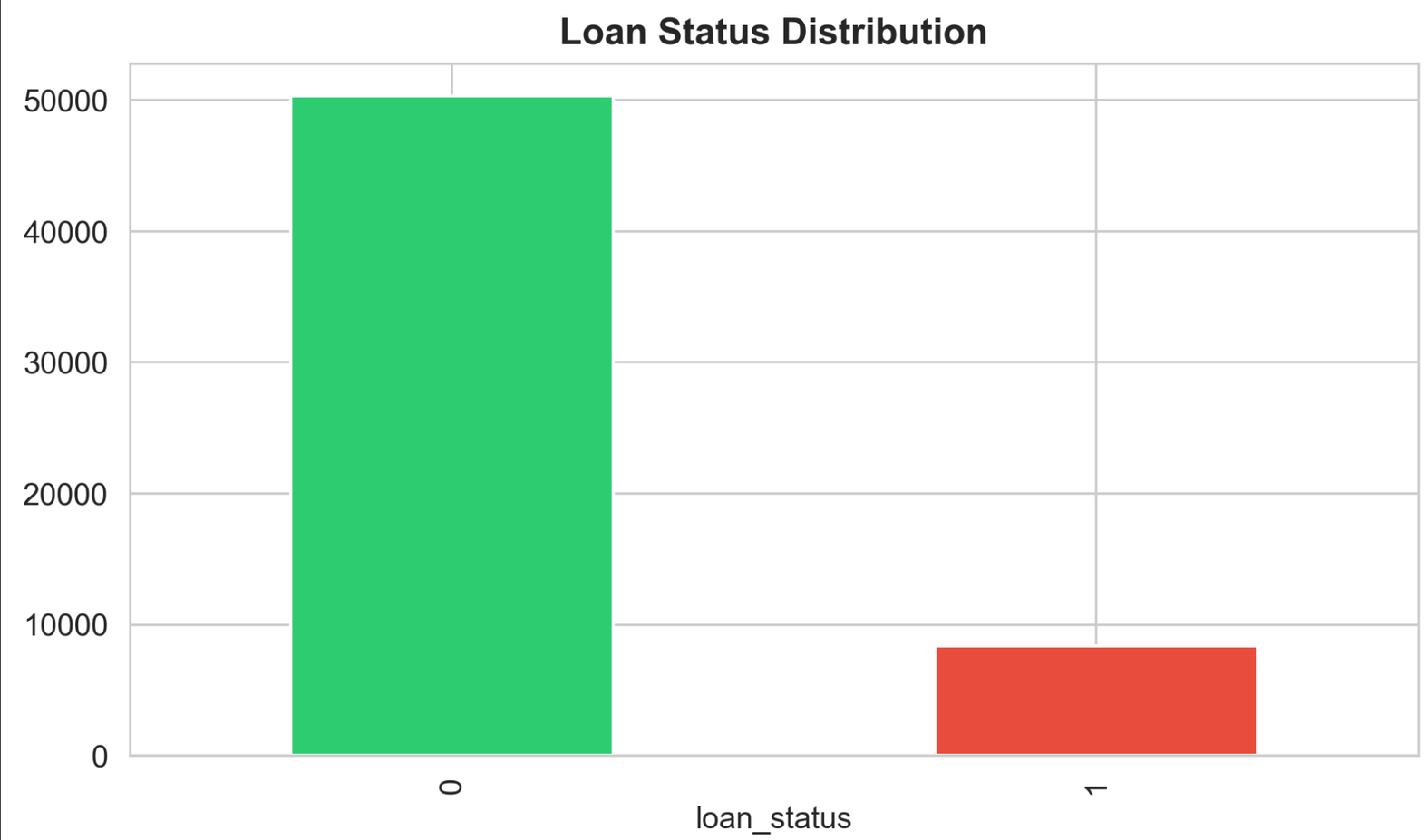
Dataset Evolution

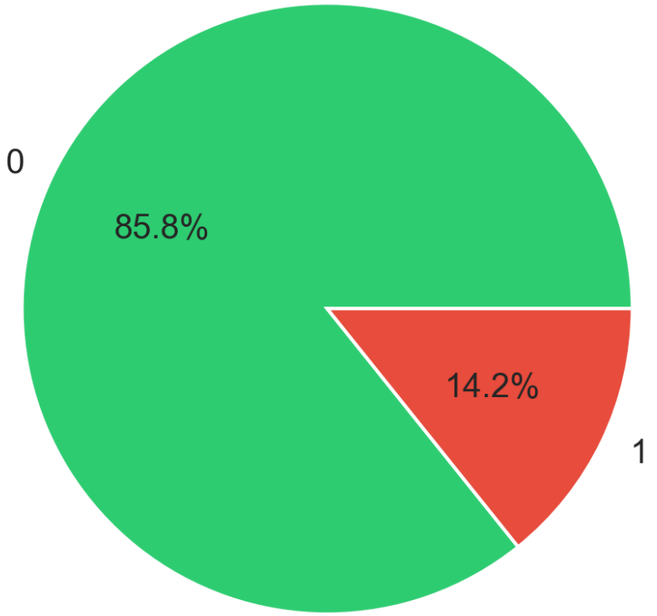
Pitch dataset: 4,269 samples (initial plan).

Updated dataset (Checkpoint 1): Kaggle Playground dataset, 58,645 samples × 13 features.

Features: demographic, employment, loan details, credit history.

Train/Test split: 80/20.





Numerical Feature Exploration

Age

Concentrated in 20s–30s.

Income

Skewed with large outliers.

Loan Amount

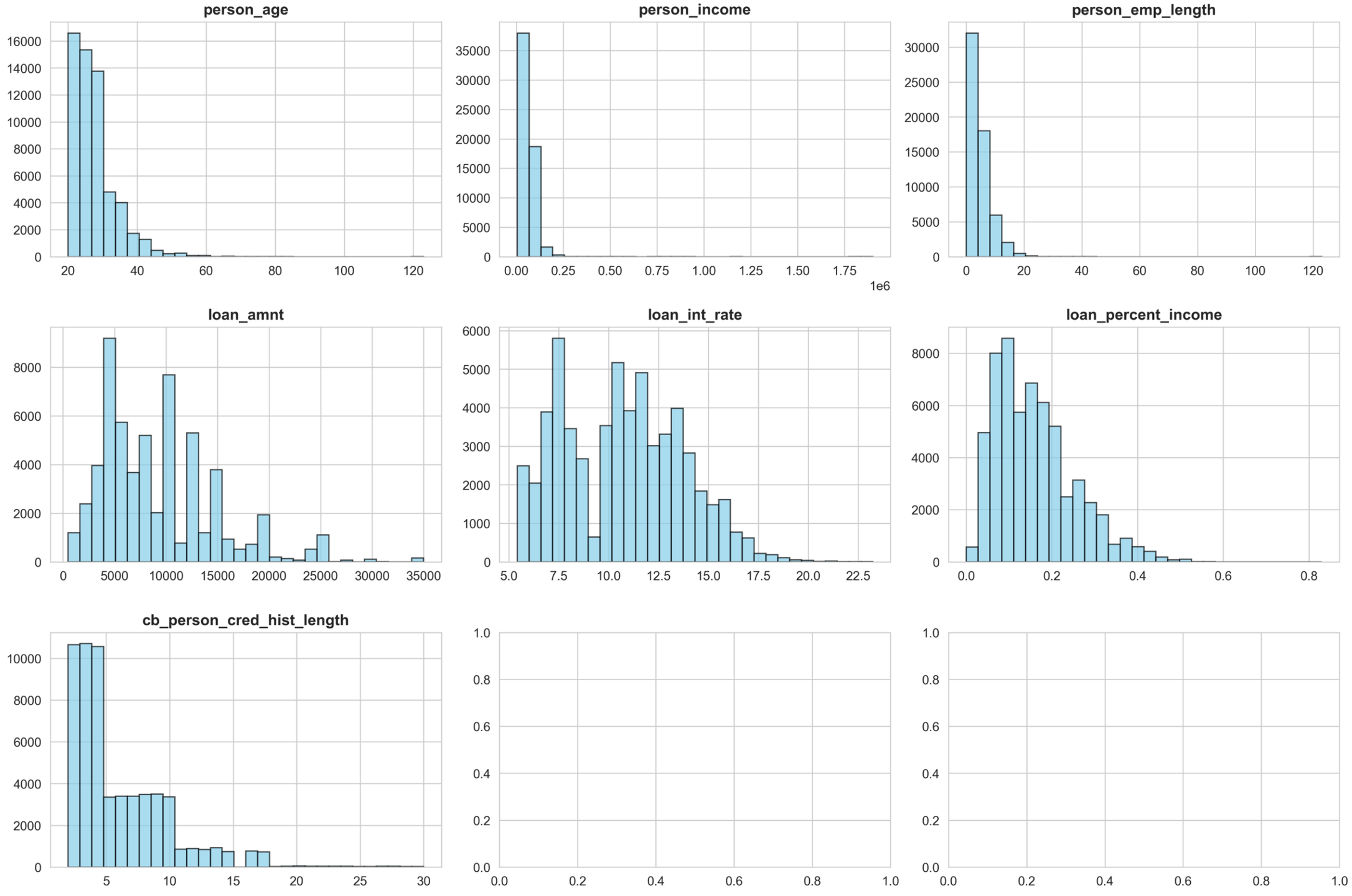
Most between 5k–15k.

Interest Rate

7–15% range dominates.

Credit History

Majority < 10 years.


person_age
person_income
person_emp_length
loan_amnt
loan_int_rate
loan_percent_income
cb_person_cred_hist_length
1e6

Categorical Feature Exploration

Home Ownership

Rent and Mortgage are common.

Loan Intent

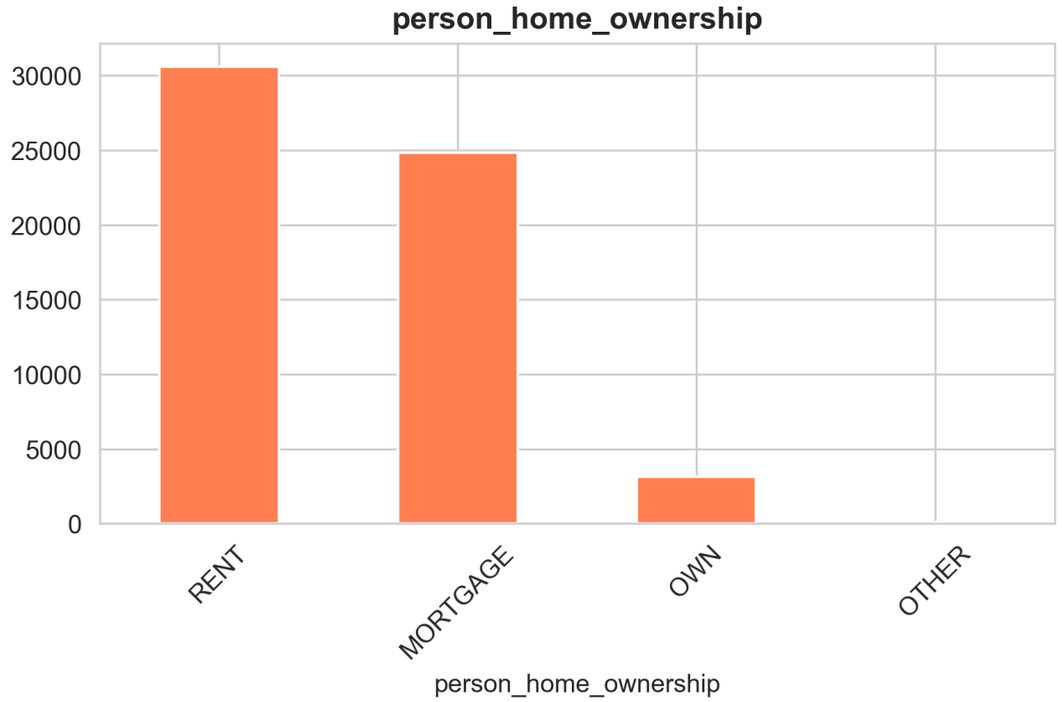
Education and Medical dominate.

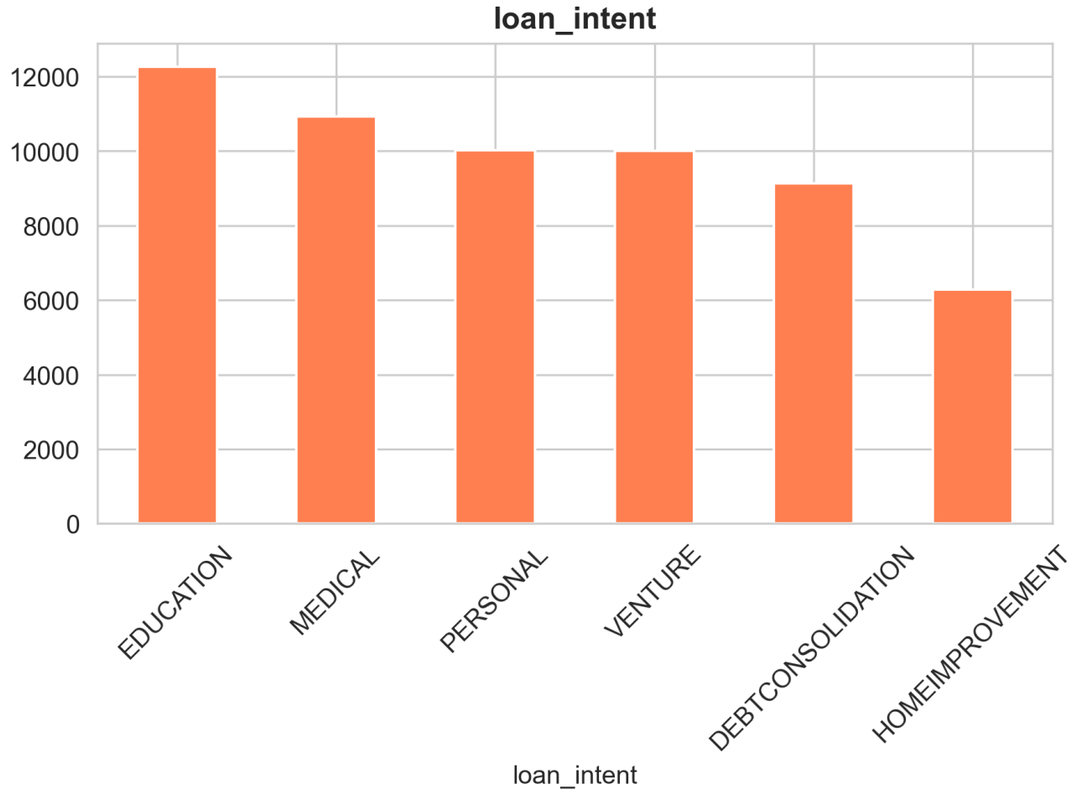
Loan Grade

Lower grades (D–G) have higher default risk.

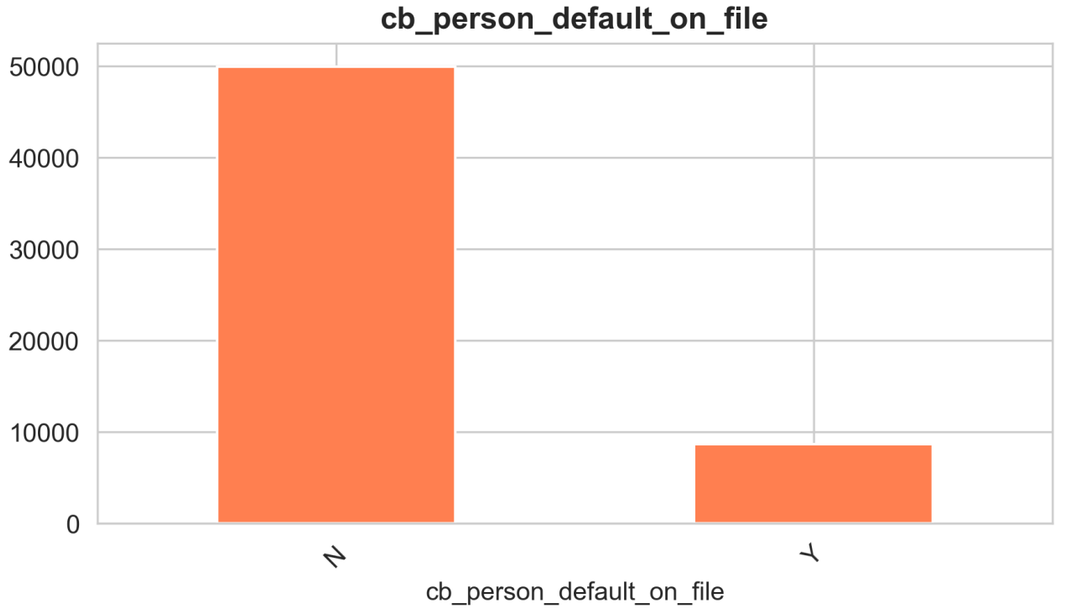
Default History

Strong predictor of loan status.


person_home_ownership
30000
25000
20000
15000
10000
5000
0
RENT
MORTGAGE
OWN
OTHER
person_home_ownership


loan_intent
12000
10000
8000
6000
4000
2000
0
EDUCATION
MEDICAL
PERSONAL
VENTURE
DEBTCONSOLIDATION
HOMEIMPROVEMENT
loan_intent


loan_grade
20000
17500
15000
12500
10000
7500
5000
2500
0
A
B
C
D
E
F
G
loan_grade


cb_person_default_on_file
50000
40000
30000
20000
10000
0
N
Y
cb_person_default_on_file

Correlations & Relationships

1

Strong Correlation

Loan Percent Income ↔ Loan Amount.

2

Age Relationship

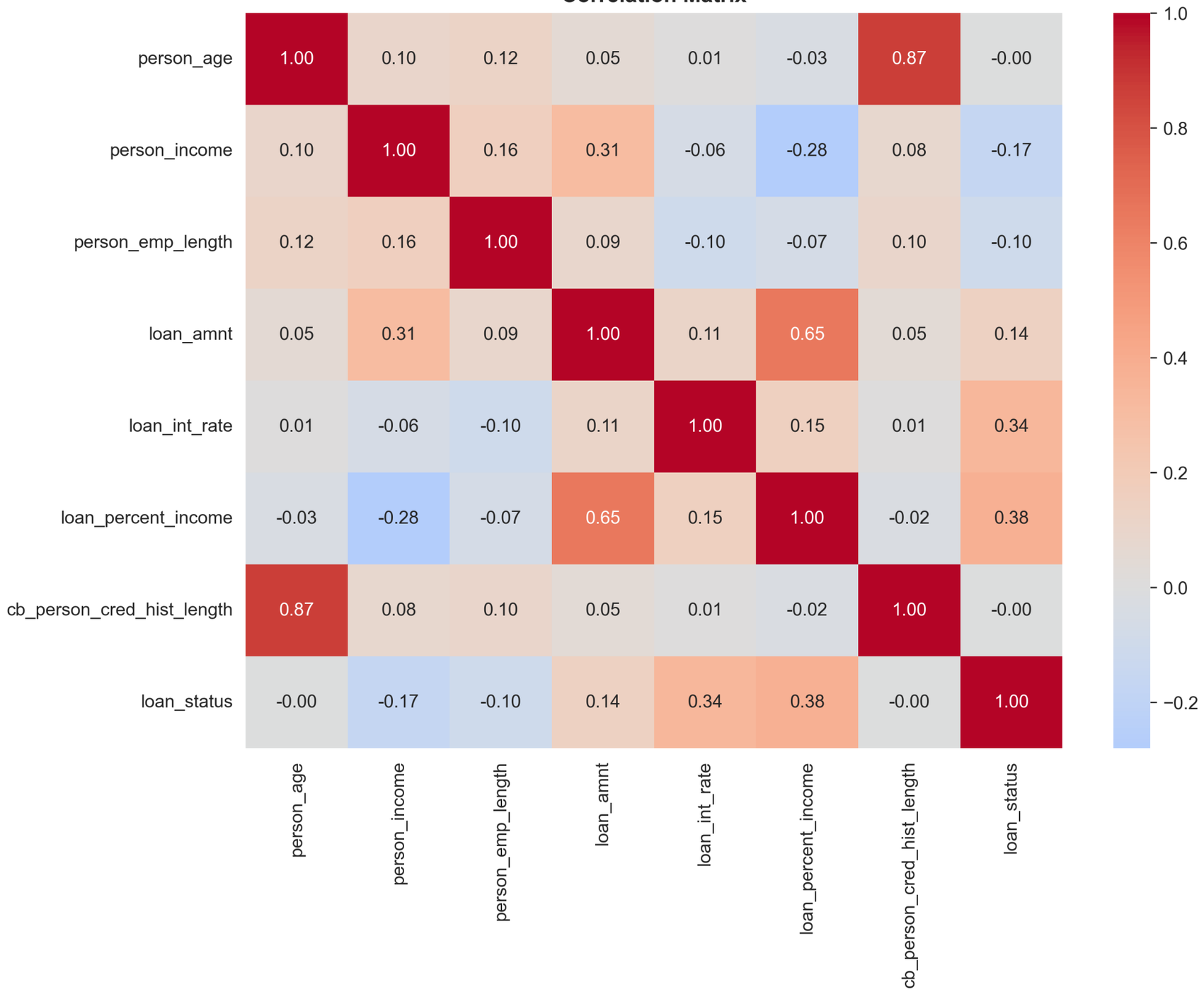
Age ↔ Credit history length also related.

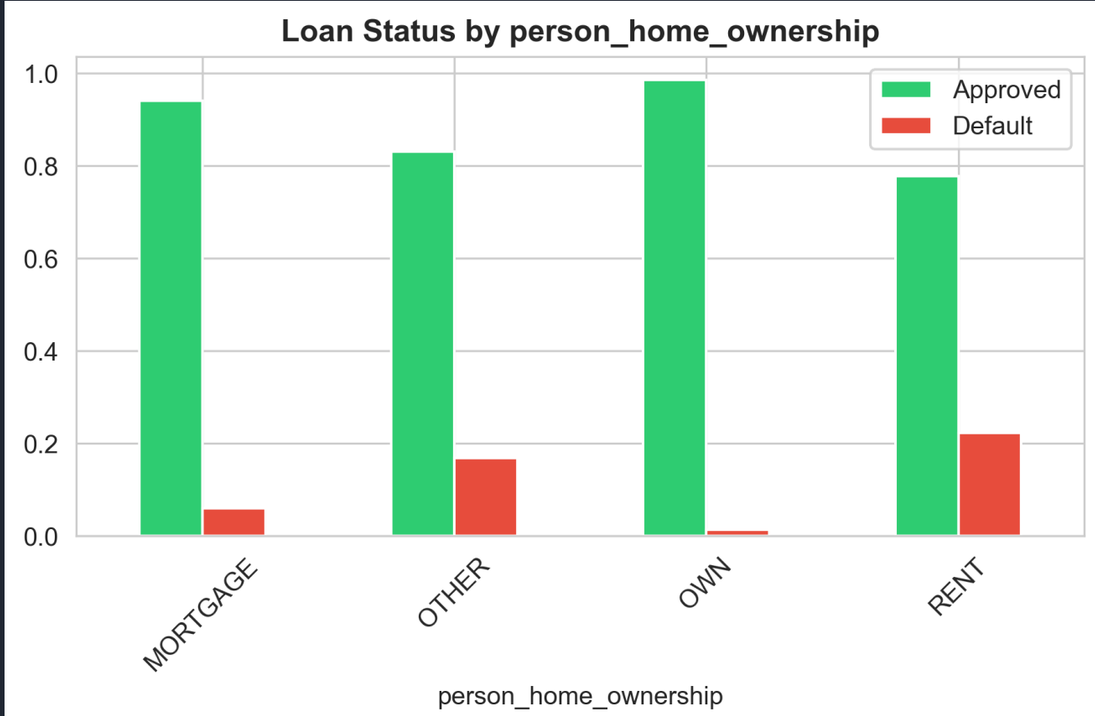
3

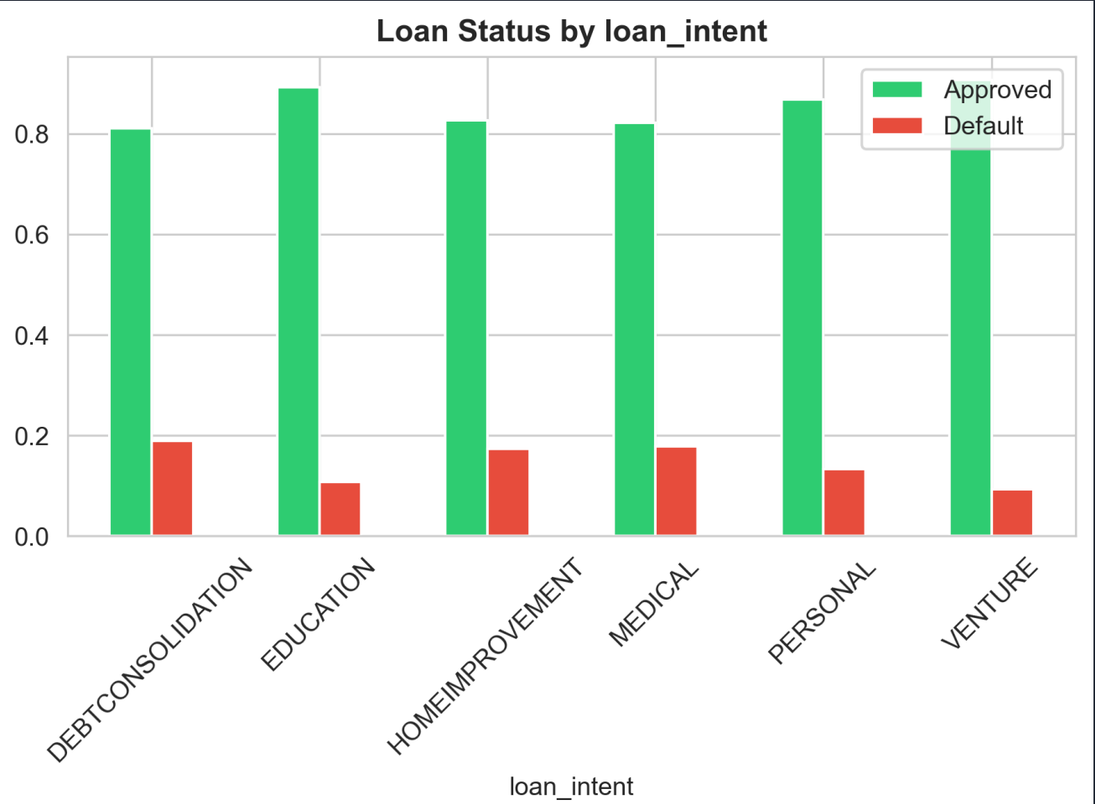
Status Influence

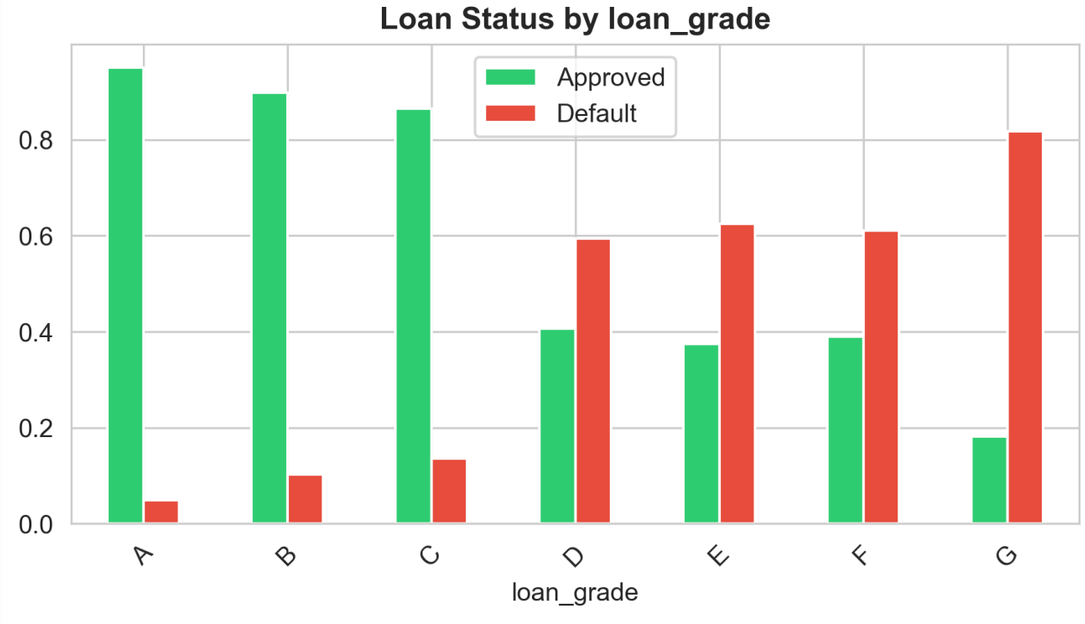
Loan Status influenced by loan grade & percent income.

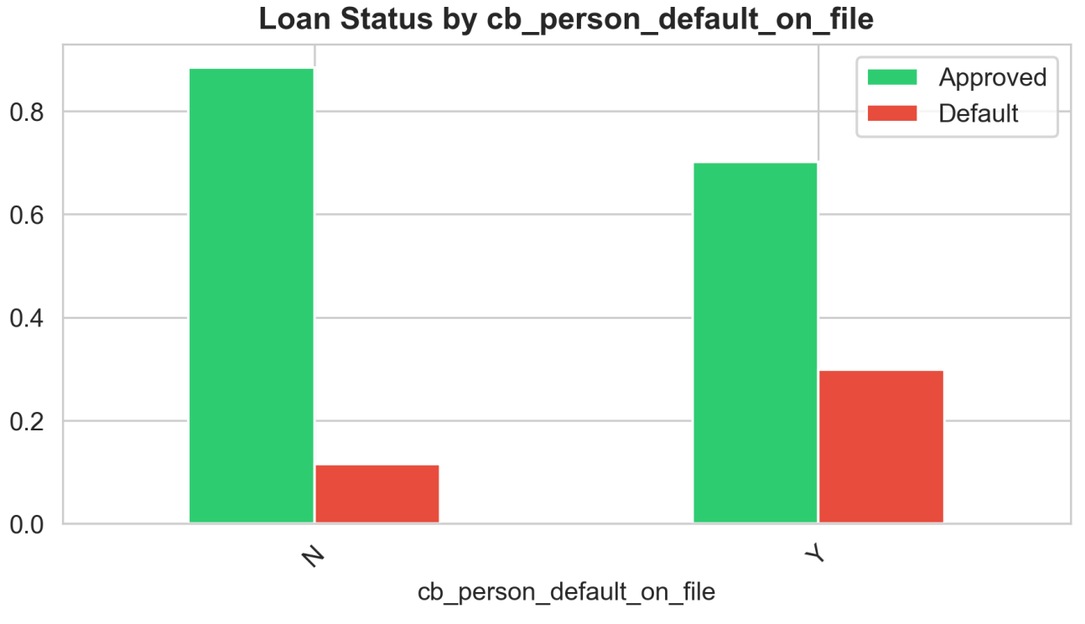
Correlation Matrix




Loan Status by person_home_ownership
Approved
Default
1.0
0.8
0.6
0.4
0.2
0.0
MORTGAGE
OTHER
OWN
RENT
person_home_ownership


Loan Status by loan_intent
Approved
Default
0.8
0.6
0.4
0.2
0.0
DEBTCONSOLIDATION
EDUCATION
HOMEIMPROVEMENT
MEDICAL
PERSONAL
VENTURE
loan_intent


Loan Status by loan_grade
Approved
Default
0.8
0.6
0.4
0.2
0.0
A
B
C
D
E
F
G
loan_grade


Loan Status by cb_person_default_on_file
Approved
Default
0.8
0.6
0.4
0.2
0.0
N
Y
cb_person_default_on_file

Data Preprocessing

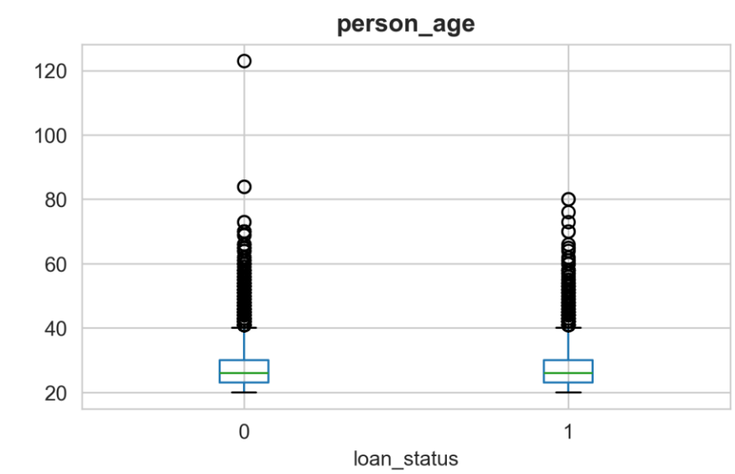
Data Cleaning & Transformation

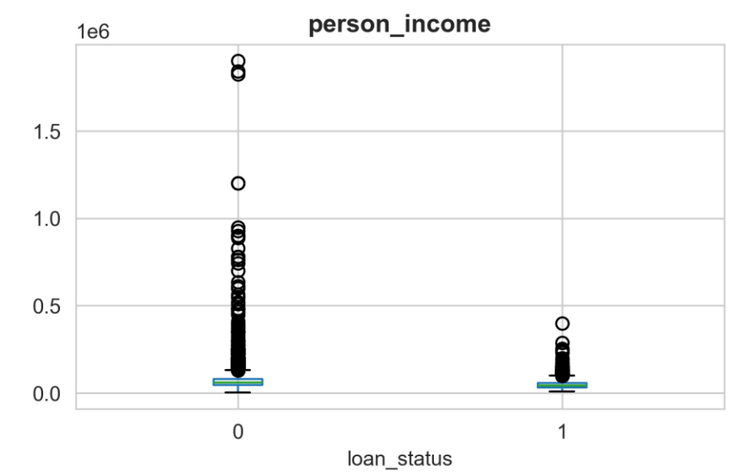
- Dropped IDs and irrelevant fields.
- Label encoded categorical features.
- Standardized numerical features.

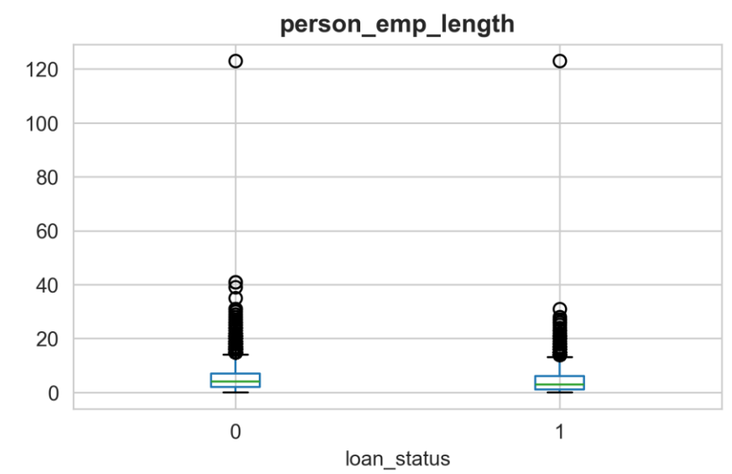
Feature Engineering:

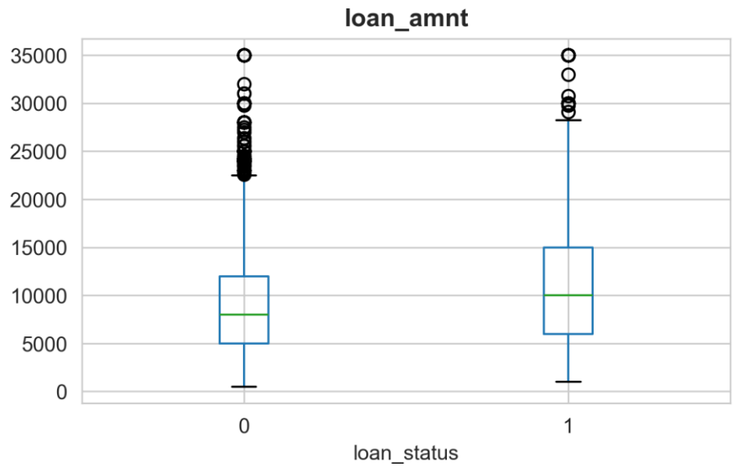
Income-to-loan ratio

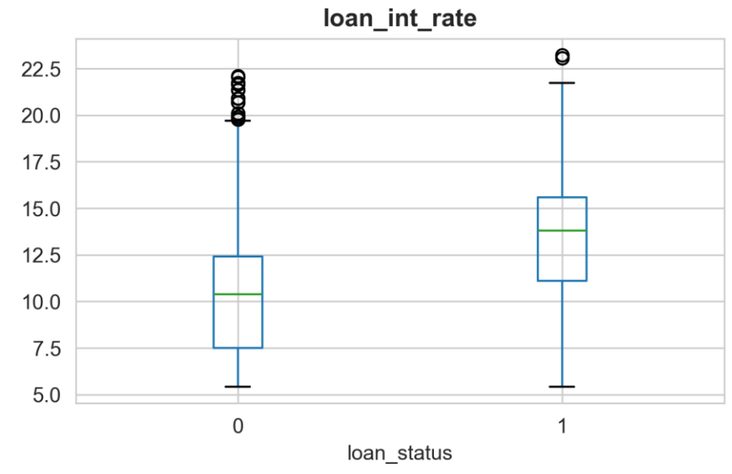
Age-to-employment ratio

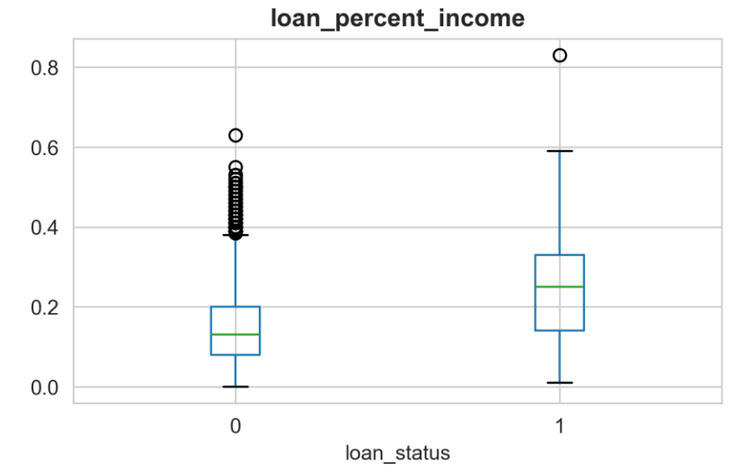

person_age
120
100
80
60
40
20
0
1
loan_status

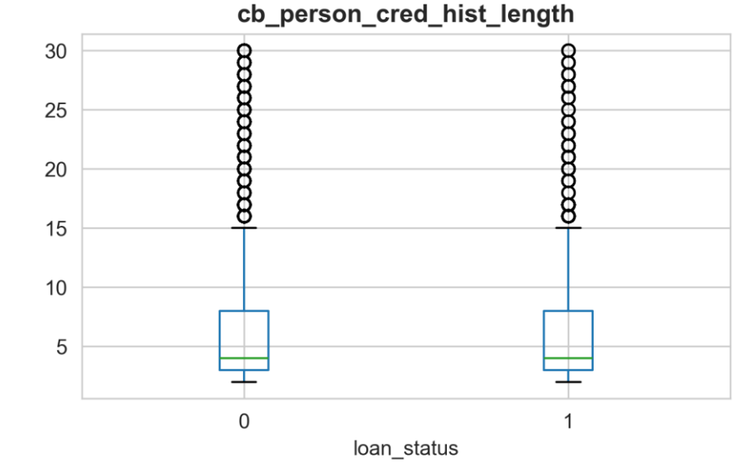

person_income
1e6
1.5
1.0
0.5
0.0
0
1
loan_status

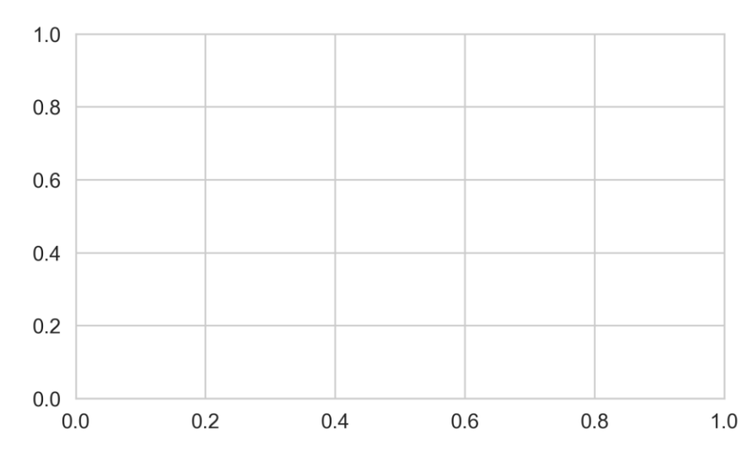

person_emp_length
120
100
80
60
40
20
0
0
1
loan_status

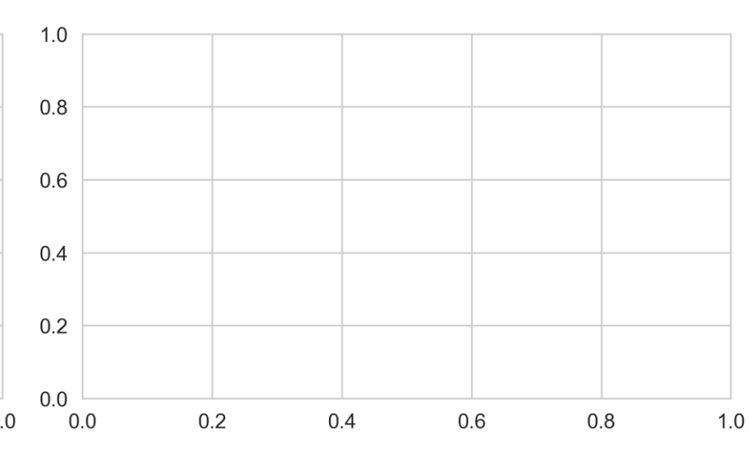

loan_amnt
35000
30000
25000
20000
15000
10000
5000
0
0
1
loan_status


loan_int_rate
22.5
20.0
17.5
15.0
12.5
10.0
7.5
5.0
0
1
loan_status


loan_percent_income
0.8
0.6
0.4
0.2
0.0
0
1
loan_status


cb_person_cred_hist_length
30
25
20
15
10
5
0
1
loan_status


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0


1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0

Model Training & Evaluation

Models: Logistic Regression, Decision Tree, Random Forest.

Metrics: Accuracy, Precision, Recall, F1, ROC-AUC.

Results:

Logistic Regression

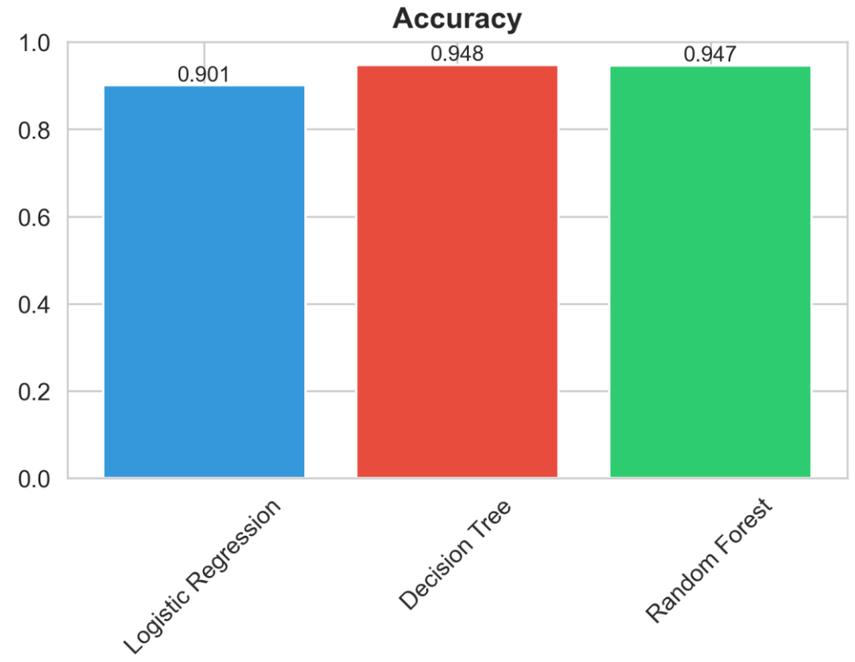
Accuracy 0.90, Recall 0.44.

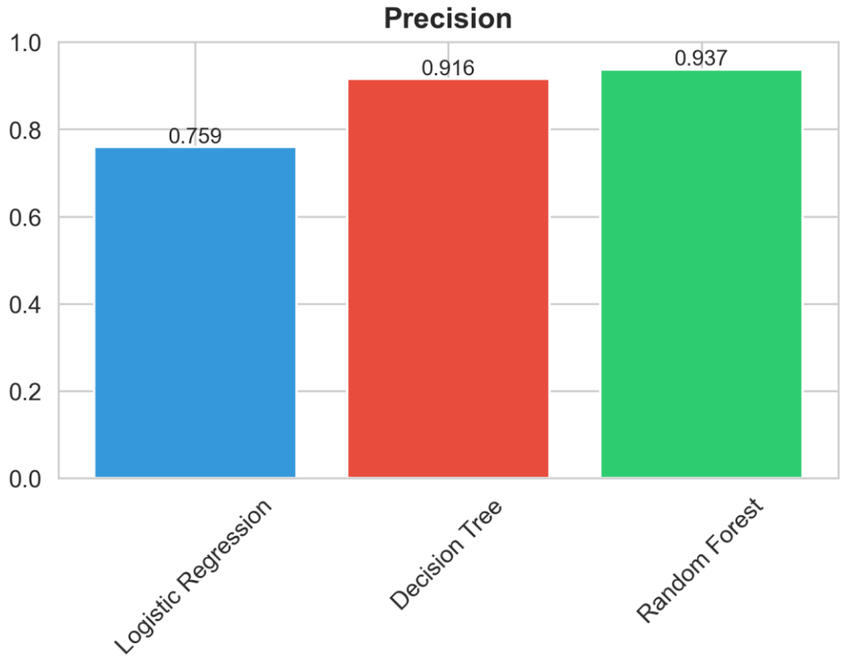
Decision Tree

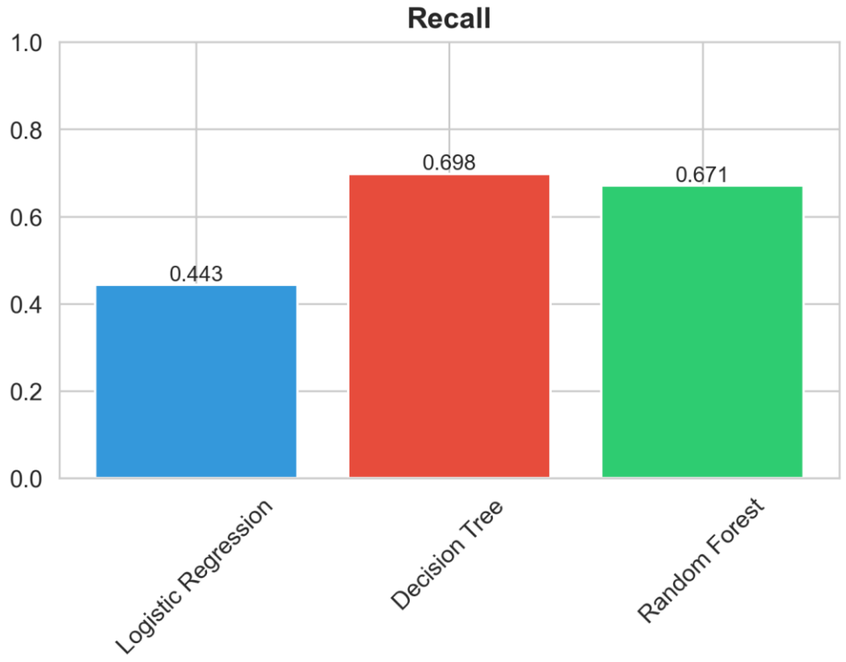
Accuracy 0.95, Recall 0.70.

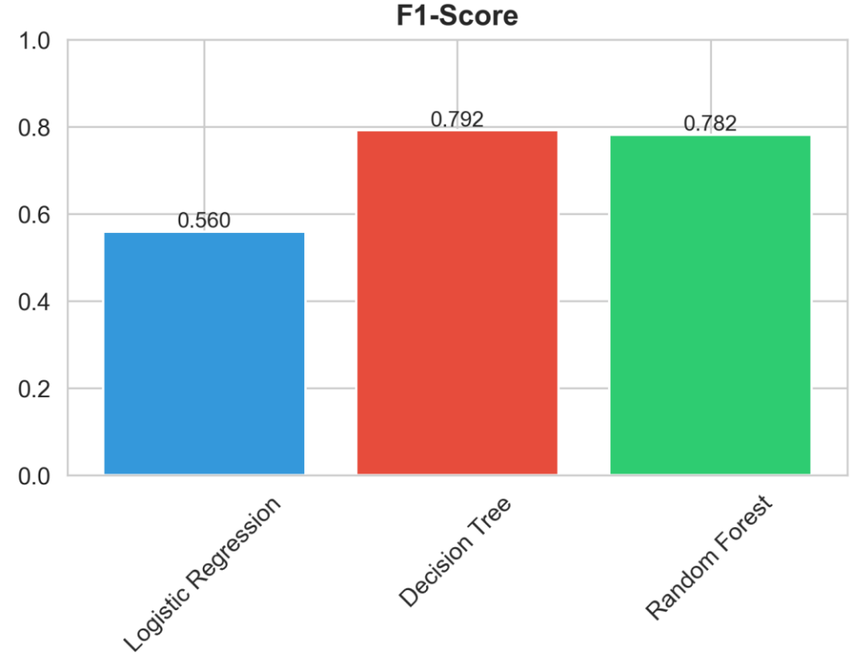
Random Forest

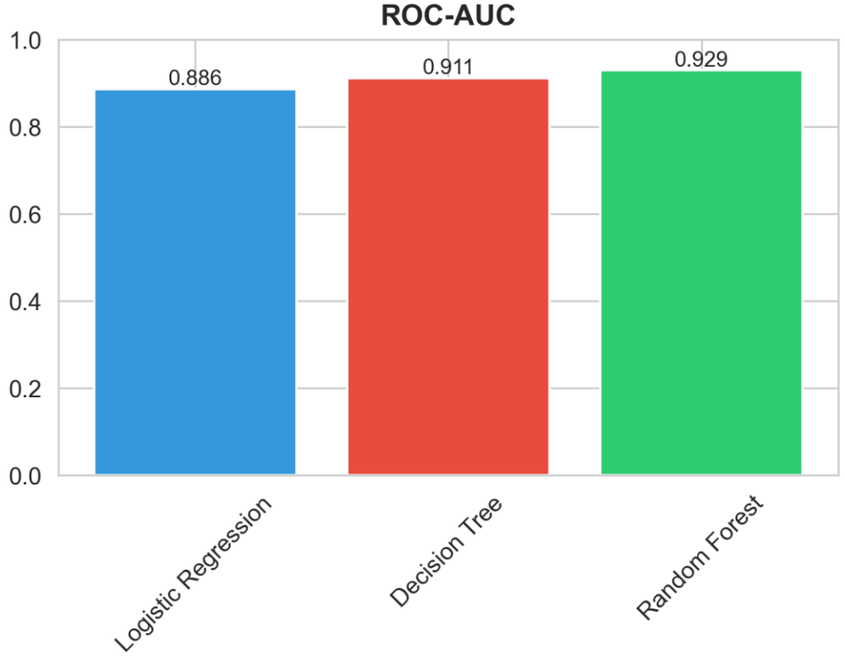
Accuracy 0.95, ROC-AUC 0.93.

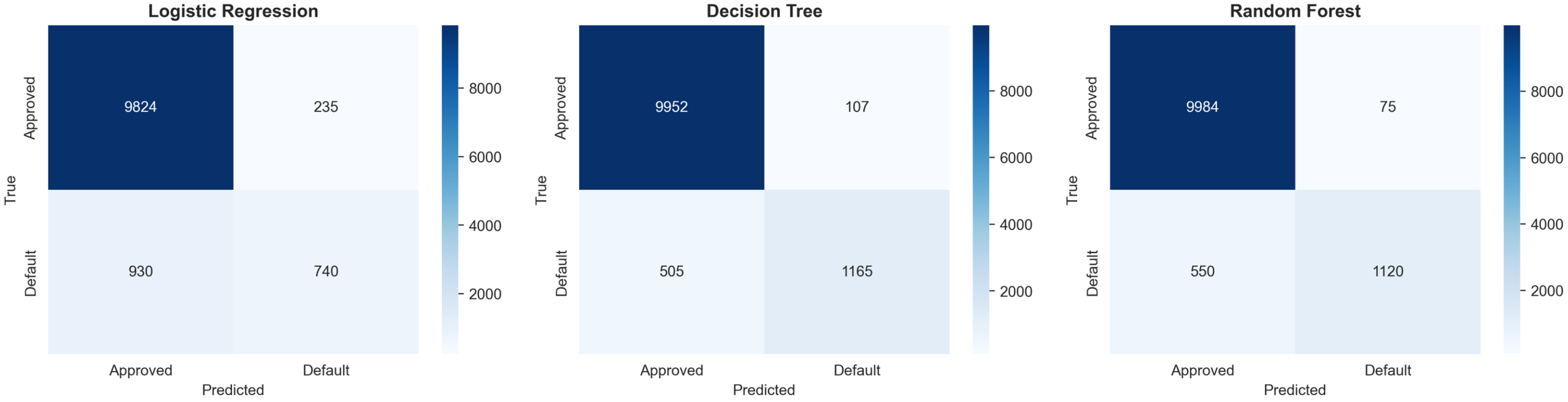

Accuracy
1.0
0.8
0.6
0.4
0.2
0.0
0.901
0.948
0.947
Logistic Regression
Decision Tree
Random Forest


Precision
1.0
0.8
0.6
0.4
0.2
0.0
0.759
0.916
0.937
Logistic Regression
Decision Tree
Random Forest


Recall
1.0
0.8
0.6
0.4
0.2
0.0
0.443
0.698
0.671
Logistic Regression
Decision Tree
Random Forest


F1-Score
1.0
0.8
0.6
0.4
0.2
0.0
0.560
0.792
0.782
Logistic Regression
Decision Tree
Random Forest


ROC-AUC
1.0
0.8
0.6
0.4
0.2
0.0
0.886
0.911
0.929
Logistic Regression
Decision Tree
Random Forest



Feature Importance

Top predictors:

01

Loan Percent Income

02

Loan Grade

03

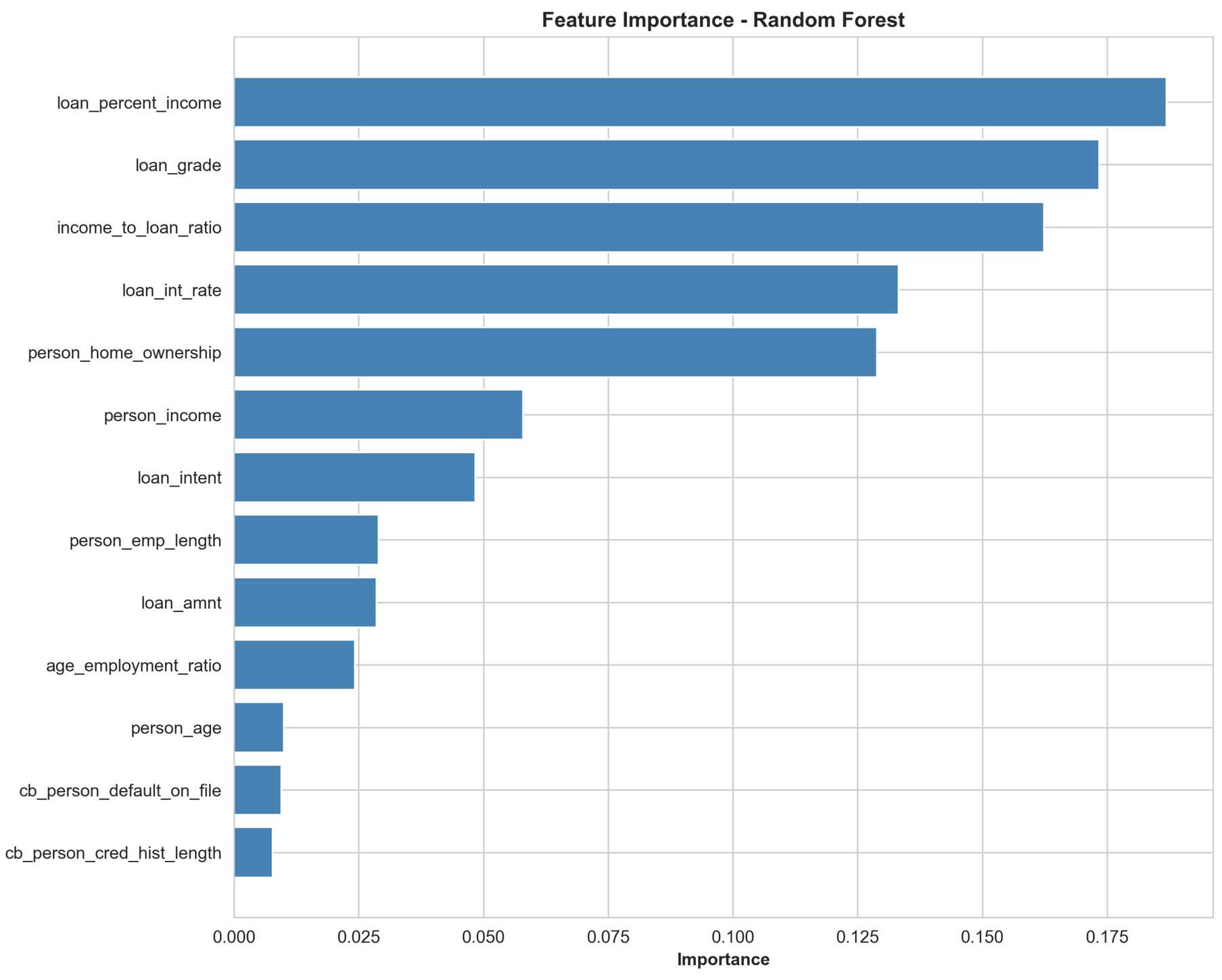
Income-to-loan ratio

04

Interest rate

05

Home ownership



Insights & Next Steps

Insights:

- Tree-based models perform best.
- Class imbalance hurts recall for defaults.
- Key predictors: income ratios, loan grade.

Next Steps:

- Apply SMOTE/undersampling for balance.
- Hyperparameter tuning.
- Try XGBoost/LightGBM.
- Add interpretability (SHAP, LIME).

Thank You

Video Link: https://drive.google.com/file/d/1DM88fetLoI78E1IDGRsYyeiD0g5knKzl/view?usp=sharing