PREDICTIVE DEFAULT RISK MODELING FOR SMARTER LOAN APPROVALS

Jed Jones
Amineh Farzannia
Olorunfemi Adeyemo

BACKGROUND & DATA SCIENCE PROBLEM

Loan defaults can be a significant problem for banks, impacting their profitability and overall stability.

- 1. Reduced profitability: When borrowers default, banks are unable to collect the principal and interest payments they expected, leading to significant losses.This can result in lower earnings for shareholders and a reduction in dividend payments. .
- NPAs represent assets that are not generating income and can negatively impact a bank's financial health.
- Impact on credit ratings
- A high level of loan defaults can negatively impact a bank's credit ratings, says FasterCapital.

Data Science Problem:

- A bank needs to approve as many high-quality loans as possible while minimizing defaults. Using 45000 historical loan applications, we will build a regularized logistic regression classifier that:
- Estimates each applicant's probability of default via .predict_proba() (to support both binary decisions and flexible interest-rate/term adjustments)

Methods:

Data :

Leverages applicant features:

Demographics: age, gender, education

Financial profile: income, employment tenure, home-ownership

Credit history: credit score, length of credit history, past defaults

Loan details: amount, purpose, interest rate, payment-to-income ratio

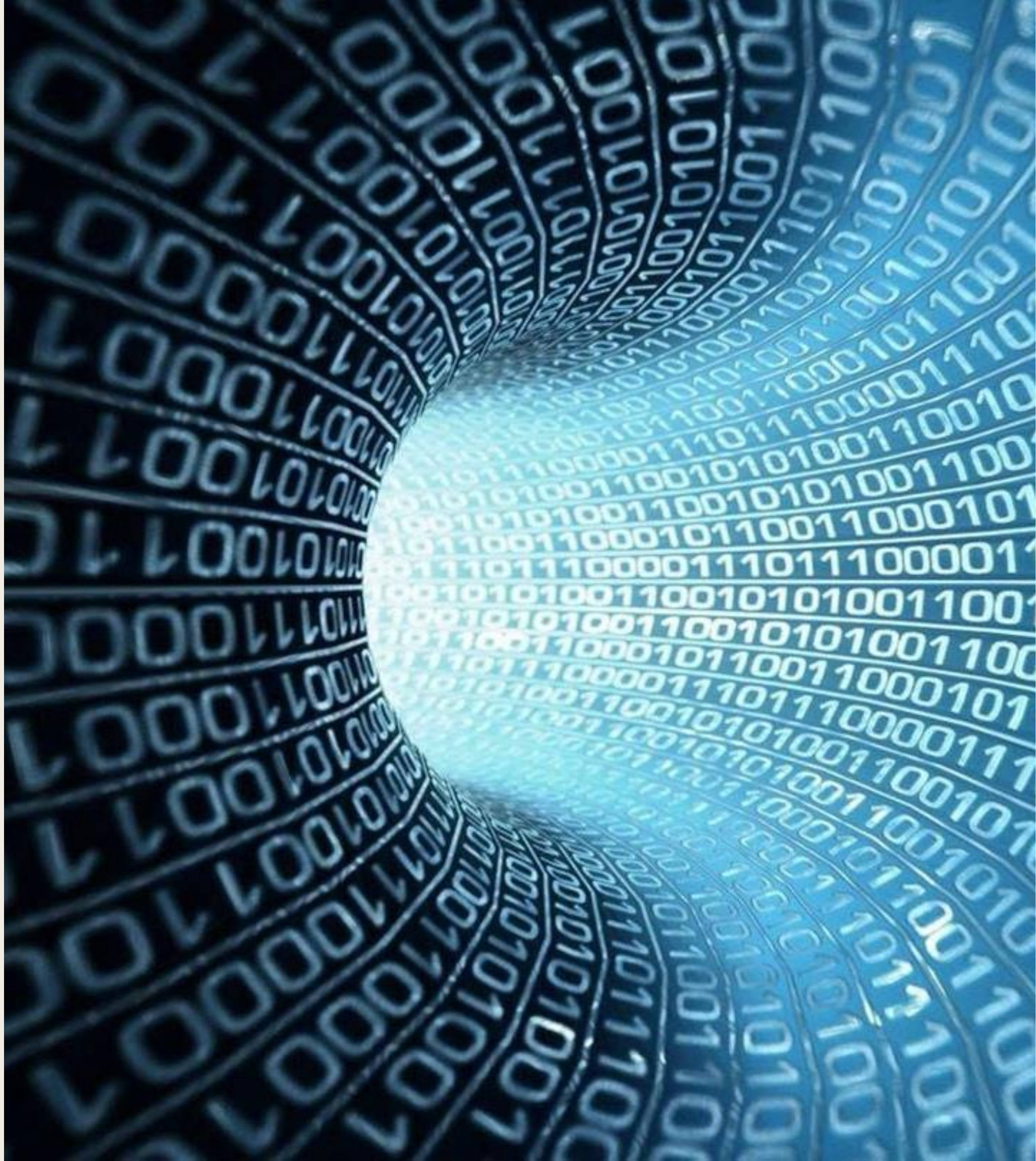
Model : Logistic Regression , Random Forests

Evaluation:

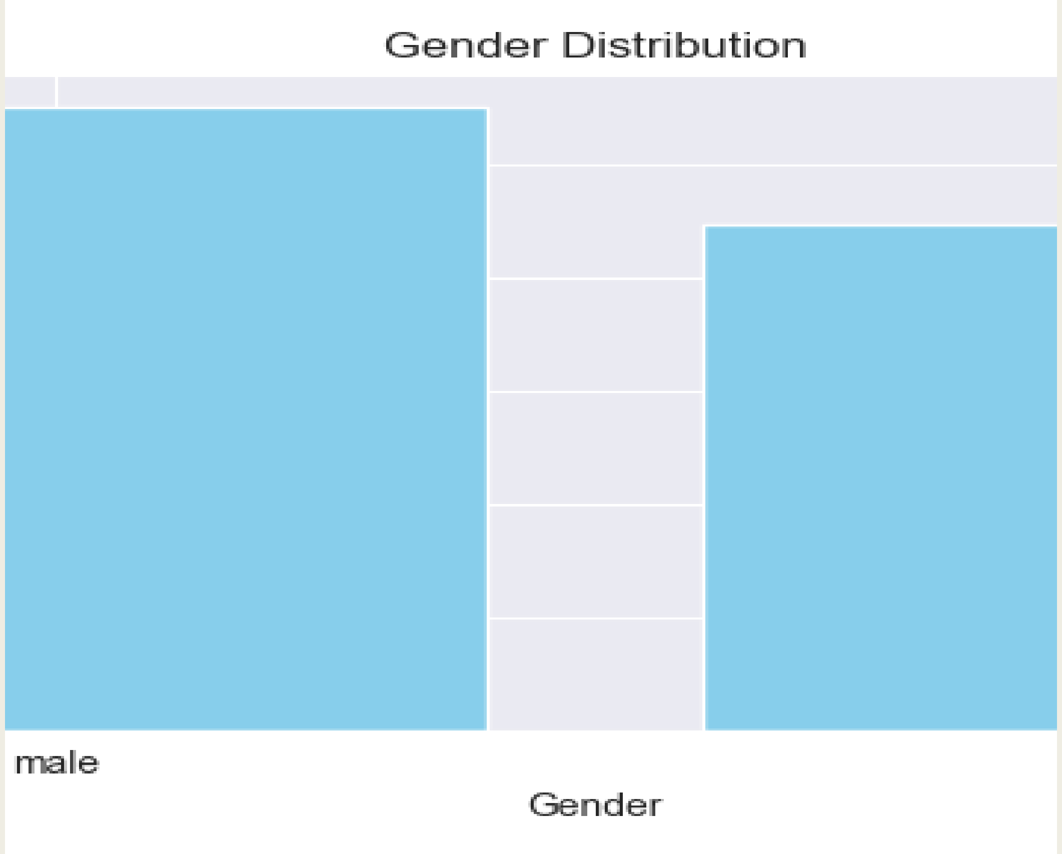
F1 score, Recall and Precision

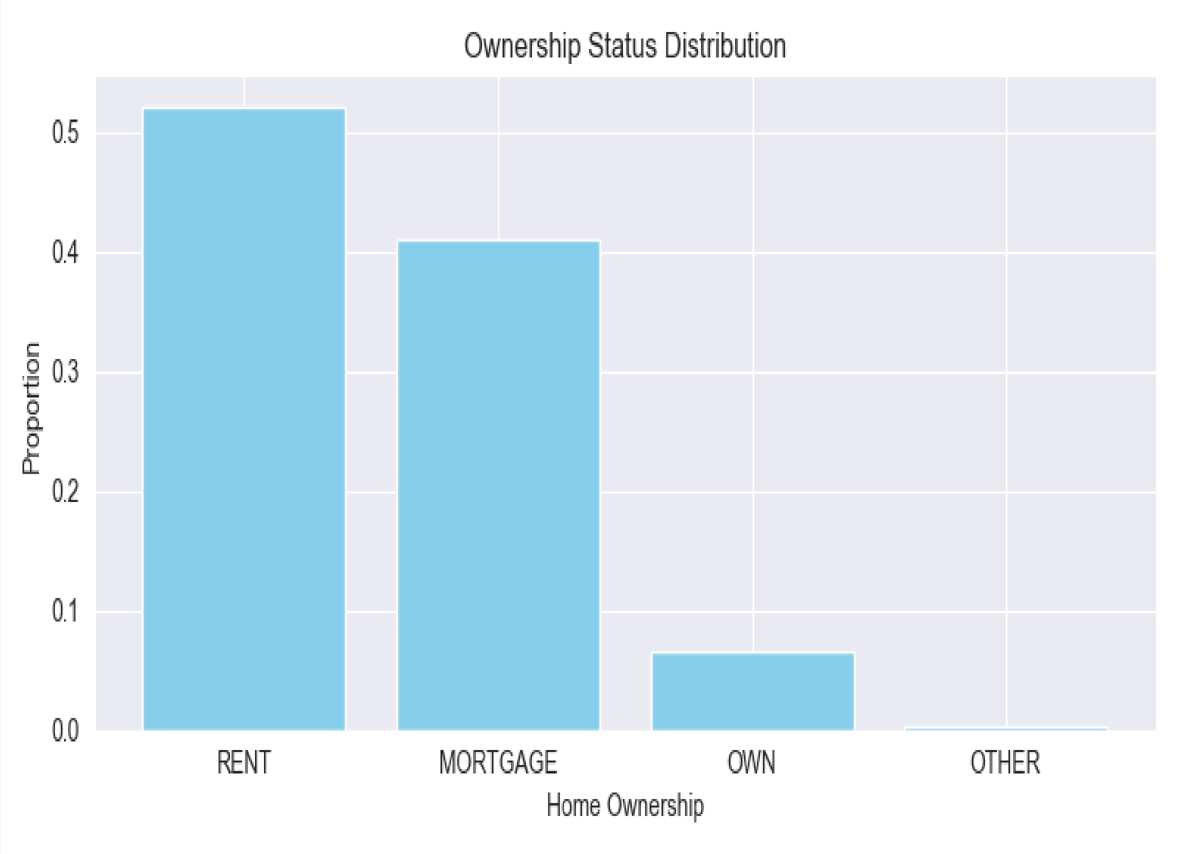
Interpretation:

DATA

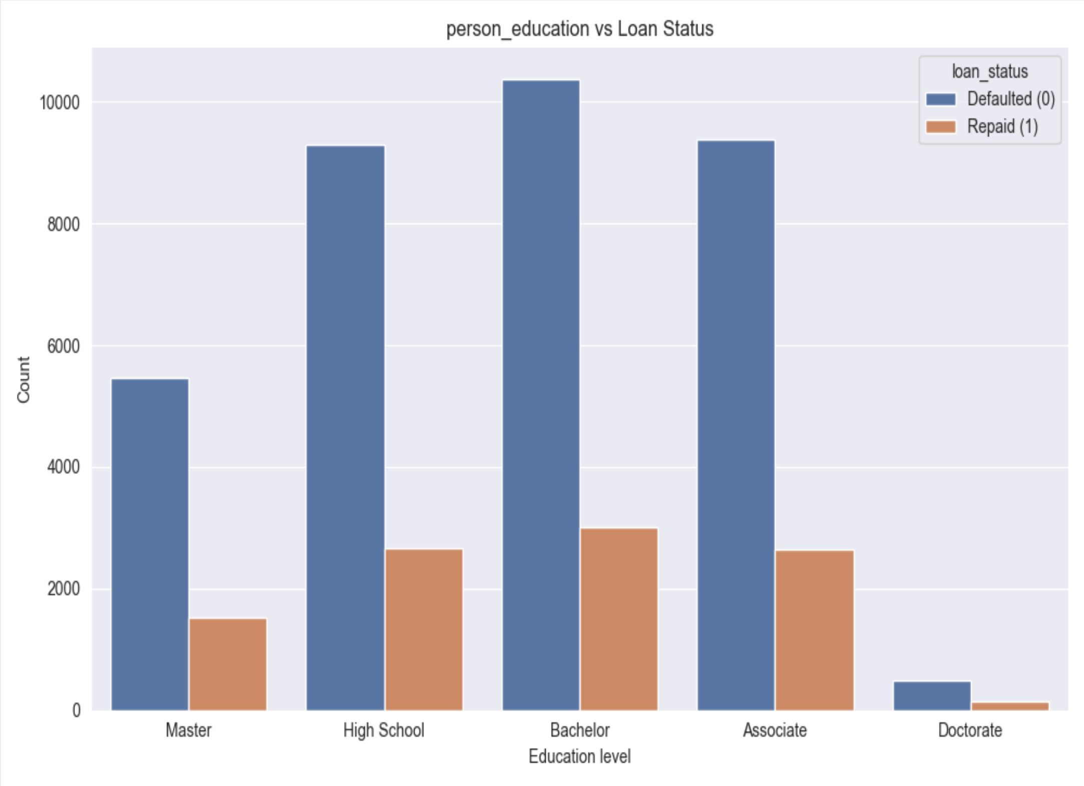


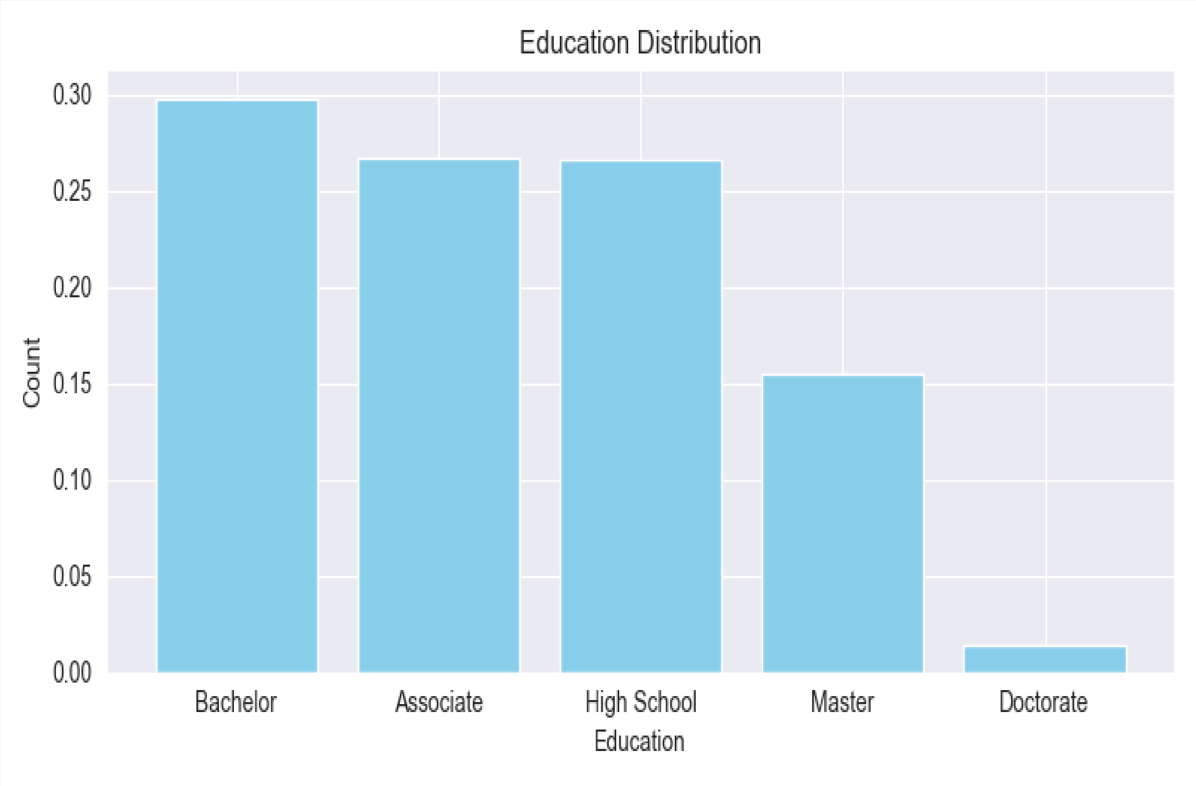
Gender and Ownership



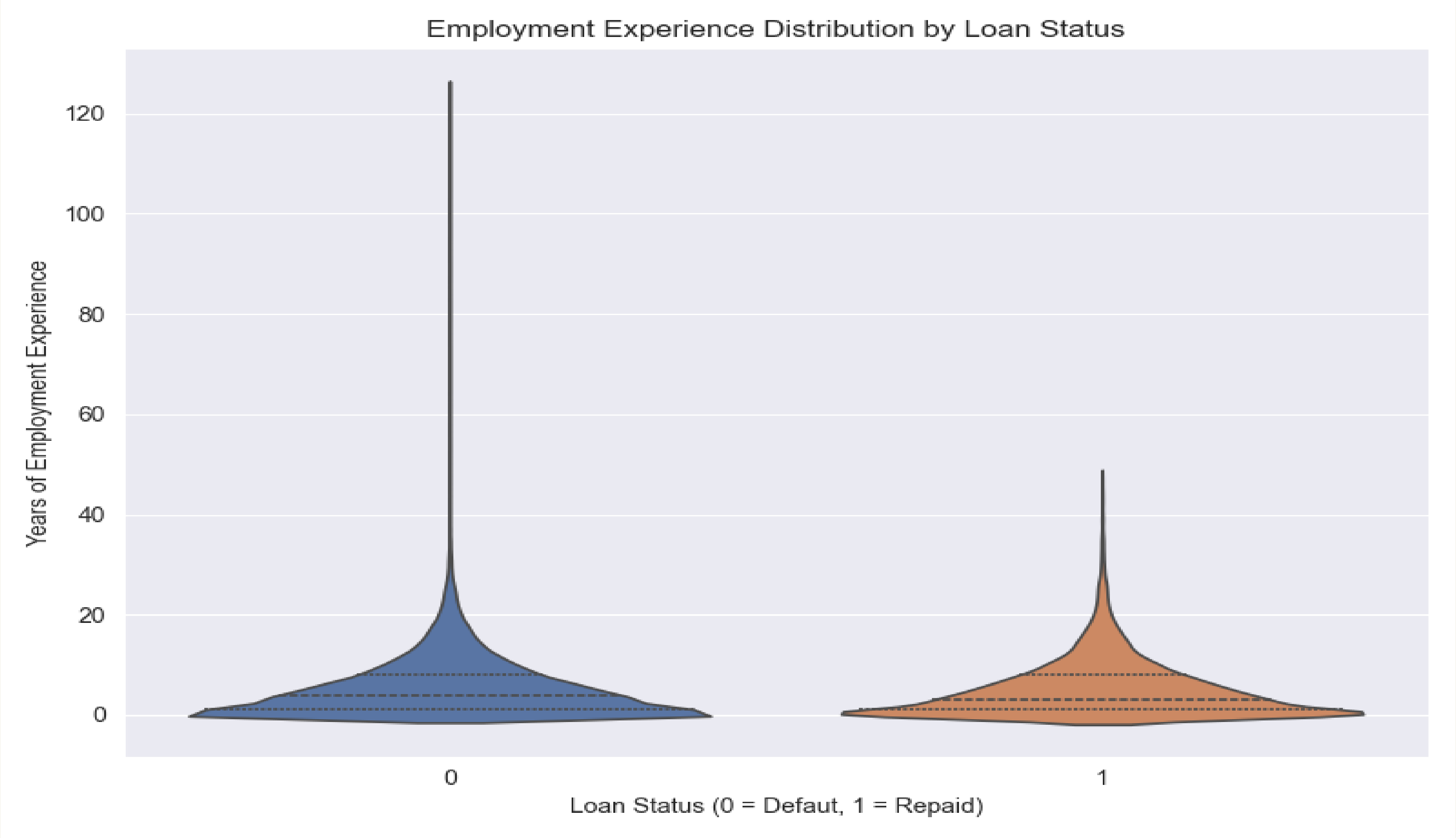


DATA AND EDUCATION AND AGE:



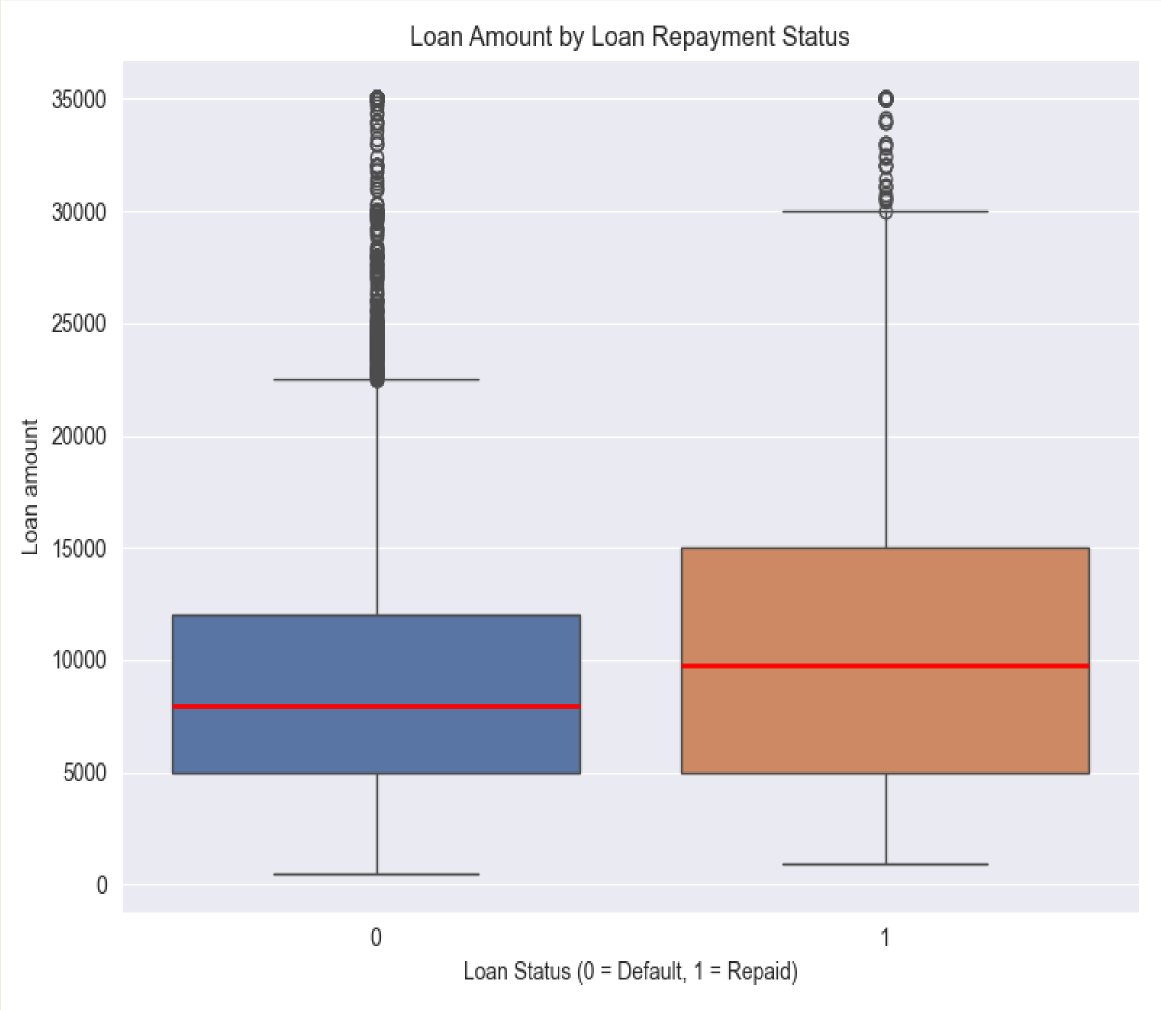


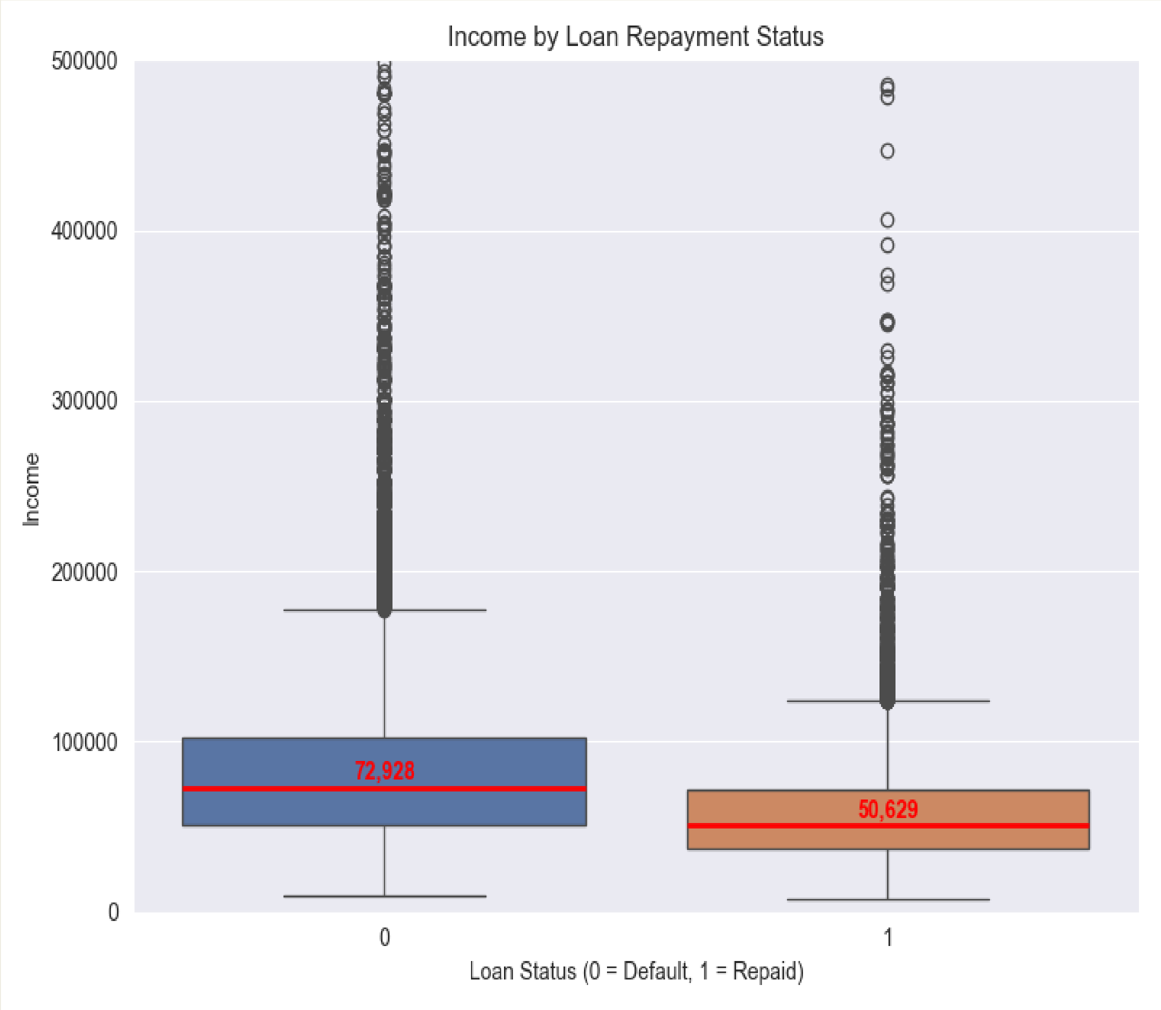
Loan Status



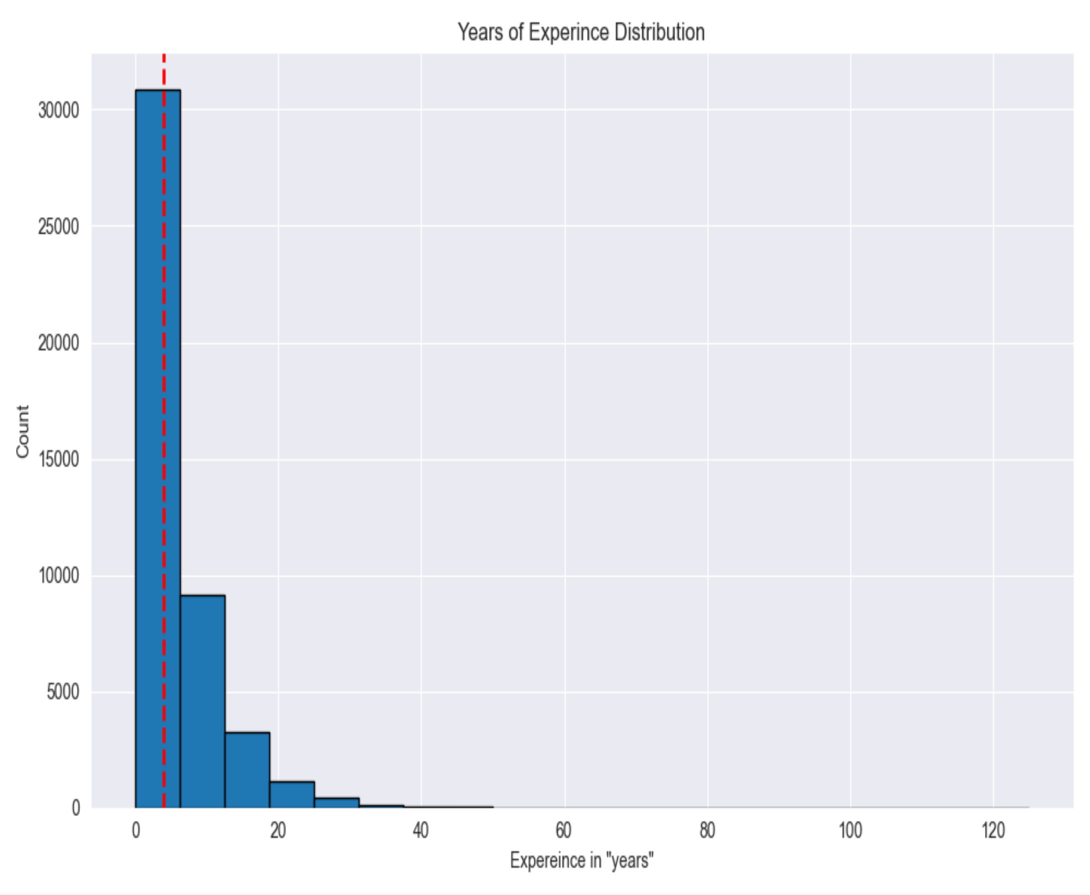
Employment Experience Distribution by Loan Status
Years of Employment Experience
120
100
80
60
40
20
0
0
1
Loan Status (0 = Defaut, 1 = Repaid)

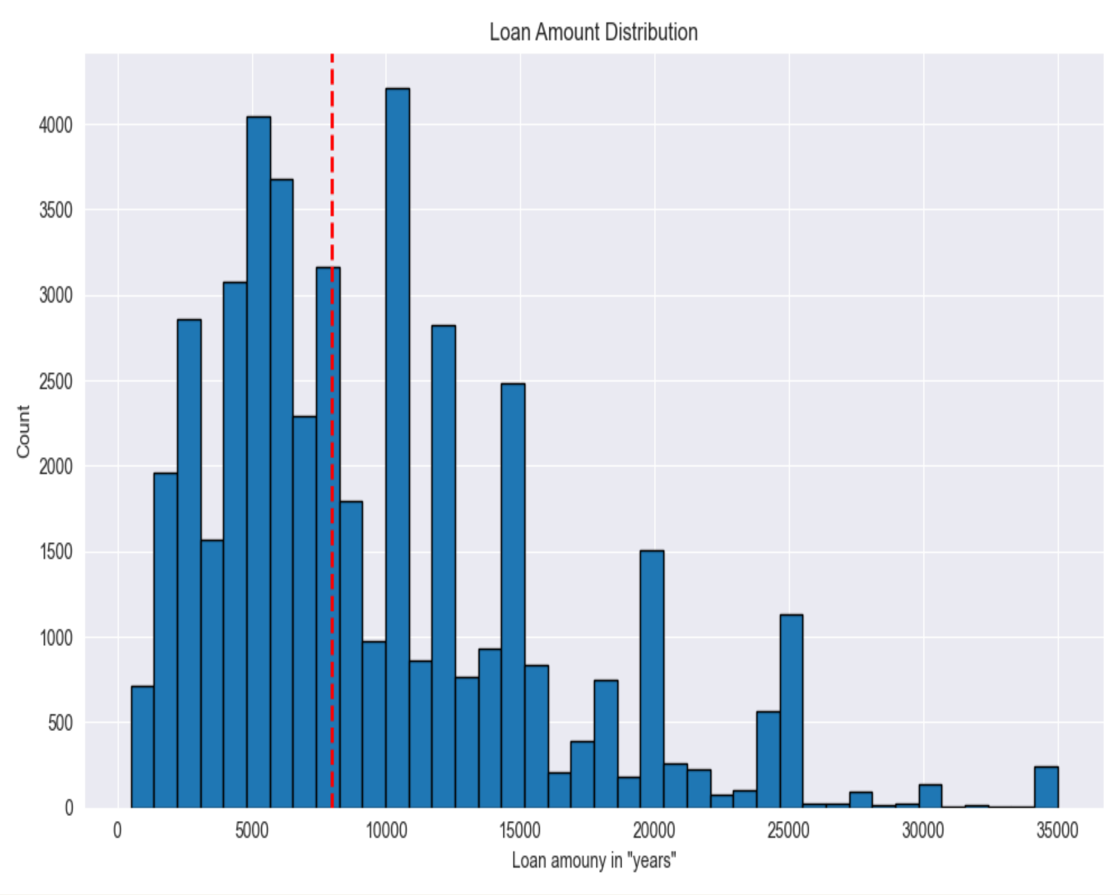
Income and Loan repayment



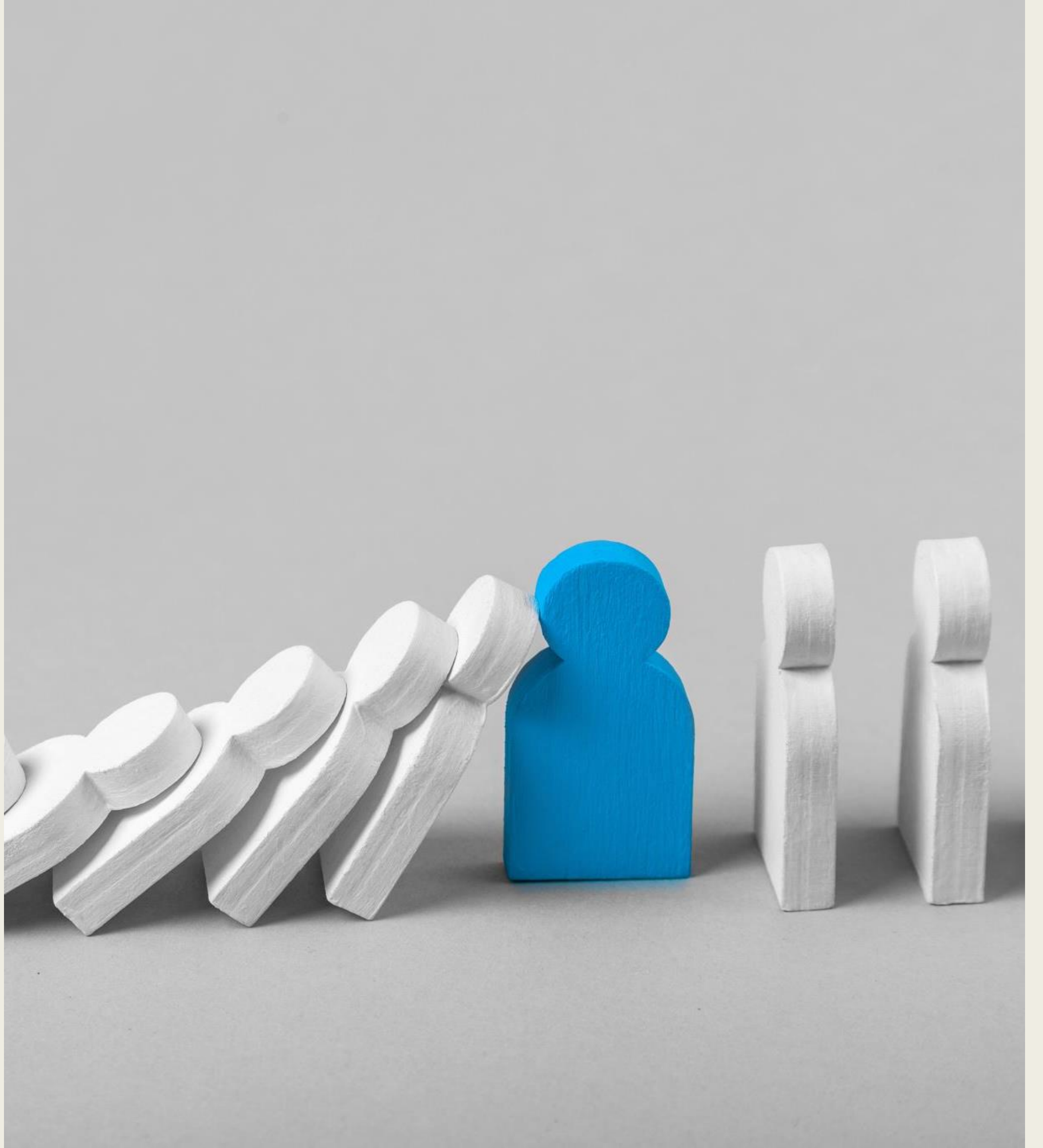


Loan amount


Years of Experince Distribution
30000
25000
20000
15000
10000
5000
0
Count
0
20
40
60
80
100
120
Expereince in "years"


Loan Amount Distribution
4000
3500
3000
2500
2000
1500
1000
500
0
Count
0
5000
10000
15000
20000
25000
30000
35000
Loan amouny in "years"

MODEL AND PREDICTION



Baseline: Null Model

- **Dropped features:** Age, loan_int_rate, loan_percent_income, Peron_epm_exp
- **Baseline Model:** logistic Regression
- **Baseline Model Evaluation Metrics:**

- Precision : 0.9145

- Recall: 0.9201

-Accuracy : 0.8710

-F1 score : 0.9173

- **Interpretation:**
- model is **strong at both identifying positives and avoiding false positives**.
- The **F1 score is high**, indicating good balance between precision and recall.
- **Accuracy is a bit lower than precision**
- **recall**, which might suggest slight class imbalance or errors in the negative class.

what does the Model tell us?

Evaluation:

- **Precision:** Of all the positive predictions the model made, **91.45% were actually correct.** The model makes few false positive errors.
- **Recall:** Of all actual positive cases, the model correctly identified **92.01%.**The model misses few actual positive cases (low false negatives).
- **Accuracy:** Overall, **87.10% of all predictions (both positive and negative) were correct**.
- **Implication:** Good general performance, but accuracy can be misleading if the classes are imbalanced.
- **F1 Score:** **0.9173.** This is a **balanced measure**—your model does well in both catching positives (recall) and being correct about them (precision).
- Minority Class:

  F1 score: 0.71

  Indicates the model is less effective at identifying non-defaults.

Model: Logistic Model with best Hyperparameter C

- **Model:** Logistic regression with best Hyperparameter C
- **Baseline Model Evaluation Metrics:**

  Precision : 0.91

  Recall : 0.92

  F1 Score : 0.92

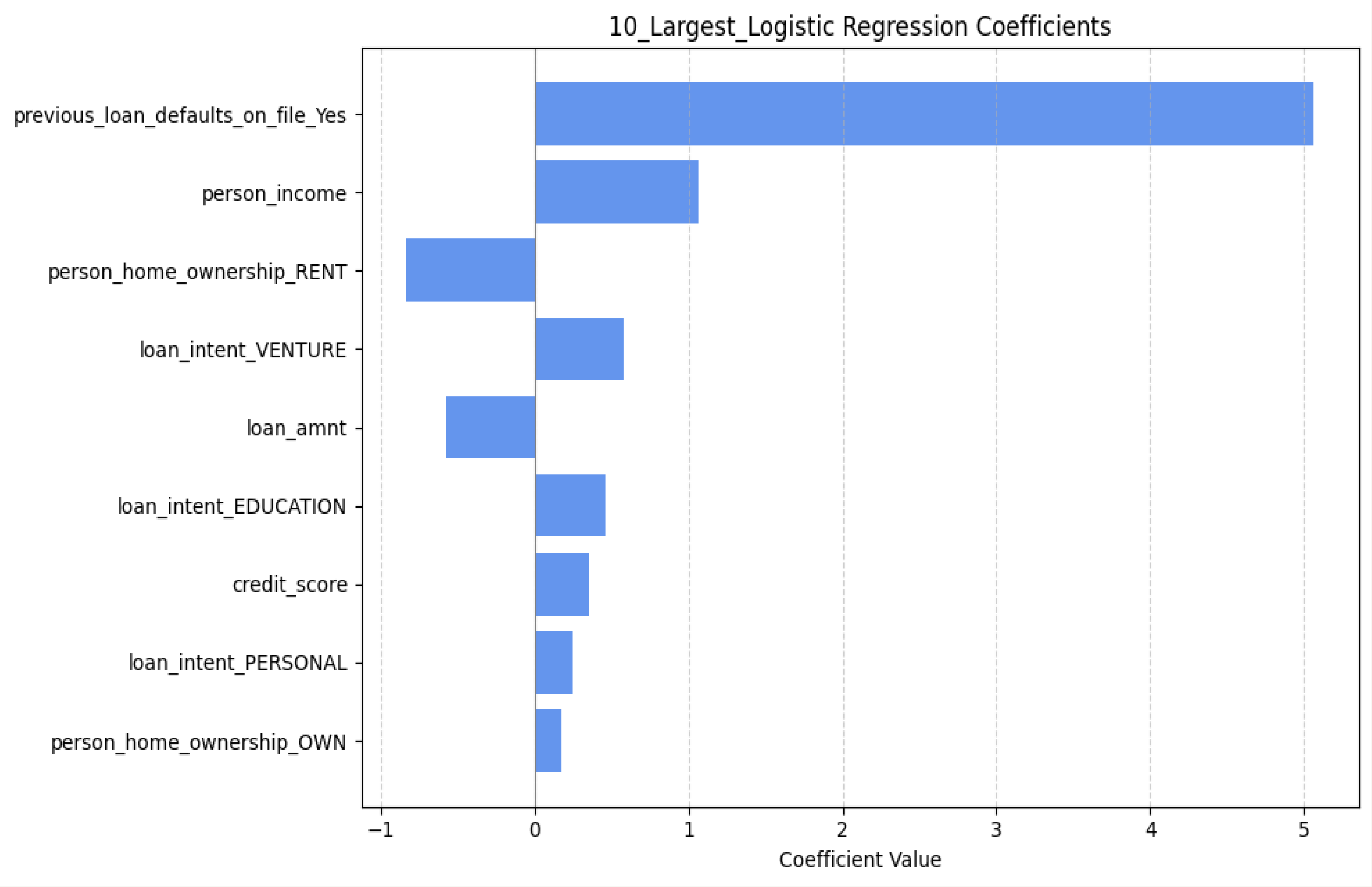
  Accuracy: 87.00%

- **Interpretation:**
- 91% of predicted positives are correct (low false positives)
- 92% of actual positives were identified (low false negatives)
- **Accuracy (87%)** is still good but slightly lower than precision and recall, possibly due to:**Class imbalance**, or Model making more mistakes on the negative class.
- The F1 score (0.92) confirms your model has excellent balance between identifying true positives and avoiding false alarms.

What does this Model tell us ?

Evaluation:

- Minority Class:
- **F1 score:** 0.70:Indicates this model is also less effective at identifying non-defaults.
- **Conclusion:** Changing C didn't improve performance (stayed essentially the same)



Model: Random Forest

Evaluation:

- **Precision :** 0.91

  91% of predicted defaulters were correct.

- **Recall :** 0.96

  96% of actual defaulters were correctly identified — **very strong!**

- **F1 Score :** 0.94

  Excellent balance between precision and recall.

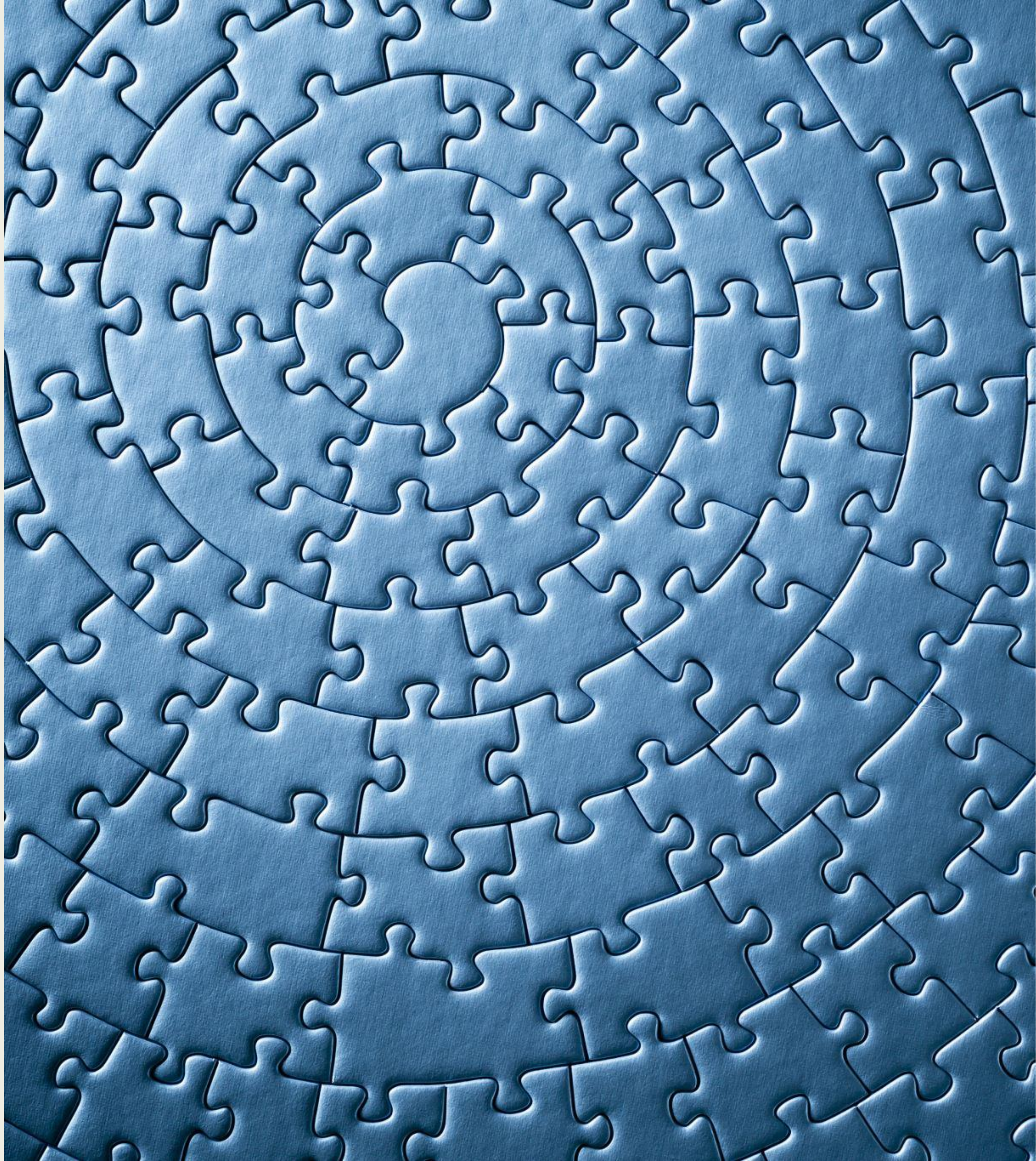
- **Accuracy:** 90.0%

  Overall, 90% of predictions were correct — higher than your logistic regression.

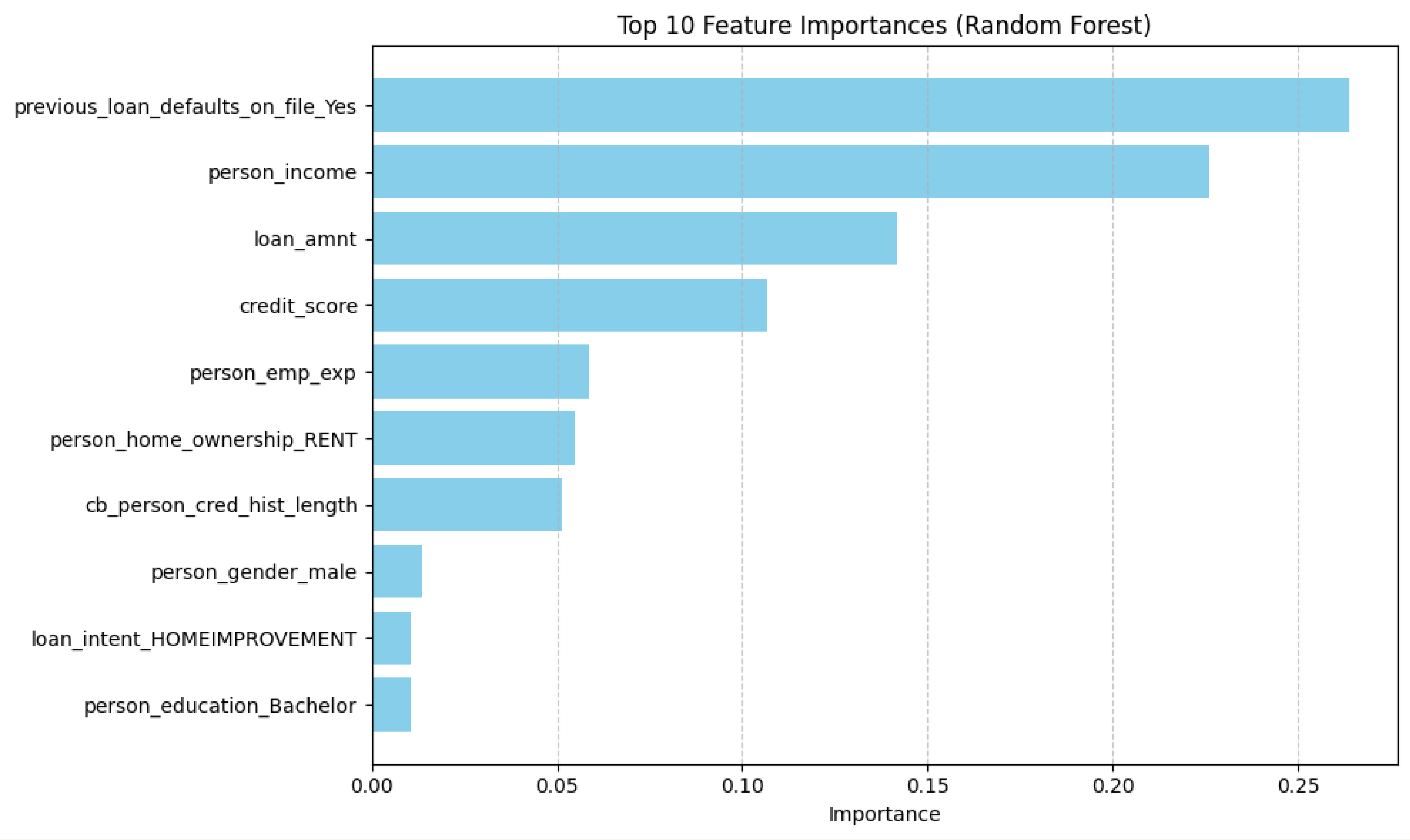
Interpretation

- **Random Forest is outperforming logistic regression** in all metrics **except precision (same).**
- The **boost in recall (from 0.92 → 0.96)** means the random forest **misses fewer defaulters** — important in financial applications where false negatives can be costly.
- **F1 score and accuracy are both improved**, suggesting a better general classifier.

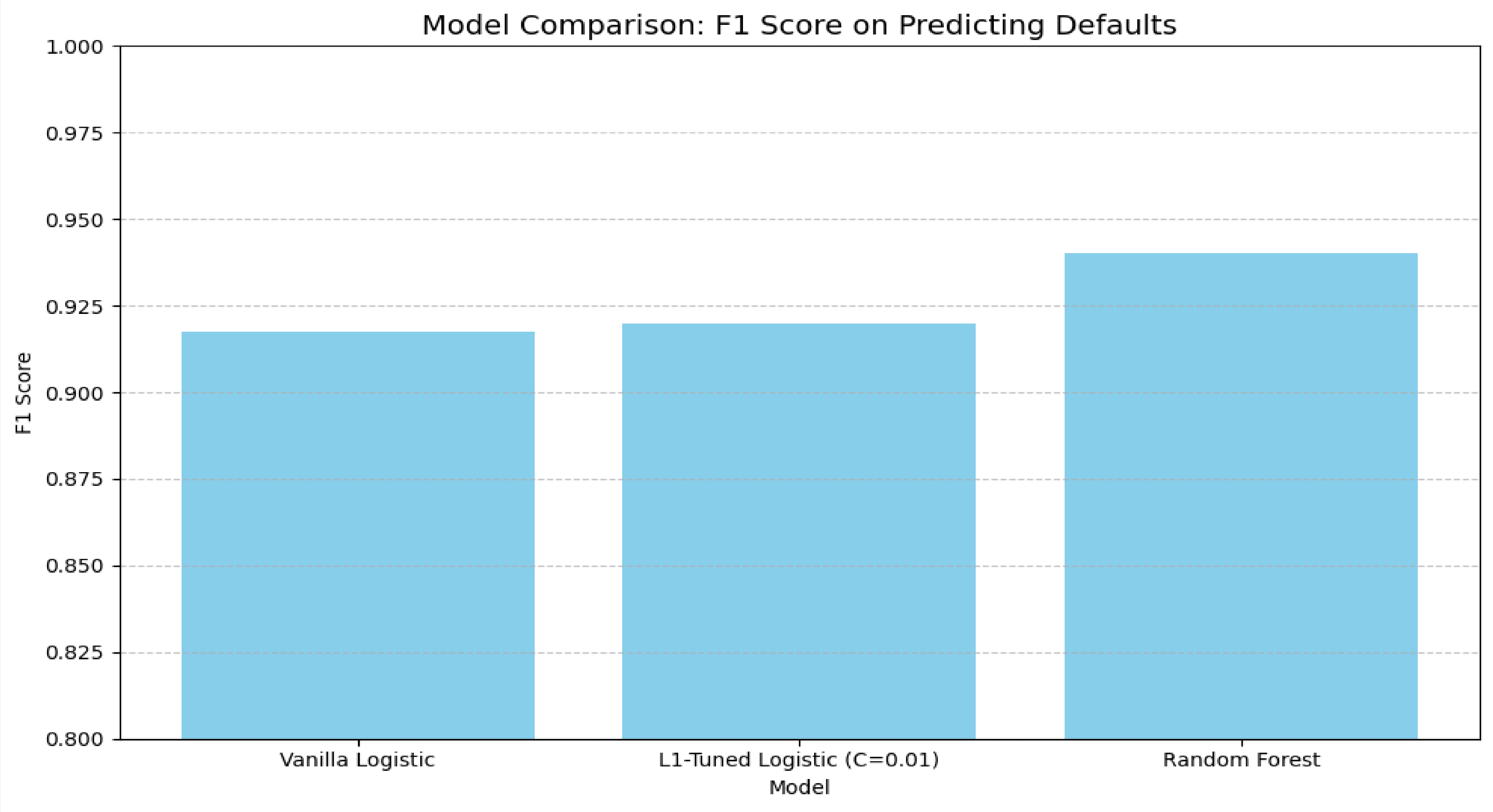
FINAL INTERPRETATION



Random Forest Coefficients



F1 score and Model Comparison



Conclusion:

- The Random Forest model is more accurate than our Logistic Regression model; however, it is far less interpretable.
- We recommend using the models as appropriate: Logistic Regression for interpretability, and Random Forest for predictive accuracy.