ECE_GY_6143 Machine Learning
Group Member: Tianqi Zhen, Jenny Chen

Loan Repayment Prediction — Machine Learning Project

This project predicts loan repayment outcomes using KNN, Random Forest, and Logistic Regression. Although all models share the same preprocessing pipeline, Logistic Regression is trained on the full dataset, while KNN and Random Forest are trained on stratified subsets due to computational constraints. This approach ensures efficient training while maintaining fair performance comparison across models.

Dataset comes from:

Predicting Loan Payback | Kaggle

1. Problem Description

Given loan-related financial and demographic features, the goal is to predict:

loan_paid_back → 1 (repaid) or 0 (not repaid)

This is a binary classification problem. The dataset is highly imbalanced, with the majority of samples belonging to class 1. Therefore, good evaluation requires going beyond simple accuracy.

2. Implementation

- **Data**
    - A unified preprocessing pipeline was applied to all three models
    - Remove non-informative features:

      id is removed because it is a random identifier and harms distance-based models like KNN.

    - Handle missing values:

    Put in missing values, used median imputation for numeric and the most frequent imputation for categorical

  - OHE applied to 6 categorical variables::

    gender, marital_status, education_level, employment_status, loan_purpose, grade_subgrade

  - Feature scaling:

    StandardScaler applied only to numeric columns.

    After preprocessing:

    Training features: 593,994 × 60

    Test features: 254,569 × 60

- **Training Strategy**

  Different models use different splits based on computational needs:

  Logistic Regression:

  Standard 75/25 train-validation split.

  KNN and Random Forest:

  Trained on a 10% stratified subsample of the training data to reduce computation time.

3. Evaluation

- **Accuracy**
  - Logistic Regression: 89.99%
  - KNN: 88.34% with k =21
  - Random Forest: 89.68%
- **ROC**
  - Logistic Regression: 90.77%
  - KNN: 87.45% with k =21
  - Random Forest: 90.86%
- **Interpretation of confusion matrix**

    - True Positive: correctly predicted repaid loans, the majority class
    - True Negative: correctly predicted defaulted loan
    - False Positive: it predict repaid for a borrower who actually default
    - False Negative: reject a borrower who have repaid

4. Model Comparison

Logistic Regression

- Interpretable
- Performs surprisingly well accuracy about 89.99%
- Useful as linear benchmark
- Fastest prediction time
- ROC about 90.77%

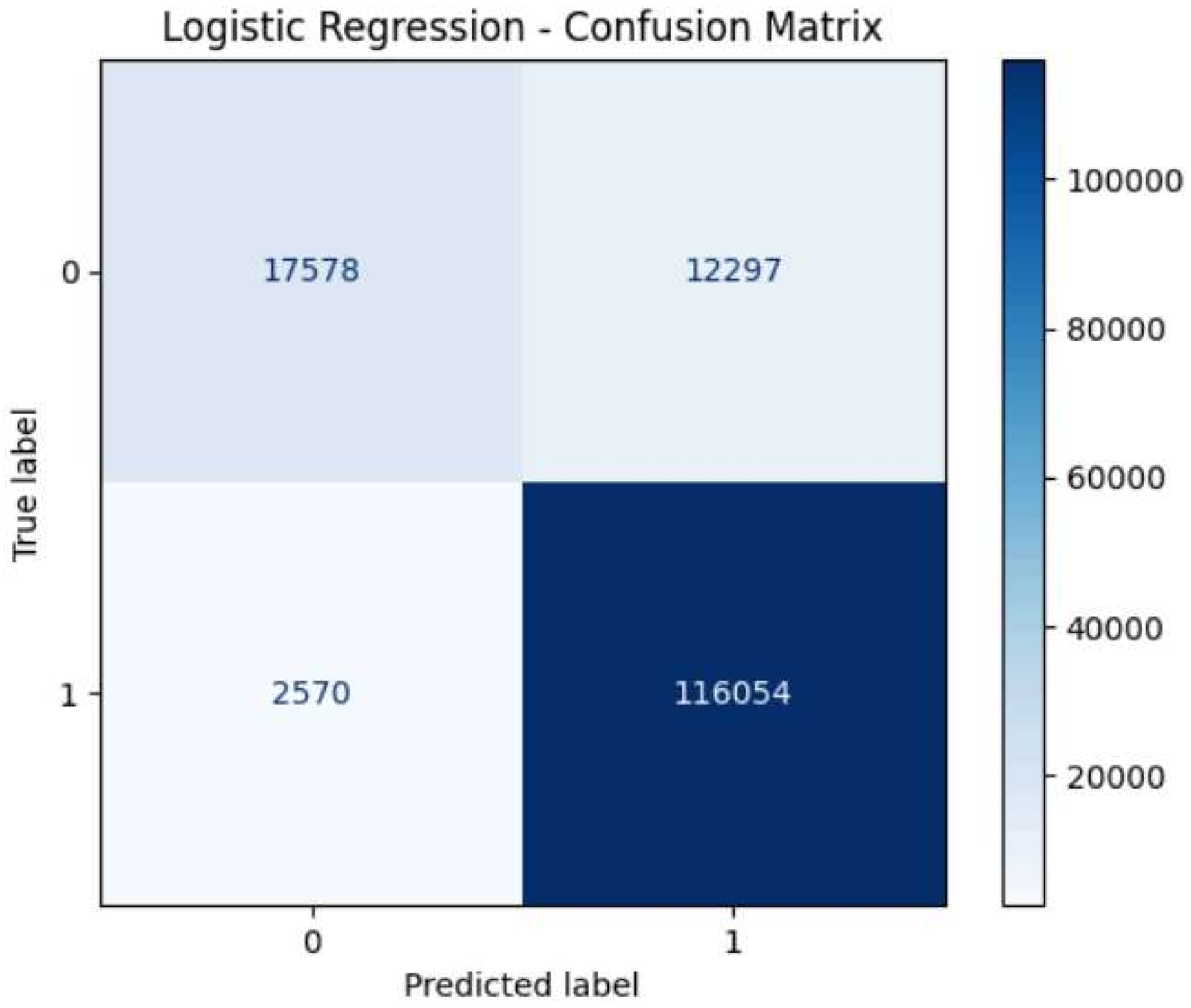
KNN

- Good baseline
- Performance limited by high-dimensional OHE features
- Super slow for large-scale predictions
- Accuracy about 88.34%
- ROC about 87.45%

Random Forest

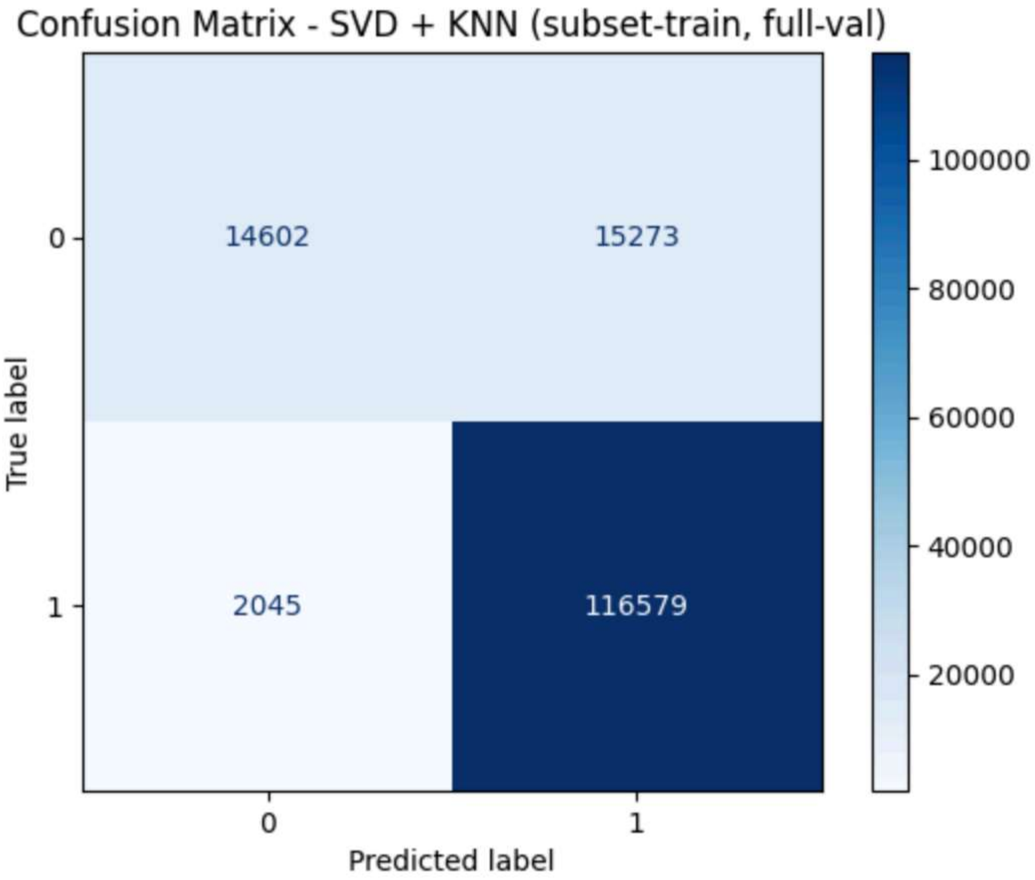
- good performing model accuracy about 89.68%
- Handles nonlinear patterns well
- Deals better with imbalanced data
- Fast prediction time
- ROC about 90.86%

5. Confusion Matrix



Logistic Regression

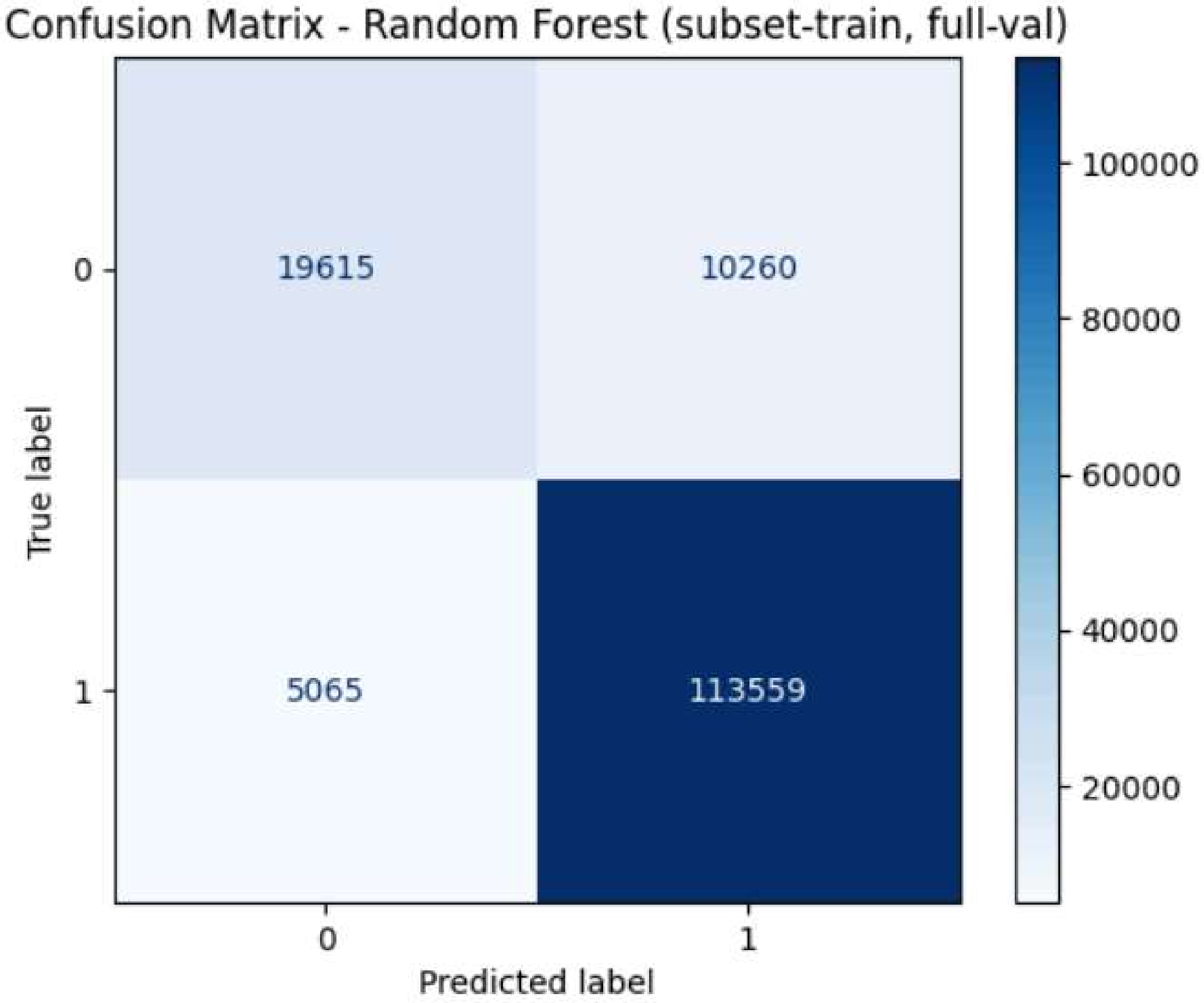
- FN = 2,570 → Lowest among the 3 models
- FP = 12,297 → Moderate
- Logistic Regression is conservative, it tends to classify borderline cases as *positive* only when sufficiently confident
- Good at correctly identifying borrowers who will repay
- Misses fewer actual repayments, meaning fewer rejected “good customers”
- Moderate FP means some risky borrowers are incorrectly classified as safe



KNN

- FP = 15,273 → Highest among models
- FN = 2,045 → Lowest overall
- KNN sacrifices FP control to reduce FN, it predicts *repayments* more often, increasing FP
- Best at identifying borrowers who truly repay
- Worst at avoiding risky approvals

- In financial contexts, high FP is dangerous, making KNN less suitable despite high accuracy



Random Forest

- FP = 10,260 → Lowest among all models
- FN = 5,065 → Highest among the three
- Random Forest is more conservative, predicting more *non-repayment* than the other models
- Best at reducing risky approvals
- More likely to reject borrowers who would actually repay
- This trade-off is often desirable in credit risk, where avoiding defaults is more important than maximizing approvals

6. Conclusion

The project demonstrates that loan repayment can be predicted effectively using classical machine learning models.

A unified preprocessing pipeline ensures consistency and fair comparison.

Among the tested algorithms:

- **Random Forest** performs the best overall.
- **Logistic Regression** provides a strong and efficient benchmark.
- **KNN** is conceptually simple but computationally expensive.

Random Forest achieved the strongest overall performance, obtaining the highest validation accuracy and the most favorable confusion matrix profile. It demonstrated the best ability to reduce false positives which is an essential property in loan repayment prediction, where incorrectly approving high-risk borrowers leads to financial loss.

Logistic Regression trained on the full dataset, performed competitively and provided a solid linear baseline. It offered a balance between false positives and false negatives, making it stable and interpretable, though slightly less effective than Random Forest in capturing nonlinear patterns.

KNN, despite respectable accuracy, showed the weakest risk control performance. Its higher false-positive rate indicates a tendency to approve borrowers who may default, which makes it less suitable for credit risk applications.

Based on ROC-AUC evaluation, Logistic Regression outperformed both KNN and Random Forest in probability ranking ability. Since loan risk assessment emphasizes ranking borrowers by default risk rather than maximizing accuracy at a single threshold, Logistic Regression was selected as the final model.

These findings show how different ML techniques behave on the same real-world financial dataset and provide insights useful for credit scoring and risk assessment. Although Logistic Regression achieved slightly higher accuracy, Random Forest was selected due to its superior control of false positives, which is critical in credit risk applications.