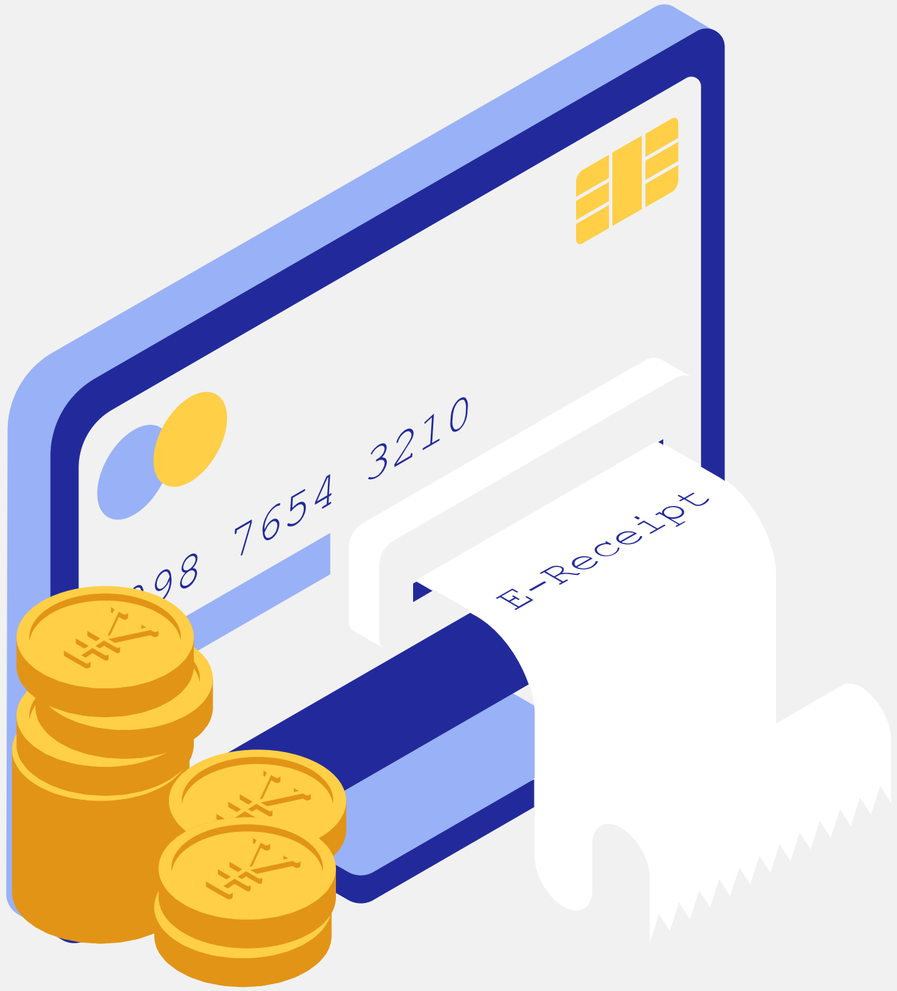


Yose Elvis Saputra - President University

Credit Risk Prediction Model

File
https://github.com/ElvissYo/Credit-Risk-Model

Background

- Loan defaults are a major challenge for lending companies
- Historical loan data can be leveraged for prediction
- Objective: build a credit risk prediction model

Dataset

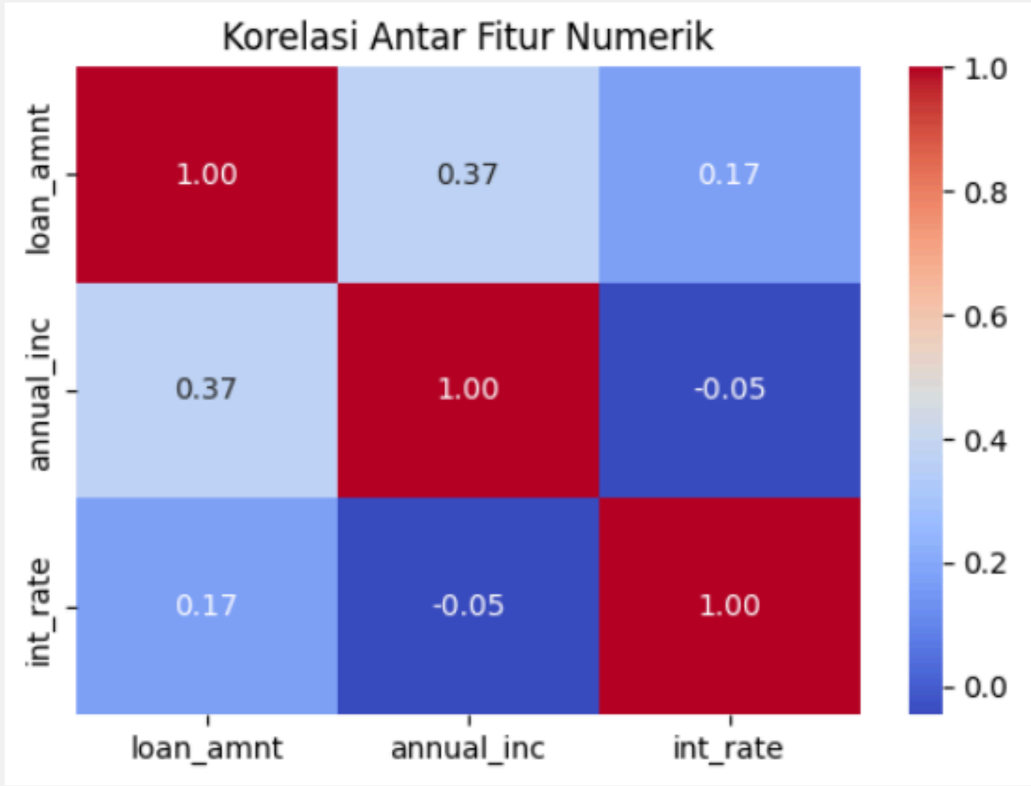
- Source: Loan Data 2007-2014
- Size: ~ 466,285 records, 75 features
- Key features:
  - Loan amount, Interest rate, Income
  - Employment length, Home ownership
  - Loan status (target variable)

EDA

- The target variable is highly imbalanced.
- Most loans belong to a few dominant categories (e.g., Fully Paid), while other statuses appear much less frequently.
- This imbalance may cause the model to be biased towards the majority class and perform poorly on minority classes.
- To address this issue, SMOTE (Synthetic Minority Oversampling Technique) is applied during preprocessing.

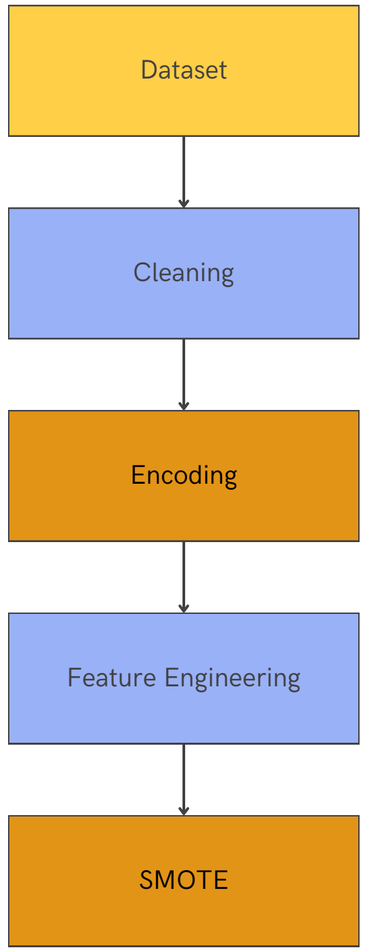
EDA

Correlation Analysis:
- Loan Amount is moderately correlated with Income (0.37).
- Interest Rate has little correlation with Income (–0.05).
- Loan Amount is negatively correlated with Interest Rate (–0.39).
- No strong multicollinearity detected among key features.



Data Preprocessing

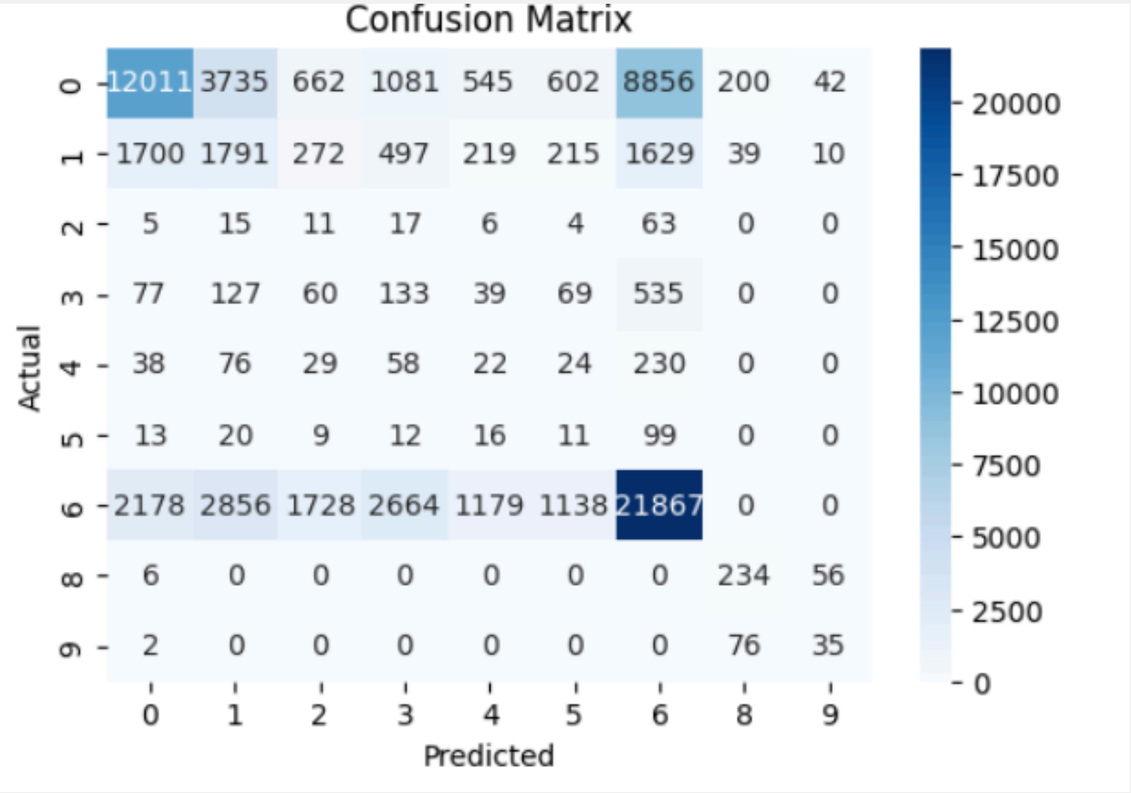
- Feature selection
- Handle missing values
- Encode categorical variables
- Feature engineering: credit history age
- Handle imbalance with SMOTE



Modeling

- Algorithm: Random Forest Classifier & Logistic Regression
- Train-test split: 80% / 20%
- Pipeline: Preprocessing → SMOTE → Model

Random Forest



Classification Report (Precision, Recall, F1)

Macro precision 0.26, recall 0.31, and F1-score 0.26, higher than Logistic Regression.

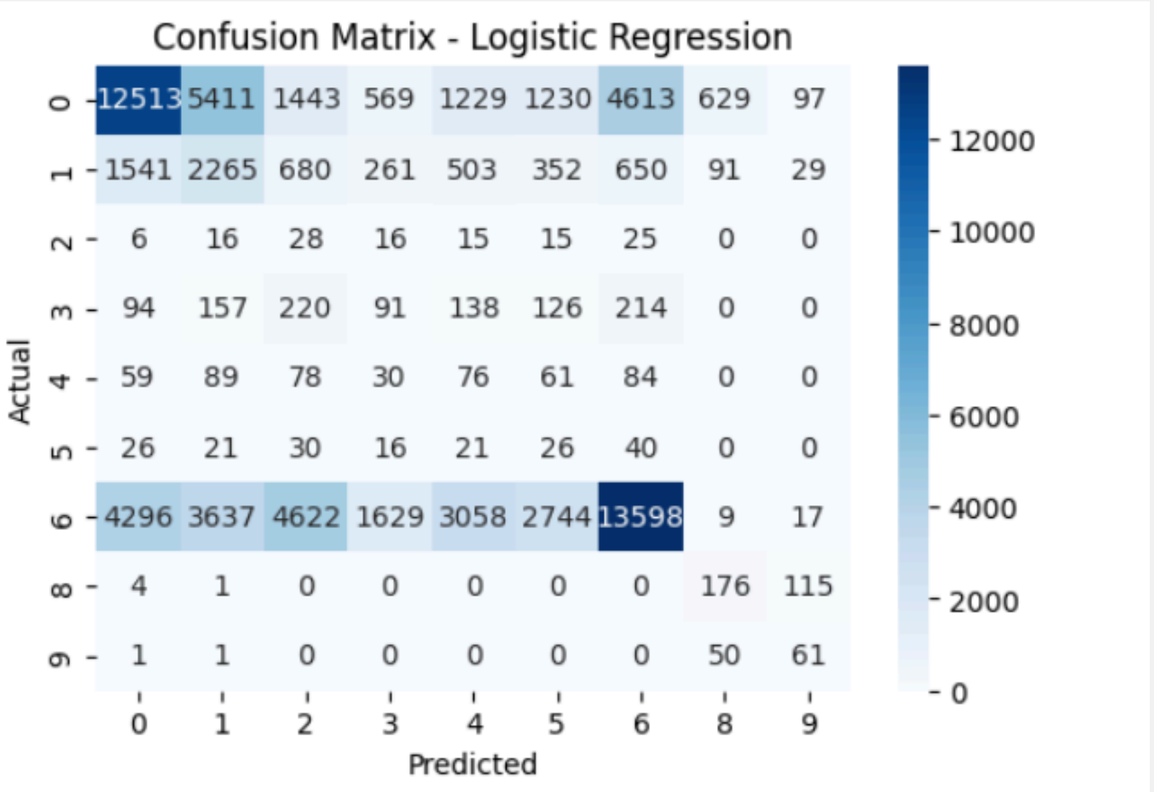
Confusion Matrix

More balanced predictions across classes, improved recall compared to Logistic Regression.

ROC-AUC Score: 0.734

slightly lower than Logistic Regression but more consistent across classes.

Logistic Regression



Classification Report (Precision, Recall, F1)

Overall macro precision 0.22, recall 0.33, and F1-score 0.22, showing weak performance on minority classes.

Confusion Matrix

High misclassifications on rare classes, but better performance for class 0 and 6.

ROC-AUC Score: 0.739

indicating moderate class separation ability.

Results & Insights

- Overall Model Performance

Random Forest achieved an accuracy of 52% and ROC-AUC of 0.734, while Logistic Regression reached 41% accuracy and ROC-AUC of 0.739. Random Forest performed better overall in terms of balanced precision, recall, and F1-score.

- Confusion Matrix Findings

Both models predict majority classes (0 and 6) more accurately, while minority classes still show high misclassification. Random Forest provided more balanced results across classes compared to Logistic Regression.

- Key Insights

-Logistic Regression serves as a baseline but struggles with minority classes.
-Random Forest offers improved recall and precision, making it more effective for this dataset.-
-Class imbalance remains a challenge, especially for rare categories.

- Business Implication

The Random Forest model is more suitable to support loan risk assessment, providing reliable predictions for the majority of loan cases. Logistic Regression can still be used for interpretability, while further re-sampling or model tuning is needed to improve detection of minority risk classes.

Conclusion

01 Credit risk model successfully developed

02 Promising evaluation results

03 High potential for real-world use