MACHINE LEARNING • STATISTICAL ANALYSIS

Loan Default Prediction Using Machine Learning Algorithms

A detailed academic case study covering exploratory analysis, hypothesis testing, feature selection, class-imbalance treatment, model evaluation and sample borrower prediction.

Project Objective

To study borrower and loan-related characteristics associated with loan default, develop classification models and compare their ability to distinguish default from non-default cases. The project also demonstrates how a trained model can generate a prediction for a new borrower record.

Key objectives

- Explore patterns in borrower data through Exploratory Data Analysis (EDA).
- Test associations between categorical variables and loan default using the Chi-Square test.
- Examine numerical relationships using a correlation matrix and Variance Inflation Factor (VIF).
- Identify influential predictors using Random Forest feature importance and Partial Dependence Plots.
- Compare Logistic Regression, Random Forest and LightGBM with SMOTE and SMOTEENN.
- Demonstrate borrower-level prediction using the trained models.

Dataset & workflow

The project used a Kaggle dataset with 18 borrower-related attributes, including age, income, loan amount, credit score, employment type, interest rate, loan term and debt-to-income ratio. The target variable was loan default (default / non-default).

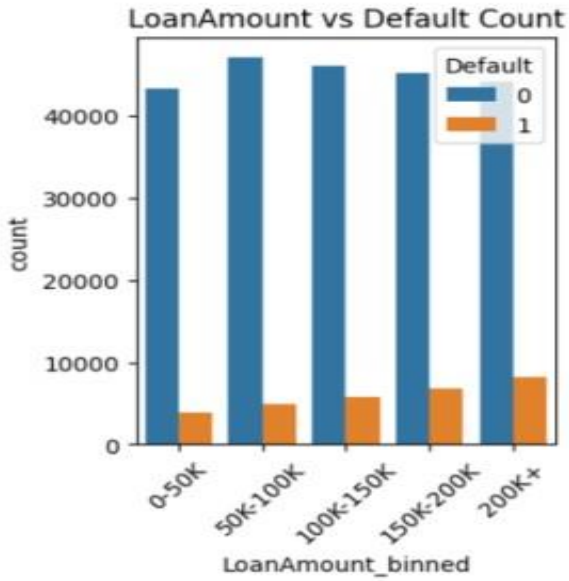
DATA Kaggle dataset	**PRE-PROCESS** Check imbalance; SMOTE / SMOTEENN	**ANALYSE** EDA and statistical tests	**MODEL** Train, evaluate, predict

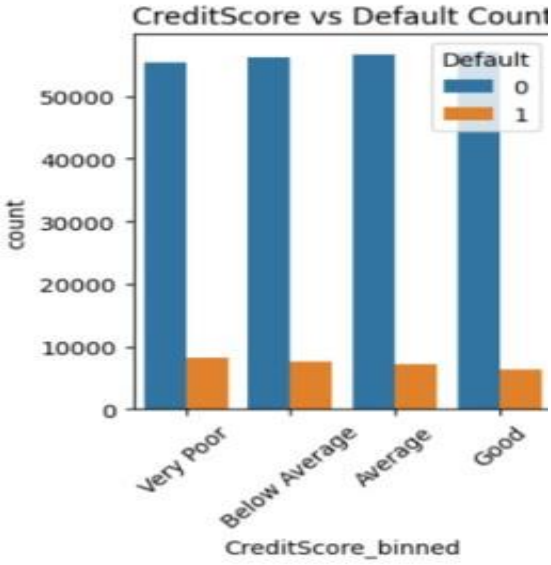
Project author: Shruti Nivasrao Tiware-Patil
Institution: Rajarshi Chhatrapati Shahu College, Kolhapur | Department of Statistics

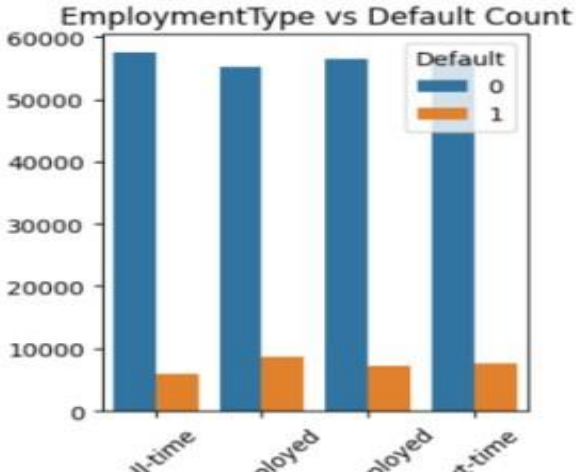
01 | Exploratory Data Analysis

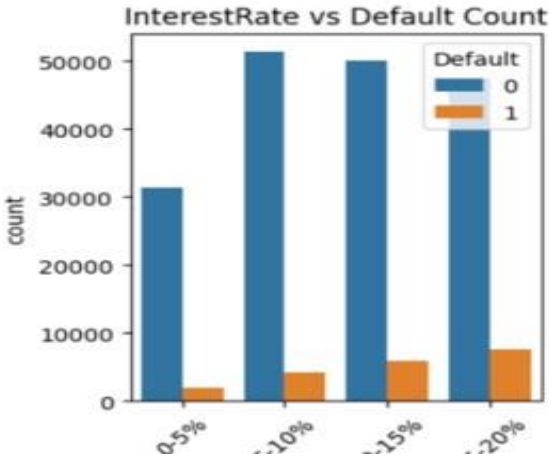
The project used binned bar charts to compare default and non-default counts across selected borrower characteristics.

These plots were used to explore patterns before fitting classification models.









Key observations reported in the project

• Loan amount: default counts increased across higher loan-amount groups in the displayed chart.

• Credit score: the project interpretation reported lower default counts for better credit-score categories.

• Employment type: unemployed applicants were described as having the highest default count in the analysis.

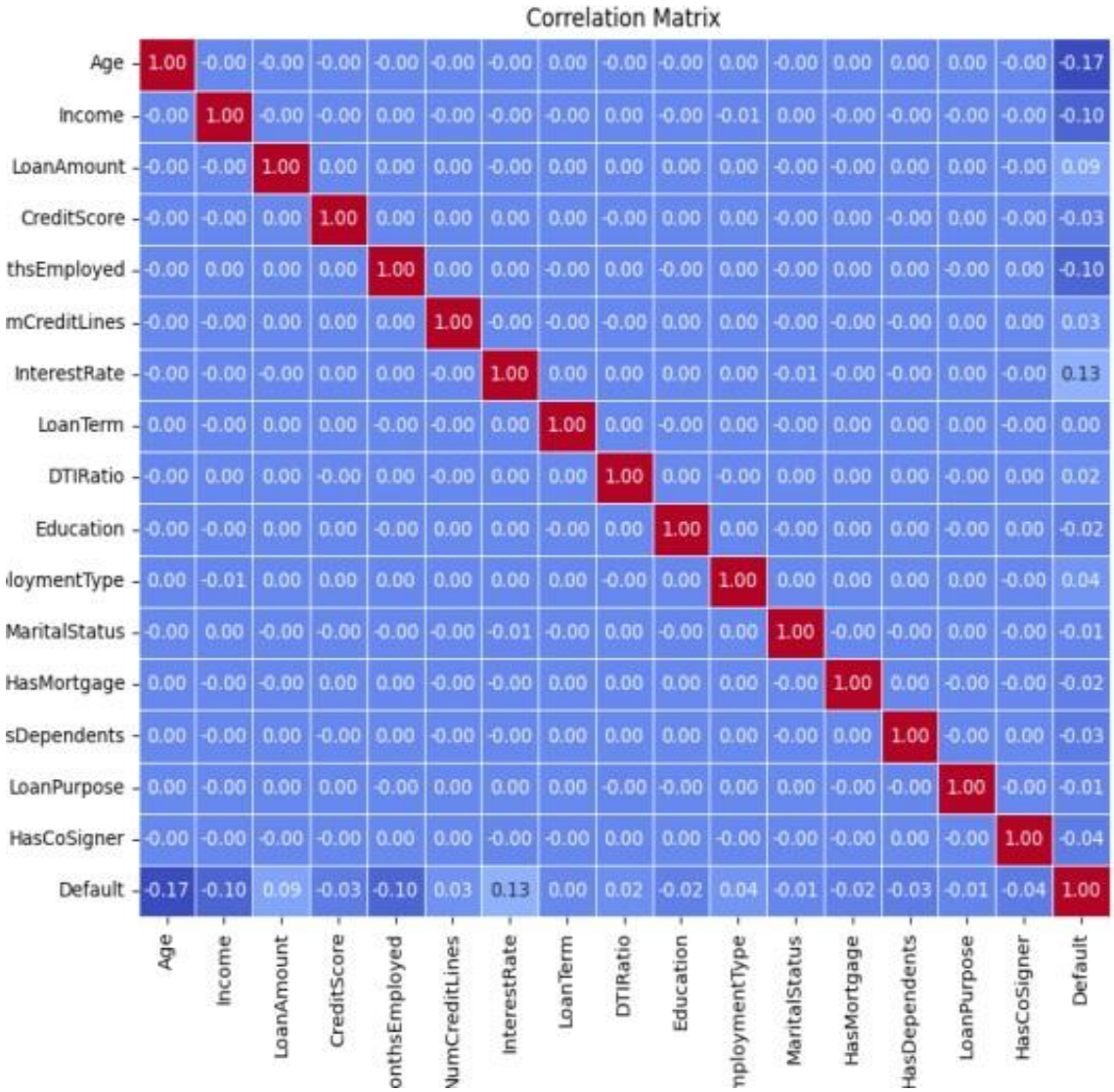
Interest rate: default counts increased across higher interest-rate groups.

These are observed patterns in the analysed dataset. They describe associations and do not establish that any single factor causes loan default.

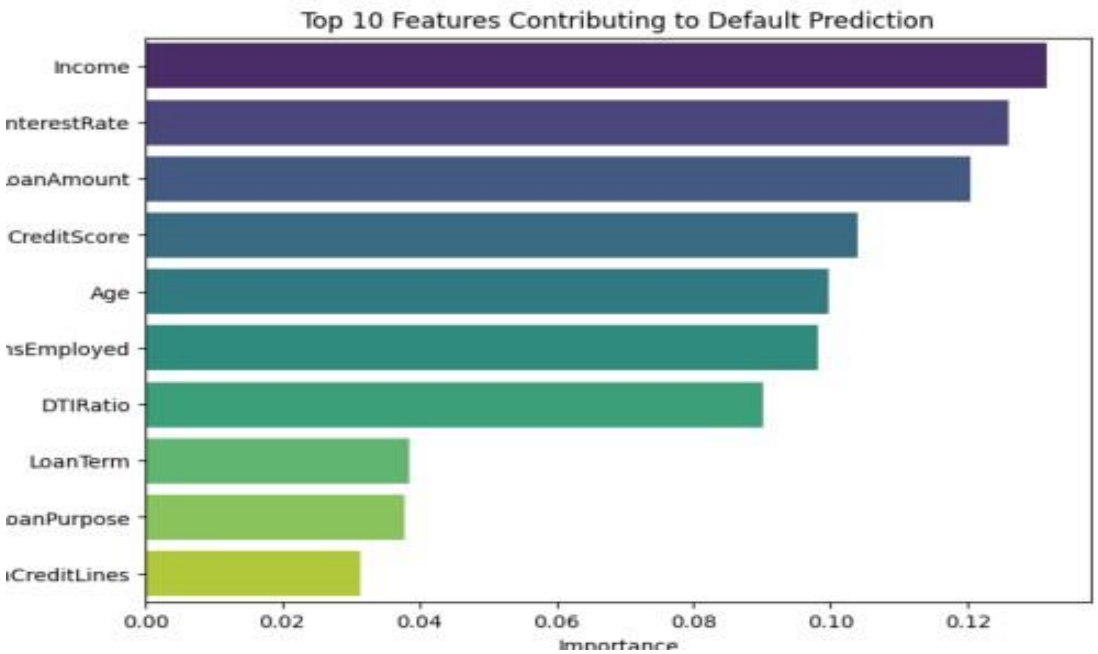
02 | Correlation, Feature Selection & Statistical Tests

Correlation analysis

The correlation matrix showed no strong pairwise linear correlation between the predictors. The reported correlation between Default and individual numerical variables was also relatively weak; the largest absolute value shown was approximately 0.17 for Age.



Feature selection using Random Forest



Income and Interest Rate were ranked as the two most influential features in the Random Forest feature-importance plot, followed by Loan Amount and Credit Score.

03 | Hypothesis Testing & Multicollinearity Check

Chi-Square tests

The project conducted Chi-Square tests for seven categorical features. All seven reported p-values were below 0.05; therefore, each test showed a statistically significant association with the Default variable at the 5% significance level.

Feature tested	Chi-square	p-value	Result (5%)
Education	214.019	3.94 × 10■■■	Significant
Employment Type	529.745	1.71 × 10■[11]■	Significant
Marital Status	200.361	3.11 × 10■■■	Significant
Has Mortgage	133.252	7.96 × 10■[31]	Significant
Has Dependents	306.850	1.06 × 10■■■	Significant
Loan Purpose	127.935	1.08 × 10■[2]■	Significant
Has Co-signer	390.305	7.10 × 10■■■	Significant

Variance Inflation Factor (VIF)

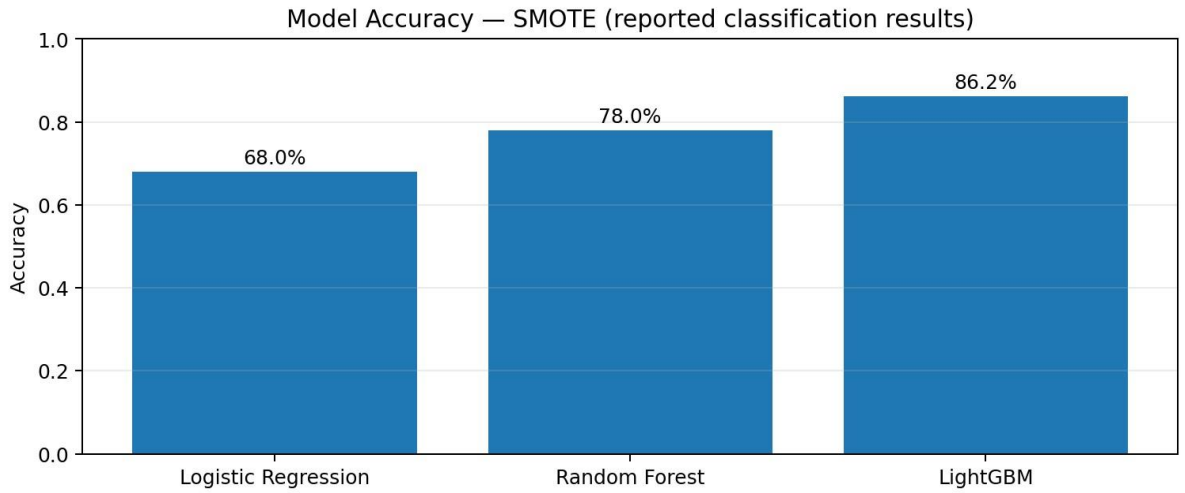
The report listed VIF values close to 1 for the numerical predictors: Age, Income, Loan Amount, Credit Score, Months Employed, Number of Credit Lines, Interest Rate, Loan Term and DTI Ratio. This indicates no high multicollinearity among these listed predictors in the reported VIF check.

Test outcome summary

Successful / significant: all 7 Chi-Square tests were significant ($p < 0.05$).
VIF check: all listed predictor VIF values were approximately 1, so the report did not indicate a multicollinearity concern.
Correlation: no strong pairwise linear relationships were reported in the correlation matrix.

04 | Model Evaluation — SMOTE

SMOTE was used to create synthetic samples for the minority class. The three classifiers were then evaluated using classification metrics.



Model with SMOTE	Accuracy	Default precision	Default recall	Default F1	AUC-ROC
Logistic Regression	0.68	0.21	0.65	0.32	0.724 (CV mean)
Random Forest	0.78	0.21	0.31	0.25	0.674
LightGBM	0.862	0.33	0.17	0.23	Not stated

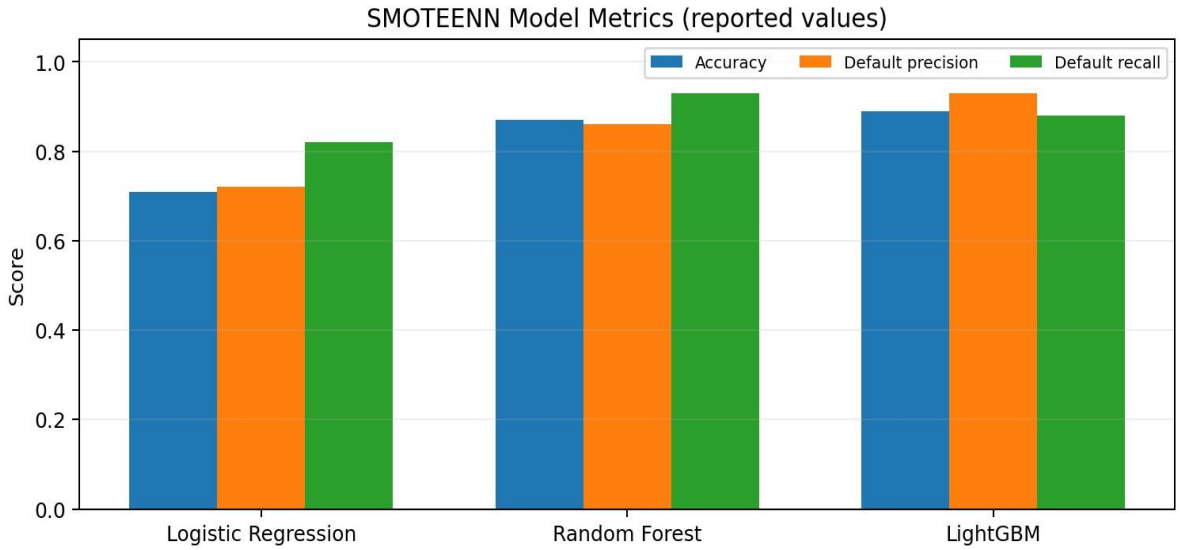
What worked and what did not

- LightGBM recorded the highest overall accuracy in the SMOTE results (86.23%). However, its reported Default recall wasonly 17%, meaning it identified only a small share of actual defaulters. Therefore, high accuracy alone did not make it effective for detecting the minority default class.
- Logistic Regression had 65% Default recall but low Default precision (21%), so many predicted defaulters were falsepositives.
- Random Forest had 31% Default recall and 21% Default precision, showing weak default-class detection in this run.

Interpretation: SMOTE-only results showed that overall accuracy did not adequately describe performance on the minority default class. Precision, recall and F1-score were important for judging default detection.

05 | Model Evaluation — SMOTEENN

SMOTEENN combines minority-class oversampling with Edited Nearest Neighbours cleaning. The project reported improved default-class precision and recall for all three models compared with the SMOTE-only runs.



Model with SMOTEENN	Accuracy	Default precision	Default recall	Default F1	AUC-ROC
Logistic Regression	0.71	0.72	0.82	0.77	0.772
Random Forest	0.87	0.86	0.93	0.89	0.944
LightGBM	0.89	0.93	0.88	0.91	0.957

Which models were successful?

- **LightGBM + SMOTEENN:** highest reported accuracy (89%), Default precision (93%), and AUC-ROC (0.957). Its Default recall was 88%. This was the strongest overall result in the report.
- **Random Forest + SMOTEENN:** Default recall was highest at 93%, with 86% precision, 89% F1 and 0.944 AUC-ROC. It was strong when identifying actual defaulters was a priority.
- **Logistic Regression + SMOTEENN:** improved substantially over its SMOTE run, with 82% Default recall and 72% precision, but lower overall scores than the other two models.

Best overall model in the submitted results: LightGBM with SMOTEENN. **Highest default recall:** Random Forest with SMOTEENN.

06 | Sample Prediction & Project Conclusion

A sample borrower profile was passed to the three trained models. The report used the following input values:

Age: 35 **Income:** 50,000 **Loan amount:** 15,000 **Credit score:** 700
Months employed: 24 **Interest rate:** 5 **Loan term:** 36 **DTI ratio:** 0.4

Model used	Predicted class	Default probability
LightGBM	Likely defaulter	0.92 (92%)
Random Forest	Likely defaulter	0.90 (90%)
Logistic Regression	Not a defaulter	0.28 (28%)

Overall findings

1. All seven Chi-Square tests reported statistically significant associations with Default.
2. VIF values close to 1 indicated low multicollinearity among the listed numerical predictors.
3. SMOTE-only models had limitations in identifying the default class, particularly LightGBM despite its higher accuracy.
4. SMOTEENN improved the reported minority-class performance.
5. LightGBM with SMOTEENN achieved the best overall reported result (89% accuracy; 0.957 AUC-ROC), while RandomForest with SMOTEENN had the highest default recall (93%).

Limitations and responsible interpretation

The project results are specific to the dataset and evaluation setup described in the academic report. The different model predictions for the same sample borrower show that predictions can vary by algorithm. These outputs are demonstrations only; they should not be used as real loan approval decisions without further validation, calibration, fairness checks and independent testing.

Project skills demonstrated

Exploratory Data Analysis • Statistical Hypothesis Testing • Correlation & VIF • Feature Importance • Class-Imbalance
Handling • Classification • Model Evaluation • Probability-based Prediction

Academic project | M.Sc. Statistics | 2024–2025
Author: Shruti Nivasrao Tiware-Patil