Predicting Credit Card Default Using Logistic Regression

By: Shivesh Raj Sahu

Executive Summary

This report presents a comprehensive analysis of credit card default prediction using logistic regression, leveraging the UCI Credit Card dataset. Through extensive data cleaning, feature engineering, and two modeling approaches—standard and class-balanced logistic regression—we highlight the critical importance of addressing class imbalance. Handling imbalance via class weighting increased defaulter recall from 23 percent to 63 percent with no loss in ROC-AUC. This work demonstrates the trade-offs between precision and recall in high-stakes financial settings and outlines next steps for further improvement.

Introduction

Credit card default prediction plays a critical role in managing financial risks for institutions. Using the UCI Credit Card Default dataset containing 30,000 records and 25 features, this analysis employs logistic regression to model the likelihood of credit card payment defaults. Two modeling approaches are evaluated—standard logistic regression and logistic regression with class weighting to handle data imbalance.

Data Preparation and Feature Engineering

Data Loading and Inspection

- Dataset: 30,000 entries with attributes such as credit limit (LIMIT_BAL), age (AGE), payment statuses (PAY_), billing amounts (BILL_AMT), and payment amounts (PAY_AMT*).
- ID column removed as irrelevant.
- No missing values detected.

Categorical Feature Handling

- Ambiguous categories (e.g., EDUCATION values 0, 5, 6; MARRIAGE value 0) grouped into “other.”
- Applied one-hot encoding for categorical variables (SEX, EDUCATION, MARRIAGE).

Engineered Features

UTILIZATION_RATIO (BILL_AMT1/LIMIT_BAL), AVG_PAY_AMT, AVG_BILL_AMT to capture detailed financial behavior insights.

Exploratory Data Analysis (EDA)

Class Imbalance

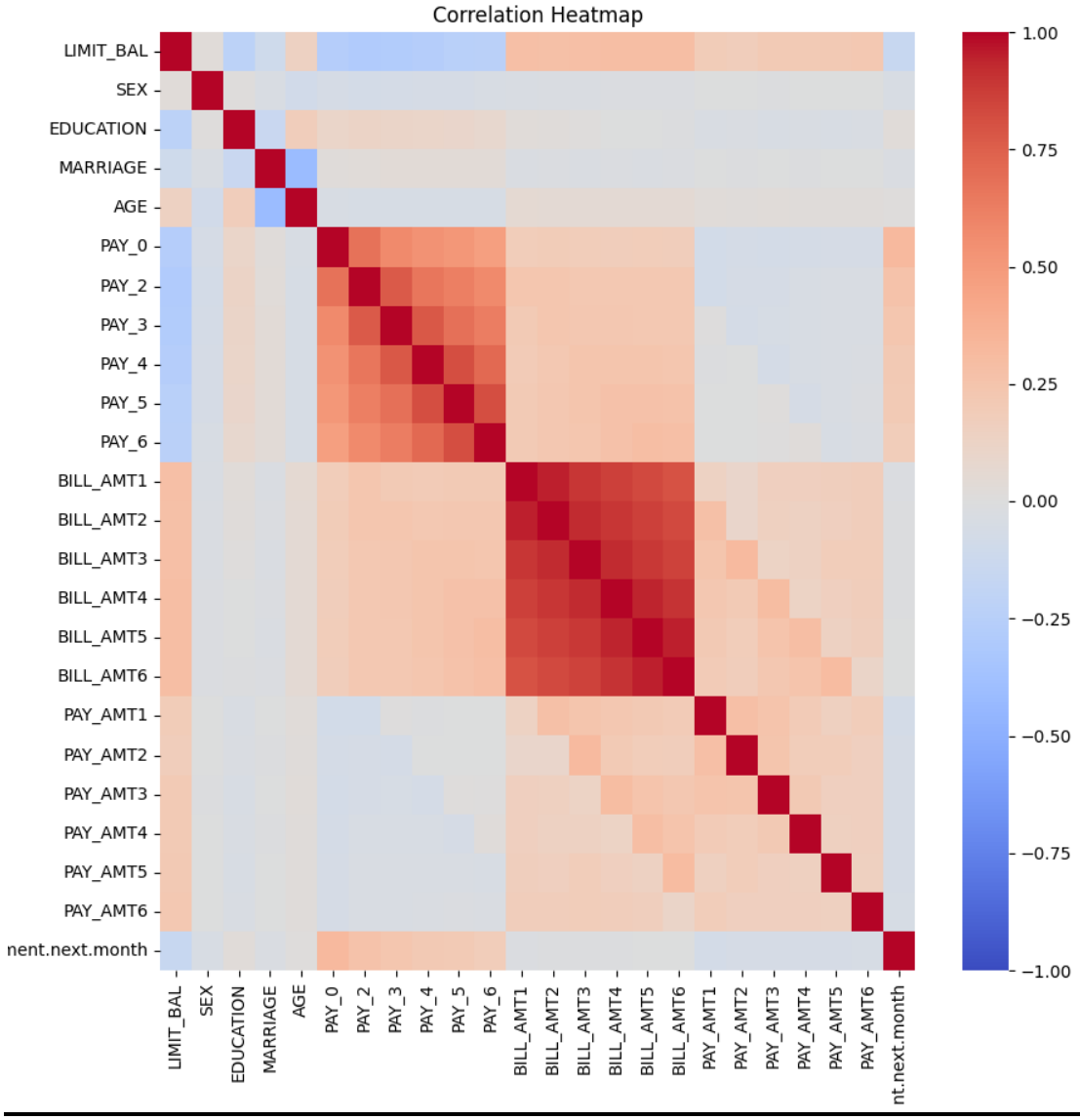
- Default rate: 22% defaults vs. 78% non-defaults, indicating significant class imbalance.
- Implication: Strong class imbalance may bias model to favor “no default” predictions.

Outlier and Distribution Analysis

- Histograms and boxplots highlight skewness and outliers, especially in BILL_AMT and PAY_AMT features.
- Correlation heatmap indicates moderate correlation but no perfect multicollinearity among features.

Predicting Credit Card Default Using Logistic Regression

By: Shivesh Raj Sahu



Predicting Credit Card Default Using Logistic Regression

By: Shivesh Raj Sahu

Modeling Approaches

Standard Logistic Regression

- Data split into 80% train and 20% test, stratified by class.
- Features standardized for uniform scale.
- Logistic Regression trained (max_iter=1000, random_state=42).

Results:

- Accuracy: 81%
- Confusion Matrix:
- True negatives: 4542, False positives: 145
- False negatives: 1005, True positives: 308
- Classification report:
- Precision (default): 68%, Recall (default): 23%, F1-score: 35%
- ROC-AUC: 0.73
- Model identifies non-defaulters effectively but struggles significantly to detect defaulters.

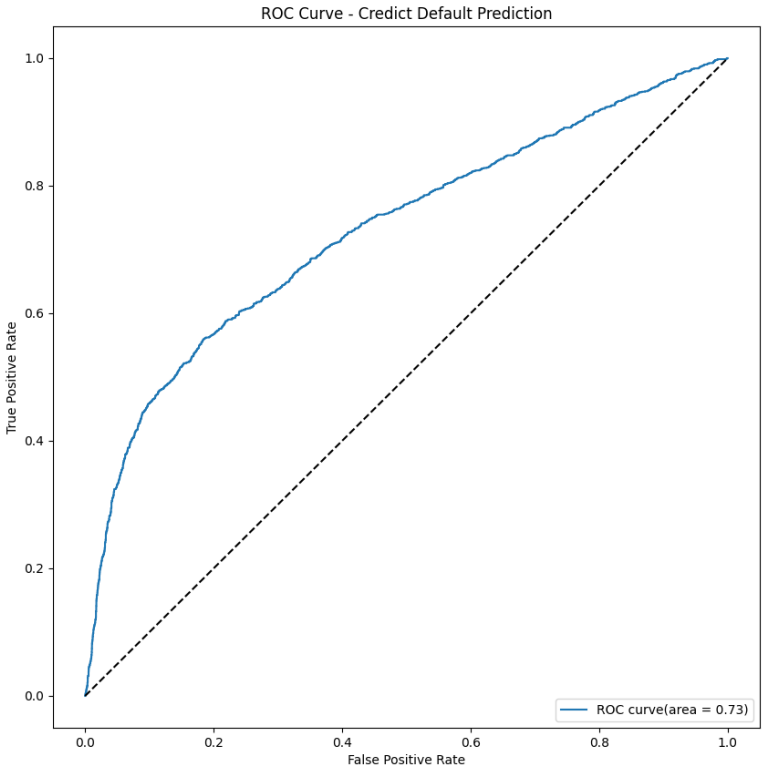
Interpretation: The model correctly identifies most non-defaulters, but misses the majority of true defaulters, highlighting the effect of class imbalance.

Figure 1. ROC Curve (Standard Logistic Regression)

Shows model's ability to distinguish defaulters from non-defaulters. Area under the curve (AUC) of 0.73 indicates moderate discriminative performance.

Predicting Credit Card Default Using Logistic Regression

By: Shivesh Raj Sahu



Logistic Regression with Class Weight='Balanced'

- Adjusted logistic regression to prioritize minority class detection (max_iter=1000, random_state=42, class_weight='balanced').

Results:

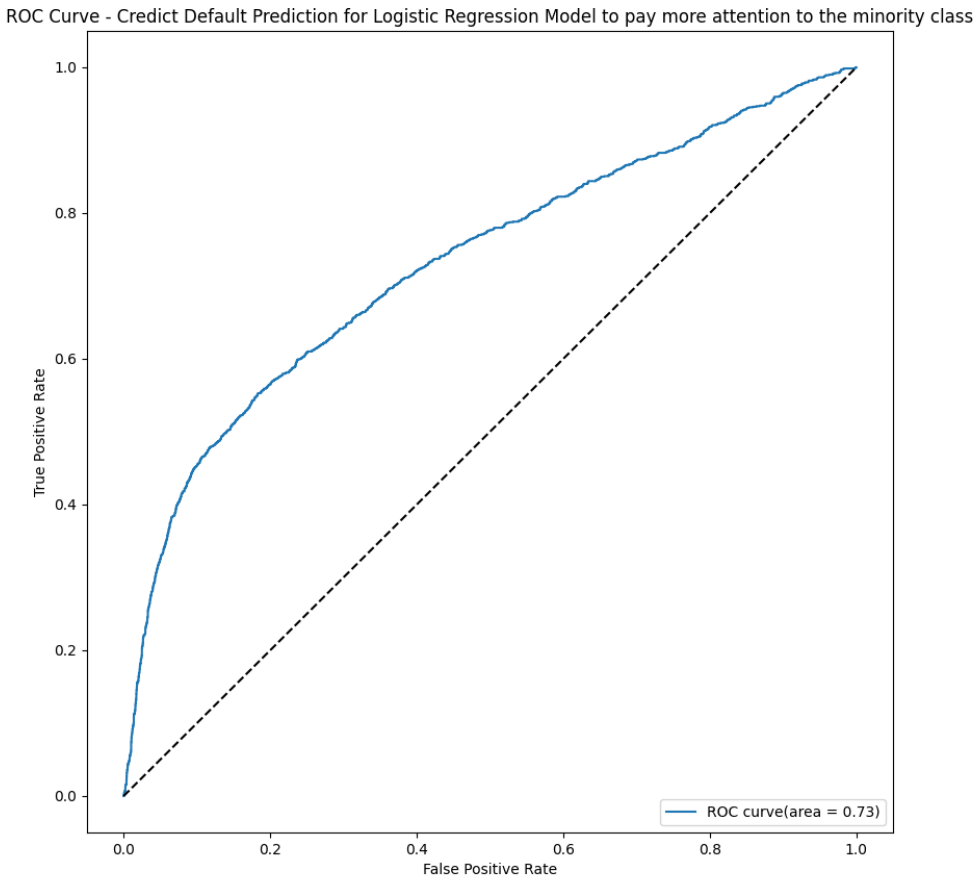
- Accuracy: 70%
- Confusion Matrix:
- True negatives: 3379, False positives: 1308
- False negatives: 492, True positives: 821
- Classification report:
- Precision (default): 39%, Recall (default): 63%, F1-score: 48%
- ROC-AUC: 0.73

- The model now captures significantly more actual defaulters at the cost of increased false positives.

Interpretation: Recall for defaulters improves significantly (23% → 63%), meaning the model flags far more risky clients. False positives among non-defaulters increase, but this may be an acceptable trade-off in risk-averse settings.

Figure 2. ROC Curve (Class-Balanced Logistic Regression)

Class balancing increases recall for defaulters without harming AUC.



Predicting Credit Card Default Using Logistic Regression

By: Shivesh Raj Sahu

Detailed Analysis of Key Graphs

Class Distribution

- Clear depiction of class imbalance emphasizing the importance of handling class imbalance during modeling.

ROC Curves

- Both models exhibit ROC-AUC ~0.73, indicating similar discriminative performance.
- Balanced model improves recall significantly.

Correlation Heatmap

- Demonstrates relationships and moderate correlations between features, useful for feature selection decisions.

Model Comparison Summary

Metric	Standard Logistic Regression	Balanced Logistic Regression
Accuracy	0.81	0.70
Precision (Default)	0.68	0.39
Recall (Default)	0.23	0.63
F1-score (Default)	0.35	0.48
ROC-AUC	0.73	0.73

The balanced model shows significantly improved recall for defaults, critical in financial risk contexts.

Predicting Credit Card Default Using Logistic Regression

By: Shivesh Raj Sahu

Rationale for Presenting Both Models

- Demonstrates natural bias towards majority class by default models.
- Highlights practical improvement via balanced logistic regression.

Business Impact

Failing to identify a true defaulter is costlier for financial institutions than occasionally flagging a reliable customer as risky. By improving recall, the balanced model supports a more conservative, risk-averse business strategy—reducing losses from unpredicted defaults.

Detailed Evaluation and Graphical Interpretation

Class Distribution Plot:

Clearly shows the dataset is highly imbalanced. This justifies the use of class weighting.

ROC Curves:

Both models achieve an AUC of 0.73, indicating reasonable ability to separate classes. The class-balanced model shifts the threshold to favor recall for defaulters, as seen by a steeper rise in the ROC curve.

Boxplots and Distributions:

Outliers and skewness are present—especially in financial features. This suggests potential future improvement via robust scaling or transformation.

Predicting Credit Card Default Using Logistic Regression

By: Shivesh Raj Sahu

Correlation Heatmap:

Some redundancy among bill and payment features, but no features are perfectly collinear.

Error Analysis

- False Negatives: Standard model misses 77 percent of true defaulters—an unacceptable business risk.
- False Positives: Balanced model increases false positives (labeling safe clients as risky), but flags nearly three times as many true defaulters.
- Pattern Analysis: Investigate whether specific age groups, utilization levels, or repayment statuses are disproportionately represented among misclassified samples for targeted business intervention.

Limitations and Considerations

- Outliers & Skewness: Logistic regression can be sensitive to extreme values. Robust scaling or winsorizing may help.
- Data Generalizability: This dataset is historical and region-specific; model may require adaptation before being deployed elsewhere.
- Imbalanced Learning: While class weighting improves recall, more advanced approaches (resampling, threshold optimization) may offer additional benefits.

Recommendations and Next Steps

- Employ feature importance analysis for selection and pruning.
- Experiment with non-linear models (Decision Trees, Random Forests, XGBoost).
- Optimize classification thresholds to meet specific business criteria.
- Implement advanced balancing techniques (SMOTE, ADASYN).
- Manage outliers through robust scaling or winsorization.

Predicting Credit Card Default Using Logistic Regression

By: Shivesh Raj Sahu

Conclusion

This project demonstrates the practical process of building and evaluating a credit default prediction model, with a special focus on class imbalance. The balanced logistic regression model delivers a substantial boost in defaulter recall, a key objective in risk-sensitive financial environments. Reporting both the baseline and improved approaches provides transparency and supports more informed decision-making.

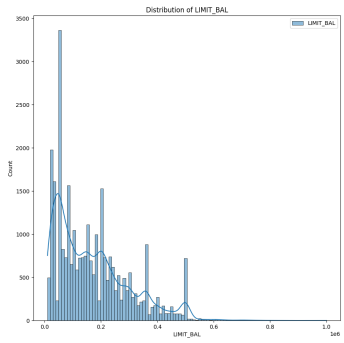
Appendices:

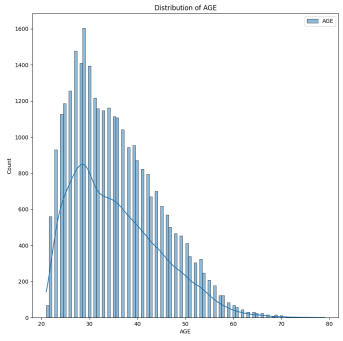
Detailed visualizations (class distributions, ROC curves, heatmaps, boxplots, and histograms).

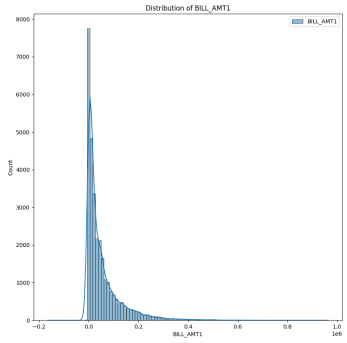
Additional EDA Plots:

Class Distribution Plot

Demonstrates strong class imbalance—only 22% default rate.

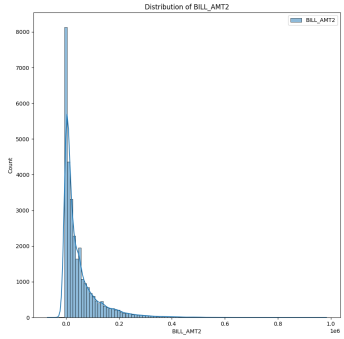


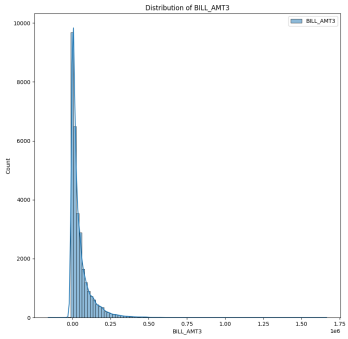


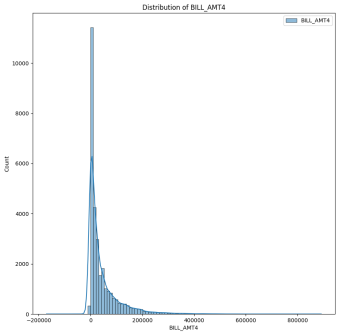


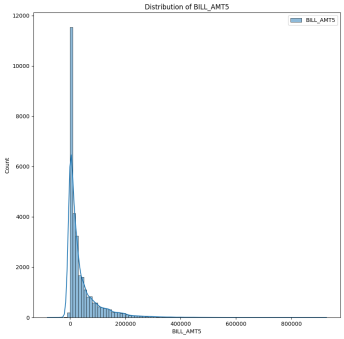
Predicting Credit Card Default Using Logistic Regression

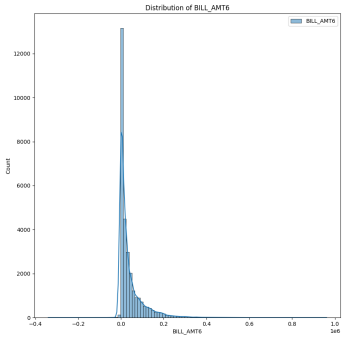
By: Shivesh Raj Sahu

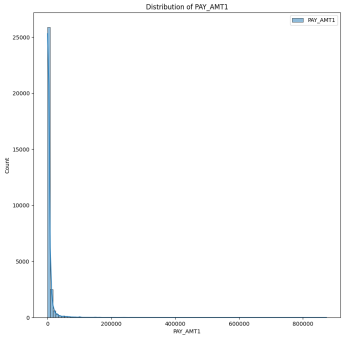


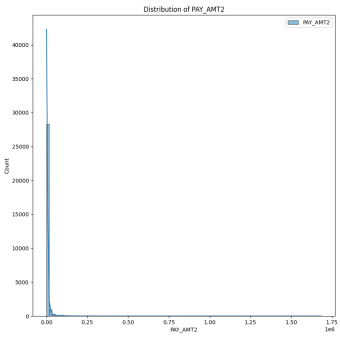


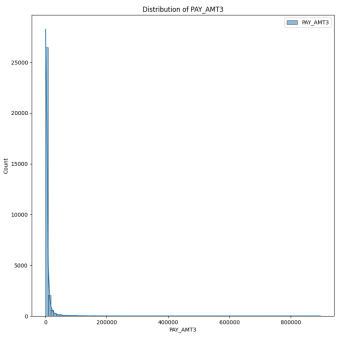


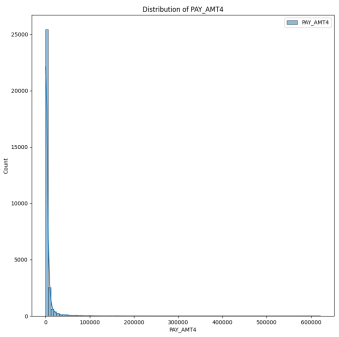


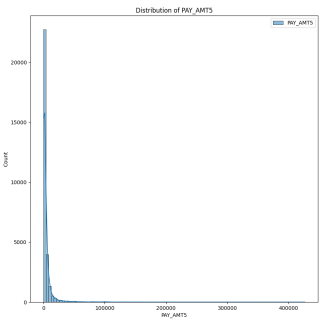










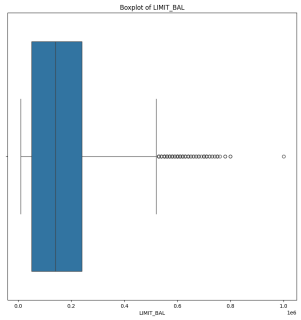


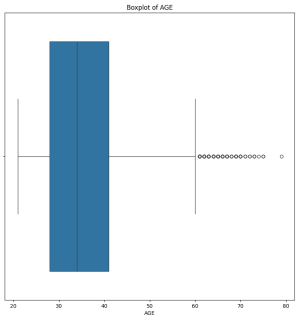
Predicting Credit Card Default Using Logistic Regression

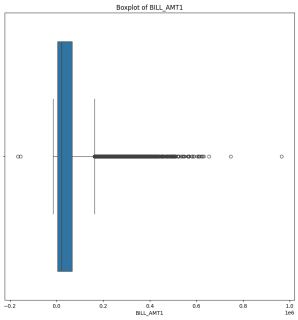
By: Shivesh Raj Sahu

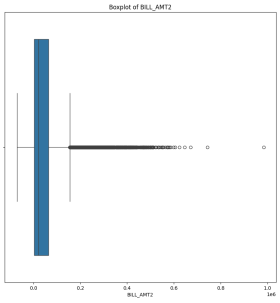
Boxplot of PAY_AMT

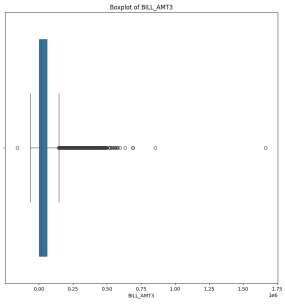
Reveals presence of outliers/skew in payment amount data.

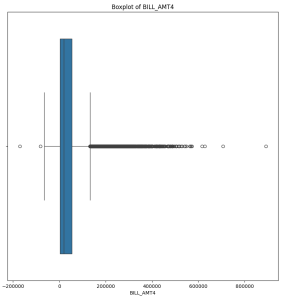


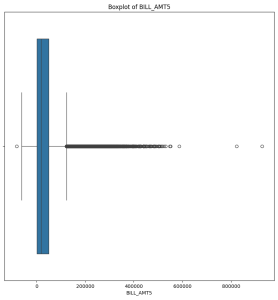


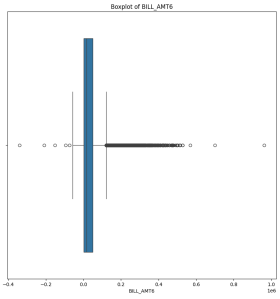


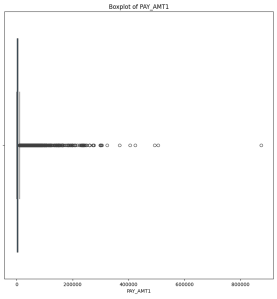


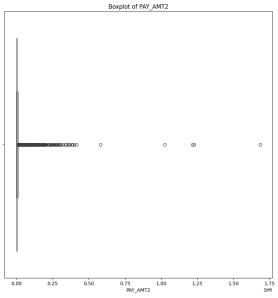


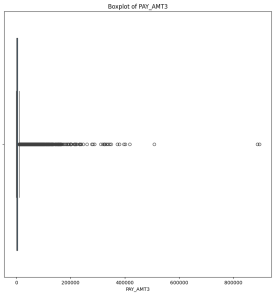


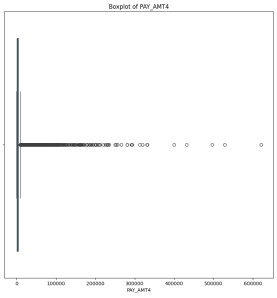












Predicting Credit Card Default Using Logistic Regression

By: Shivesh Raj Sahu