

PREDICATING CREDIT CHARGE-OFFS

A MACHINE LEARNING APPROACH FOR RISK ASSESSMENT

INTRODUCTION



The financial sector faces growing challenges from rising credit risks and fraudulent activities, especially in lending domains like credit cards and personal loans. To address these issues, this study presents a predictive model that generates a Credit Risk Score (0 to 1000) for each customer, identifying high-risk accounts and potential fraud. By analyzing key financial and behavioral patterns, the model enhances decision-making, reduces risks, and ensures regulatory compliance, promoting a secure and reliable lending environment.

METHODS

- Data Cleaning And Missing Value Imputation:**
 - KNN for missing value imputation, and addressed outliers through capping and normalization.
- Count plot and Pairplot:**
 - Understand charge-off distribution and feature relationships with charge_off_status.
 - Findings:
 - 26.1% charge-offs and 73.9% non-charge-offs highlight a class imbalance.
 - fico_score** strongly predicts charge-offs
 - Delinquency_status** correlates with charge-offs
- Correlation Matrix:**
 - Purpose: Analyze numerical feature relationships
 - Findings:
 - Strong negative correlations:(-0.65) fico_score with delinquency_status
 - Weak correlations with age and income_level
- Chi-Square Test:**
 - Purpose: Analyze categorical feature relationships with charge-offs
 - Findings:
 - unusual_submission_pattern** and **payment_methods_high_risk** show strong relationships
- Feature Engineering (Pandas):**
 - Purpose: Create **account_age** and **credit_tenure** features
 - Findings: No significant relationship with charge-off status
- Model Training:**
 - Method: Train-Test Split (80-20 ratio)
 - Algorithms: Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM
 - Best Model: **Logistic Regression** (86% accuracy)
- Model Evaluation:**
 - Tools: Accuracy Score, Classification Report, ROC-AUC Score
 - Results:
 - Logistic Regression:** 86% accuracy
 - Other Models:** 77% accuracy (Decision Tree, Random Forest, XGBoost, LightGBM)

RESULTS

- Accuracy:**
 - Logistic Regression: **86%**
 - Decision Tree, Random Forest, XGBoost, LightGBM: **77%**
 - ROC-AUC Score: **0.82**
- Feature Importance:**
 - fico_score** and **delinquency_status** are key predictors of charge-offs, with **fico_score** showing the highest absolute coefficient in the logistic regression model.
- Risk Score Calculation:**
 - Predicted probabilities were scaled to generate **Risk Scores** (0-1000) for each customer. Higher scores indicate higher risk of charge-off.

86
LOGISTIC
REGRESSION

77
DECISION
TREE

82
ROC-AUC
SCORE

CODING
LIBRARIES



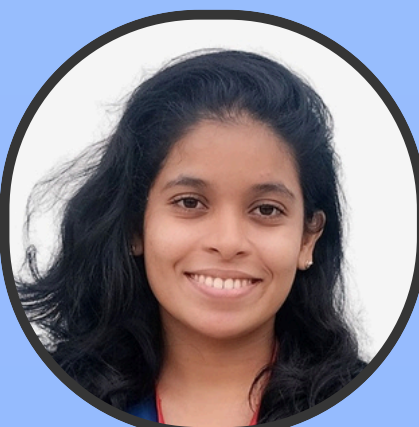
Pandas



CONCLUSION

This analysis highlights critical insights into predicting charge-off statuses, emphasizing key features such as fico_score and delinquency_status. The Logistic Regression model demonstrated robust performance with an accuracy of 86%, showcasing its effectiveness. These findings provide a solid foundation for enhancing decision-making in credit risk management, aiding in proactive strategies to mitigate financial risks.

"PREDICTING CREDIT RISK
AND DEVELOPING A CREDIT
RISK SCORE TO DETECT
FRAUD"



W.T. Naduni Wimalarathna
G.G. Piyumi Imasha
W.A.D Rooma Upenya Perera

PROUD PRODUCT BY,
FACULTY OF IT,
UNIVERSITY OF MORATUWA.