Mortgage Default and Prepayment Model Analysis

Prepared by: Dallen Huang

Table of Contents

---

1. Introduction

The purpose of this analysis is to explore and evaluate machine learning models for predicting mortgage defaults and prepayments. The models were tested to assess predictive performance and determine the optimal approaches for risk classification, clustering, and deep learning applications.

Data Source: https://1drv.ms/u/s!Amfo1lixPzv-j3k_230kv-58OmZW?e=x5WIz8

---

2. Model Summaries

2.1 Default and Prepayment Model Performance

Default Model ROC AUC Scores

- Logistic Regression: 0.657
- Decision Tree: 0.654
- Naive Bayes: 0.649
- Stochastic Gradient Descent: 0.500
- K Nearest Neighborhood: 0.629

- Light Gradient Boosting: 0.667
- XGBoost: 0.654
- Random Forest: 0.630
- Ensemble Learning (Weighted Average): 0.630

Prepayment Model ROC AUC Scores

- Logistic Regression: 0.712
- Decision Tree: 0.698
- Naive Bayes: 0.679
- Stochastic Gradient Descent: 0.505
- K Nearest Neighborhood: 0.653
- Light Gradient Boosting: 0.734
- XGBoost: 0.723
- Random Forest: 0.704
- Ensemble Learning (Weighted Average): 0.710

Model Selection
The Light Gradient Boosting model demonstrated the best predictive performance for both default (0.667) and prepayment (0.734) risks. Boosting and ensemble methods generally outperformed simpler models, capturing complex data patterns more effectively.

2.2 Clustering Analysis Findings

The clustering analysis segmented the mortgage data into 7 distinct clusters representing various risk profiles. This segmentation allows for:

- Targeted resource allocation
- Risk-based loan terms and pricing
- Proactive monitoring for high-risk clusters

These findings aid in optimizing risk management and portfolio segmentation strategies.

2.3 Deep Learning Analysis

Model Performance

- CNN Model: Achieved 88% accuracy and an AUC-ROC of 0.92.
- ANN Model: Achieved 85% accuracy and an AUC-ROC of 0.89.

Model Selection
The CNN model outperformed ANN, with improved predictive accuracy and AUC-ROC. However, the ANN remains valuable where interpretability is prioritized, with CNN better suited for more complex mortgage risk predictions.

---

3. Key Insights and Conclusions

3.1 Model Selection: Light Gradient Boosting provided the best results for both default and prepayment predictions, suggesting that ensemble learning is highly effective for this domain.

3.2 Clustering Benefits: The seven clusters highlight distinct borrower risk profiles, offering actionable insights for tailored loan management.

3.3 Deep Learning Utility: CNN is the optimal model for precision, while ANN is valuable for interpretability. Combining the two could enhance predictive performance further.