Modeling Mortgage Loan Default Prediction: Methods and Techniques

In this study, we applied various machine learning models to predict mortgage loan defaults using a set of features related to both the borrower's financial situation and the equity in the property. The primary models employed were Random Forest, XGBoost, LightGBM, Histogram-Based Gradient Boosting, and Neural Networks (MLP). To address class imbalance, we used the Synthetic Minority Over-sampling Technique (SMOTE), which is a commonly used technique for generating synthetic data points for the minority class in an imbalanced dataset.

Random Forest is an ensemble learning method that constructs a collection of decision trees and combines their predictions. It is robust to overfitting and is well-suited for classification tasks involving a mixture of categorical and continuous data.

XGBoost is a gradient boosting model known for its efficiency and performance. It constructs an ensemble of decision trees where each tree attempts to correct the mistakes of its predecessor. XGBoost is particularly effective in handling imbalanced datasets when tuned correctly.

LightGBM (Light Gradient Boosting Machine) is another gradient boosting method. It is optimized for speed and efficiency in handling large datasets. Like XGBoost, it uses decision trees but optimizes the tree-building process, leading to faster training times.

Histogram-Based Gradient Boosting is a variant of gradient boosting, where the training process uses histogram-based approximations for continuous feature values, allowing for faster training on large datasets.

Neural Networks (MLP) were used for non-linear classification tasks. These networks consist of multiple layers of neurons and are able to model complex relationships within the data.

To address the issue of class imbalance, SMOTE was applied. SMOTE works by generating synthetic samples for the minority class by interpolating between existing minority class instances. This improves the model's ability to predict the minority class and reduces bias towards the majority class.

For each model, we evaluated performance using various metrics, including accuracy, precision, recall, F1-score, and ROC AUC. SMOTE was found to improve performance in certain models, especially Random Forest, where recall for the minority class was significantly enhanced. However, its impact on other models like XGBoost, LightGBM, and Histogram-Based Gradient Boosting was more marginal.

For more detailed insights and the mathematical underpinnings of the models and SMOTE, please refer to the following resources:

- **Random Forest** - Breiman, L. (2001). Random Forests. *Machine Learning, 45*(1), 5-32.
- **XGBoost** - Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.*
- **SMOTE** - Chawla, N. V., et al. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research, 16*, 321-357.
- **LightGBM** - Ke, G., et al. (2017). LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances in Neural Information Processing Systems, 30.*