Enhanced Credit Risk Acquisition Scoring via Noise-Augmented Feature Selection and Bayesian Hyper-parameter Tuning

Abstract

Effective credit risk acquisition scoring is crucial for financial institutions, where gradient-boosted models, such as XGBoost and LightGBM, are increasingly prevalent. However, their performance is contingent upon appropriate feature selection and hyper-parameter tuning. Feature selection identifies pertinent predictors, reducing complexity and enhancing interpretability, while hyper-parameter tuning optimizes model configurations, mitigating overfitting and maximizing predictive accuracy. This study investigates advanced feature selection and hyper-parameter optimization techniques to improve gradient-boosted model performance in credit risk acquisition.

A novel feature selection method, employing random noise variables, was explored. The method involves generating artificial noise features and training the gradient-boosted models on the augmented dataset. By comparing the feature importance of the original features with that of the noise features, less informative features are identified and eliminated. This technique offers an improvement over recursive feature elimination (RFE), a widely used method, by providing a more robust assessment of feature importance through comparison with noise variables, thus aiding in filtering out features that may appear important but do not truly contribute to the model's predictive power. This was tested on an acquisition model development with an initial set of 6000 bureau features. The noise-based technique yielded a more optimal feature subset, achieving a significant dimensionality reduction from over 6,000 features to fewer than 200 in a single step, whereas RFE required significantly more computational time to reach a comparable, yet less refined, outcome. This direct comparison underscores the proposed method's ability to identify truly relevant features, improving model performance, interpretability and efficiency.

Subsequently, Bayesian hyper-parameter tuning, utilizing Optuna, was investigated. Optuna leverages past trial results to intelligently explore the hyper-parameter space, accelerating convergence to optimal configurations, unlike grid search's exhaustive approach. To empirically validate this advantage, a stratified 5-fold cross-validation scheme was implemented, maximizing average AUC-ROC, with overfitting controlled by cross-validation standard deviation and a custom overfit index. Results confirmed Optuna's efficacy in rapidly identifying optimal hyper-parameters, supporting its superiority over grid search.

The paper will show how the combined application of the proposed robust feature selection method and Bayesian hyper-parameter optimization enhances predictive accuracy, robustness, and computational efficiency of gradient-boosted models in credit risk acquisition scoring.

Authors & Affiliations

Mr Arijit Ganguly[1]
[1]Revolut Group Holdings Ltd, Bangalore, India