Abstract

In this study, we explore the Home Credit Default Risk dataset to predict loan repayment capabilities among underbanked individuals. Utilizing alternative data sources, such as telecommunications and transaction records, we apply advanced machine learning techniques, including Random Forest and Linear Regression classifiers. These methods are enhanced through rigorous hyper-parameter tuning using Grid Search and Bayesian Optimization. Our approach also includes extensive feature engineering to enrich the dataset, focusing on the creation of interaction variables and the integration of external data to augment the models' predictive power. Additionally, we assess the performance of these models by calculating and comparing the log loss, providing a quantitative measure of model accuracy and confidence in predictions. This is visualized through comparative plots, which highlight the relative effectiveness of each model in our predictive framework. This research aims not only to improve prediction accuracy but also to promote financial inclusion by providing insights into how machine learning can be effectively optimized to assess creditworthiness.

-Mohammed