Loan Default Prediction Project | ML Model Results (Milestone 5)

Executive Summary Report

Prepared by: PhD Aleksandar Osmanli

⟩ ISSUE / PROBLEM

I'm currently developing a data analytics project aimed at decreasing overall loan default among borrowers by predicting which client attributes most contribute their loan default. For the purposes of this project, loan default means borrowers failed to pay their loans to the financial institution. The ultimate goal for this project is to develop a machine learning (ML) model that predicts loan default. **This report offers details and key insights from Milestone 5, which could impact the future development of the project, should further work be undertaken.**

⟩ IMPACT

→ **The ML models developed for Milestone 5 demonstrate a critical need for additional data in order to more accurately predict loan default.**

→ **This modeling effort confirms that the current data is insufficient to consistently predict churn. It would be helpful to have additional historical information for each client.**

→ **Since engineered features are a proven valuable tool for improving the performance of ML models, it is recommended more iterations of the Loan Default Prediction Project.**

⟩ RESPONSE

● **To obtain a model with the highest predictive power, four different tree-based models are developed to cross-compare results: LightGBM, XGBoost, CatBoost and HistGBM. Also, SVEnsemble and STEnsemble models were used to combine the predictive power of the individual models.**

● To prepare for this work, the data was split into training, validation, and test sets. Splitting the data three ways means that first the results from the training and the validation sets were compared, and then the model was trained on the test set. **The champion model had the best ROC_AUC score on the test set after optimizing results on the training and on the validation sets.**

⟩ KEY INSIGHTS

Model	ROC_AUC Score on the training set	ROC_AUC Score on the validation set	ROC_AUC Score on the test set
LightGBM	0,6949	0,7558	0,7577
CatBoost	0,6891	0,7552	0,7567

● **After extensive testing, the LightGBM model ouperformed the CatBoost model on the training, validation and on the test sets.**