Video Streaming Churn 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 increasing overall growth by preventing monthly user churn on the video streaming platform. For the purposes of this project, churn quantifies the number of users who have canceled the subscription with the video streaming service. The ultimate goal for this project is to develop a machine learning (ML) model that predicts user churn. **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 user churn.**
- ➔ **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 user engagement.**
- ➔ **Since engineered features are a proven valuable tool for improving the performance of ML models, it is recommended more iterations of the User Churn 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.**

- To prepare for this work, the data was split into training, validation, and test sets. Splitting the data three ways means that there is less data available to train the model than splitting just two ways. However, **performing model selection on a separate validation set enables testing of the champion model by itself on the test set, which gives a better estimate of future performance than splitting the data two ways and selecting a champion model by performance on the test data.**

KEY INSIGHTS

Model	ROC_AUC Score on the validation set	ROC_AUC Score on the test set
LightGBM	0,7468	0,747
XGB	0,7484	0,6989
CatBoost	0,7477	0,7487
HGBM	0,7354	0,7475
SVEnsemble	0,7481	0,7019
STEnsemble	0,7481	0,7022

- **After extensive testing, even the XGBoost model was the "champion" model on the validation set, the CatBoost was the "champion" model on the test set.**

- **On the final run, the models were tested on extensive feature set of 40 features to capture the strongest predictive power as possible on the user churn**.