Submission Model Documentation

Name: Guilherme Righetto **Location**: Brazil **Email**: guilhermerighetto02@gmail.com
Competition: Predict Future Sales **Background**: I work as a data analyst in the fraud detection area.

Summary

The work was developed in four parts. The first step was the analysis of the features, data leakage and the creation of new features. The second step was the execution of the feature selection using shap values. The third stage was the tuning of the hyperparameters of each regressor using Bayesian optimization. Finally, the last step was used stacking ensemble using the LGBMRegressor, CatBoostRegressor and SGDRegressor algorithms on the first level and the SGDRegressor algorithm on the second level.

Features Selection/Engineering

Several features were created mainly using the features categories of the dataset, in which the mean enconding technique was used and also features based on time (lag). While to select the best features, shap values were used using a tree-based algorithm.

Training Methods

In this work the staking ensemble method was used using the out of fold technique, using 3 folds. The first level models were LGBMRegressor, CatBoostRegressor and SGDRegressor. And the second level model using the meta-features was the SGDRegressor.

Dependencies

All libraries used in this work are in the file requirements.txt.

References

How to win Kaggle Competitions - Coursera