A Study on Gradient Boosting Classifiers

Juliano Garcia de Oliveira
Advisor: Prof. Dr. Roberto Hirata Jr.

Department of Computer Science – Institute of Mathematics and Statistics
University of São Paulo

Introduction

Gradient Boosting is a machine learning technique widely used due to its efficiency, accuracy, and interpretability that has been achieving state-of-the-art results in a range of machine learning challenges. A common step when building machine learning models is to choose the *hyperparameters* of the model, i.e. the values the algorithm doesn't directly learn from the data. Moreover, the most used implementations of gradient boosting – XGBoost and LightGBM libraries – have several hyperparameter options, which can take a long time to optimize in the typical hyperparameter tuning procedure.
To expand on the knowledge of how hyperparameters impact machine learning model performance, in this work a large-scale experiment with 70 datasets from the OpenML platform is conducted. The experiment takes into account the datasets' characteristics, LightGBM classification models, a set of hyperparameters and three evaluation metrics for classification tasks. In classification, the objective is to correctly predict the labels of unlabeled data.

Gradient Boosting Machines

GBM sequentially construct *weak learners* to form an additive ensemble model. The LightGBM binary classifier uses decision trees as the weak learners, and in the boosting process, the gradient of the logarithmic loss is used to improve the model at each iteration step. Given a dataset $\boldsymbol{X}$, the GBM model takes the form of a composition of tree models $\boldsymbol{h(X;\gamma)}$:

$$\boldsymbol{F_M(X)} = \sum_{\boldsymbol{m=1}}^{\boldsymbol{M}} \beta_{\boldsymbol{m}} \boldsymbol{h(X;\gamma_m)}$$

$\boldsymbol{M}$ denotes the number of estimators, while $\beta_{\boldsymbol{m}}$ and $\gamma_{\boldsymbol{m}}$ are the weight and parameters of the $\boldsymbol{m}$th tree.

Study Methodology

Using the OpenML API, 70 datasets of binary classification tasks were used in the training of multiple LightGBM models, with different hyperparameter combinations. The **max_depth**, **num_estimators** and **learning_rate** were analyzed, and the models were evaluated in their respective test set using AUC, Logloss and Brier Score metrics.

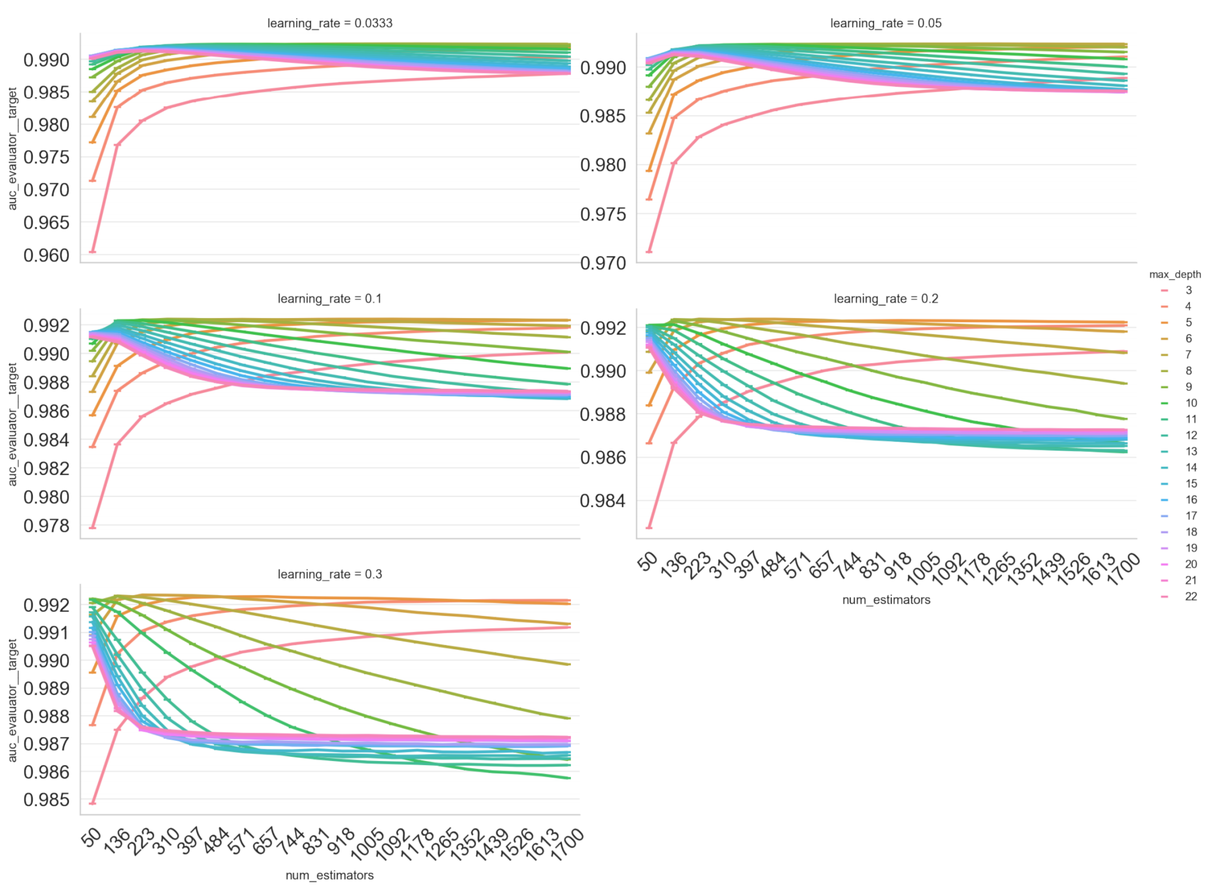

Fig. 1: AUC results for a single dataset experiment. The datasets were clustered together using K-means based on the descriptive statistics of each dataset, like the ratio of categorical features, mean skewness of numerical columns, cardinality, etc.
To measure the sensitivity of the LightGBM models to each hyperparameter value, a $\boldsymbol{\delta_{metric}}$ was defined as the difference of ***metric*** (AUC, Logloss or Brier Score) obtained with a new hyperparameter value from the ***metric*** obtained with the default value of the hyperparameter as defined in the library.

Analysis of Variance

Analysis of variance was applied to the results of the experiments, considering the hyperparameter values as different treatment levels and the $\boldsymbol{\delta_{metric}}$ the observed outputs. Since most of the results failed to meet the assumptions of ANOVA, the nonparametric **Kruskal-Wallis test** was used.

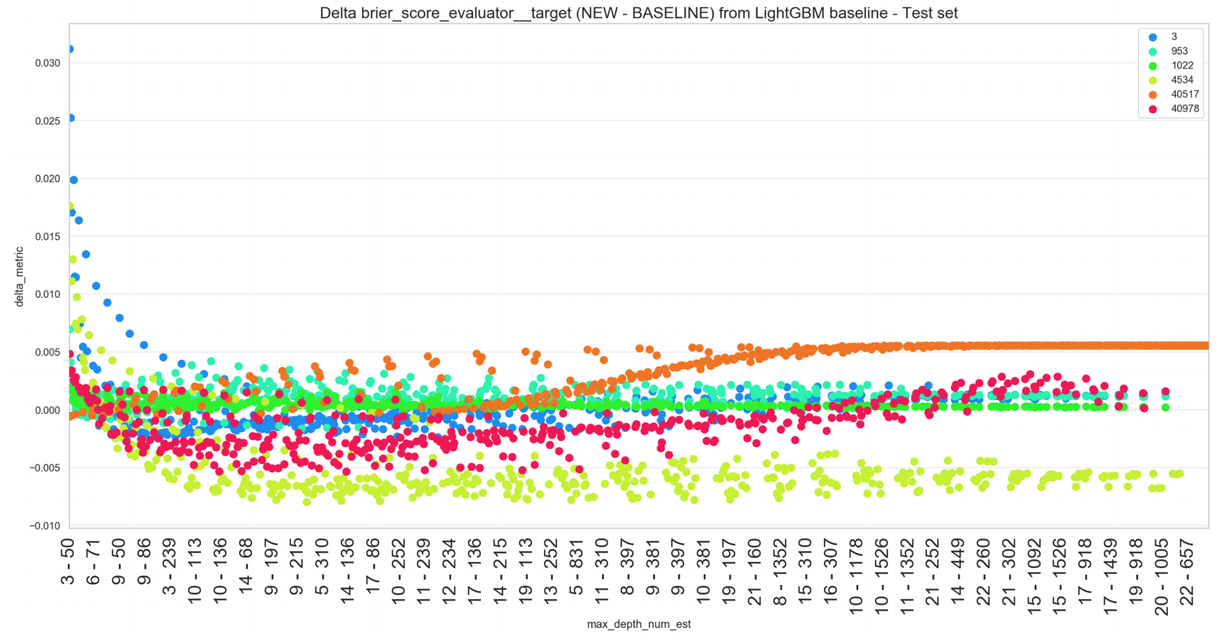

Fig. 2: Observed δ_{Brier} when changing max_depth and num_estimators in Cluster 3. Each statistically significant experimental result was modeled as a single-factor fixed effects model, where each observation $\boldsymbol{y_{ij}}$ is decomposed in the $\boldsymbol{i}$th treatment effect $\tau_{\boldsymbol{i}}$, the overall experiment mean μ and an error term $\epsilon_{\boldsymbol{ij}}$:

$$\boldsymbol{y_{ij}} = \mu + \tau_{\boldsymbol{i}} + \epsilon_{\boldsymbol{ij}}$$

Results and Single-Factor Models

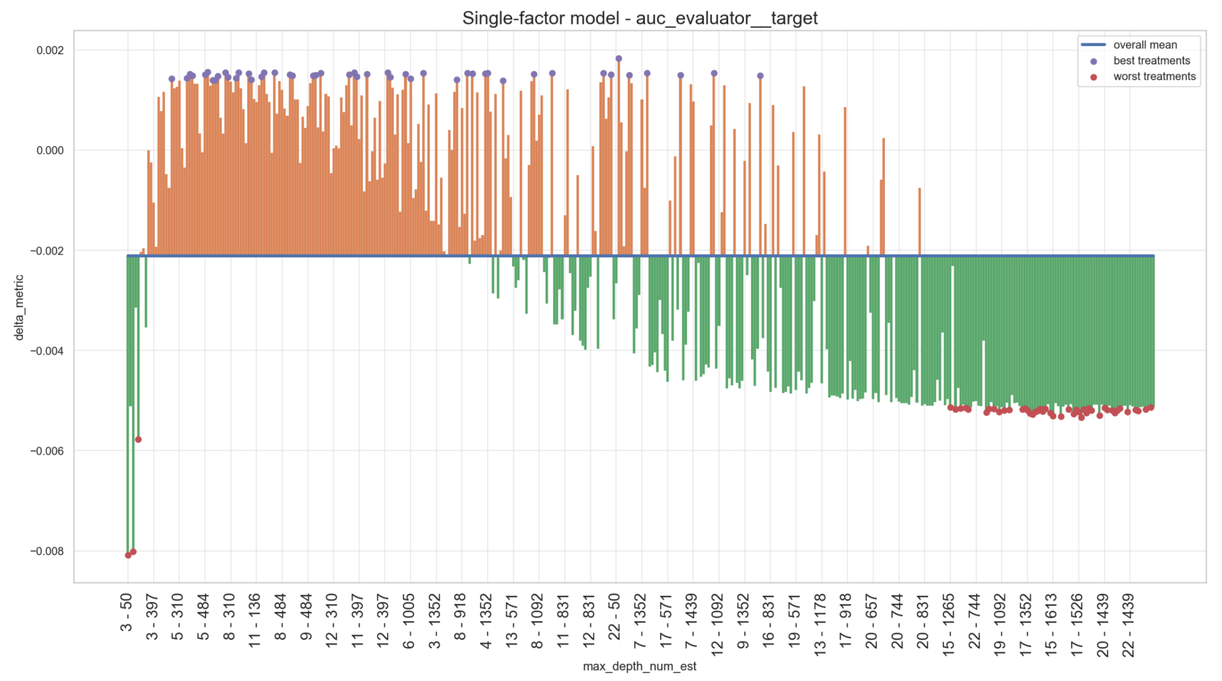

Fig. 3: Single-Factor model example: each vertical line is an estimated treatment effect. The results can be summarized into three categories: results by hyperparameter, results by dataset characteristics and by performance metric. The results display different insightful behaviors of the hyperparameter effects, like how relatively small values of maximum depth and estimators impact negatively on the AUC, as exemplified in Fig. 3.

Conclusion

The experiment provided enough results to observe how each combination of hyperparameter affected the metrics in the classification models. Using a solid analysis of variance statistical framework, the experimental results of this work also take into account the statistical significance of each experimental scenario. It was found that some combinations of hyperparameters highly impacted the machine learning models, while others had a small impact.
This work can help machine learning practitioners to quickly build baseline classification models in LightGBM, by analyzing characteristics of the dataset and the hyperparameter impacts reported in the study. Furthermore, this work could easily be extended to different hyperparameters, metrics, and algorithms.

juliano.garcia.oliveira@usp.br