A STUDY ON GRADIENT BOOSTING CLASSIFIERS

A large-scale experimental analysis of hyperparameter effect on binary classification models

Juliano Garcia de Oliveira

Advisor: Prof. Dr. Roberto Hirata Jr.

Motivation and Objectives

In the model building process hyperparameter tuning can take a long time, even with the available optimization procedures.

Using LightGBM algorithm, this study objectives are:

- How hyperparameters affect the performance?
- How different characteristics of a dataset affect the hyperparameter impact?
- How the performance metrics of classification metrics behave?

1.

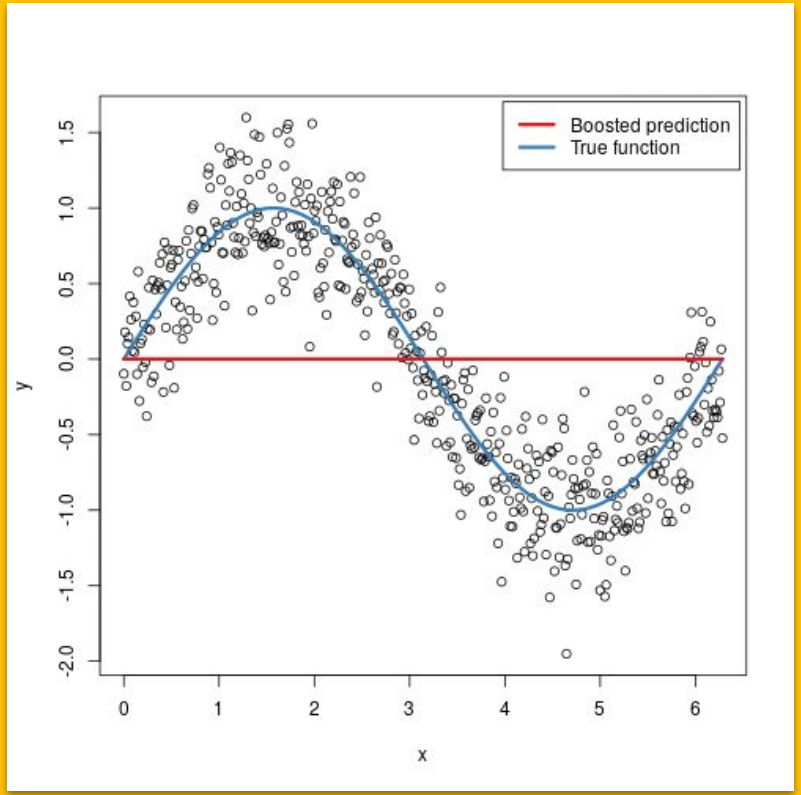
GRADIENT BOOSTING MACHINES

WHAT IS GBM?

- Additive ensemble model
- Multiple *estimators* (shallow trees)
- Sequential procedure: each new learner corrects the last one:

$$F_m(\boldsymbol{X}) = F_{m-1}(\boldsymbol{X}) + \eta\Delta_m(X)$$

- XGBoost and LightGBM



GBM HYPERPARAMETERS

Three LightGBM hyperparameters considered:

- **`num_estimators`** — the total number of boosting iterations, i.e. the total number of trees.
- **`max_depth`** — maximum depth each estimator can have;
- **`learning_rate`** — the *weight* of each new estimator;

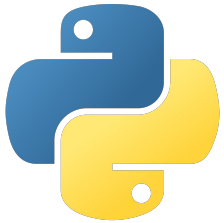
2.

STUDY STRUCTURE

TOOLS AND DATASETS

DATASETS

- OpenML Platform;
- Binary classification;
- Filters for consistency (e.g. minimum of 1000 samples);
- 70 datasets.



python™

fklearn

Microsoft LightGBM

scikit learn

pingouin

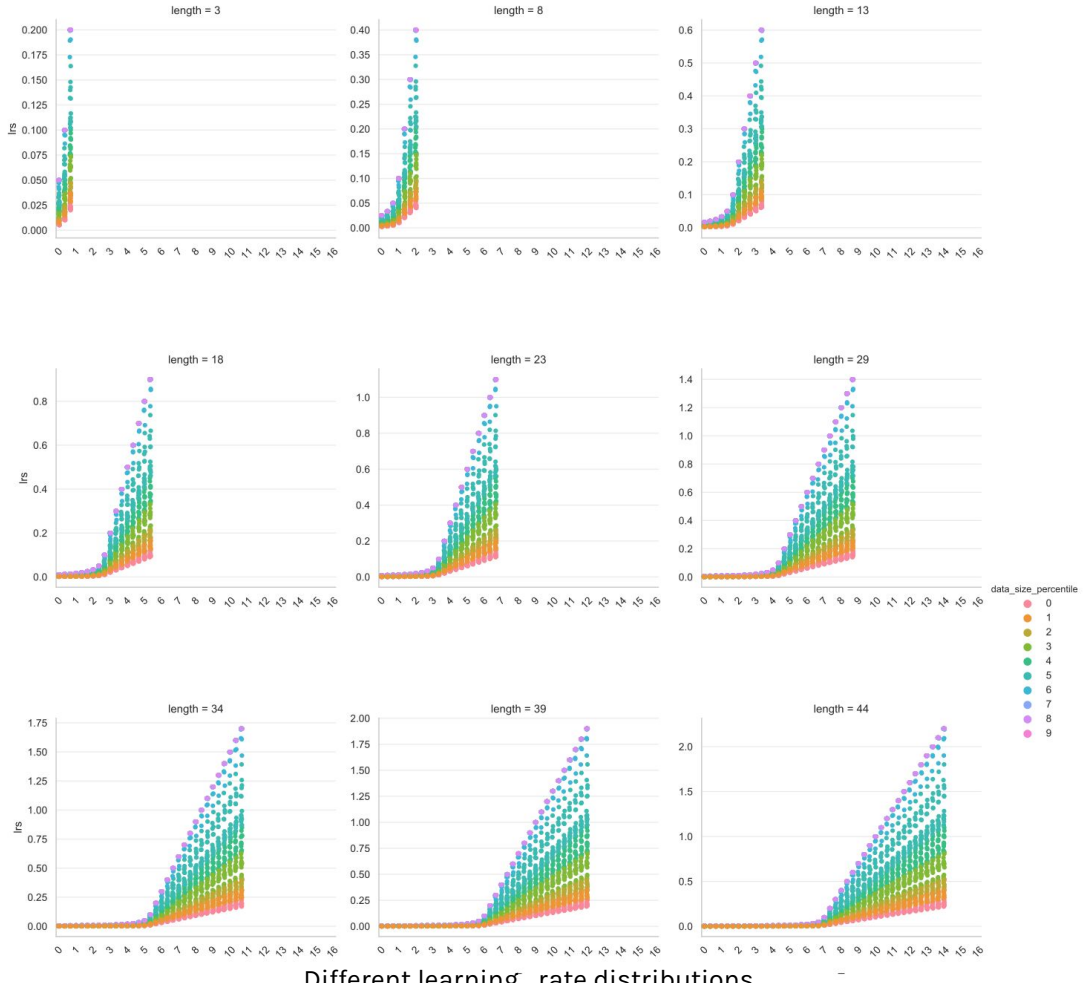
SciPy

DATASET'S DESCRIPTIVE STATISTICS

- To categorize similar datasets, some specific characteristics were calculated for each one.
- These characteristics include the number of total features, categorical features, cardinality, skewness, etc.
- After the experiment, similar datasets were analyzed together according to their characteristics.

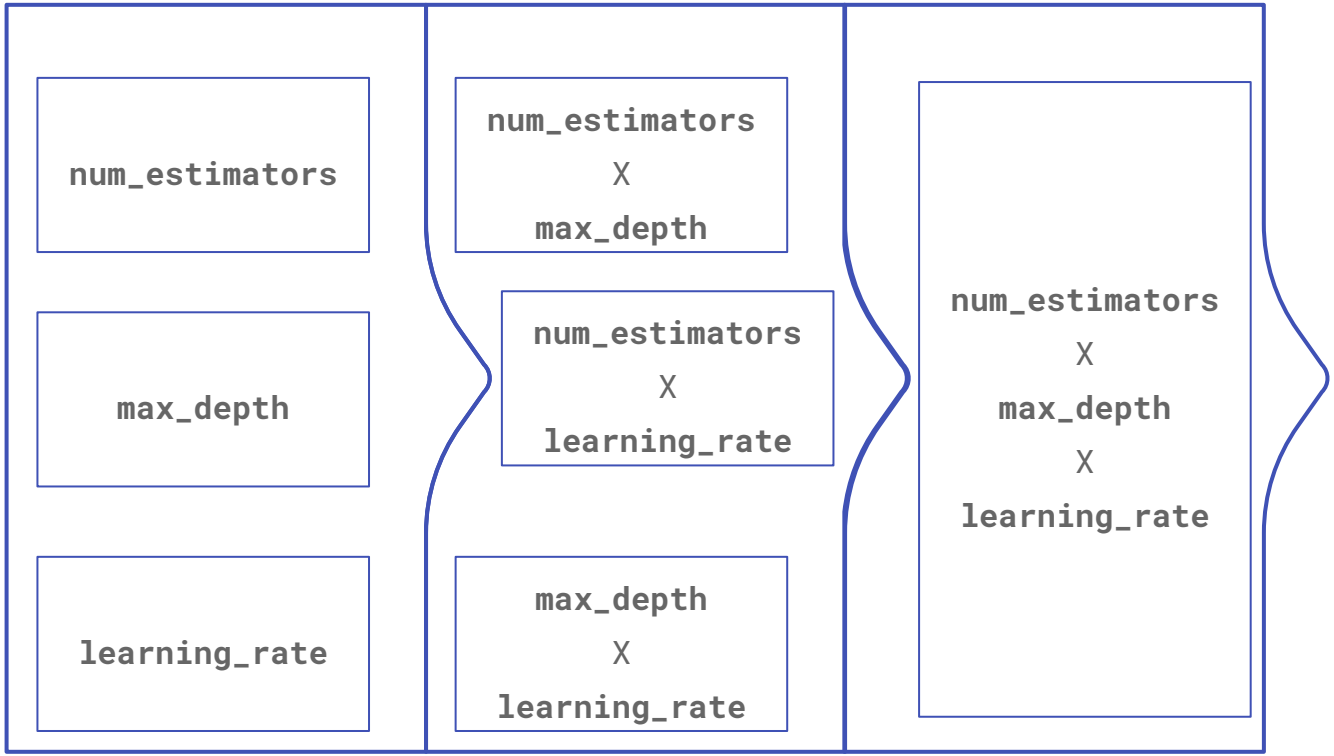
HYPERPARAMETER SPACE

- Hyperparameters values depend on the dataset size;
- Specific rules to generate hyperparameter values;
- Each one has a set of values that will be tested in the experiment.



Different learning_rate distributions

HYPERPARAMETER SPACE



MODEL PERFORMANCE METRICS

- **AUC** — Area Under the ROC curve, measures the model ordering capacity;
- **Logloss** — Logarithmic Loss, measures the probability accuracy;

$$Logloss = -\frac{1}{N}\sum_{i=1}^{n}[y^{(i)}\log\hat{y}^{(i)} + (1-y^{(i)})\log(1-\hat{y}^{(i)})]$$

- **Brier Score** — Mean Squared difference between the predictions and actual labels:

$$Brier = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}^{(i)} - y^{(i)})^2$$

FINAL STUDY PIPELINE

For each dataset:

Calculate descriptive statistics

Generate hyperparameter space

Split train and test set (70/30)

Train all LightGBM models

3.

EXPERIMENTAL ANALYSIS

DATASETS CLUSTERING

- Each dataset experiment results has multiple metrics;
- The experiments were aggregated with **K-means** using the characteristics of its features and the descriptive statistics:
  - `Num_rows, num_features, mean_skewness, mean_variance, num_categorical, sum_cardinality_over_categorical, categorical_ratio, numeric_ratio, boolean_ratio, constant_ratio`

DATASETS CLUSTERING

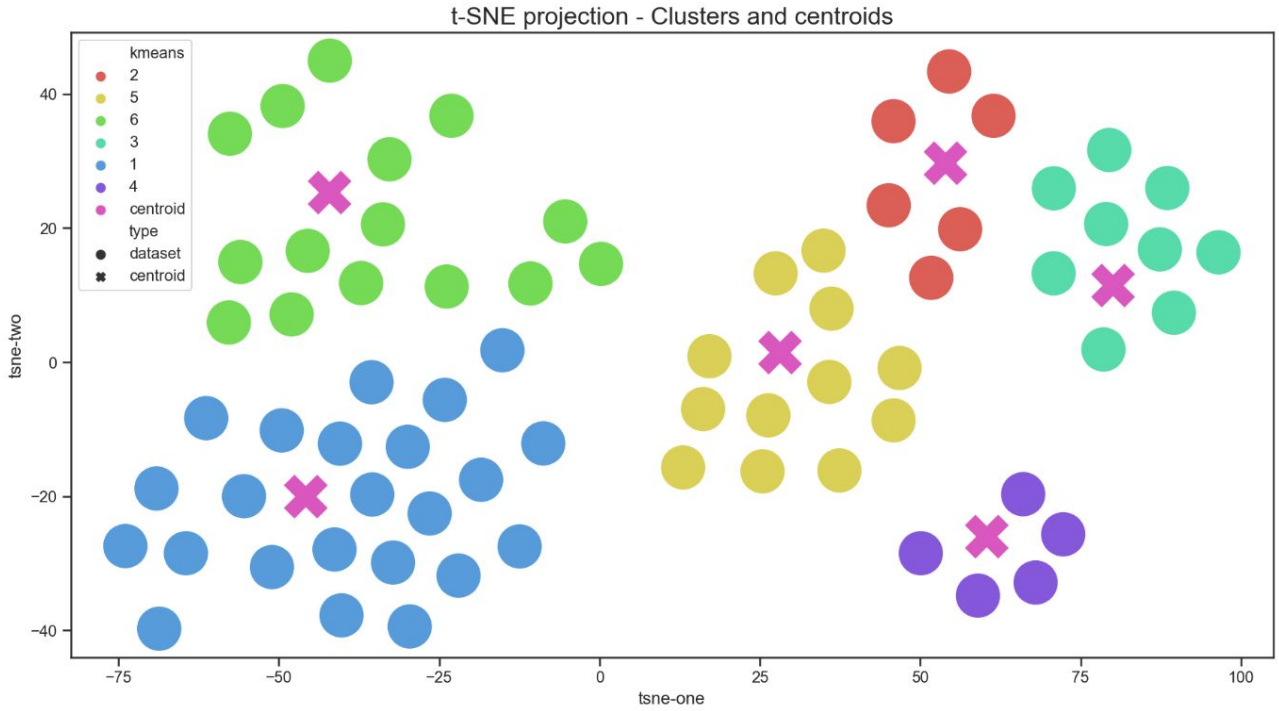
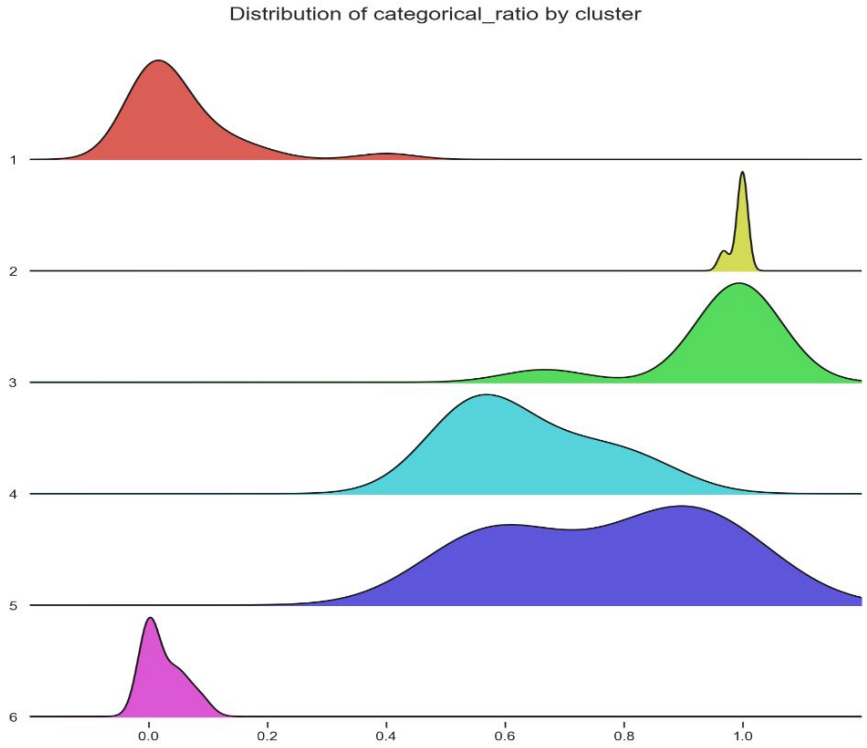


Figure 5.11: *t-SNE projection with the assigned clusters and the centroids*

DATASETS CLUSTERING



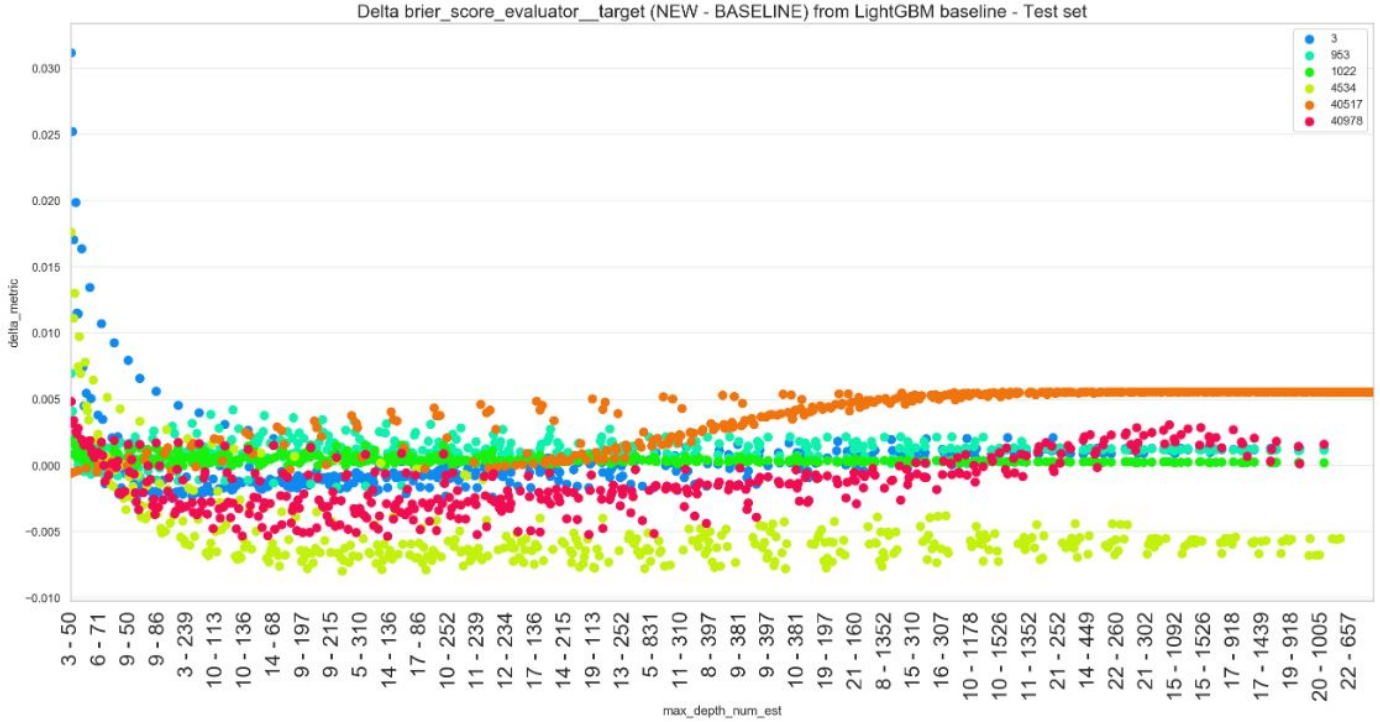
STATISTICAL ANALYSIS

- To measure hyperparameter sensitivity, the performance metrics were converted to a relative change from baseline metric;
- In experimental analysis terminology:
  - The hyperparameter values are the treatment levels;
  - The metrics are the observed outcomes;
  - Each dataset is an experimental unit;

$$\mathcal{S}(C_k, \eta_Q^{(k)}, m)$$

- Nonparametric analysis of variance;

STATISTICAL ANALYSIS



4.

RESULTS AND CONCLUSION

SINGLE-FACTOR MODELS

- A **Kruskal-Wallis** one-way analysis of variance test was applied to every experiment;
- Statistically significant experiments were used to calculate a single-factor effects model:

$$y_{ij} = \boldsymbol{\mu} + \tau_i + \epsilon_{ij}$$

- The treatment effects are interpreted as the effect a single hyperparameter value has on a metric;

Cluster	Metric	η_{NE}	η_{MD}	η_{LR}	$\eta_{MD,LR}$	$\eta_{MD,NE}$	$\eta_{LR,NE}$	$\eta_{NE,MD,LR}$
1	δ_{AUC}	×	×	✓	✓	×	×	✓
	δ_{Brier}	×	✓	×	✓	×	✓	✓
	$\delta_{Logloss}$	✓	✓	×	×	✓	✓	✓
2	δ_{AUC}	×	✓	×	✓	×	✓	✓
	δ_{Brier}	×	✓	×	✓	×	✓	✓
	$\delta_{Logloss}$	✓	✓	×	✓	✓	✓	✓
3	δ_{AUC}	×	×	✓	✓	×	×	×
	δ_{Brier}	×	×	×	✓	×	×	×
	$\delta_{Logloss}$	×	×	×	×	×	×	×
4	δ_{AUC}	×	×	×	×	×	×	×
	δ_{Brier}	×	×	×	✓	×	✓	✓
	$\delta_{Logloss}$	×	×	×	✓	×	✓	✓
5	δ_{AUC}	×	×	✓	✓	×	✓	✓
	δ_{Brier}	×	✓	✓	✓	×	✓	✓
	$\delta_{Logloss}$	×	✓	✓	✓	×	✓	✓
6	δ_{AUC}	✓	×	×	✓	×	×	✓
	δ_{Brier}	×	×	×	✓	×	×	×
	$\delta_{Logloss}$	×	×	×	✓	×	×	×

Statistical test results for all experimental scenarios

EFFECT BY HYPERPARAMETER COMBINATION

Combinations	Metric		
	δ_{AUC}	δ_{Brier}	$\delta_{Logloss}$
Individual	22.2%	22.2%	38.8%
Pair	38.8%	55.5%	50%
Triple	66.6%	66.6%	66.6%

Table 6.2: *Percentage of statistically significant results for each comparison and metric*

EFFECT BY HYPERPARAMETER

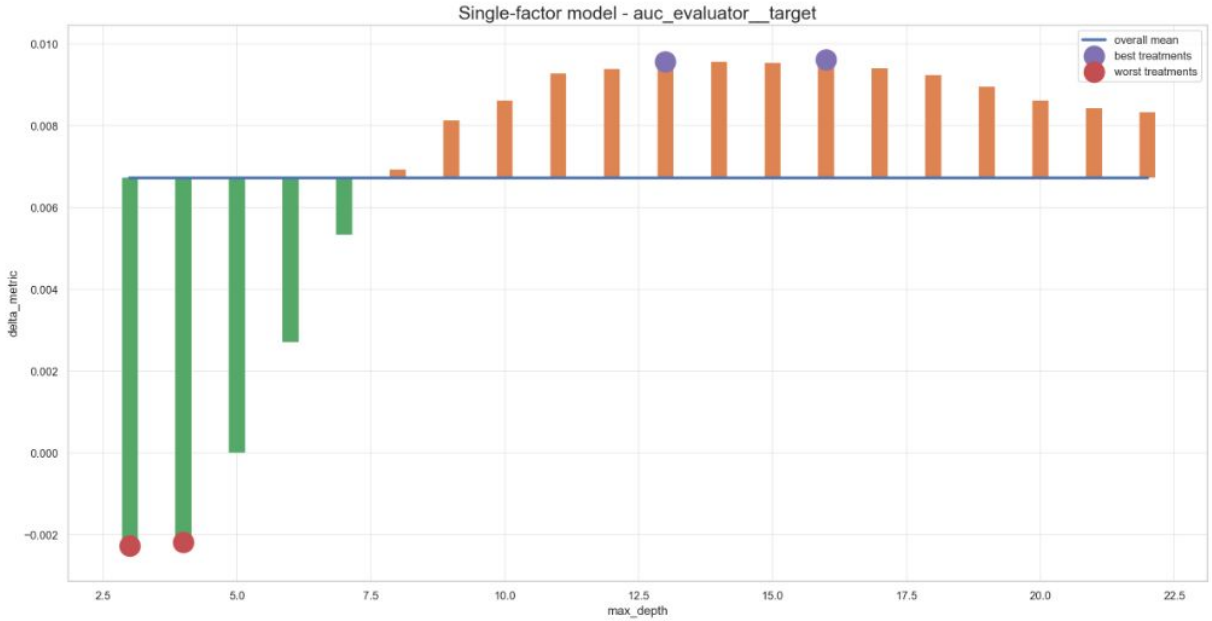


Figure 6.6: *SFM plot for* $\mathcal{S}(C_3, \eta^{(3)}_{MD}, AUC)$

EFFECT BY HYPERPARAMETER

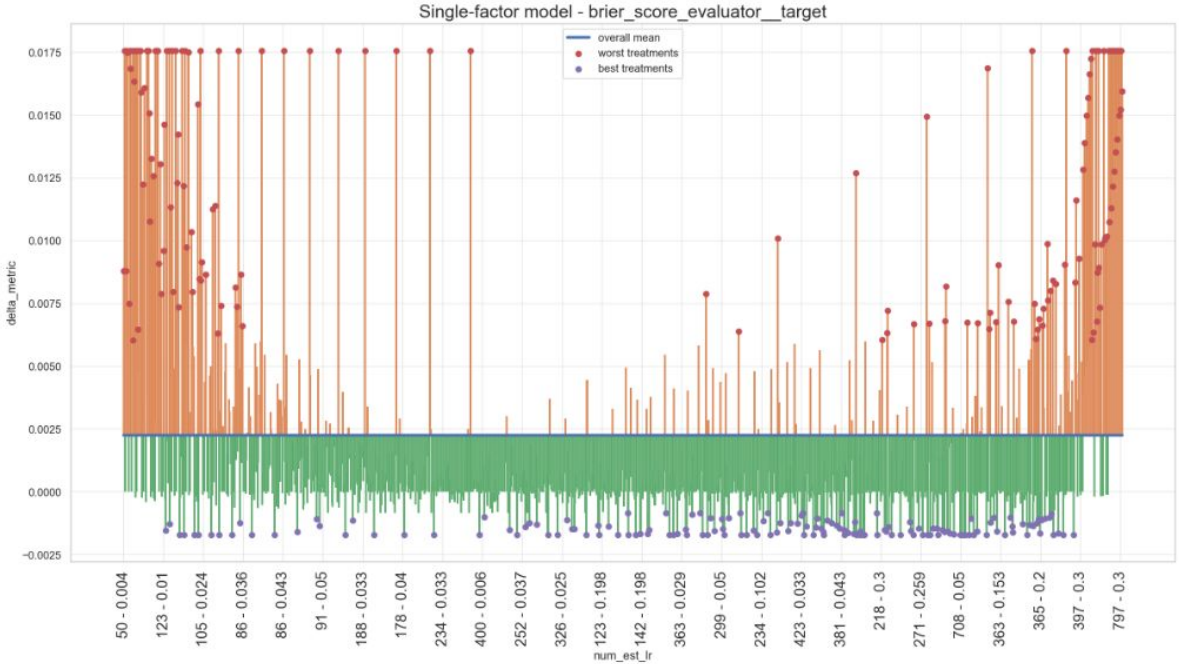


Figure 6.13: *SFM plot for* $\mathcal{S}(C_1, \eta^{(1)}_{LR,NE}, Brier)$

EFFECT BY CLUSTER AND METRICS

Metric	Cluster 1	2	3	4	5	6
δ_{AUC}	42.8%	57.1%	28.5%	0.0%	57.1%	42.8%
δ_{Brier}	57.1%	57.1%	14.2%	42.8%	71.4%	14.2%
$\delta_{Logloss}$	71.4%	85.7%	0.0%	42.8%	71.4%	14.2%
Overall	57.1%	66.6%	14.2%	28.5%	66.6%	23.8%

Table 6.3: *Percentage of statistically significant results in each cluster, by metric*

δ_{AUC}	δ_{Brier}	$\delta_{Logloss}$
38.1%	42.7%	47.6%

Table 6.4: *Percentage of statistically significant results of each metric*

Thanks!

The full thesis with all details and results is available at:
https://linux.ime.usp.br/~robotenique/mac0499/full_tcc.pdf