Report: Predict Bike Sharing Demand with AutoGluon Solution

Nandini Mundhra

Initial Training

What did you realize when you tried to submit your predictions? What changes were needed to the output of the predictor to submit your results?

Initially, I had to fix some syntax errors and review the documentation in order to run the model as intended. I had some trouble starting the tasks on my local environment and I decided to do everything in the SageMaker studio and managed to improve my results and get a score of below 0.5.
CHANGES ADDED-> Kaggle refuses the submissions containing negative predictions values obtained from the predictor. Therefore, all such negative outputs from respective predictors are replaced with 0.

What was the top ranked model that performed?

My top score was achieved by the completing the third run with optimization of hyperparameters. The score was 0.48912 and was better than the previous two runs.

Exploratory data analysis and feature creation

What did the exploratory analysis find and how did you add additional features?

I managed to derive from the distribution that the temperature categories are normally distributed. For the extra features I divided the datetime in month, day, year and hour. Also it was useful to transform the season and weather features to categorical

How much better did your model preform after adding additional features and why do you think that is?

Additional features can be good predictors to estimate the target value. So, It will be good to separate the date because it helps the model to analyze seasonality patterns in the data which can be useful for a regression model. The best improvement is because of the split of the datetime field into year, month, day and hour.

Hyper parameter tuning

How much better did your model preform after trying different hyper parameters?

Model demonstrated notable enhancement compared to the original version but fell slightly short of the performance achieved with just the features. Despite following Autogluon's recommendations for hyperparameter improvement in tabular data, I encountered difficulties in further tuning the parameters to achieve better results. .

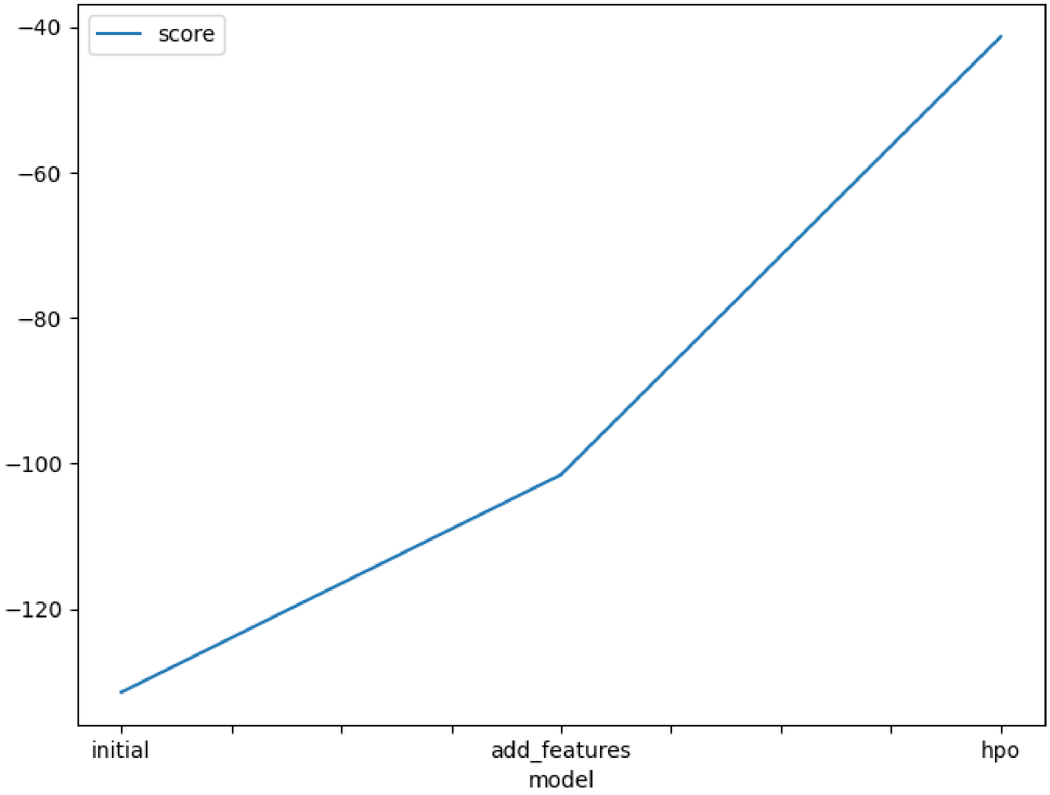
If you were given more time with this dataset, where do you think you would spend more time?

I will do more extensive data analysis in order to get more information about this dataset and also do more research about the hyperparameters. I will also try to figure out more ways to improve on the features as they seem to be the ones that lead to improvement of the score.

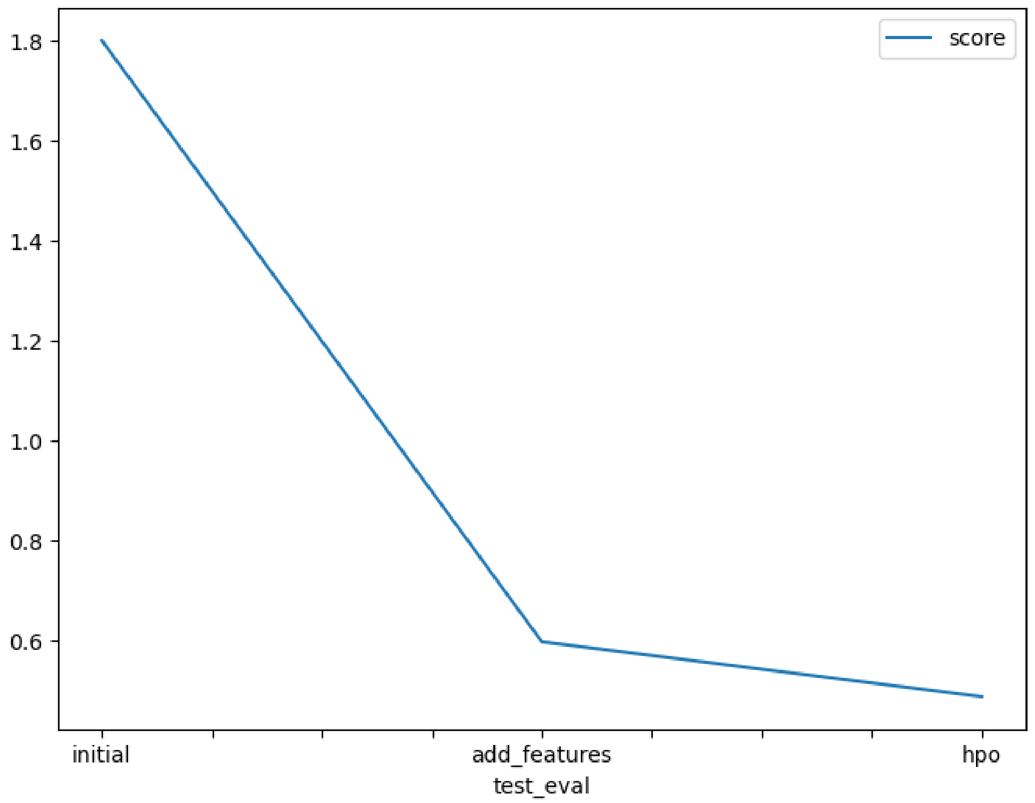
Create a table with the models you ran, the hyperparameters modified, and the kaggle score.

model	hpo1	hpo2	hpo3	score
initial	time_limit=600	presets=best_quality	none	1.80059
add_features	time_limit=600	presets=best_quality	problem_type=regression	0.59878
hpo	time_limit=600	presets=best_quality	hp-method: tabular autogluon	0.48912

Create a line plot showing the top model score for the three (or more) training runs during the project.



Create a line plot showing the top kaggle score for the three (or more) prediction submissions during the project.



Summary

- The AutoGluon AutoML framework for Tabular Data was thoroughly studied and incorporated into this bike sharing demand prediction project.
- The top-ranked AutoGluon-based model improved results significantly by utilizing data obtained after extensive exploratory data analysis (EDA) and feature engineering without hyperparameter optimization.

The most benefit is gain by working with the features and one can gain great insights from the EDA