Report: Predict Bike Sharing Demand with AutoGluon Solution

Samin Poudel

## 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?

Kaggle is not accepting the submission if any predicted values are negative. Therefore, I converted the negatively predicted values to zero and then submitteed the results.

What was the top ranked model that performed?

Weighted Ensemble L3 performed best among the 19 trained models in AutoGluon.

## Exploratory data analysis and feature creation

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

The exploratory analysis of the training data based on histogram revealed that the count of bike sharing is about twice during the working days compared to that of non-working days. And, it is also seen that the count of bike sharing has inverse relationship to the windspeed. Also, we found that weather has significant affect on the bike sharing counts.

The additional features were added by separating the data in the datetime columns to month, day and hour columns.

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

The addition of additinal features improved the performance of the model signficantly. The assumption behind the imporvement in the performance is that the separated columns month, day and hour has signficant correlation with the bike sharing count.

## Hyper parameter tuning

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

The performance of the model was slightly improved as a result of trying different hyper parameters.

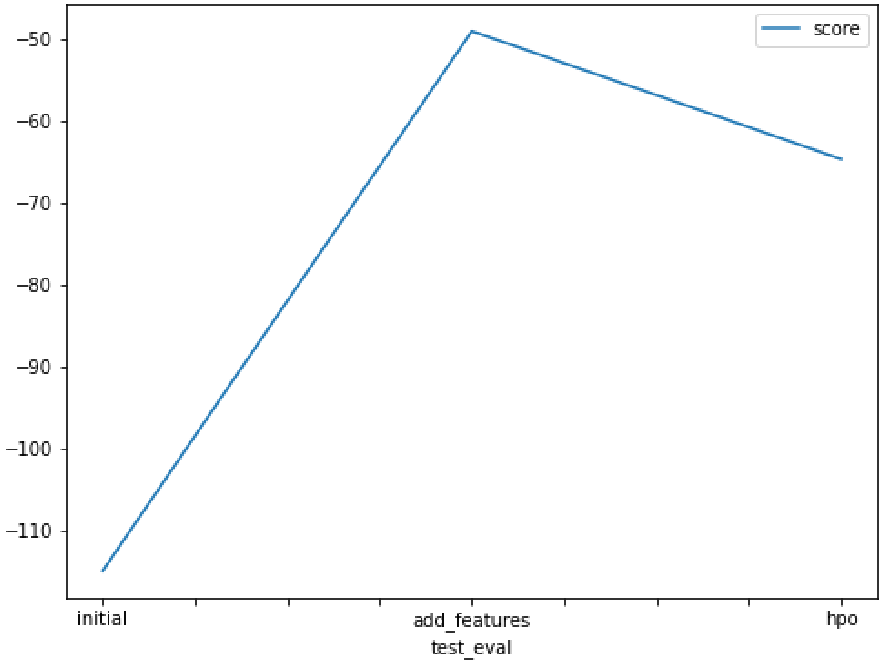
If you were given more time with this dataset, where do you think you would spend more time?

I would have spend time with exploratory data analysis to find the redundant features that are not useful to train the models and drop them during training. I would have also applied other machine learning regression algorithms and compared the results with the ones from neural networks.

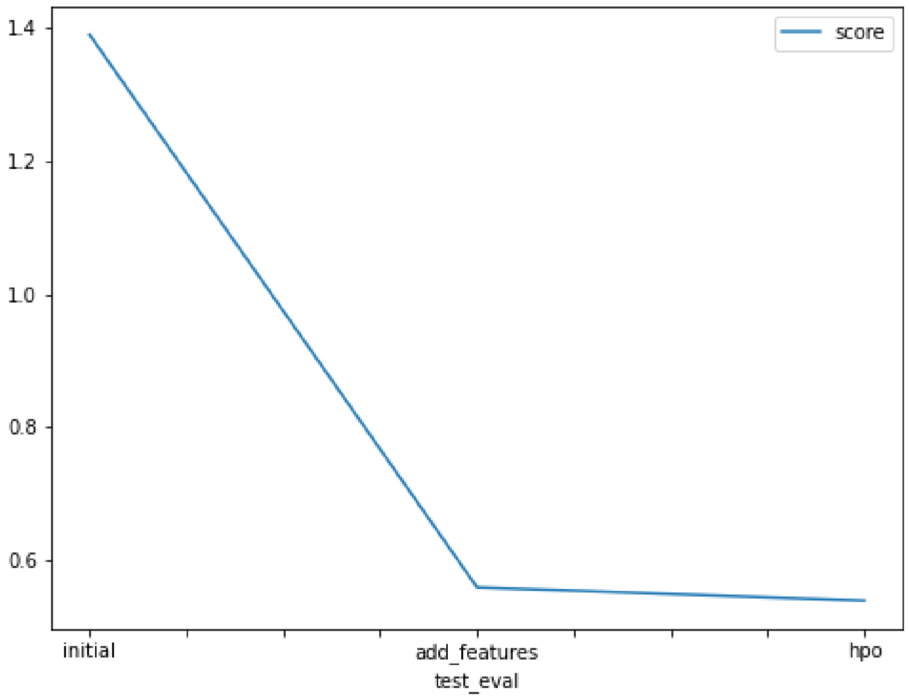
### Create a table with the models you ran, the hyperparameters modified, and the kaggle score.

Model	learning_rate	dropout_prob	num_leaves	score
initial	0.0005	0.1	36	1.39
add_features	0.0005	0.1	36	0.56
hpo	(0.0001,0.01)	(0.1,0.5)	(26,66)	0.54

### 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

This project was started by utilizing the capability of AutoGluon library to train the multiple deep learning algorithms in an efficient way. The exploratory analysis of the bike sharing data revealed that the count of bike sharing is about twice during the working days compared to that of non-working days, bike sharing counts has inverse relationship the windspeed, the weather significantly affects the bike sharing count. This study on the "Predicting Bike Sharing Demand with AutoGluon" showed that the performance of a machine learning model can be improved significantly by deriving the features from the existing features in the available data. Also, tuning of hyperparameters are worthy to improve performance of a model, however the tradeoff between the performance improvement and increased training time is to be considered.

I would like to thank the Kaggle for providing the dataset and thank the Udacity "AWS Machine Learning Engineer Nano Degree Program" team for giving me the opportunity to do this project.