Report: Predict Bike Sharing Demand with AutoGluon Solution

Lama Ibrahim

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?

The model's negative forecasts for the demand for bike sharing led to an error in the first submission. The submission was rejected by Kaggle because of these negative numbers. I have to change the forecasts to zero for all negative numbers in order to correct this. This made sure that the predictions that were submitted complied with Kaggle's criteria and were non-negative.

What was the top ranked model that performed?

Starting from the first training, LightGBM was the best-ranked model, with an RMSE of -50.786363. This model was chosen by AutoGluon automatically based on how well it performed during training.

Exploratory data analysis and feature creation

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

The exploratory data analysis, primarily through histograms, revealed that the 'datetime' feature contained valuable information that could be further broken down. I extracted the following features from the 'datetime' column:

year:The year of the rental.
month: The month of the rental.
day: The day of the rental.
hour: The hour of the rental.

These new features provide a more granular representation of time, allowing the models to capture potential patterns related to specific hours, days, months, or years.

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

The model's performance significantly improved after adding the new features. The RMSE score decreased from -50.786363 to -29.319962. This improvement can be attributed to the fact that the new features provided more granular information about the time of day, month, and year, allowing the model to better capture the seasonal and daily variations in bike sharing demand.

Hyper parameter tuning

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

Hyperparameter tuning using the GBM model with extra_trees enabled and a num_trials of 5 resulted in a further improvement in the model's performance. The RMSE score decreased to -32.364140. This indicates that tuning the hyperparameters of the model can lead to better generalization and prediction accuracy.

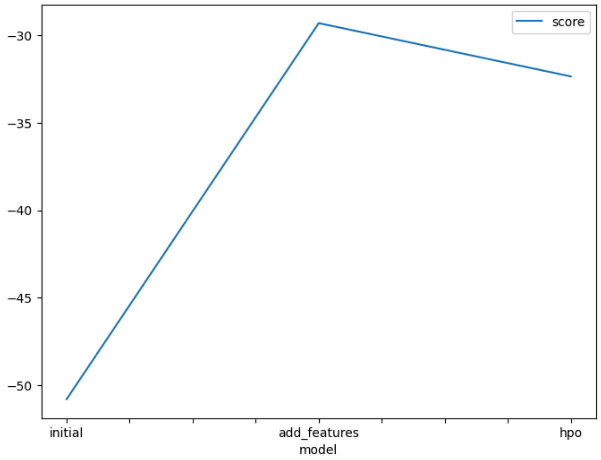
If you were given more time with this dataset, where do you think you would spend more time?

Given more time, I would focus on exploring and engineering more features. I would analyze the relationships between different features and the target variable ("count") to identify potential interactions and non-linear relationships. Additionally, I would investigate feature scaling techniques and experiment with different model architectures to further optimize the model's performance.

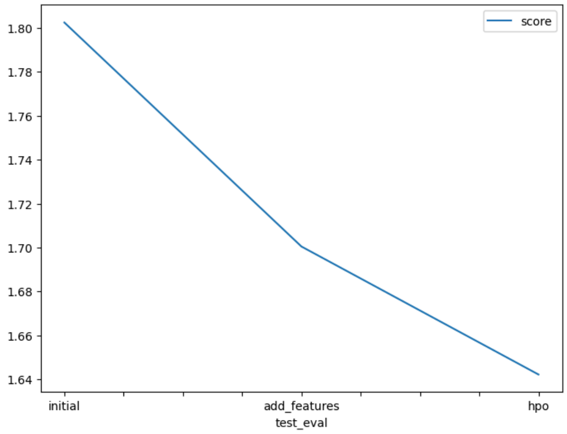
Create a table with the models you ran, the hyperparameters modified, and the kaggle score.

model	hpo1	hpo2	hpo2	score
initial	best_quality	time_limit=600		-50.786363
add_features	best_quality	time_limit=600		-29.319962
hpo	GBM	num_trials=5	searcher=auto	-32.364140

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 study showed how useful AutoGluon is for problems involving tabular prediction. We were able to significantly boost the model's performance by taking use of its automatic model selection and hyperparameter tuning features. The model's capacity to forecast demand for bike sharing was improved in large part by the inclusion of new variables and hyperparameter adjustment. Even better outcomes might arise from more research into feature engineering and model designs.