Seoul Bike Data

William K Davis III Pei-Yin Yang Max Kutschinski

2022-07-04

Results

Baseline

The MAE of the baseline model is evaluated using TSCV and puts a lower bound of 255.5 on our metric.

Random Forests

Random Forests (RF) is implemented using the ranger package for increased computational speed. Note that RF is an algorithm that is known to provide good results in the default settings (Fernández-Delgado et al., 2014). The arguably most influential hyperparameter is *mtry*, the number of randomly drawn features that are available at each split. In the regression case, p/3 is the default setting. When executing ranger via caret it automatically performs a grid search of *mtry* its entire parameter space. By default, the algorithm evaluates 3 points in the parameter space (smallest and largest possible *mtry*, as well as their mean) with 25 bootstrap iterations as an evaluation strategy and chooses the value with the lowest MSE.

The hyperparameters of our model are evaluated using TSCV, yielding an optimal *mtry* value of 53. Such a high value is usually indicative of a high number of relevant predictors (Probst et al., 2019). On the test set, this model results in a MAE of 166.2 outperforming the baseline by a reasonable margin.

XGBoost

There are a variety of booster parameters in XGBoost that can be optimized via TSCV. Rather than tuning all parameters simultaneously, it often helps to make small changes incrementally (Banerjee, 2020). The general idea is to start with a high learning rate and a small number of base learners, then tune other parameters, and finally decrease the learning rate while proportionally increasing the number of trees. The other adjusted parameters are the maximum depth of a tree *max_depth*, the minimum required loss reduction *gamma*, the fraction of columns to be subsampled *colsample_bytree*, the minimum sum of weights of all observations required in a child *min_child_weight*, and the fraction of observations to be randomly sampled per tree *subsample*. Using the outlined tuning approach, our best model produces a MAE of 131.6 on the test set.

References

Fernández-Delgado, M., Cernadas, E., Barro, S. and Amorim, D. (2014). "Do we need hundreds of classifiers to solve real world classification problems?" Journal of Machine Learning Research, 15, 3133–3181.

Probst, P., Wright, M. N., and Boulesteix, A. L. (2019). "Hyperparameters and tuning strategies for Random Forest". WIREs Data Mining and Knowledge Discovery, 9(3). https://doi.org/10.1002/widm.1301

Banerjee, P. (2020). "A Guide on XGBoost Hyperparameters Tuning". Kaggle, https://www.kaggle.com/code/prashant111/a-guide-on-xgboost-hyperparameters-tuning/notebook.