David Müller **davmuell@yahoo.de**

Prediction with Random Forest Regressor and Bayesian Optimiziaton Tuning

A Random Forest Regressor (RFR) is built. The dataset is splitted 80/20 in order to ensure that the final model is tested on data it has never encountered in the training process. The objective function for the Bayesian Optimization is based on the mean absolute error of the estimated test error of the RFR determined by 5 fold cross validation. The predictions of the tuned RFR yield a mean absolute error of 20.79.

Feature Preprocessing and Selection

For Data Storage a medallion architecture is chosen, saving the original data file in the Bronze folder. Several features are typecasted into proper datatype, e g. 'dteday' into datetime, 'weekday', 'month', 'workingday',. . . , into categorical features. The hourly Bike Rental Counts 'cnt' are shown in Figure 1. We see that the minimum amount of bike rentals had been 1 and the highest demand had been 977 bikes, while on average approx. 189 bikes are rented (see Table 1). The table also suggests that the distribution is right skewed which is confirmed by Figure 2. In Figure 3 the Bike Rentals are plotted against several categorical features. It seems that hours and weather conditions have a noticeably impact on the Bike Rentals, whereas the demand is only slightly varying by the weekdays. On a working day the demand for bikes is clearly increasing, where it also does not matter which season it is (Figure 4). Hour has cyclic structure, so feature 'hr' is projected into 2 dimensions per cyclic transformation which ensures that distance between succeeding hours is remained, see Figure 5. It is not clear from the the context how dynamic the prediction tool shall be. Thus, it is assumed that the stock of bikes can be adjusted between hours. Therefore, a lag variable of 'cnt' is added to the features (for instance between different rental stations). We check the correlations between the numerical features and see that features 'temp' and 'atemp' basically carry the same information (Figure 6). As 'atemp' has some unusual values we keep 'temp', see Figure 7. Furthermore, season and month are very similar, so month is dropped. Moreover, the years are not reproducible and with only 2 years there is not enough evidence that the demand for bike rentals is increasing in years, so it is also dropped from the dataset. Hours is dropped, since the 2D transformation is kept. If colleagues would like to keep or drop other features they could work with the dataset in folder silver. The dataset ready for prediction is saved in the Gold folder.

Model Selection and Results

As the model should be used for a business case, a Random Forest Regressor is selected. RFR are known to be powerful predictors, often even outperforming GBM or Deep NN. Besides, the model will only predict nonnegative values for this dataset, which is meaningful. A customer could be sceptical concerning the benefits of a model which predicts negative rentals. Such scenario could for example happen with powerful prediction models such as XGB Regressor. For this business case it seems that accurate prediction are way more important than the c.p. impact of individual features, so RFR is preferred to e. g. Bayesian Regression. RFR can be trained in parallel and they avoid overfitting by combining many weak learners and learning on subsets of the features. The parameters of the RFR are chosen via Bayesian Optimization for searching the parameter space by using gained information. With the tuned model the generalization error is estimated by 5 fold cross validation which yield on average 21.24. The predictions of the tuned RFR on the unseen test data (20% of the entire data) yield a mean absolute error of 20.79. The predictions are shown in Figure 8. The script for prediction is separated

from the EDA script, such that colleagues could easily make changes in one or the other[1]. Also the code for the prediction script is mainly written in functional programming in order to allow colleagues to rerun predictions with slight changes of the non-tuned parameter or to adjust the feasibility set and parameters of the Bayesian Optimization process.. In order to provide the customer with more information, the feature importance are approximated with the SHAP explainer. According to the results, the rented bikes in the preceded hour and the 2D transformed hours as well as the feature workingday contribute most to the predictions of the model, see Figure 9.

Mean Absolute Errors of Tuned Random Forest Regressor

Fold 1 : 20.65, Fold 2 : 21.34, Fold 3 : 21.8, Fold 4 : 21.28, Fold 5 : 21.11, Test Error : 20.79.

[1]I did not create a git repository which would I have done in a teamwork of course. However, I included the requirements files.

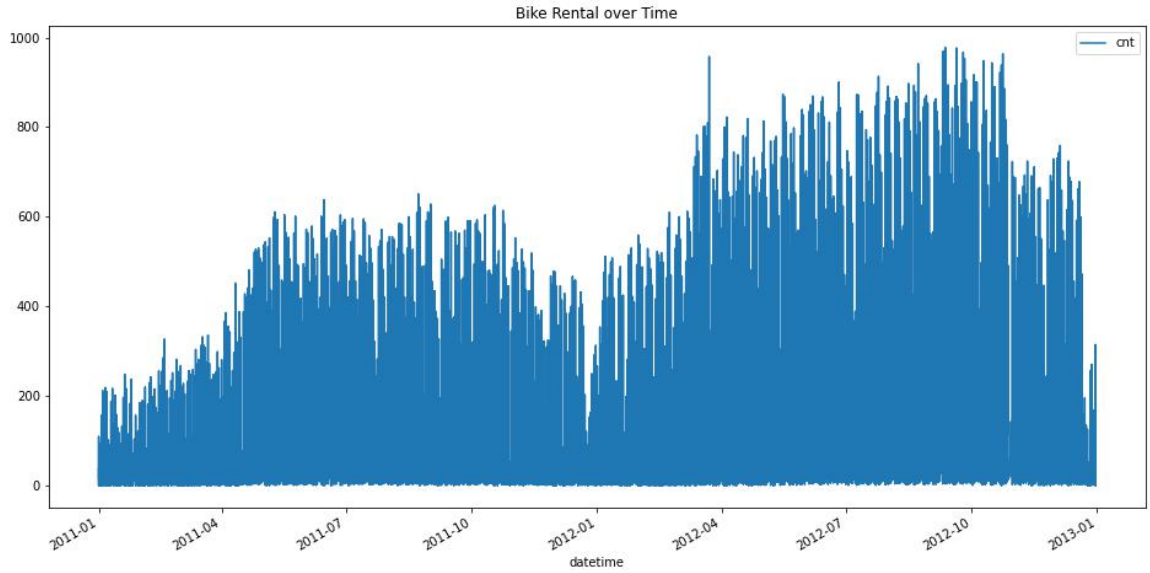

Figure 1

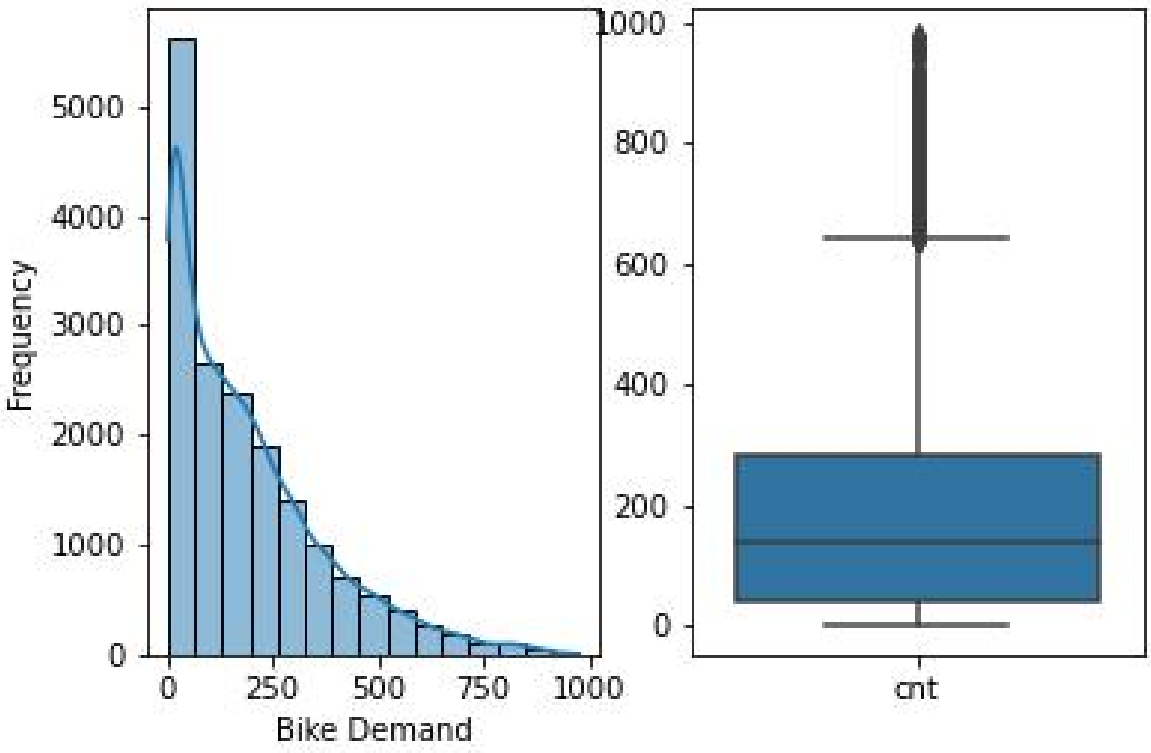

Figure 2

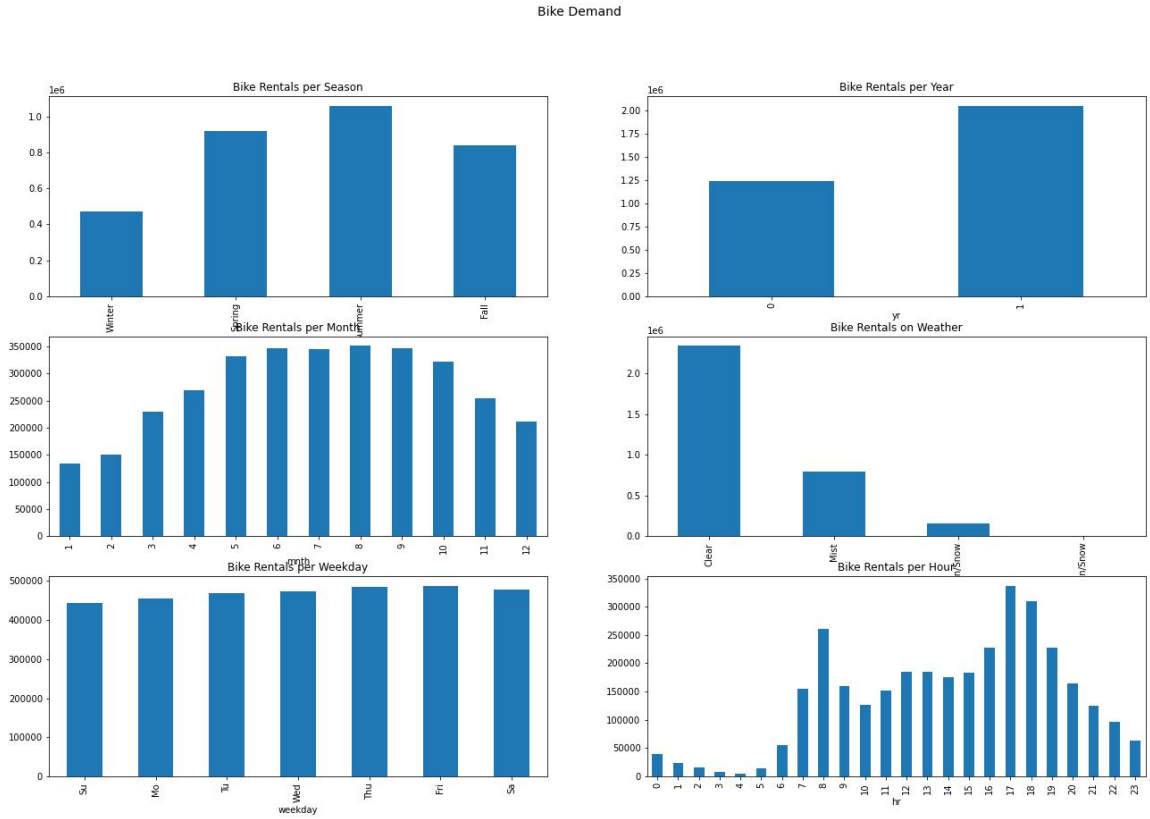
Bike Demand
Bike Rentals per Season
Bike Rentals per Year
Bike Rentals per Month
Bike Rentals on Weather
Bike Rentals per Weekday
Bike Rentals per Hour
Winter
Spring
Summer
Fall
yr
mnth
weekday
hr
Clear
Mist

Figure 3

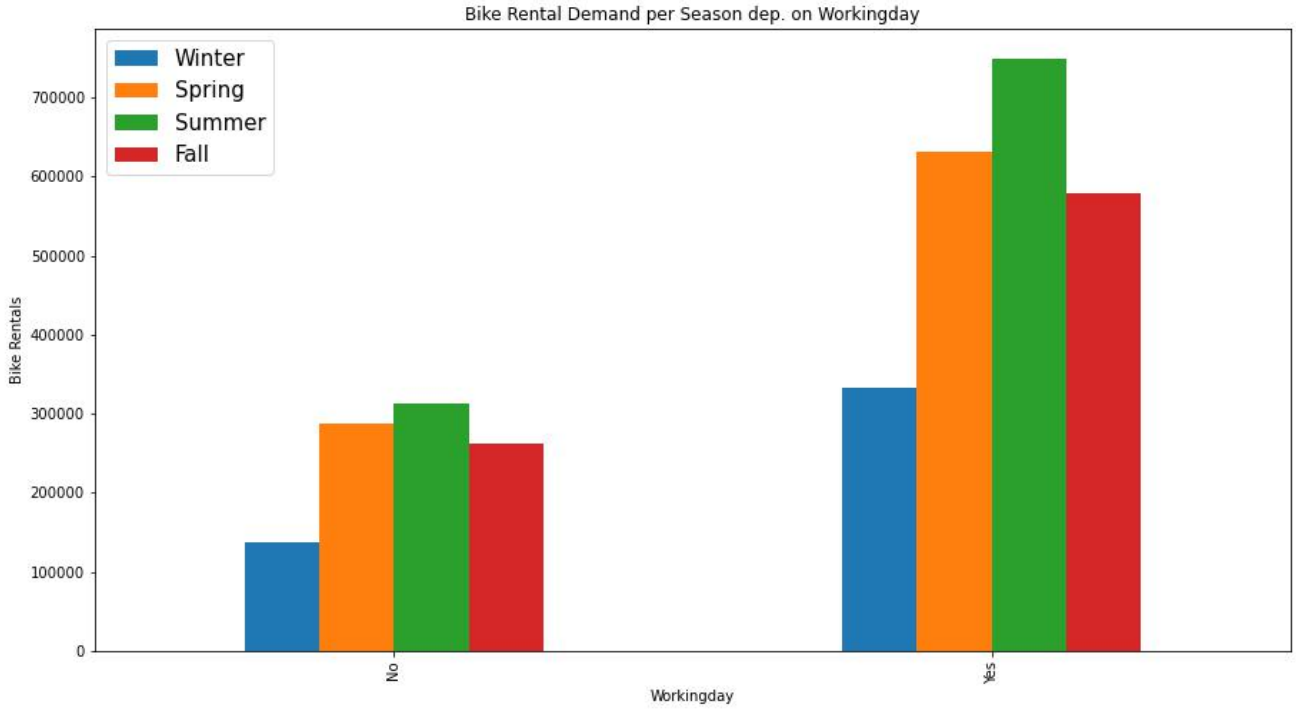
Bike Rental Demand per Season dep. on Workingday
Winter
Spring
Summer
Fall
Bike Rentals
0
100000
200000
300000
400000
500000
600000
700000
No
Yes
Workingday

Figure 4

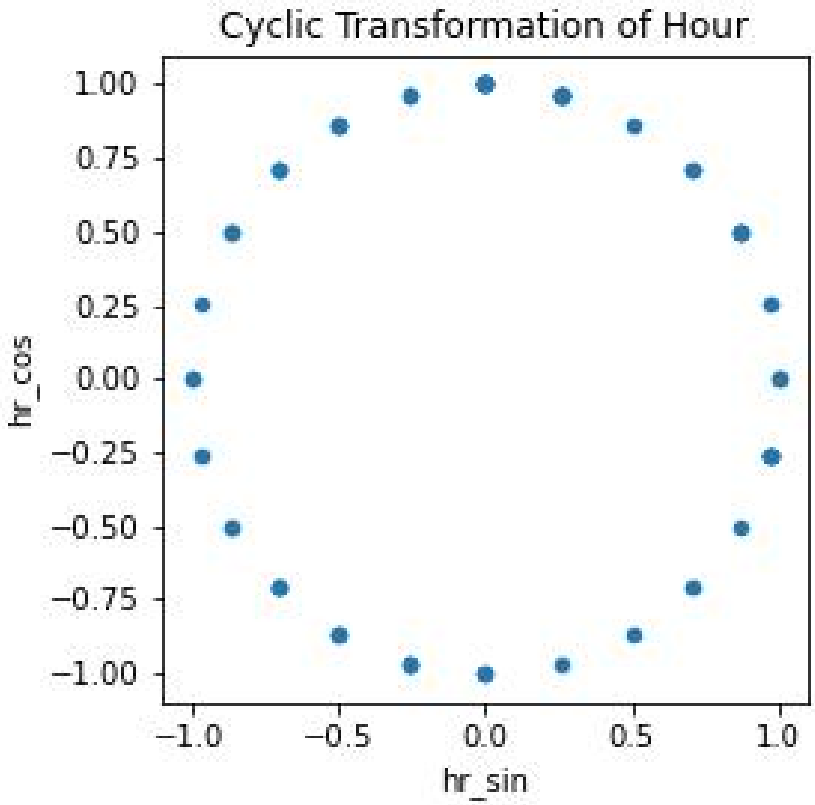
Cyclic Transformation of Hour
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
hr_cos
−1.0
−0.5
0.0
0.5
1.0
hr_sin

Figure 5

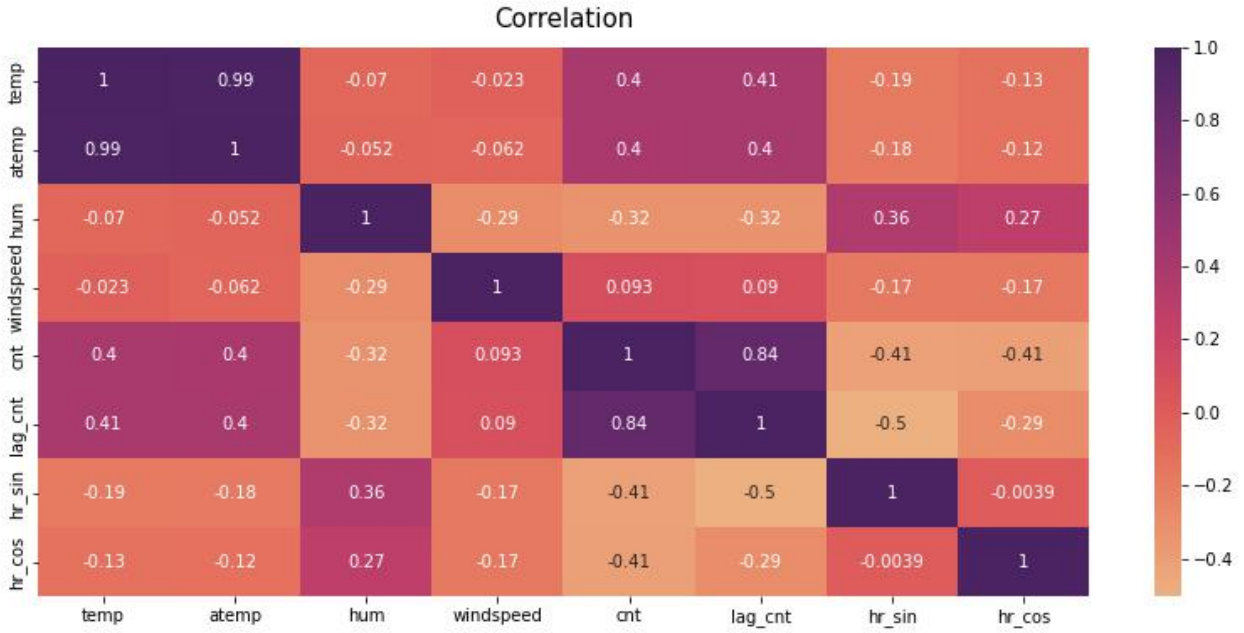
Correlation
temp
atemp
hum
windspeed
cnt
lag_cnt
hr_sin
hr_cos
1 0.99 -0.07 -0.023 0.4 0.41 -0.19 -0.13
0.99 1 -0.052 -0.062 0.4 0.4 -0.18 -0.12
-0.07 -0.052 1 -0.29 -0.32 -0.32 0.36 0.27
-0.023 -0.062 -0.29 1 0.093 0.09 -0.17 -0.17
0.4 0.4 -0.32 0.093 1 0.84 -0.41 -0.41
0.41 0.4 -0.32 0.09 0.84 1 -0.5 -0.29
-0.19 -0.18 0.36 -0.17 -0.41 -0.5 1 -0.0039
-0.13 -0.12 0.27 -0.17 -0.41 -0.29 -0.0039 1
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Figure 6

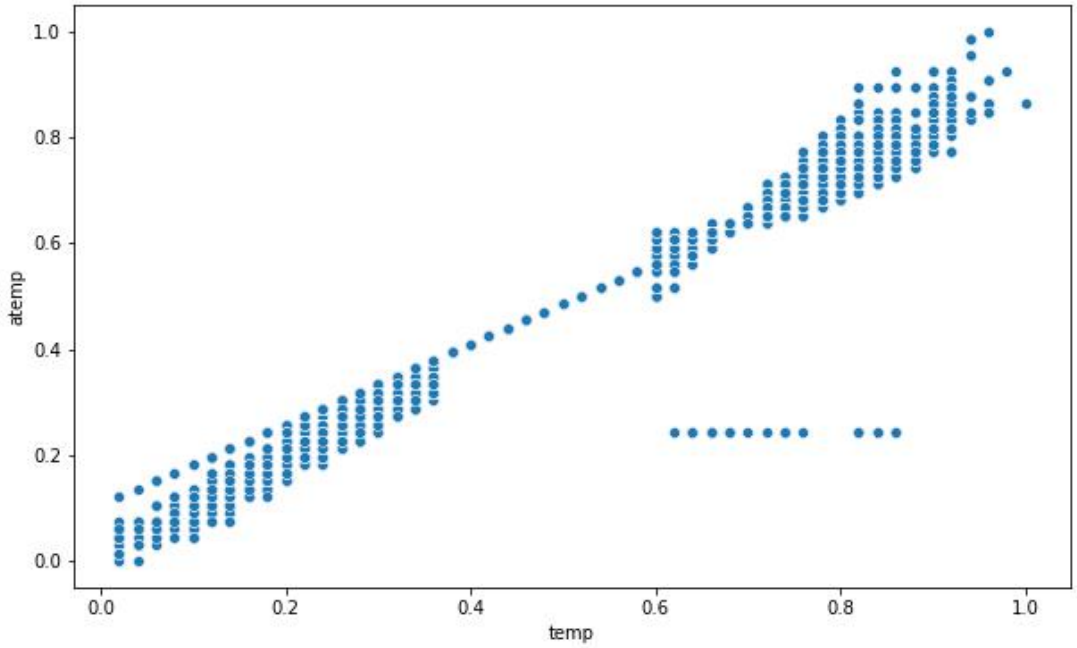

Figure 7

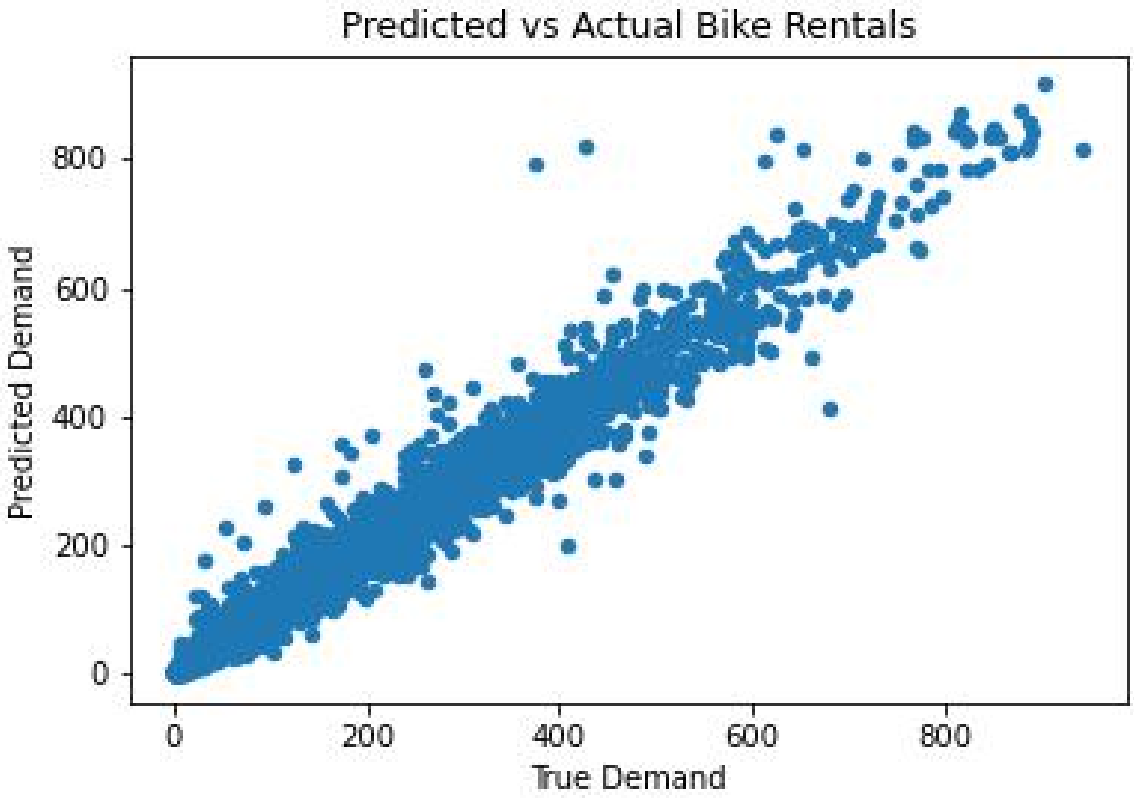
Predicted vs Actual Bike Rentals
Predicted Demand
True Demand
0
200
400
600
800

Figure 8

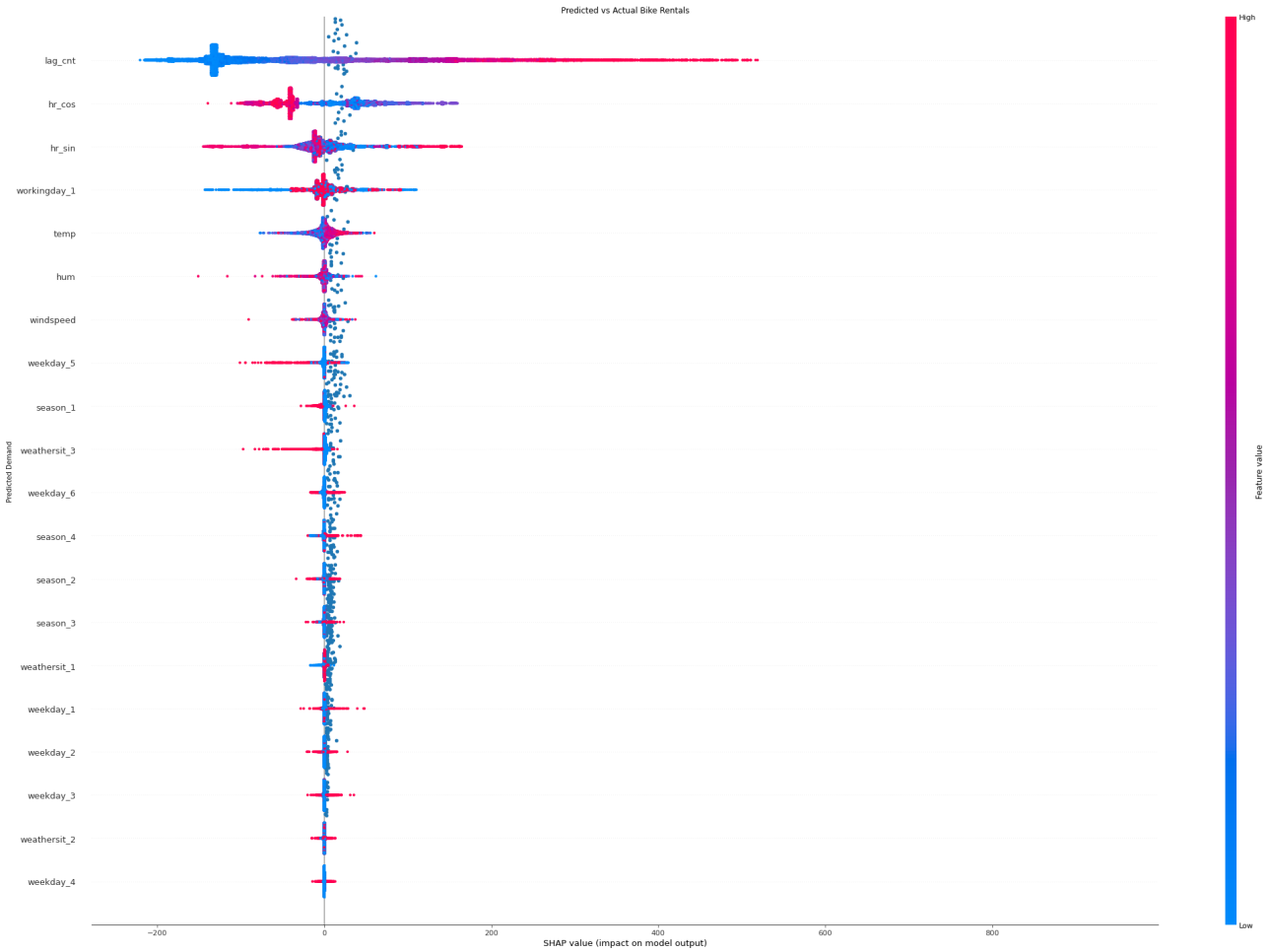
Predicted vs Actual Bike Rentals
lag_cnt
hr_cos
hr_sin
workingday_1
temp
hum
windspeed
weekday_5
season_1
weathersit_3
weekday_6
season_4
season_2
season_3
weathersit_1
weekday_1
weekday_2
weekday_3
weathersit_2
weekday_4
Predicted Demand
−200
0
200
400
600
800
SHAP value (impact on model output)
High
Feature value
Low

Figure 9

Statistic	Value
Mean	189.463
Std	181.288
min	1
25%	40
50%	142
75%	281
max	977

Table 1: Summary Stats of 'cnt'.