Bikeshare Project with Git / Github

In this Bikeshare project with github project we run analysis on Bike sharing data on daily and hourly datasets. This Bike sharing is a new generation of traditional bike rentals where the whole process has become automatic from membership and rentals. Through these systems, users are able to easily rent a bike from a particular position and return back to another position. Currently, there are about over 500 bike-sharing programs around the world which are composed of over 500 thousands bicycles.

The DATASETS (daily and hourly):

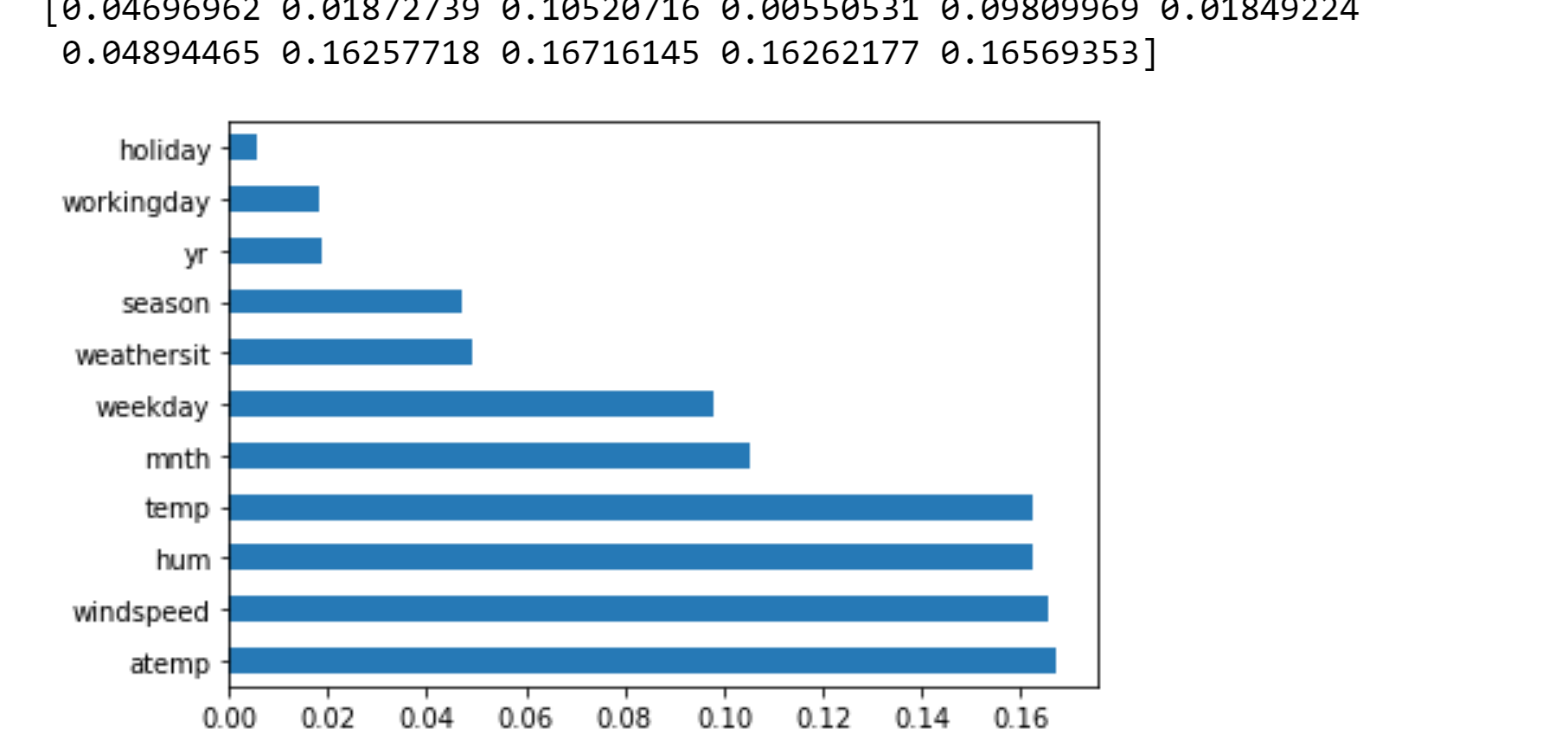
Daily dataset contains (731, 16) 731 rows and 16 columns. The column names are 'instant', 'dteday', 'season', 'yr', 'mnth', 'holiday', 'weekday','workingday', 'weathersit', 'temp', 'atemp', 'hum', 'windspeed', 'casual', 'registered', 'cnt'.

Hourly dataset contains (17379, 17) 17379 rows and 17 columns. Column names are 'hr','instant', 'dteday', 'season', 'yr', 'mnth', 'holiday', 'weekday','workingday', 'weathersit', 'temp', 'atemp', 'hum', 'windspeed', 'casual', 'registered', 'cnt'.

The date contains the hourly and daily count of rental bikes between years 2011 and 2012. It contains the external factors that affect people riding bikes like 'weathersit', 'temp', 'atemp', 'hum' and 'windspeed' and days like holiday and weekday that increase and decrease the number of riders.

In my analysis we made a linear regression model and Random Forest model to predict the accuracy of different columns on no. of count daily or hourly. The idea of linear regression is to examine whether a set of columns do a good job in predicting an target variable (“cnt”) and Which variables in particular are significant predictors of the target variable, and we consider accuracy of the model to decide whether to consider the variables or not for the model.

To select the columns that have more relation on the target variable first, we used the decision Tree algorithm to select the columns that are more related to the target variable.



So, here we can see temp, humidity, windspeed and atemp are more related to the target variable. So, we built a linear regression model with x= 'mnth','weekday','temp', 'atemp',"hum","windspeed" and y = ‘cnt’. The accuracy core is 48%. And we built another linear regression model with x= 'temp', 'atemp' and y = ‘cnt’ and the accuracy score is 36%. So, the first model has the better accuracy score then the second model.

Additionally we made Random Forest Regression on the Hourly dataset. Random forest regression is an ensemble method and other tasks that operate by constructing a multitude of decision trees at training time. We did data preprocessing and used One-Hot-encoding and created dummy columns. We excluded ‘casual’,’registered’,’instant’ and ‘dteday’. We divided the dataset into 80% Train and 20% Test. We run the model by setting the no. of estimators to 300 and max features to ‘auto’. The accuracy score of the model is 92% and the mean absolute error is 31.17.

Instead of making many linear Regression models to predict the target variable which is not suitable for this dataset. Random Forest model give much better result than Linear Regression models. The accuracy rate of Random forest is higher and best for predicting the target variable.