**Assignment Part 2**

**Linear and Non-Linear Regression with Multiple Features**

ASSIGNMENT

Use Orange to perform a multi-feature regression on the bikeshare-daily.xlsx dataset. The dataset is available for download from GitHub at the following URL (you’ll see a “Download” button on the upper right of the page):

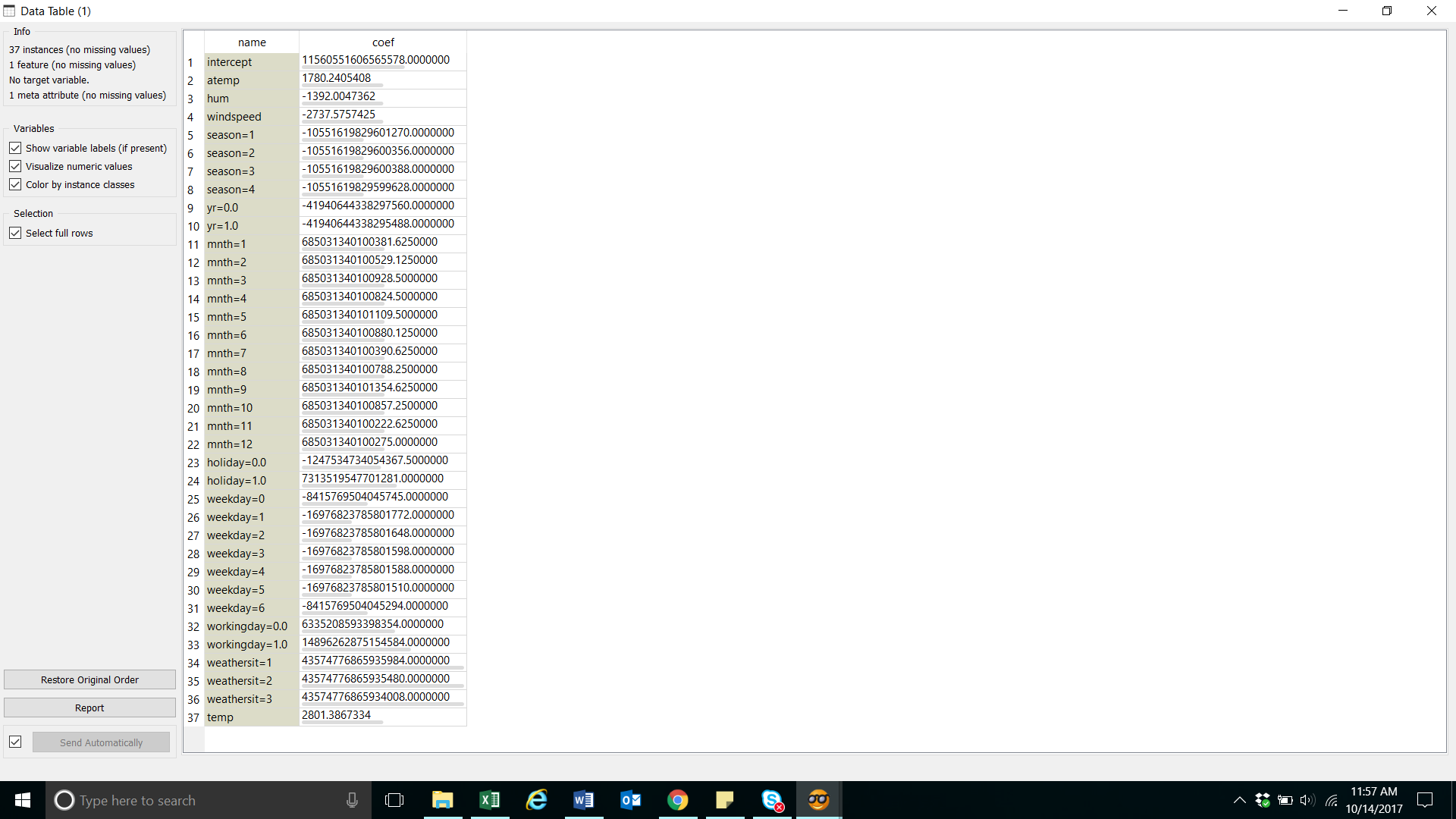
https://github.com/jsub10/MLCourse/blob/master/Data/bike-share-daily.xlsx

Specifically, carry out the following steps:

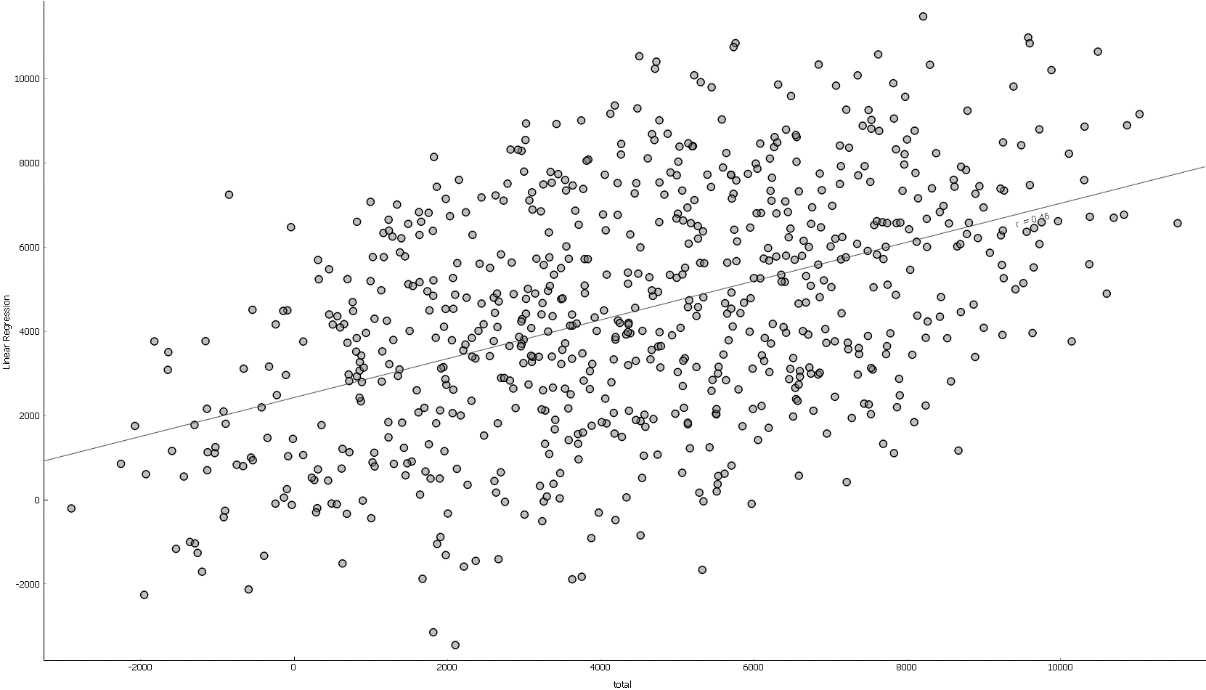
1. **Use the feature named “total” as the target feature. This is the output that you’re interested in predicting using regression with multiple features and then non-linear regression. (Note: This is just a step to take and there are no questions to answer)**
2. **Use Orange’s File widget to designate each feature as categorical or numeric. Make “instant” a meta feature. Skip the “casual” and “registered” features because they add up to the “total”. (Note: This is just a step to take and there are no questions to answer)**
3. **Visualize the features using a scatter plot in a systematic way. Select categorical-categorical, categorical-numeric, and numeric-numeric comparisons. In particular, create 6 scatter plots using your choices of features to display and write a sentence or two describing each scatter plot. Make sure you capture the images and paste them into your Word document.**

|  |  |  |
| --- | --- | --- |
| Numerical-Numerical |  |  |
|  | This scatter plot shows how the total number of rides are affected by the windspeed. There seems to be a weak negative correlation between the two (i.e. as windspeed increases number of rides drop) | This scatter plot shows how the total number of rides are affected by the tempruture. It can be obsereved that as temprtures increase, the number of rides increase. However, there must be a cap to this (i.e. at a certain temprature it will be too hot and rides will drop) |
| Categorical-Numerical |  |  |
|  | This scatter plot shows how many rides happen during the workday vs. a weekend. As expected the frequency of datapoints are higher on the workingdays. Also, it can be obsered that the highest number of rides happen on the weekend. | This scatter plot shows how many rides happen during each season. As expected the frequency of datapoints are higher during summer (season 2). |
| Categorical - Categorical |  |  |

1. **Create a linear regression model using all the features you’ve selected using the default regression settings (“No regularization”). (Note: This is just a step to take and there are no questions to answer)**
2. **What are the optimal values of the parameters? In Orange w0 is called the “intercept”. The names of the other coefficients will be the same as the feature names. Report these values in the Word document and save them values in an Excel file. (Use Orange’s Save Data widget.) [Please see my Orange file for the list of features and the optimal parameter values associated with each feature.]**



1. **Use the model and data to make predictions on your data. In other words, for each set of feature values in the dataset, find the predicted total number of bikes rented. Compare the predicted values with the original totals in the dataset using a scatter plot. Save these values in an Excel file.**



With an ideal model (where the predicted data match the actual data perfectly) all the dots should be on a straight line. However, as seen in the scatter plot, there is quit a difference between the two.

1. **Now simplify the dataset and create a dataset that only includes the features “atemp”, “hum”, “windspeed”, and “total”. Use “total” as the target feature. (Note: This is just a step to take and there are no questions to answer).**
2. **Split this simplified dataset – put 80% into a dataset we’ll call the “training” dataset and 20% into a dataset we’ll call the “test” dataset. Use Orange’s Data Sampler widget to do the 80-20 split. (Note: This is just a step to take and there are no questions to answer).**
3. **For the *training dataset*, create the following regression models (use the default “No regularization” setting): (Note: This is just a step to take and there are no questions to answer)**
   1. **y = (w1 \* atemp) + (w2 \* hum) + (w3 \* windspeed)**
   2. **y = (w1 \* atemp^2) + (w2 \* hum/windspeed) – (w3 \* windspeed^3)**
   3. **y = (w1 \* atemp \* windspeed) + (w2 \* hum \* windspeed^(1/2)) – (w3 \* windspeed^5)**
4. **For each model, capture the coefficients of the model in an Excel file. Use these coefficients to make predictions on the *test dataset*. You should end up with 3 Excel files – one each for predictions on the test dataset using models 9a, 9b, and 9c above.**
5. **Compare the predictions of the models using side-by-side scatter plots of the actual versus predicted values in the dataset. How do the non-linear models compare when it comes to prediction? Describe your results.**

|  |  |  |
| --- | --- | --- |
|  |  |  |
| **9.a.** | **9.b.** | **9.c.** |

A perfect model would make the dots be on a straight line (i.e. exact prediction). However, as shown in all three scatter plots each model contains a certain amount of error in predicting the total number of rides. It can also be observed that model 9.a. is a better predictor than 9.b and 9.c therefore, it would be wise to revisit the models built in 9.b. and 9.c.

1. **Suppose you are the head of business strategy for the bike share company. List 3 business questions you would use this dataset (the entire original one you started with) to answer.**
2. Do we have a model that has a high prediction power? It is crucial to understand if we have a good predicting model as this would allow for better availability of bikes and not losing potential business for not having bikes available.
3. Are there certain days / seasons / weather conditions that shows enough drop in bike rides that it would justify shutting down operations on those days ?
4. What features result in lower number of rides? Would it make sense to transfer the none used bikes to a second location to generate more revenue at the second location that does not have those certain features ?