Bike Share Mini-Project

Abstract—Bikeshare has become an increasingly popular transportation method in the United States. This project is to create some models that predict bike usage. The data used for this project is collected from Capital Bikeshare trips during 2011 and 2012. This project includes Exploratory Data Analysis, Data Wrangling, Variable Selection, Modeling, Plots and Conclusion.

Dataset Selection

There are two csv data sets "hourly" and "daily." "Hourly" has information about bike sharing counts on an hourly basis while "daily" contains counts daily. The former data set contains 17,379 rows while the latter only has 731 rows. This project only uses the "hourly" data set as it contains more data.

Exploratory Data Analysis (EDA)

In this process, there will be initial investigations of the data set to explore the features and the data types of the features. There is a total of 17 columns. The *instant* column will not be used as it is the record index. The independent variables are 1) *dteday*, 2) *season*, 3) *yr*, 4) *mnth*, 5) *hr*, 6) *holiday*, 7) *weekday*, 8) *workingday*, 9) *weathersit*, 10) *temp*, 11) *atemp*, 12) *hum*, 13) *windspeed*, 14) *casual* and 15) *registered*. The dependent variable is *cnt*, which refers to total number of rental bikes including both *casual* and *registered*. There are no missing values in the data set. However, some of the columns' data type will require conversion.

Data Wrangling

Since the numerical columns are already normalized, scaling them is not needed in this case. Data type conversion is necessary as some of the columns' types are integer, but they are indeed categorical features. These columns are *season*, *yr, mnth*, *holiday*, *weekday, workingday* and *weathersit*. Next, the data set is split into training and test sets randomly. The training set contains 80% of the data while the test set contains 20%.

Variable Selection

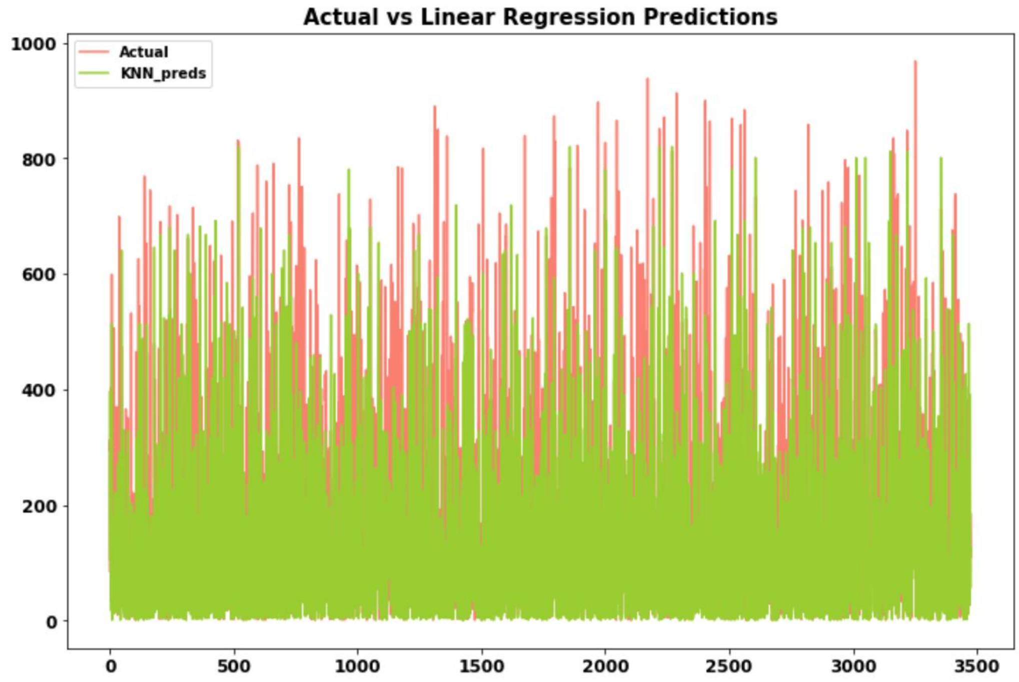
Since *cnt* is the sum of *casual* and *registered, casual* and *registered* will not be used. The variable *dteday* will be excluded also as the other columns such as *mnth* and *yr* contain similar information. Correlation analysis is used on all other numerical variables to find out which are significant to the target column and whether any of the pairwise numerical variables are highly correlated to each other. From the correlation matrix, it shows that only *hum*, *temp* and *atemp* are significant to *cnt*. Additionally, *temp* and *atemp* are highly correlated, so only *temp* will be used. Then, Ordinary Least Squares regression (OLS) is performed to check if *temp*, *hum* and all categorical variables are significant to *cnt* if used together. The features that would be used for prediction are *temp, hum, season, yr* and *hr.* After selecting the variables, one-hot encoding is performed for the categorical variables.

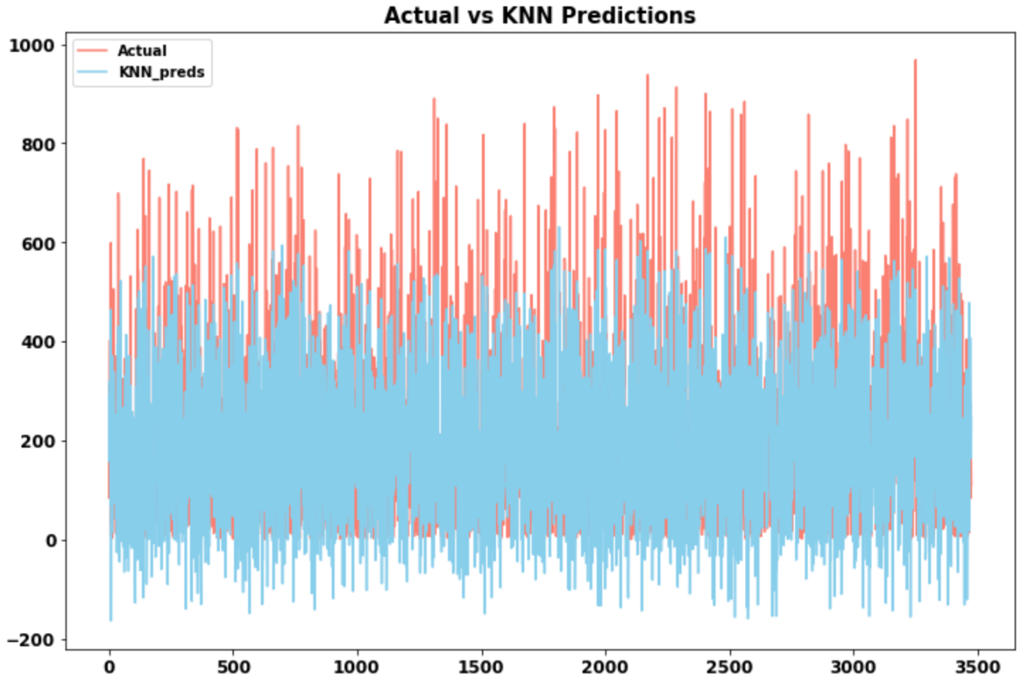
Modeling

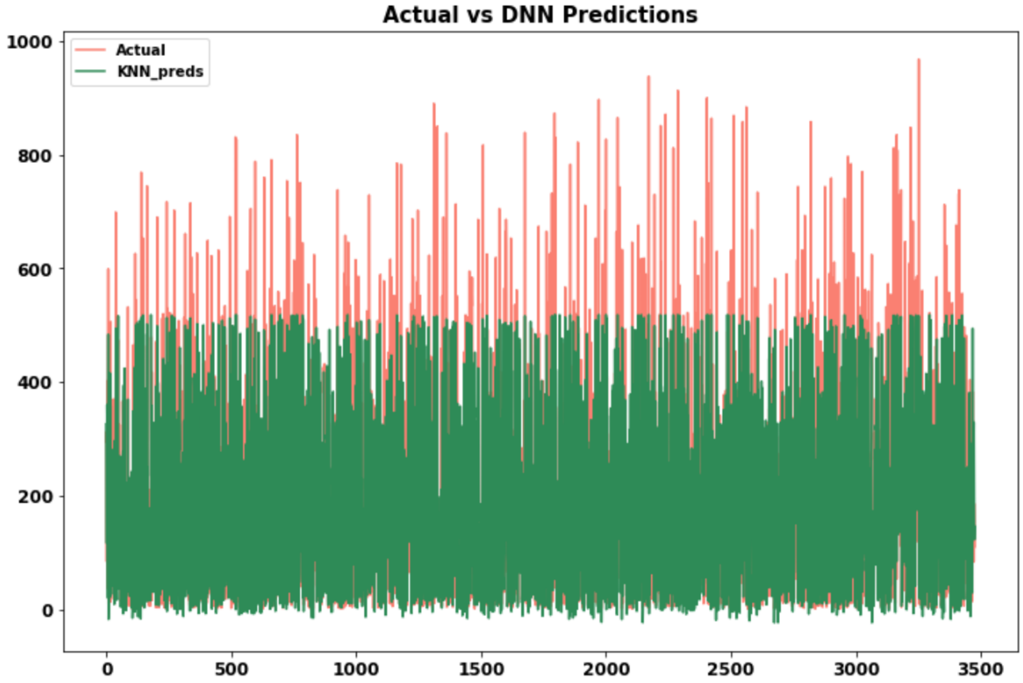
Three modeling techniques are used in this project and they are Linear Regression (LR), K-Nearest Neighbor (KNN) and Deep Neural Network (DNN). LR is useful in this case since it helps to explore the associations between multiple categorical and/or numerical variables and a continuous dependent variable. KNN can also be used to make predictions when the target variable is continuous. Since this data set has over 17,000 rows, DNN can also be an effective method. Python has built-into functions available for these three methods. After the models are trained on training data, they will be used to make predictions on the test set.

Plots

The Actual versus Predicted plots for all three models are shown below:







Conclusion

The plots for the actual values and the predicted values are displayed above to show how close the predicted values are when compared to the actual values. However, from the plots, it is hard to visually tell which model performs the best. Thus, the Root Mean Square Error (RMSE) is used to evaluate the performance of the models. RMSE measures how close the observed data points are to the models' predicted values. The RMSE for the DNN model is 95.05, for the LR 105.66, and for KNN is 143.74. DNN gives the lowest RMSE, so it is the best model among the three.