Elinor Velasquez
Capstone Two Problem Statement: Seoul, South Korea Bike Sharing Counts

Figure 1. Rental bikes in Seoul, South Korea

The problem to be solved is the prediction of bike count required per day and at a given collection of locations for the stable supply of rental bikes. The client is Seoul Bike, which participates in a bike share program in Seoul, South Korea. An accurate prediction of bike count is critical to the success of the Seoul bike share program. It is important to make the rental bikes available and accessible to the public at the right time as it lessens the waiting time: Providing the city with a stable supply of rental bikes is a major concern for the program. The goal of the project is to predict the required number of rental bikes per week using machine learning and data mining.

The data set is concerned with rental bikes as a ride sharing program. There are 8760 instances and 14 attributes (features). The features are aspects of the weather, and the number of bikes rented per hour and date.

The criteria for success is the following: The score to rate the prediction will be the R^2 M.A.E. and other types of error will be computed to rate the success of the model used to predict bike needs. The critical part is the prediction of bike count per day or week for a stable supply of rental bikes. The stakeholders are affiliated with the Seoul bike share program: the program operators and the bicyclists who use the program to ride bicycles in Seoul.

The data was acquired from the Machine Learning Repository, Center for Machine Learning and Intelligent Systems, U.C. Irvine. Histograms were plotted to see if the data is normally distributed or skewed.

A Pearson correlation heatmap was generated to see which features are linked to bike usage. The date and hour was combined by invoking datetime. Then, the time was

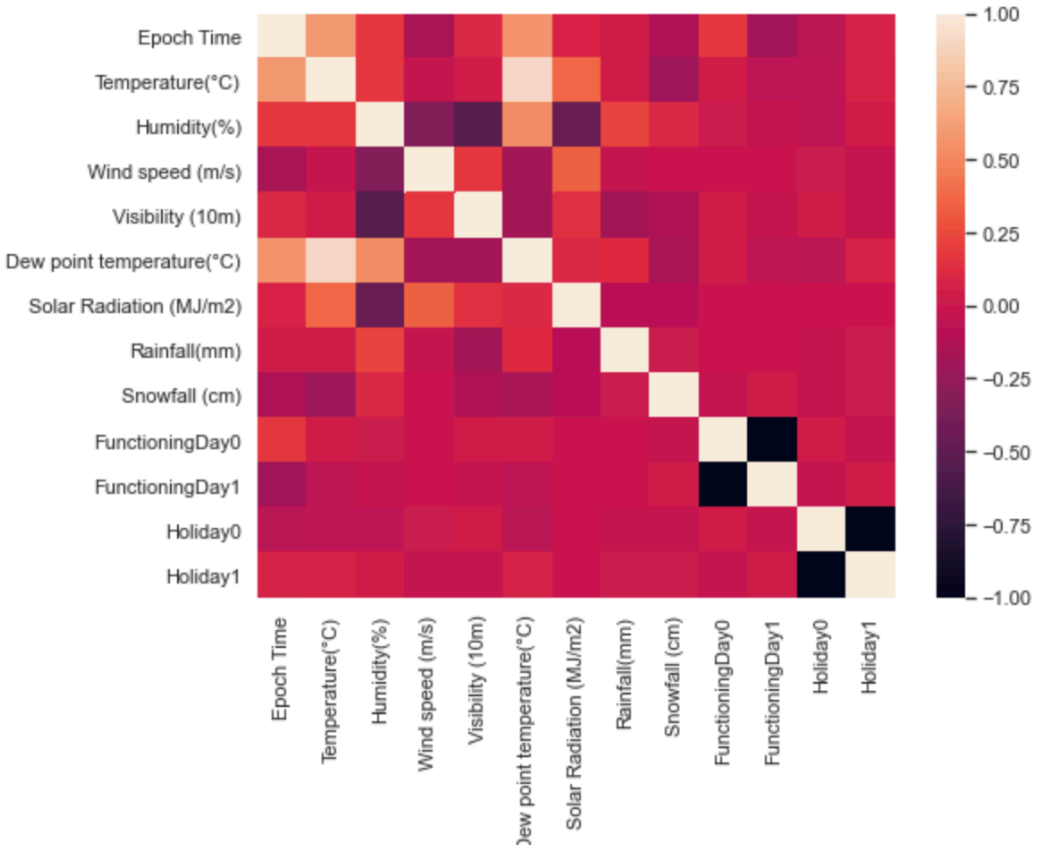

Figure 2. A heat map of the bike share features

converted to Epoch Time (January 1, 1970 served as a reference point for the Epoch Time conversion). After inspecting the data as a time series, the prediction was formulated as: Can we predict weekly rented bike count?

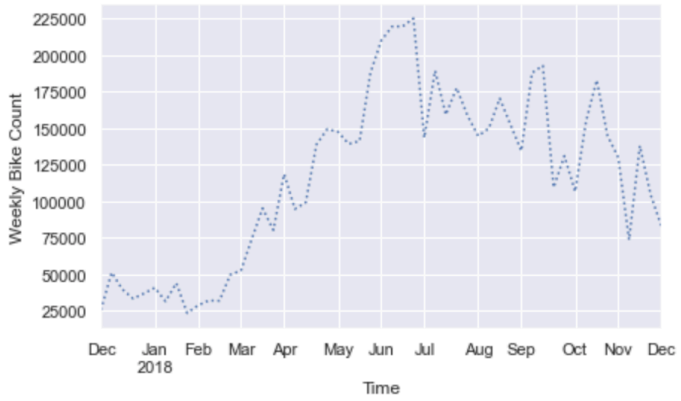

Figure 3. Weekly Bike Count as a function of Time

The first predictive modeling used was ordinary least squares as the simplest regression map to get a baseline for the problem. R-squared = 0.46 for linear regression. Next, hyperparameters were optimized for linear regression: The number of features were optimized using GridSearchCV.

The optimal number of features to select was asserted to be 10, using recursive feature elimination. That is, the metric, R-square, was computed by first by fitting the training data and the target training data to an instance of linear regression, and then use an instance of recursive feature elimination to predicting the target values using the optimal number of features equal to ten, and lastly computing R-square on the test target values and the predicted target values.

A support vector machine (SVM) regressor was used to predict bike usage. The kernel for the SVM Regressor was a radial basis function. GridSearchCV was again used to optimize the SVM Regressor. The hyperparameters to be optimized in this case were C, epsilon, gamma, and the kernel. The best hyperparameters were then used to compute R-square for the test (actual) target values and the predicted target values. For the SVM Regressor model, R-square equaled 0.68.

A neural network model (Multi-Layer Perceptron Classifier) was then applied. Both the training data (X_train) and the testing data (X_test) were normalized using the StandardScaler algorithm. The maximum precision and recall were 0.82 and 0.98, respectively. Since this was a classifier algorithm, the CatBoost Regressor was used as a final model. The R-square for the CatBoost Regressor was 0.75, as illustrated by the following curve:

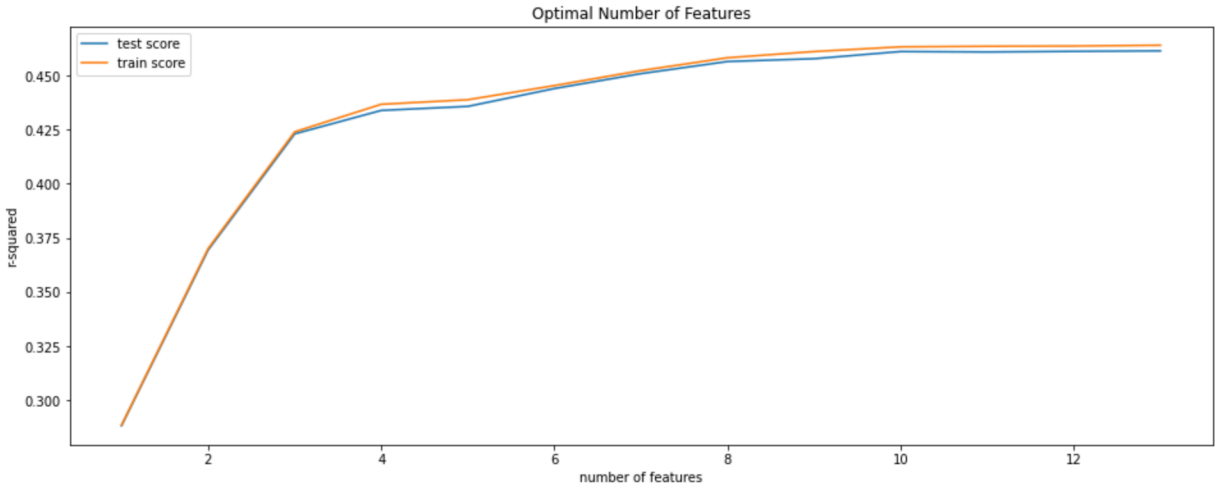

Figure 4. The Number of Features is a hyperparameter for linear regression.

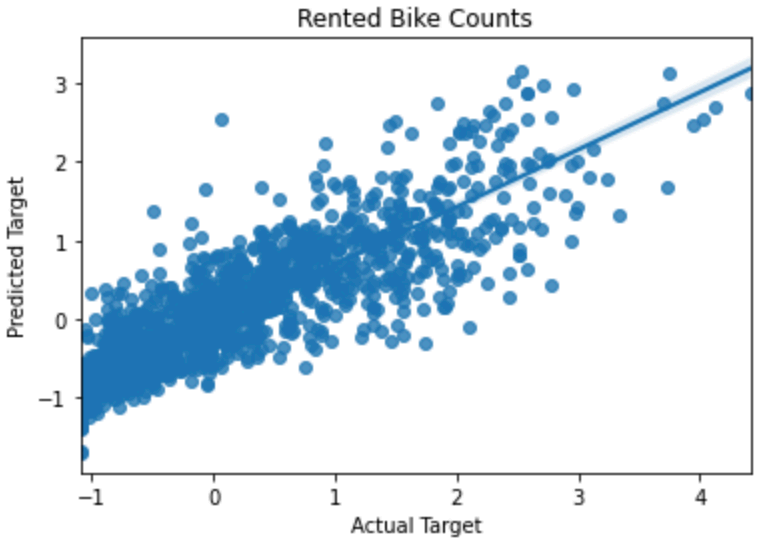

Figure 5. The predicted target correlates well with the actual target (R-square = 0.75, CatBoost Regressor)

Lastly, CatBoost Regressor was optimized using the best hyperparameters obtained by Bayesian Optimization. The two hyperparameters that were tuned were “depth” and “bagging temperature.”

To summarize, the data to be explored had features of time and weather-related. The data was normalized, and split into training (0.30) and test sets (0.70). The variable to be predicted was Rental Bike Count. The time (date joined with hour) was recast as Epoch Time, and then normalized, so that it could be used as a feature. Linear Regression, Multi-linear Regression, Support Vector Machine Regressor, Neural Network Classifier, and CatBoost Regression were models that were used to predict Rental Bike Count. The hyperparameters associated with each model were either estimated by GridSearchCV or Bayesian Optimization. The highest performing model was Cat Boost Regressor with R-square of 0.75.

For further research, it may be helpful to collect spatial data, that is, data of bike share locations, annotated by time. New models could be designed with spatial and temporal features, and with additional features that are weather related. It also may be helpful to collect data about the users, namely, if users are new or repeat customers. It may be helpful to join the snowfall and rainfall features into one data feature. There is already a temperature feature, so snowfall-rainfall feature would complement the temperature feature. There is already a humidity feature, so it is possible to join snowfall/rainfall (100% humidity) into one feature.