MDA 9159A Data Analysis Project

**Seoul Bike Sharing Demand Data Set**

By: Debduti Sengupta

Parthavi Madaan

Shivanika Shah

**Introduction: Background/Motivation of the Dataset**

Currently, Rental Bikes are introduced in many major cities for the enhancement of mobility, comfort, and eco-friendly transportation. It is important to make the rental bike available and accessible to the public at the right time as it reduces the waiting time. Eventually, providing the city with a stable supply of rental bikes is a major concern. There are many factors like temperature, time of day, holidays, seasons, etc that might affect the availability of bikes. The crucial part is the prediction of bike count required at each hour for the stable supply of rental bikes.

**A rule-based model for Seoul Bike sharing demand prediction using weather data** [1]uses the two datasets - Seoul Bike and Capital Bikeshare program. It presents results of 5 statistical models that were trained with optimized hyperparameters using a repeated cross validation approach and testing set is used for evaluation: (a) CUBIST (b) Regularized Random Forest (c) Classification and Regression Trees (d) K Nearest Neighbour (e) Conditional Inference Tree. Multiple evaluation metrics such as R2, Root Mean Squared Error, Mean Absolute Error and Coefficient of Variation were used to measure the prediction performance of the regression models. The results show that the rule-based model CUBIST was able to explain about 95% of the Variance (R2) in the testing set of Seoul Bike. An analysis with variable importance was carried to analyse the most significant variables for all the models developed which show that Temperature and Hour of the day are the most influential variables in the hourly rental bike demand prediction.

The Cubist model is introduced on the basis of Quinlan’s M5 model tree that creates a series of “if-after-after” rules where each rule has an associated linear multivariate model. The benefit of the Cubist model is that it is a viable method of regression and can be applied to a variety of issues.

In **Regression Model to Predict Bike Sharing Demand** [2]they applied regression model on the Seoul biking dataset which gave them a value of 0.567 suggesting that the linear regression was able to determine 56.7% of changes in the Rental Bike Count. The research paper states that regression models with low R-squared values can be perfectly good models for several reasons as some fields of study have an inherently greater amount of unexplainable variation. In these areas, R2 values are bound to be lower.

If one has a low R-squared value but the independent variables are statistically significant, one can still draw important conclusions about the relationships between the variables. As observed in the paper, 0.56 is a relatively low value but statistical significance aids to understand the factors affecting the Rental Bike Count better. Further scope to extract better results and patterns from the datasets involves implementation of advanced algorithms like Classification Trees, Random Forest, K Nearest Neighbours.

In this project, we aim to use statistics to determine which factors lead to optimal supply of bikes rented at each hour. For example, if the temperature is too high and visibility is poor, we will supply a smaller number of bikes as the demand will be less. If our model is successful, it will allow bike owners and managers to decide the count of bikes to be rented at any given time. Bike users will also experience a friendly and comfortable service.

**Data Description**

The data used in the report was acquired through UCI Machine Learning Repository [3].

The dataset contains weather information (Temperature, Humidity, Windspeed, Visibility, Dewpoint, Solar radiation, Snowfall, Rainfall), the number of bikes rented per hour and date information.

It has 8760 instances and 14 attributes. Rented Bike count which is the Count of bikes rented at each hour is our response variable and our 13 predictors are –

Date - year-month-day

Hour - Hour of the day

Temperature-Temperature in Celsius

Humidity - %

Windspeed - m/s

Visibility - 10m

Dew point temperature – Celsius

Solar radiation - MJ/m2

Rainfall – mm

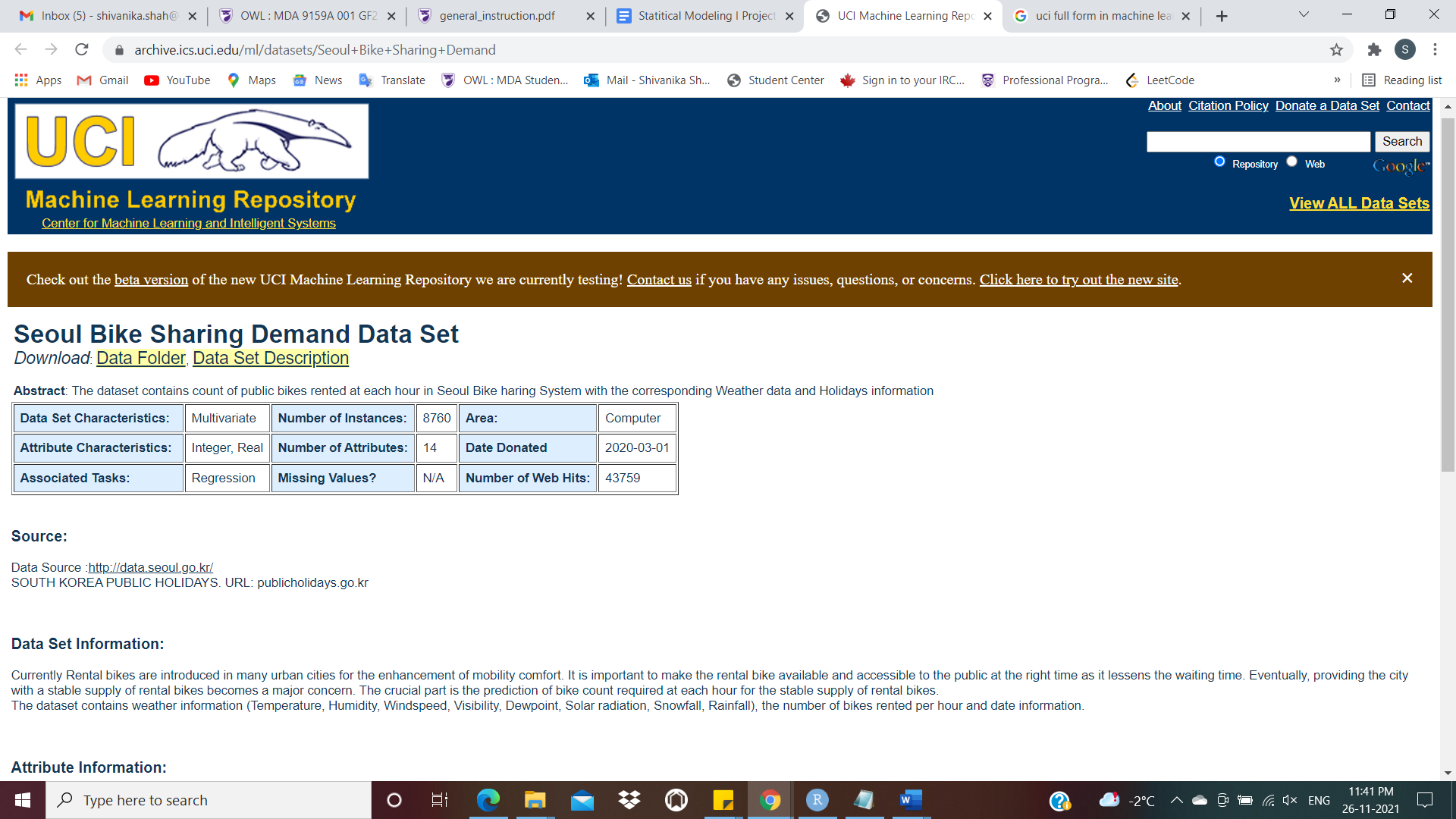
Snowfall – cm

Seasons - Winter, Spring, Summer, Autumn

Holiday - Holiday/No holiday

Functional Day - NoFunc(Non Functional Hours), Fun(Functional hours)

The dataset is of multivariate characteristic with real and integer type values.



**Data Preprocessing**

Since the Date is not useful in making the analysis, the column, “DATE” is removed. The dataset does not contain any missing or NA values and there are no duplicate rows in the dataset either. There are categorical columns, “Seasons”, “Holiday” and “FunctionalDay”, these columns are factor predictor variables. We have further renamed our column names for simplifying the operations in analysis.

Next, we created a boxplot of our numeric predictors after normalizing our data to analyze the outliers in our data. We analyze that the data has some outliers, but we don’t remove them as they are all valid as the weather parameters can be of any range.



**Data Exploration**

Several plots are created to obtain a better understanding of the dataset. First, to understand the relationship between the response and predictor variables, we plotted the scatter plot.

Diagram

Description automatically generated

From the plots, there does not seem to exist clear relationship between the response variable and the predictor variables, however, slightly exponential relation can be observed of response variable with snowfall and rainfall. In general, it is quite hard to determine the exact nature of the relationship between the predictors.

Then, to avoid high VIF values that would cause us to draw poor conclusions about our beta coefficients, it is important to understand which predictor variables are highly correlated. The correlation plot pictured below help us to attain this goal.

Chart, bubble chart

Description automatically generated

From the plot, it can be seen that the correlation between certain terms like “DewPointTemp” and “Temperature” is very high whereas for other predictors, the correlation isn’t that significant. Since the dew point is the temperature to which air must be cooled to become saturated with water vapor and its relationship can be clearly represented with temperature, we will remove “DewPointTemp” from our model, but before that we will use variable selection process to choose the optimum variables for our model.

Graphical user interface, application, table, Excel

Description automatically generatedTable 1: Correlation Values

**Variable Selection Process for Modelling**

First, we split the data into a training set and a testing set. 80% of the data is put in the training set and is used for training the model and the other 20% is put in the testing set and is used to check the performance of the model on unseen data.

Then we started our analysis over the full model consisting of all the predictors. This model had an R-square **0.5502** and Adjusted R square of **0.5493.**

In order to choose the best subset of predictor variables for our final model to predict the count of rented bikes in Seoul, we applied methods like Forward AIC, Backward AIC, Forward BIC, Backward BIC along with Stepwise selection methods

The model that gave us the best trade-off between adjusted R-squared and simplicity was the one that was created by stepwise selection method with AIC as validation. Here are its performance metrices:

Multiple R-squared**: 0.5502**, Adjusted R-squared**: 0.5494**

We choose this model as our base model and then try to improve upon it.

**Model Building and Comparisons**

We further explore a model that suits the dataset the best and might be one that has interaction terms between different predictors and from the initial correlation plot it is evident that there exists some collinearity among the predictor variables. If there is high multicollinearity between the predictor variables, then it can cause problems when we fit the model and interpret the results.

Next, we created a model with two-way interaction terms between some predictors that we think might be important to consider together in order to better predict the outcome. We introduce 2 way interactions between Temperature:Humidity, Visibility:Hour ,Seasons:WindSpeed + SolarRad:Temperature. After introducing the interaction terms, we find that the variables Snowfall and Windspeed lose significance. We confirm this with an ANOVA test and drop them for our model. This model had an R-squared of **0.58** and an Adjusted R-squared of **0.5789.**

Next, we introduce a 3-way interaction term between Temperature,Solar Radiation and Humidity. This model has an R-squared of **0.5823** and Adjusted R-squared of **0.5811.**

In this model, even though all the interactions are significant, the VIF(Variance Inflation Factor) values for Temperature, Seasons and DewPoint is very high.

GVIF Df GVIF^(1/(2\*Df))

Temperature 128.553235 1 11.338132

Hour 3.379461 1 1.838331

FunctioningDay 1.082236 1 1.040306

Humidity 23.051869 1 4.801236

Seasons 116.354320 3 2.209516

Rainfall 1.133475 1 1.064648

SolarRad 7.056754 1 2.656455

Holiday 1.023534 1 1.011698

DewPointTemp 142.330380 1 11.930230

Temperature:Humidity 22.742137 1 4.768872

Hour:Visibility 4.121795 1 2.030220

Seasons:WindSpeed 90.585514 4 1.756436

Temperature:SolarRad 29.032960 1 5.388224

Temperature:Humidity:SolarRad 19.984112 1 4.470359

We create a model which does not have the variables having extremely high VIF values.

For this model, the R-squared**: 0.5557**, Adjusted R-squared**: 0.5549.** The VIF values have also gone down and are in the acceptable range.

GVIF Df GVIF^(1/(2\*Df))

Temperature 3.153255 1 1.775741

Hour 3.207939 1 1.791072

FunctioningDay 1.057036 1 1.028122

Humidity 2.324459 1 1.524618

Rainfall 1.086260 1 1.042238

SolarRad 5.251761 1 2.291672

Holiday 1.013793 1 1.006873

Hour:Visibility 4.001838 1 2.000459

Seasons:WindSpeed 4.174282 4 1.195564

Temperature:Humidity:SolarRad 5.042972 1 2.245656

We now take this as the base model and validate model diagnostics. Following are the plots for residuals vs fitted values, and the Q-Q plot. None of the assumptions- Linearity, Equal Variance and Normality hold. We confirm the same doing bptest and Shapiro test.

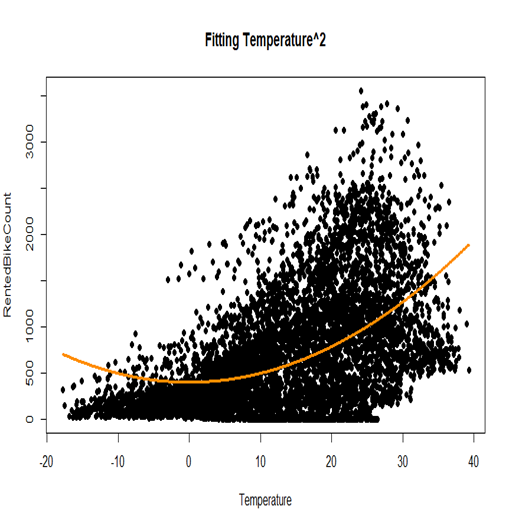
Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

We decide to transform the response variable (Bike Counts) using BOX COX transformation to see if the model meets assumptions of linear regression. Note, we added a very small constant to the response variable for this transformation because there were quite a number of 0 values for this variable.

After transformation, the model does not meet any of the linear regression assumptions.

When further analyzing the data, we wanted to determine which variables would benefit from having higher-order terms. We made plots of each variable and tested whether adding polynomial terms would improve the model fit.Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generated

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

We found that the above higher order terms fit the pattern of the data better than a linear model, so we decided to include these terms in our model, along with the interaction terms included earlier. For this model Multiple R-squared: **0.5714**, Adjusted R-squared**: 0.5702.**

The VIF is very high for seasons, solar^2 and solar ^3.

GVIF Df GVIF^(1/(2\*Df))

I(Temperature^3) 7.636749 1 2.763467

Hour 3.366128 1 1.834701

FunctioningDay 1.081136 1 1.039777

Humidity 3.495862 1 1.869722

Seasons 119.937865 3 2.220714

I(exp(-Rainfall)) 1.243638 1 1.115185

SolarRad 52.064587 1 7.215579

I(SolarRad^2) 277.835863 1 16.668409

I(SolarRad^3) 114.101750 1 10.681842

Holiday 1.020182 1 1.010040

Humidity:Temperature 7.932875 1 2.816536

Hour:Visibility 4.186435 1 2.046078

Seasons:WindSpeed 89.696060 4 1.754271

SolarRad:Temperature 34.319335 1 5.858271

Humidity:SolarRad:Temperature 18.016814 1 4.244622

We remove these predictors, and our new model has much better VIFs, although Interaction terms with temperature and solar radiation still have quite VIF. R-squared: **0.5432,** Adjusted R-squared**: 0.5423.**  This model has all significant predictors and acceptable VIFs.

> vif(model\_transform1\_noVIF)

GVIF Df GVIF^(1/(2\*Df))

I(Temperature^3) 6.746134 1 2.597332

Hour 3.226740 1 1.796313

FunctioningDay 1.058692 1 1.028928

Humidity 3.354377 1 1.831496

I(exp(-Rainfall)) 1.233268 1 1.110526

SolarRad 7.520018 1 2.742265

Holiday 1.013020 1 1.006489

Humidity:Temperature 5.532854 1 2.352202

Hour:Visibility 4.077289 1 2.019230

Seasons:WindSpeed 5.129066 4 1.226746

SolarRad:Temperature 22.833466 1 4.778438

Humidity:SolarRad:Temperature 14.443592 1 3.800473

The model assumptions are still violated.

Chart, scatter chart

Description automatically generated Chart, line chart

Description automatically generated

Considering the goodness of fit and the VIF of predictors, we consider the model with interaction terms and no higher order terms to be the best performing so far. We decided to transform the response variable using Log transformation. Although there is an improvement in goodness of fit, the model assumptions are still violated.

We ran LASSO and RIDGE regressions to check if any of the variables from our full base model would be dropped or the coefficients will be minimised, but there was no remarkable improvement.

For LASSO, the best value of Lambda was found to be at **0.3880575.** No variables were dropped.

A picture containing graphical user interface

Description automatically generated

Since we do not have a lot of predictors, we also tried the RIDGE regression approach, the best Lambda was found to be 34.54546, but again there was no significant improvement in model metrics, so we do not build on this approach.

**Results and Inferences**

So, at this point, we have 4 different models -

1. Full Model (with no interactions or variable transformation)
2. Model with only interactions
3. Model with Interactions and Higher Order Predictor variables
4. Model with Log transformed response and interaction between predictors.

We do a comparative analysis on different metrics of these models and here are the results:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Regression Results** | | | | |
|  | **Before Interactions** | **With Interactions** | **Interactions + Polynomial** | **Log Transformed** |
| **AIC** | 105020.35 | 104979.56 | 105158.22 | 15608.81 |
| **BIC** | 105123.17 | 105082.38 | 105274.75 | 15711.63 |
| **Adjusted R Sq** | 0.54764 | 0.5502649 | 0.5387834 | 0.7797208 |
| **R Sq** | 0.5484793 | 0.5510993 | 0.5397707 | 0.7801295 |
| **Train Error (RMSE)** | 433.465 | 432.2055 | 437.6252 | 0.7353292 |
| **Test Error (RMSE)** | 428.8569 | 424.9944 | 432.3553 | 953.6677 |
| **PRESS** | 434.4708 | 433.2034 | 438.5918 | 0.7393001 |

Table 2: Results

We see that although the log transformed model performs better in almost all aspects, from goodness of fit and Training Error- it also has extremely **high Test Error**, which makes it **unsuitable for prediction purposes,** as there is clearly some overfitting going on in the model.

Based on this, we choose the **Model with Interactions** as our final model, with the caveat that it has moderate goodness of fit and does not meet model assumptions.

**Variable Importance**

The following plot shows the most important variables in a descending order for our final model.

Chart, funnel chart

Description automatically generated

For all our other models too, the selected variables are in line with the variables selected by our final model.

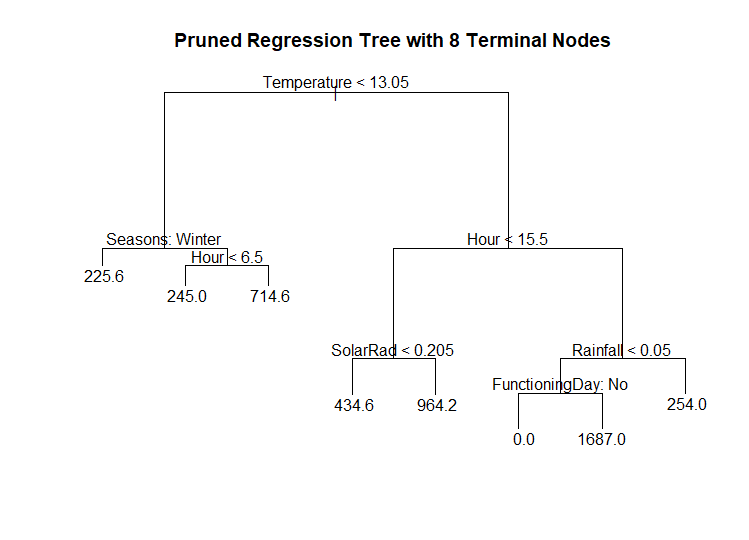
**Approach to Improve the Results: Addendum-Tree based Model**

We found that we were not able to create a model that meets all model assumptions, so we decided to try a tree-based approach. The initial unpruned tree uses 14 terminal nodes and 7 predictors. We prune the tree and find that from the graph, the best RMSE is given at 8 and 13 leaf nodes. At tree size = 13, it will lead to overfitting so we choose 8 leaf nodes.

Chart, line chart

Description automatically generated

The pruned tree with 8 leaf nodes is as follows:



With this tree-based approach, we obtain a Train Error of **389.0871** and a Test Error of **395.0405** which we think is quite an improvement from our previous models.

**Limitations**

The major limitation or drawback that our model has is that it does not meet the assumptions of a linear model. We believe that had we had more data (a larger sample) we would have been able to make a higher quality model.

Also, there were many observations where the bike count was 0, we believe that if we had more samples of data that had non-zero values for bike counts, the model would have had better prediction power.

**Appendix**

1. Data File:



1. R File:



**References**

1. Sathishkumar V E & Yongyun Cho (2020) A rule-based model for Seoul Bike sharing demand prediction using weather data, European Journal of Remote Sensing, 53:sup1, 166-183, DOI: 10.1080/22797254.2020.1725789
2. Aditya Singh Kashyap & Swastika Swastik (2021), “Regression Model to Predict Bike Sharing Demand”, International Journal of Innovative Science and Research Technology, Volume 6 Issue 3, ISSN No:2456-2165
3. <https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand>