Caitlin Alano

Professor Mentch

STAT 1361

19 April 2022

Technical Report

Introduction/Exploratory Data Analysis

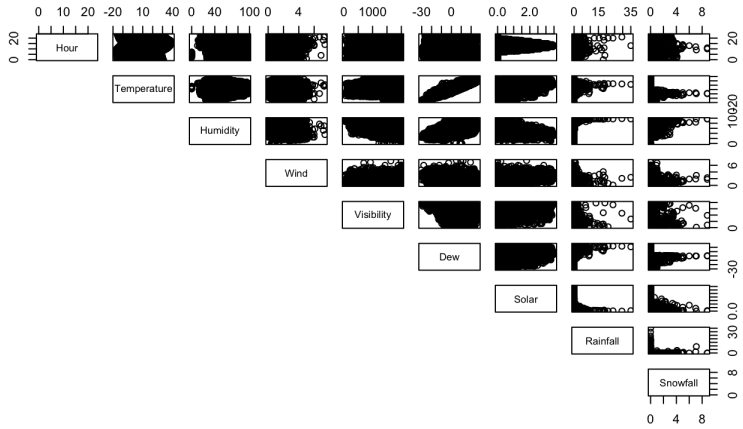
Introduction

The goal of this project is the predictive modeling of demand in the form of the response variable of rental bike counts for the bike sharing system, using all or a subset of the other variables in the dataset to be determined through exploratory data analysis and model selection techniques. The dataset that will be used for training and building the predictive model contains 6552 observations including the numeric type response variable rented bike count, as well as five variables of character type. Seventy percent of this dataset will be used for training the models and thirty percent will be used for testing the models.

Exploratory Data Analysis

Primarily, a scatterplot of the possible combinations of numeric-type predictors were obtained along with a correlation matrix. From looking at both the scatterplot matrix and correlation matrix, it appears that the Temperature and Dew variables could exhibit collinearity if both are present as explanatory variables in a predictive model, as they have a strong positive correlation of 0.9158 between them. Thus, it is best to only include one of them in the model in order to avoid this. See the Scatterplot matrix in Figure 1 below.

Figure 1: *Scatterplot Matrix of Possible Numeric Predictors: Illustrating Collinearity*



As mostly the numeric-type variables were able to be analyzed via these matrices, these nine numeric variables will be further analyzed during model selection to determine the best combination of these variables in explaining rented bike counts. The character-type variables will also be analyzed through other methods. See Table 1 below for a detailed breakdown of summary statistics for the numeric-type variables in this dataset.

Table 1: *Breakdown of Summary Statistics for Numeric Variables: Identifying Data Oddities*

Variable	Minimum	Q1	Median	Mean	Q3	Maximum
Count	0	189	492	702.9	1062	3556
Hour	0	5.75	11.5	11.5	17.25	23
Temperature	-17.8	3	13.6	12.59	22.30	39.40
Humidity	0	42	57	58.15	74	98
Wind	0	0.9	1.55	1.764	2.400	7.400
Visibility	27	970.8	1708	1448.8	2000	2000
Dew	-30.600	-5.425	4.2	3.799	14.5	26.8
Solar	0	0	0.01	0.5727	0.9400	3.5200
Rainfall	0	0	0	0.1572	0	35
Snowfall	0	0	0	0.08365	0	8.8

Data Oddities

From the summary statistics displayed in Table 1 above, it initially seems that all of the numeric variables are fairly normal due to the median values for each variable being relatively

close to their respective mean values. However, it appears that a lot of observations contain zeroes for the variables. Solar, Rainfall, and Snowfall, as their first quartile and median values are zero. This is something to keep in mind for future reference later on.

Methods Overview/Details

Best Subset Selection

First, the method of best subset selection will be performed by fitting separate least square regressions for each possible combination of the nine numeric-type predictors that were analyzed in the previous section to predict rented bike count in the training dataset as well as for the twelve predictors including character-type predictors. The best model for each subset size will be obtained, and the lowest test MSE model for this method is obtained by using the which.min() function to obtain the best sized model.

Ridge Regression

An alternative method that utilizes all predictors is one that regularizes the coefficient estimates by shrinking them towards zero, which can significantly reduce their variance. The first shrinkage method of ridge regression will be performed by minimizing the sum of the residual sum of squares (RSS) and a shrinkage penalty that is small when the coefficient estimates for predictors approach zero. Cross validation will be used to choose the tuning parameter, lambda, which serves to control the relative impact of these two terms on the regression coefficient estimates. This method will first be performed on only the numeric-type predictors, but afterwards, it will be performed on all the predictors, including character-type ones with the exception of Date and ID, as these are irrelevant in the form of dummy variables.

Lasso: Least Absolute Shrinkage and Selection Operator

The second shrinkage method, Lasso, will be performed in a similar manner to ridge regression but instead the lasso penalty forces some of the coefficient estimates to be exactly equal to zero when the tuning parameter, lamda, is large enough and thus is similar to best subset selection in terms of variable selection. Cross validation will also be used to choose the tuning parameter, lambda. This method will first be performed on the numeric-type predictors, but after, it will be performed on all predictors, including character-type ones except Date and ID.

Random Forests

Because decision trees lack stability and could change structure almost completely from slightly differing training data, a random forest, which is a group of randomly made decision trees, will be utilized. This will be done by making a group of decision trees by performing bootstrapping, or sampling with replacement, on the training dataset. The predictions for all the decision trees will be averaged to obtain that of the random forest. For interpretation purposes, variable importance will be measured in relation to the mean squared error (MSE).

Summary of Results

In the models obtained from best subset selection, Visibility and Dew were removed, while in models obtained through ridge regression and lasso, Hour and Temperature had the largest coefficient values. In ridge regression with all variables, Functioning has the largest coefficient value, while in random forests Functioning has the second highest percent increase in MSE. More detail on outputs is presented in Table 2 below.

Table 2: *Model, Test MSE (Best is Bolded), Model Size and Measures of Variable Importance*

Model (Predictor Type)	Test MSE	Model Size	Largest Coefficient Values	Removed Variables	% Inc MSE
Best Subsets (Numeric Only)	208619.4	6	N/A	Wind Visibility Dew	N/A
Ridge Regression	208514.2	9	Hour: 26.146	N/A	N/A

(Numeric Only)			Temperature: 21.093		
Lasso (Numeric Only)	208500.9	9	Hour: 26.475 Temperature: 32.113	N/A	N/A
Best Subsets (All)	173011.6	12	N/A	Visibility Dew	N/A
Ridge Regression (All)	172482.2	12	Hour: 26.731 No Holiday: 98.668 Functioning: 912.815	N/A	N/A
Lasso (All)	172865.7	12	Hour: 27.027 Temperature: 26.806	N/A	N/A
Random Forests (All)	**47374.86**	12	N/A	N/A	Hour: 108.781 Functioning: 91.209

Conclusions/Takeaways

Due to the consistent results across models, it appears that random forests is the best model, with its test MSE value of 47374.86 being significantly lower than all of the other models. Since Hour and Temperature often had the largest coefficient values, it is safe to conclude that these two variables are very important in the prediction of rental bike counts. Functioning also appears to be an important predictor for rental bike counts due to its outputs from random forests and ridge regression with all predictors. From the initial exploratory data analysis that revealed a correlation between Temperature and Dew, as well as its removal from both final best subsets selection models, it is fair to conclude that Dew does not have a significant influence when predicting rental bike counts. For future analysis, it is important to investigate Holiday and Functioning more to see if they truly have a significant influence on rental bike counts.