Technical Report
Chris Park

Introduction and Description of Exploratory Analysis:

After being tasked to predict and model rental bike counts for the bike sharing system (to properly balance the available supply), a thorough analysis on the data (train.csv) provided was necessary. First, the data set was checked to see if there were any missing values because it can be influential in making inaccurate assumptions about the data (there were no missing values). Next, it was searched to see if every instance/row of the data was unique so that there were no redundant data points that could leverage our predictions (there were no identical points). Once the data set was assumed to be full and of quality, the data types of the columns were explored. After realizing the categorical predictors were character types, they were converted into factor types because they have a fixed and known set of possible values. However, the Date column was converted from a character type to a Date type.

A numerical summary of all the predictors was run to see any data oddities. There seemed to be outliers in Count, Visibility, Solar, Rainfall, and Snowfall. However these instances were kept because they did not seem like high leverage points which will be explained in further sections of this report. Next, the correlations of the predictor variables to the response were explored in which Temperature (0.5374), Hour (0.4049), and Dew (0.3768). As a hypothesis, these three predictors have a high chance of being the more important predictors compared to others. Additionally, there were some instances of possible collinearity between Dew and Temperature (0.9158), Visibility and Humidity (-0.5582), and Dew and Humidity (0.5429).

Scatter plots were utilized to visualize continuous relationships with Count and boxplots were used to visualize categorical data. In the scatter plots, there were certain predictors that displayed non-linear relationships such as Rainfall and Snowfall. As for the boxplots, it displayed higher Count on Non-holidays, Summer, and Functioning days. Lastly, histograms were used to visualize the shape of the data. Predictors that showed to be approximately normal were Temperature and Humidity while Wind, Solar, Rainfall, and Snowfall were skewed

to the right and Visibility and Dew were skewed to the left. Also, Date displayed a uniform distribution overall.

Before creating models that would be best at making predictions, the data from train.csv was split into another training and test set with a 70-30 split. Also, a multiple linear regression model including all the columns except ID was run to see if more assumptions about the data could be made through the diagnostic plots. This model yielded a test MSE of 184986.3, adjusted R^2 value of 0.5569, and showed that all predictors were significant except Visibility and Dew. The residual vs. fitted plot showed that the relationship between the predictors and response have non-linear relationships because it did not display a horizontal line. The normal Q-Q plot displayed linearity (normal distribution) until it reached the theoretical quantile of ~1.7 which means that the residual points are not normally distributed. The scale-location plot showed heteroscedasticity (data points have standardized residuals greater than other data points). Lastly, the residuals vs. leverage plot showed that there are no observations with unusually high leverage points because none of the data points were above the Cook's distance threshold.

To further assess the non-linearity of the data, polynomial regression was done to compare to the normal linear model. All the categorical variables were to the 5th degree as anything above the 5th degree is dangerous of overfitting. There were instances where a degree higher than one was more significant than the original such as Temperature, Humidity, Wind, Visibility, Dew, Solar, and Rainfall. Next, an ANOVA test was done which showed that the second model (model with polynomials) was significant. The ANOVA for nonparametrics effects show that Hour, Temperature, Humidity, Wind, Visibility, Dew, Solar, and Rainfall are not adequate with a linear function. The information gained by these plots and the ANOVA test made it reasonable to transform the predictors and response variables as the next step.

Predictors that were skewed and the response variable were transformed. Square root transformation was applied to Count, Wind, Solar, Rainfall, and Snowfall while square transformation was applied to Visibility. After applying these transformations, some variables seemed to have more of a normal distribution while others seemed to still be skewed (but less skewed than before transformation).

Methods Overview/Details

Predictive accuracy for all models was measured by the test MSE because it displays how well a model predicts using unseen data. The first model that was done was a multiple linear regression model because it is easy to interpret and implement. The model summary showed that all the predictors were significant except for Visibility, Dew, Solar, and Snowfall. Also, the adjusted R^2 value was 0.6835 and the test MSE was 49.8523 which is a huge improvement from the initial model from the exploratory data analysis. A permutation test was done to further explore the significance of the predictors through test MSE. In this test, the predictors that came up to be insignificant in the model were sampled. It was discovered that these predictors are actually significant which can be seen by the majority of the data in the histogram being greater than the calculated test MSE from the previous model. Also, the p-value for this test was 0.01 which is evidence to reject the null that these predictors are insignificant. This gave insight into the possibility of some predictors having an interaction effect.

An identical model was run with interaction terms of Dew*Visibility, Solar*Snowfall, Solar*Visibility, and Dew*Solar. These interaction terms came out to be overall significant and yielded a test MSE of 49.16153 and adjusted R^2 value of 0.6855 which is barely better than the model before. Not only that, but it uses more degrees of freedom for only an insignificant amount of improvement (more prone to overfit).

Best subset selection was utilized to see which predictors are better at predicting the response variable. When it was done on the training set, a subset of 13 predictors was best on

the basis of adjusted R^2 and a subset of 11 predictors was best on the basis of Mallow Cp and BIC. For background information, there are three components of the Seasons variable with one other being the baseline, two components of the Holiday variable with one other being the baseline, and two components of the Functioning variable with one other being the baseline. However, when basing the subsets on the test MSE, the subset of predictors chosen were Date, Hour, Temperature, Humidity, Wind, Visibility, Dew, Solar, Rainfall, Seasons, Holiday, and Functioning. In other words, it chose all the predictors except Snowfall. A final multiple linear regression model was done without the Snowfall predictor which yielded a test MSE value of 49.83781 and adjusted R^2 value of 0.6836 which is only a slight improvement from the original model.

Next, shrinkage/regularization methods were used to see if accepting a bit more bias will greatly reduce the variance. Ridge regression was executed even though the number of observations is way larger than the number of predictors. First, cross-validation was done to find the best tuning parameter (lambda) which could possibly give up some bias for a reduction in bias. The best lambda value came out to be 0.6568708 and the test MSE was 50.87662 which is slightly higher than the previous multiple linear regression model. Lasso regression was done as well with cross-validation choosing the best tuning parameter (lambda) again. The best value was 0.004222414 and the test MSE was 49.93659. Although lasso regression performed better than ridge regression, both did not perform as well as the multiple linear regression model. This makes sense as both tuning parameters were very close to zero meaning that the impact of shrinking penalty is small. Ultimately, accepting more bias for reducing bias was not worth it.

Next, dimension reduction methods were used (good if there is collinearity in the model). First, principal component analysis was done to see if the direction of max variation in the data is the most informative about the response variable. Through cross-validation, it was found that using 13 components was the best which yielded a test MSE value of 49.78935. This is a slight

improvement from the multiple linear regression model after best subset selection. Next, partial least squares was done to see if giving more weight to variables that are more strongly correlated with the response variable will improve the fit. Through cross-validation, it was found that using 8 components was the best which yielded a test MSE value of 50.1541. This is worse than the principal component analysis and also worse than the multiple linear regression model after best subset selection.

Regression splines were fit to have more of a “smooth” continuous model with continuous first and second derivatives (introduce more flexibility and good for polynomial relationships which was confirmed). Specifically, cubic splines with three knots at uniform quantiles were used. This is because cubic models are fairly robust and tend to cover most relationships in practice. The resulting test MSE was 35.48157 and adjusted R^2 value was 0.7728 which is a solid improvement from the previous best model which was the principal component analysis. After, an even more flexible model (GAM) with functions of splines were utilized. This method was used because it does a good job at balancing predictive accuracy with flexibility and interpretability (bias/variance). First, cross-validation was done to select the best degree of freedom value (df = 5). All the predictors were found to be significant and the test MSE was 32.13517 which is an improvement from the regression spline method.

Lastly, tree based models were used. First, a normal regression tree was fitted which utilized Temperature, Hour, Seasons, Functioning, Humidity, Solar, and Rainfall in the tree construction. Also, the resulting test MSE was 41.81327 which is higher than the GAM method used before. A pruned tree was also used after cross-validation which showed that having 11 terminal nodes was best. However, the test MSE was the same as the normal regression tree. Since predictive accuracy is the primary goal, other ensembles of decision trees such as bagging and random forests were executed as well. Not only that, but they provide a reduction in variance and are robust to non-linear data and outliers. Random forest with the default mtry

value of predictors/3 deemed Hour, Functioning, and Humidity as the top three most important predictors and had a test MSE of 14.09744. This is a huge improvement from any of the previous models. A bagged approach was done as well to obtain estimates/predictions with a lower variance (more stable). This method utilized a mtry value that equaled the number of predictors (12). The test MSE was 14.09615 which was only a slight improvement from the normal random forest approach. Also, the top three most important predictors were identical to the normal random forest approach. With bagging, trees are built in the same way with bootstrap samples which means that trees and predictions are highly correlated. The randomness in forests helps to reduce this correlation (reduce the variance at the averaging step. As a result, the bagged approach was selected as the best model to make predictions.

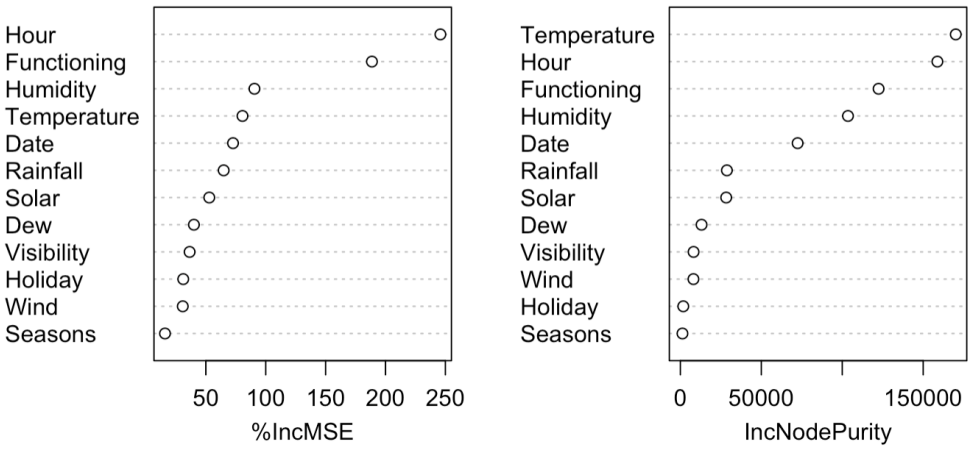

Figure 1: Variable Importance of Best Model

Summary of Results

Overall, the order of the models in terms of predictive accuracy was bagging approach (random forest), random forest, GAM, regression spline, regression tree, multiple linear regression with interaction terms, principal component regression, multiple linear regression with best subset, original multiple linear regression, lasso regression, partial least squares, and ridge regression.

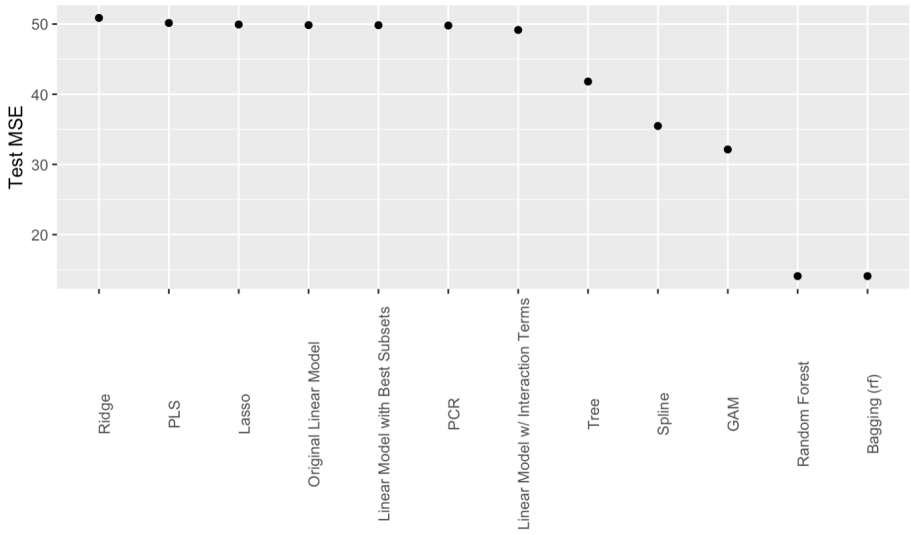

Figure 2: Test MSE of Models Used

It seems that tree-based models performed best, then linear models, then dimension reduction models, then shrinkage/regularization models (Fig. 2).

In terms of variable importance, best subset selection and cross-validation was utilized. Overall, it seemed that Temperature, Hour, Functioning, and Rainfall were the most important predictors. Temperature, Functioning, Rainfall, and Hour had the highest frequency of being selected in the model in best subset selection in regards to adjusted R^2, Mallow Cp, and BIC (Fig. 3).

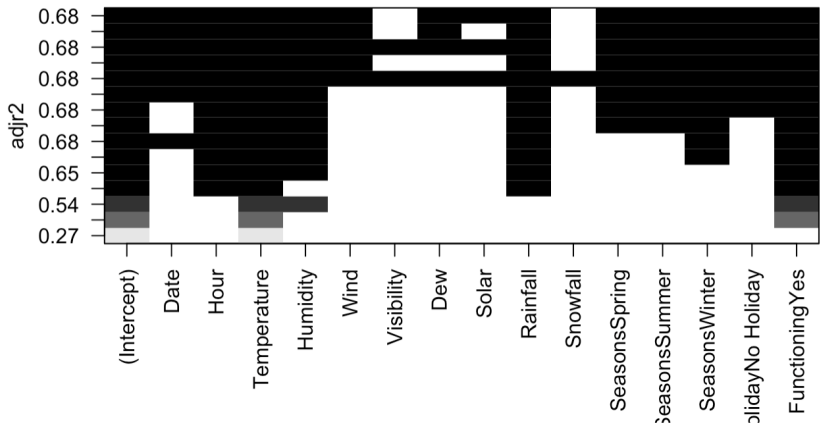

Figure 3: Adjusted R^2 Example

Similarly, the regression tree utilized Temperature, Hour, Functioning, and Rainfall in the construction of the tree. Not only that but the most important predictors in random forest and bagging approach were Hour, Functioning, and Temperature. Lastly, GAM deemed these predictors to be significant in the model. This accepts the initial hypothesis in the exploratory data analysis that the predictors that had the highest correlation with the response variable (Temperature and Hour) would be more important than others.

Conclusion/Takeaways

It can be safely concluded that the predictors provided in this data set does an overall good job in predicting the response variable (Count). Specifically, the higher the temperature, the later into the evening, during functioning hours, and the less rainfall will equate to a higher rental bike count. The most challenging aspect of the dataset was that the predictors were not normal to do certain methods such as linear regression models. This issue was mitigated after the data was transformed to follow certain assumptions that were desired to be kept. Another challenge was to remove outliers or not. In this project, no outliers were removed because it was thought to bring value into the data. This issue was mitigated by utilizing certain models that are robust to outliers. Although the data was pretty well prepared to run models, it could always be better by attempting different splits in the data. For example, the data was split into 70-30. Utilizing 65-25 or even 85-15 splits could have been beneficial to gather more accurate predictions. If there were other information that would improve the final model/predictions further would be to get more data/predictors such as gas prices or traffic. For example, more rental bike counts may be a result of the availability of different transportation methods. As for models/tests, bayesian additive regression trees would be another option to utilize because tree based models seem to work best with this data set. Also, it is a flexible approach to fitting regression models while avoiding strong parametric assumptions.