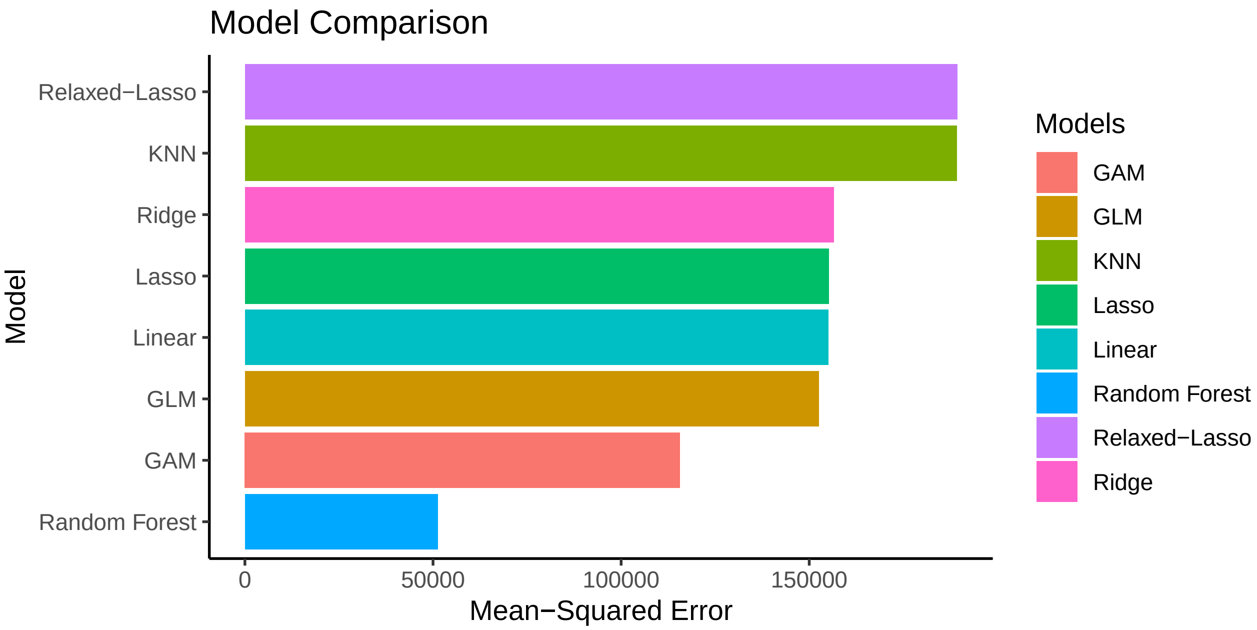
Technical Report

4/14/22

For this project, my goal was to predict and model bike rentals in the city Ourra in order to better understand the demand for bike rentals so that a proper balance of supply could be obtained. The dataset provided consisted of 15 different variables and 6552 observations that each represented the hourly statistics provided by the company. In my exploratory data analysis I first wanted to visualize the two categorical variables I thought might play the biggest role in explaining the bike rental sales, whether or not the given day was a holiday and what was the season. The average sales on a holiday and non-holiday did in fact show significant difference in sales with holidays producing less bike rental sales than non-holidays. Secondly, I found that the sales did not differ significantly between Autumn, Spring, and Summer but that Winter experienced a significant decrease in sales during that season most likely due to the colder weather. I decided that both these variables should be factored into my model but decided that breaking the seasons down into months instead would prove to be a more accurate parameter than just the four general seasons. I used a percentile range of 1% to 99% to determine extreme values and flag them as outliers in the dataset so that they should be ignored in producing the model in attempt to minimize the error on average.

My initial analysis of models produced varying results with the linear model, ridge regression, lasso, and generalized linear model all performing similarly on the testing data and producing average MSE's. Other models such as relaxed-lasso and k-nearest neighbors performed worse than all of the other models and had slightly higher MSE's. By far the most accurate models produced where the random forest models which minimized the MSE significantly compared to the liner models. The outlying model which really intrigued me was the generalized additive model (GAM) since it had a lower MSE than the linear models but also still a relatively high MSE compared to the random forest models. However this could be a more favorable model if were are strongly concerned with variable importance, since we can analyze individual variables and how some non-linear and non-parametric functions may fit with the data.



Across all the models numerous variables were consistently weighted heavier than the others: hour of the day, temperature, humidity, rainfall, holiday, month, and day of the month. There was some inconsistency with the visibility variable because both ridge regression and lasso shrunk the visibility coefficient to zero, while the random forest had a relatively high percentage increase in MSE value and a high increase in node purity value for visibility. Within the GAM model summary, I found that the variables consistent across all models also had clear non-parametric effects according to their low p-values which indicates that the majority of these variables do not have simple linear relationships and explains why the GAM (which uses non-linear trends unlike lasso and the linear model) fit better on the testing data and produced a lower MSE.

This dataset was quite similar to other datasets I've dealt with in the past and so I figured some of the odd variables such as the date and the hour of the day were going to problematic and how I handled them could effect my outcome. I chose to separate the date into three separate variables for the day, the month, and the year and since the dataset only contained data from 2017 and 2018 I decided to not include the year in any of my models, and both month and day consistently appeared as relevant across all of the models. I also suspected that there may be some overlap and colinearity between a few different variables so after further investigation I found a strong linear relationship between temperature and dew point temperature, therefore I chose to exclude the dew variable and include rainfall in the chosen models.

Since random forest was producing the most accurate model I wanted to investigate how accurate the model could be so I used cross-validation in order to determine the most optimal value of the mtry tuning parameter and found that an $mtry = 7$ produced the lowest MSE. As can be seen in the graph below:

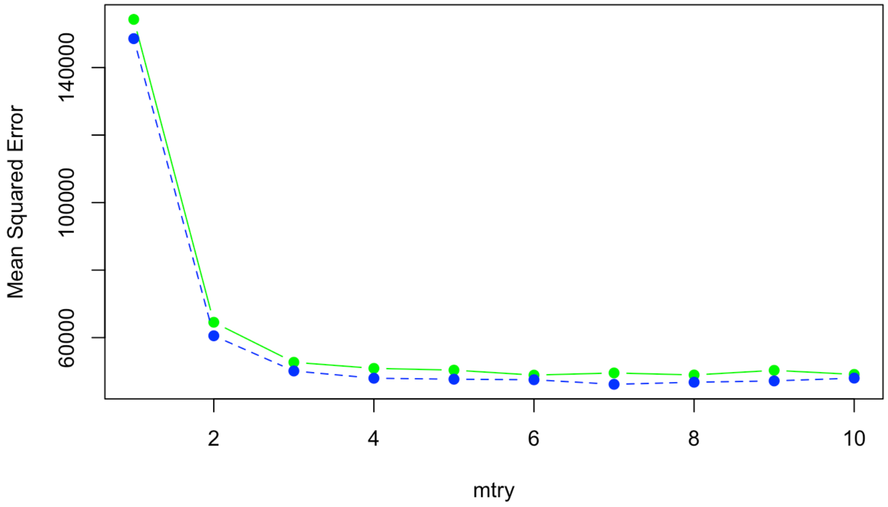
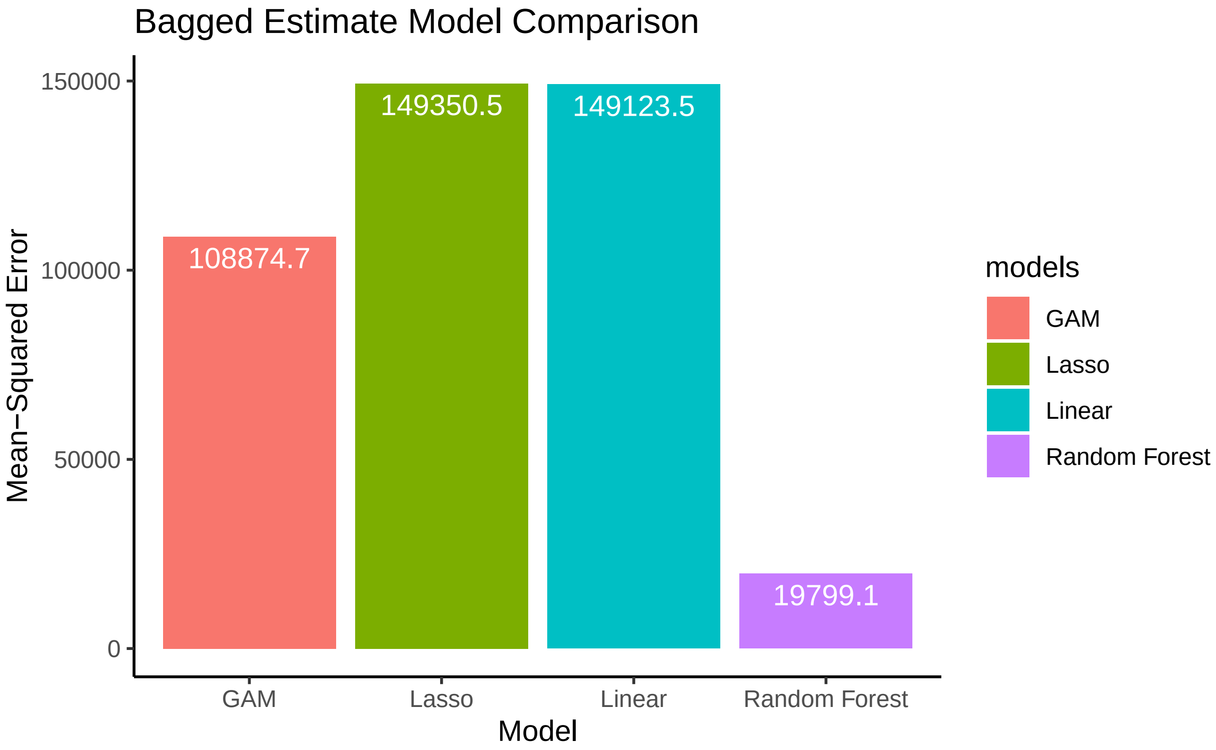


Figure 1: mtry



In my final model analysis I chose to further examine linear regression, lasso, GAM, and random forest to see if a bagged estimate of the model would yield slightly different results and hopefully add bias to the models in attempt to reduce the variance and avoid overfitting, so I calculated the MSE on multiple different testing sets to yield an average MSE for the model overall. What I found was linear, lasso, and GAM performed slightly better than their initial MSE calculation, but random forest managed to perform tremendously better on average than our initial estimate would suggest.

In conclusion, I found that random forest was by far the best in predictive accuracy, but that the GAM model may provide further assistive insight on what variables may be considered important in predicting the amount of bike rentals. The GAM model showed that a simple linear approach would not prove efficient in measuring some variables such visibility, which did not seem significant in the ordinary linear model but was shown to be significant using a fourth degree polynomial. I believe that further individual investigation could be done to try and determine the exact relationship between each predictor and the response using a comprehensive approach that tests multiple different degree polynomial on each predictor in attempt to minimize the MSE of the GAM model. I am very wary of relying the predictive accuracy of the random forest because it is a black-box, unsupervised method and for a company that is trying to learn more about the supply and demand for bike rentals this method isn't as informative as the supervised learning methods such as lasso and GAM. An interesting parameter to consider would be location data within the city and see where people are renting the most bikes and for how long, if we consider the time each person spends on the bikes we can know how many bikes are just sitting at the rental location not being used. If we have a lot of bikes available to rent and people are only using them for short rides then we may not need a large number of bikes in supply, but if people are using the bikes for extended periods of time it would be important to make sure we have sufficient supply. If my job depended on the predictive accuracy of the random forest I would feel very comfortable and if my job depended on the GAM model I would suggest further data and look for areas to improve on but I would still trust the results if we did not have a small margin for error which in this case I think is the truth.