TEAM 2 – WRITE-UP

MODEL DESCRIPTION

Based on our linear regression modeling, we found that an optimal way to come-up with the final model with low RMSE value was to make two models (one for casual and one for registered riders) and then combine those two results to get the final output. We observed that variables were differently correlated across the two models and individually making changes in variable relationship for these two models yielded better results. For example, for registered riders, impact of weather and temperature seems more dominant than that on casual riders.

CASUAL MODEL

#Code in R script. Following are the variables used in writing code for casual riders-

- Casual: This is the dependent variable.
- temp * humidity * hour * daynum * windspeed * workingday: Different variable interactions.
- hourcat: This is hour of the day. There seems to be some seasonality with the demand during the day.
- dayofweekcat: Month of casual usage peaks on Sunday, dips during the weekday, then peaks on Saturday.
- monthcat: Month of demand, casual usage peaks towards middle of the year.
- weathercat: Usage declines as weather condition changes, usage peaks in clear weather, and declines with more extreme weather.
- windspeed: We see that as the windspeed increases, number of casual users decrease.
- We found high multicollinearity among few features such as temp and atemp as well as month and season. We found that eliminating atemp and season led to better results when computing a model for the total amount of users, and this was still true when just used on casual.

REGISTERED MODEL

#Code in R script. Following are the variables used in writing code for registered riders-

- Registered: This is the dependent variable.
- Effect of variables (atemp, humidity, hourcat, holidaycat, daynum, workingdaycat): When these variables interact with each other, there is an increase in adjusted R-squared value. The reasoning behind using this variable interaction is that at a particular temperature, humidity, windspeed, hour of working day, the number of users will increase and decrease depending on their values and their multicollinearity.
- hourcat: This is hour of the day. In our analysis, we saw that for registered users, during hours 0 to 6, there were few users, from hours 7 to 10 & 17 to 19, the number of users increased, and from hours 21 to 24, the number of users decreased. Also, from hours 10 to 16, we see that the number of users is neither too high nor too low.
- dayofweekcat: For registered users, we observed that the number of users is less on weekends (Saturday and Sunday) than that on weekdays.
- weathercat: Number of registered users decline as the weather becomes more extreme.
- windspeed: Number of registered users decline as the windspeed increases.