Homework 3 - BUSN 41204

Fernando Garcia, Paula Gaviria, Victor Fuentes 02/06/2021

Setup

```
library(data.table)
library(ggplot2)
library(kknn)
library(boot)
library(MASS)
library(rpart)
library(rpart.plot)
library(kableExtra)
```

1 Loading Data

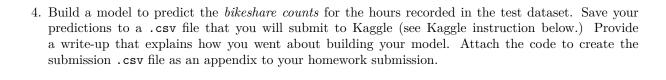
```
setwd("C:/Users/vfuentesc/OneDrive - The University of Chicago/Winter 2021/Machine Learning/Week 3/HW3-I
train = data.table(read.csv("Bike_train.csv"))
test = data.table(read.csv("Bike_test.csv"))
```

2 Questions

- 1. Data Exploration
- a. Visualize the relationship between count and each one of the following variables on a separate scatter plot: windspeed, humidity, temp, and atemp.
- b. How does count depend on the season? Consider visualizing this relationship with a boxplot.
- c. How does count depend on the time of the day (hour)? Does this relationship change depending on whether it is a workingday or not? A scatterplot could be used to visualize the relationship. You might consider coloring the observations on the scatterplot using the temperature (temp or atemp) do discern how the temperature affects hourly number of rentals.
- d. Does the relationship between count and hour change by season?
- e. Does the distribution of hourly number of rentals change between 2011 and 2012? What does this tell you about the rental business?

- 2. Fitting a Random Forest model and a Boosting model
- a. Create a partial dependence plot for predicted count vs each one of the following variables: windspeed, humidity, temp, and atemp. Do this for both the *Random Forest* and the *Boosting* model.
- b. Build a marginal model that regresses count on each one of the following variables: windspeed, humidity, temp, and atemp. You can use whichever nonlinear model you want for these regression tasks. Plot the marginal fits and compare them with partial dependence plots above. How are the plots different and why?
- c. Create variable importance plots for the two models in part 2.a. Do the two models rank the variables in the same way?

- 3. Investigate how predictive each variable is on its own. There are a number of ways to do this. The simplest way would be to regress count on each one of the variables separately and evaluate the out-of-sample MSE for each one of the models. For instance, you could fit a regression tree model.
- a. What can you say by comparing how predictive each variable is on its own vs the variable importance ranking obtained in the previous question?
- b. Why do you think is the reason for the difference?
- c. How could you use the variable importance ranking to select variables? We will talk about this in detail in Week 5. Here I want you to think a bit about the variable selection problem and how would you use the tools that you have learnt so far to identify a good set of variables.



3 Appendix