**Homework #12 (Due Nov 19 11:59 PM)**

IST 3420 - Fall 2017, Chen

**Name**: \_\_\_\_\_\_\_\_\_Adam Forestier\_\_\_\_\_\_\_\_\_\_\_

**Predicting Count of Bike Rentals (30 points)**

**Data**

The data file “DC\_bike\_rental.csv” contains data on bike rental in DC for years 2011 to 2012.

The original data were collected from Capital Bikeshare in DC.

The refined dataset includes the following 10 variables:

* + hour: hourly time
  + season: 1 = spring, 2 = summer, 3 = fall, 4 = winter
  + holiday: whether the day is considered a holiday, 1 = yes, 0 = no
  + workingday: whether the day is neither a weekend nor holiday, 1 = yes, 0 = no
  + weather: 1 = Clear, Few clouds, Partly cloudy, Partly cloudy; 2 = Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist; 3 = Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds; 4 = Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
  + temp: temperature in Celsius
  + atemp: "feels like" temperature in Celsius
  + humidity: relative humidity
  + windspeed: wind speed
  + count: number of total rentals



**Task A: Data Manipulation (10 points)**

1. Read the dataset into R.
2. Show the structure of the dataset.
3. A very important step in predictive analytics is to represent different scales of measurement correctly in the dataset. What variables in the dataset should be represented as factors? List them in the box. (5 points)

|  |
| --- |
| Variables that should be represented as factors:  Hour, season, holiday, workingday, weather |

1. Transform those variables identified in step 3 into factors. (3 points)
2. Use stargazer() function to show summary statistics of the transformed dataset. Past summary statistics in the following box. (2 points)

|  |
| --- |
|  |

**Task B: Predictive Modeling (18 points)**

1. Load caret package.
2. Create a 10-fold cross-validation configuration of control. (2 points)
3. Use the train() function in the caret package to train a multiple linear regression model that predict count based on all other variables using 10 fold cross-validation. (5 points)
4. Print out the trained model. Fill in the performance of the trained liner model in the following box. (2 points)

|  |
| --- |
| RMSE = 110.181  R Squared = 0.630384 |

1. Use the train() function in the caret package to train a GBM model that predict count based on all other variables using 10 fold cross-validation. (5 points)
2. Print out the trained model. Fill in the best performance of the trained GBM model in the following box. (2 points)

|  |
| --- |
| RMSE = 92.56742  R Squared = 0.7495542 |

1. Compare the two predictive models (linear regression and GBM). Which model is more preferable? Why? Explain in the following box. (2 points)

|  |
| --- |
| GBM is the preferable method. Across the five number summary, GBM has higher Rsquared values, and lower RMSE values.  High Rsquared numbers are better because Rsquared is s a statistical measure of how close the data are to the fitted regression line. Lower RMSE numbers are better because it is a measure of the differences between values (sample and population values) predicted by a model or an estimator and the values actually observed. Thus, GBM is the better model because it has higher Rsquared numbers and lower RMSE numbers. |

**R Markdown Code**

Paste your R code in the following box. (1 point)

|  |
| --- |
| Paste your code here:  ---  title: "Homework 12"  author: "Adam Forestier"  date: "November 19, 2017"  output: html\_document  ---  ```{r setup, include=FALSE}  knitr::opts\_chunk$set(echo = TRUE, message = F)  ```  Clear Current Environment  ```{r}  rm(list = ls())  ```  Load Piping package  ```{r}  library(magrittr)  ```  # Read CSV File data  ```{r}  bike\_df <- read.csv("DC\_Bike\_Rentals.csv", header = T)  # Remove na values  bike\_df <- na.omit(bike\_df)  # Explore data  head(bike\_df)  ```  # Show the structure of the data  ```{r}  print(str(bike\_df))  ```  # Transform hour, season, holiday, workingday, & weather into factors  ```{r}  # Choose variables to coerce to factors  factor\_cols <- c("hour", "season", "holiday", "workingday", "weather")  # Convert variables to factors  bike\_df[factor\_cols] <- lapply(bike\_df[factor\_cols], factor)  print(str(bike\_df))  ```    # Show summary statistics of the data set  Load stargazer package to improve appearance of output  ```{r}  library(stargazer)  ```    Show summary statistics output  ```{r}  stargazer(bike\_df)  ```    # Create 10-fold validation configuration of control  Load caret package  ```{r}  library(caret)  ```  Cross-validation  ```{r}  fit\_control <- trainControl(method = "cv", number = 10)  ```    # Train multiple linear regression model to predict count  ```{r}  # Train linear model  multi\_fit <- train(count ~ ., data = bike\_df,  trControl = fit\_control,  method="lm")  # Show results  print(multi\_fit)  ```      # Train GBM model that predicts count  ```{r}  # Train GBM model  GBM\_fit <- train(count ~ ., data = bike\_df,  trControl = fit\_control,  method = "gbm", verbose = F)  print(GBM\_fit)  ```    # Compare the two different predictive models  Resample and show results  ```{r}  results <- resamples(list(linear = multi\_fit, GBM = GBM\_fit))  summary(results)  ``` |

**Homework Submission**

Upload this document with your answers to “Homework 12” on Canvas.

Upload your R Markdown file (for both task 1 and task 2) to “Homework 12” on Canvas. (1 point)