**Homework 10 Programming Part (16 points)**

IST 5535 – Spring 2020, Chen

**Name**: Sayantan Majumdar

**Predicting Count of Bike Rentals**

**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 (the target we want to predict)



**Task A: Data Manipulation (3 points)**

1. Read the dataset into a data frame. Show the structure of the dataset.
2. 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 categorical variables? List them in the box.

|  |
| --- |
| Variables that should be represented as categorical variables: **season, holiday, workingday, weather** |

Transform those variables identified above into factors.

1. Use 30-70% data partition strategy, with 30% of data used as test data.

**Task B: Predictive Modeling (14 points)**

1. Train a regression tree to predict count of bike rental on the training set. Prune the regression tree by using a best parameter identified through a cross-validation. Train the pruned regression model. (3 points)
2. Train a random forest model to predict count of bike rental on the training set. Use a 10-fold cross-validation to run a parameter tuning process to find the optimal value of parameter “mtry”. Fit the final random forest model with the optimal mtry on the whole training set. (3 points)
3. Train a support vector machine model to predict count of bike rental on the training set. The svm() method in the e1071 can also be used for regression problems. Use a 10-fold cross-validation to run a parameter tuning process to find the optimal kernel among linear, RBF, and polynomial kernels. Fit the final SVM model with the optimal kernel on the whole training set. (3 points)
4. Use test dataset to evaluate the performance of your final models. Organize model comparison in a data frame. Fill in all blanks in the following table. (3 points)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model** | **Parameters** | **Performance Indicators** | | |
| RMSE | R2 | MAE |
| Pruned Regression Tree | # of terminal nodes = 13 | 105.04 | 0.65 | 74.09 |
| Final Random Forests | mtry = 9 | 71.10 | 0.84 | 47.10 |
| Final SVM | kernel = Radial (sigma=1, C=1) | 101.02 | 0.69 | 63.59 |

***Question***: What conclusion did you get according to the model comparison? (1 point)

Your Answer: The Random Forests model works best and has the least error metrics. This is followed by SVM. The pruned regression tree performs the worst.

**Homework Submission**

Upload this Word document with your answers to “Homework 10” on Canvas.

Upload your R Markdown and HTML report to “Homework 10” on Canvas.