Case 2: Energy Forecasting (cont'd)

Instructions:

- You can use either Python or R to work on this case.
- No sharing of work. You can work in your teams only
- You are expected to submit a report that summarizes the key steps in your implementation as a flow chart and also submit fully functional code.
- Deadline: 6/27/2016 11.59 PM. Late submissions lose 15% points per day.

The city of Boston has hired you to build a forecasting model to predict their energy usage. In order to do so, you are planning to use multi-linear regression to model Power usage as a function of multiple variables (temperature, day of week, month, weekday, hour of day etc.). You are expected to work on this in three parts. **You will work with the files given in Assignment 1. However, you will only use the NewData.csv and ignore the rawdata.csv.**

Part 1: Algorithm implementation

1. Data wrangling and cleansing and Multiple linear regression

You realize that there are many zeros in the **NewData.csv**. You are wondering how to handle it. Explore 3 methods:

a. Remove all the zero-entries and use only the non-zero entries for KWH to build a regression model (KWH = function of (temperature, day of week, month etc..). What is the in-sample MAPE, RMS and MAE when you apply this model to
    a. Non-zero dataset?
    b. the "raw" (including zeros) data set?
b. You build a model to fill the data. For this, you build a regression model (KWH = function of (day of week, month etc.) without temperature) with non-zero data. You use this model to replace the zeros with the model generated values. You use this data build a regression model (KWH = function of (temperature, day of week, month etc..). What is the in-sample MAPE, RMS and MAE when you apply this model to
    a. "Filled" dataset?
    b. the "raw" (including zeros) data set?
c. You come across a package zoo(https://cran.r-project.org/web/packages/zoo/zoo.pdf) that offers functions na.approx, na.fill, na.locf to fill NAs. You replace the zeros with NA and try using this package to fill the NAs. You use this data build a regression model (KWH = function of (temperature, day of week, month etc..). What is the in-sample MAPE, RMS and MAE when you apply this model to
    a. "Filled" dataset?
    b. the "raw" (including zeros) data set?

Which technique works best for you? Justify your choice. Choose one "Filled" dataset to continue with the rest of the assignment

2. Prediction

Now that you are familiar with Regression Trees and Neural Networks, use these to techniques to build models for prediction of KWH. Note that you will have to aggregate the hours to hourly since the temperature data you have is hourly. You can reuse the code you implemented for Assignment 1 to get the temperature data to create data in the format **sample format.csv. Save it as Hourly_filled_data.csv**

You should have a script that can take any input file in the format **NewData.csv** and outputs the regression tree model and neural network model. The script should also compute the RMS error, MAPE and MAE for your model and output to a text file. (See **PerformanceMetrics.csv** for formats) and save the **Hourly_filled_data.csv**

Note: Temperature must be one of the features.

3. Forecast

You are given a sample **forecastNewData.csv** and **forecastNewData2.csv** files with temperature data. You should write a script that does 2 things:

a. Convert the file to the format in **forecastInput.csv**
b. Use the **Tree model and Neural Network models** generated in step 2 and the **forecastNewData.csv** and **forecastNewData2.csv** files to predict the power usage in KWH for each hour.

Part 2: Classification

In the **Hourly_filled_data.csv file,** compute the average KWH and add a new column, KWH_Class.

If KWH > average KWH, KWH_Class = "Above_Normal" otherwise, KWH_Class = "Optimal".

a. Your goal is to build classification models to predict the KWH_Class variable for each of the following methods.
   i) Logistic Regression
   ii) Classification Tree
   iii) Neural Network
b. You should also compute the overall error rate and the confusion matrix for each method and output it to a text file. **ClassificationPerformancemetrics.csv**
c. Use the **Tree model and Neural Network classification models** and the **forecastNewData.csv** and **forecastNewData2.csv** files to predict the power usage class KWH_Class for each hour.

Submissions:

Report with flowchart and design of regression AND classification models and comments on evaluation metrics for the NewData.csv

Outputs from:

Part 1

1. Data Cleansing: Code and outputs from the three approaches to filling data.
2. Prediction:

   a. Code and prediction output files in the format **PredictionPerformanceMetrics.csv**
   b. Output from your forecast script in the format **forecastOutput_<Account No>_neuralNetwork.csv** and **forecastOutput_<Account No>_regressionTree.csv**
3. Classification:
   a. Code and Regression output files in the format **ClassificationPerformanceMetrics.csv**
   b. Output from your forecast script in the format **forecastOutput_<Account No>_neuralNetworkClassification.csv** and **forecastOutput_<Account No>_ClassificationTree.csv**