Assignment 2

Team 10: Suhas Pirankar, Zalak Shah

Course: Advances Data Science /Architecture

Case 2: Energy Forecasting(cont'd)

Abstract: The city of Boston needs to build a forecasting model to predict their energy usage. The Power usage model needs to build on the multiple parameters i.e. temperature, day of week, month, weekday, hour of day.

Propose Solution:

Built a forecasting model using R in RStudio using the Multi linear regression, Regression Tree, and Neural Networks:

1. Data wrangling and cleansing:

3 Methods are implemented on the NewData.csv for building the models:

1. Remove all the zero-entries and use only the non-zero entries for KWH to build a regression model with following datasets:

a. The values of Non-Zero dataset are as below:

MAPE = 20.40898122
RMS=64.46072535
MAE=47.11087139

b. The values of Raw dataset (inclusive of zeroes) are as below:

MAPE = Inf
RMS=71.09709748
MAE=53.99662692

2.Replace all the zero-entries with model generated values to build a regression model with the following datasets:

a. The values of "Filled" Dataset are as below:

MAPE =11.48256301
RMS = 85.93820539
MAE = 67.89393371

b. The values of "Raw" Dataset are as below:

MAPE = Inf
RMS = 82.76651099
MAE = 60.18904577

3.Use the zoo Package using the na.fill functions to fill NAs for all the zero-entries and build a regression model with the following datasets:

a. The values of "Filled" Dataset are as below:

MAPE =20.10842589
RMS = 66.08014212
MAE = 45.73483077

b. The values of "Raw" Dataset are as below:

MAPE = Inf
RMS = 66.93108274
MAE = 44.70364458

Among all the data sets we extracted and worked about we have selected the 3rd dataset as it has all the values and it is filled and has minimum MAPE, RMS and MAE values.

2. Prediction:

We have used Regression trees and Neural network technique to build our model for prediction. We have aggregated hours to hourly as we have temperature as hourly.

Below are the values for MAPE, RMS and MAE

Tree MAPE	16.45531437
Tree RMS	44.22490702
Tree MAE	25.59298563
Net MAPE	NA
Net RMS	NA
Net MAE	NA

3. Forecast.

We have used Tree model and Neural Network models generated in step 2 and the used them to predict values for kwh for each hour. Refer forecastoutput2.csv

Classification:

We have computed the average KWH and added a new column called KWH class.
If KWH> average , KWH_Class = "Above_Normal"
Else KWH_Class = "Optimal"

We have built a classification models to predict the KWH_Class variable for each of the following methods. Refer files in github

i) Logistic Regression
ii) Classification Tree
iii) Neural Network

Overall error rate and the confusion matrix for each method are stored in file ClassificationPerformancemetrics.csv. Below are the values for the same.

Logistic Error rate =0.071608356

Tree Error rate=0.074880443

We have used Tree model and Neural Network classification models to predict the power usage of KWH_Class for each hour

Flow of Implementation:

- Gathered relevant data
- Removed all the zero-entries and use only the non-zero entries for KWH to build a regression model
- Computed MAPE, RMS and MAE with the Non Zero data set.
- Computed MAPE, RMS and MAE with the Raw data set.
- Build a regression model (KWH = function of (day of week, month etc.) without temperature) with non-zero data
- Used this model to replace the zeros with the model generated values.
- Built a regression model (KWH = function of (temperature, day of week, month etc.)
- Computed MAPE, RMS and MAE with the Non Zero data set.
- Computed MAPE, RMS and MAE with the Raw data set.
- Added package for zoo that offers various functions
- Replaced the zeros with NA and try using this package to fill the NAs

- built a regression model (KWH = function of (temperature, day of week, month etc..).
- Computed MAPE, RMS and MAE with the Non Zero data set.
- Computed MAPE, RMS and MAE with the Raw data set.
- Figured out the best data set based on the values for MAE, RMS and MAPE
- Used regression Trees and Neural Networks techniques to build models for prediction of KWH.
- Aggregated the hours to hourly since the temperature data you have is hourly
- Computed MAPE, RMS and MAE and completed prediction part.
- Used basic data cleaning and formatting and Converted the file to the format in forecastInput.csv
- 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.
- Next step is Classification.
- Computed the average KWH and add a new column, KWH_Class.
- Built a classification models to predict the KWH_Class variable for Logistic Regression
- Built a classification models to predict the KWH_Class variable for Classification tree
- Built a classification models to predict the KWH_Class variable for Neural Network.
- Computed overall error rate and the confusion matrix for Logistic Regression
- Computed overall error rate and the confusion matrix for Classification Tree
- Computed overall error rate and the confusion matrix for Neural Network.
- Used 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