Case 1 - Report

Objective:

The objective of this problem statement is to build a forecasting model to predict the energy usage (Kwh) for city of Boston using multiple linear regression. This report contains the Algorithm Implementation of energy Usage in 3 parts:

1. Data Wrangling and Cleansing
2. Multiple Linear Regression
3. Forecasting

Data Wrangling and Cleansing:

Input: rawData.csv
Output: SampleFormat.csv

Data provided in Headers:

1. Account – Account Number provided, which is constant throughout the file
2. Date – Calendar year of 2014, in the format mm/dd/yyyy
3. Channel – Gives information about the energy usage, we considered Kwh as per requirement
4. X0.05-X24.00 – The raw data consists of power consumption every five min (12 intervals for an hour).This has been aggregated and presented as total consumption every hour in the day (kWh).
5. Month – 0 to 12 (derived from date)
6. Year - 2014(derived from date)
7. Days of week – 0 to 6
8. Weekday/Weekend – 1 for weekday and 0 for weekend
9. Peak Hour – 7AM to 7PM – Peak Hour (1)
10. Temperature data – This is data obtained from wunderground.

Process:

1. The data was reviewed and cleansed using R.
2. The raw data set had data power data in three different Units(kWh, Power Factor and kVarh). This has been fitered and data in kWh has been obtained.
3. energy_df_new <- energy_df[energy_df$Units!= 'Power Factor' & energy_df$Units!= 'kVARh',]

4. The day, month and year data has been obtained from the date field.
5. The power consumption data has been aggregated from an interval of 5 minutes to each hour.
6. energy_hourData <- sapply(seq(5,292,by=12),function(i) rowSums(energy_df_new[,i:(i+11)]))
7. The rbindlist has been used to convert the data in multiple columns into a single row.
8. energy_big_data = rbindlist(energy_datalist)
9. The above dataset has been written the file "Energy_Data.csv"
10. Using the above steps, the data has been filtered to make sure that there was consistency. After preprocessing, no irregularities has been found in the data.

Multiple Linear Regression

Problem Statement:

The objective of this model is to predict the power consumption data using multiple linear regression. This is a supervised learning approach where we have the power data we need to predict. We implement the linear regression technique to evaluate performance metrics and create a model that produces best regression coefficients.

Hypothesis:

$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots \beta_p x_{ip} + \varepsilon_i$ **for** $i = 1,2, \ldots n$.

The r squared value should be closer to 1.0 to ensure that the model is a better predictor of the power consumption. To ensure we get a better value of R square, we used various predictor variables in the dataset and arrived at the following variables.

Dependent variable - kWh

Independent variables – Weekday, Temperature, Peak Hour, Day of Week, Hour

lm.fit = lm(kWh ~ Weekday + Temperature + peakhour + DayOfWeek + hour , data = train)

The root mean square error gives an estimate of the accumulated error in the model.

$$RMSErrors = \sqrt{\frac{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}{n}}$$

RMSE value depends on the dependent variable.

Min – 60.555 Max – 405.97

The closer the RMSE value to the minimum value, better the prediction.

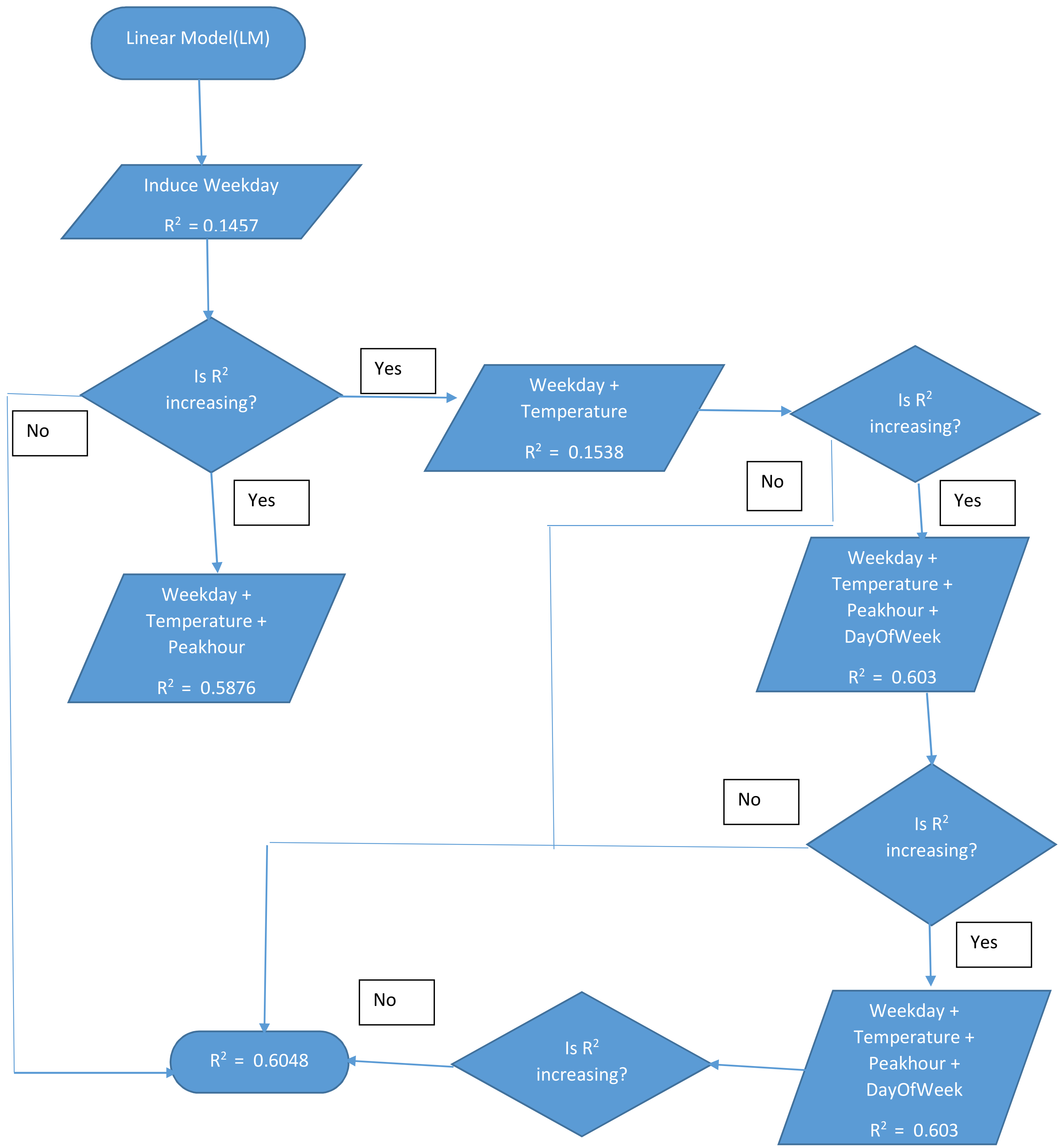
```
Call:
lm(formula = kWh ~ Weekday + Temperature + peakhour + DayOfWeek +
    hour, data = train)

Residuals:
     Min       1Q   Median       3Q      Max
-116.113  -43.334   -0.728   35.884  203.982

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.11753    2.60449  13.483  < 2e-16 ***
Weekday      68.04355    1.45755  46.684  < 2e-16 ***
Temperature   0.02868    0.03619   0.792    0.428
peakhour    110.84343    1.37065  80.869  < 2e-16 ***
DayOfWeek     5.09970    0.32975  15.465  < 2e-16 ***
hour         -0.52200    0.09825  -5.313 1.12e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 51.36 on 6076 degrees of freedom
  (487 observations deleted due to missingness)
Multiple R-squared:  0.6048,	Adjusted R-squared:  0.6045
F-statistic:  1860 on 5 and 6076 DF,  p-value: < 2.2e-16
```

RMSE

```
> predict
                 ME     RMSE      MAE       MPE     MAPE
Test set -0.6874964 50.84861 42.46997 -11.79947 35.07918
```

The same operations have been performed for the data that was provided on June 15

Regression model:



Forecasting

Part 1:

The “forecastInput.csv” has nine variables which has been transformed to three in “forecastData.csv”
Variables: Day, Hour, Temperature

Part 2:

Using the regression variables, we forecast the power data

kwh <- ((m1_weekday*newdata$Weekday) + (m2_temp*newdata$Temperature) (m3_peakhour*newdata$peakhour) + (m4_day_of_week*newdata$DayOfWeek) + (m5_hour*newdata$hour) + constant)

Conclusion:

R square and RMSE are the important predictors in the regression algorithm

Data	R square	RMSE	MAPE	MAE
Raw data	0.60	60.84	35.07	42.46
New data	0.26	87.25	21421.3	62.46