Assignment 3 - Report
Jyotirmayee Mahanandia | Aditya Shinde | Rachitha Dhanraj
INFO 7390: Advance Data Sci/Architecture

Project Proposal:

The main objective is to manage power supply efficiently for a given area. By predicting the electricity usage for the next year, the electricity providers can manage power supply for that particular region for working days and holidays.

We propose to predict the electricity usage over the next year for a particular area

-The area with maximum usage of electricity in kWh during working days

-The area with maximum usage of electricity in kWh during holidays

The behavior of electricity consumption based on the day of the year (Holiday or Working day) for all areas is analyzed by exploratory data analysis on Power BI.

Dataset:

The Datasets consisted of have many records of different areas The attributes were Account, Channel, Unit, Date and Power consumption which is recorded at every five minutes interval for 24 hours. The 3 Units recorded were kWh, Power factor and kVARh.

Data Preprocessing:

1. All csv files were imported onto RStudio and merged based on the areas. We came up with 7 areas in total. They are:
    -School Part I
    -School Part II
    -Agassiz neighborhood
    -Boston Police Department (BPD)
    -Boston Public Library (BPL)
    -Public Work Department (PWD)
    -Property Management (PROP)

2. The datasets were transposed so that we could get all the time interval data in one single column. We used the **reshape** package and **melt** function in R to execute it. Finally there were 6 columns namely Account, Date, Channel, Unit, nth minute and value.

The following is the before and after snap shot of our dataset.

Before:

	A	B	C	D	E	F
1	Account	Date	Channel	Units	nth minute	value
2	26436731017	1/1/14	507115429 1	kWh	0.05	0
3	26436731017	1/1/14	507115429 1	Power Factor	0.05	0.27049
4	26436731017	1/1/14	507115429 2	kVARh	0.05	4.36
5	26436731017	1/1/14	507115429 3	kWh	0.05	2.45
6	26436731017	1/1/14	507115429 4	kVARh	0.05	0
7	26436731017	1/1/14	AGASSIZ SCH	kWh	0.05	0
8	26436731017	1/1/14	AGASSIZ SCH	kVARh	0.05	4.36
9	26436731017	1/2/14	507115429 1	kWh	0.05	0
10	26436731017	1/2/14	507115429 1	Power Factor	0.05	0.271563
11	26436731017	1/2/14	507115429 2	kVARh	0.05	4.43
12	26436731017	1/2/14	507115429 3	kWh	0.05	2.5
13	26436731017	1/2/14	507115429 4	kVARh	0.05	0
14	26436731017	1/2/14	AGASSIZ SCH	kWh	0.05	0

After:

B	C	D	E	F	G	H
Account	Date	Channel	Units	X0.05	X0.10	X0.15
26436731017	1/1/14	507115429 1 kWh	kWh	0	0	0
26436731017	1/1/14	507115429 1 Power Factor	Power Factor	0.27049	0.265678	0.276204
26436731017	1/1/14	507115429 2	kVARh	4.36	4.3	4.28
26436731017	1/1/14	507115429 3	kWh	2.45	2.37	2.46
26436731017	1/1/14	507115429 4	kVARh	0	0	0
26436731017	1/1/14	AGASSIZ SCH 1	kWh	0	0	0
26436731017	1/1/14	AGASSIZ SCH 2	kVARh	4.36	4.3	4.28
26436731017	1/2/14	507115429 1 kWh	kWh	0	0	0
26436731017	1/2/14	507115429 1 Power Factor	Power Factor	0.271563	0.246074	0.252777
26436731017	1/2/14	507115429 2	kVARh	4.43	4.51	4.44
26436731017	1/2/14	507115429 3	kWh	2.5	2.29	2.32
26436731017	1/2/14	507115429 4	kVARh	0	0	0
26436731017	1/2/14	AGASSIZ SCH 1	kWh	0	0	0
26436731017	1/2/14	AGASSIZ SCH 2	kVARh	4.43	4.51	4.44
26436731017	1/3/14	507115429 1 kWh	kWh	0	0	0

3. The NA values were removed and replaced with blank values so that it can be handled in MS Azure ML Studio.

4. The csv files were exported to the Desktop to be uploaded onto MS Azure ML Studio.

The following is a snapshot of the R code used to merge and transpose the files.

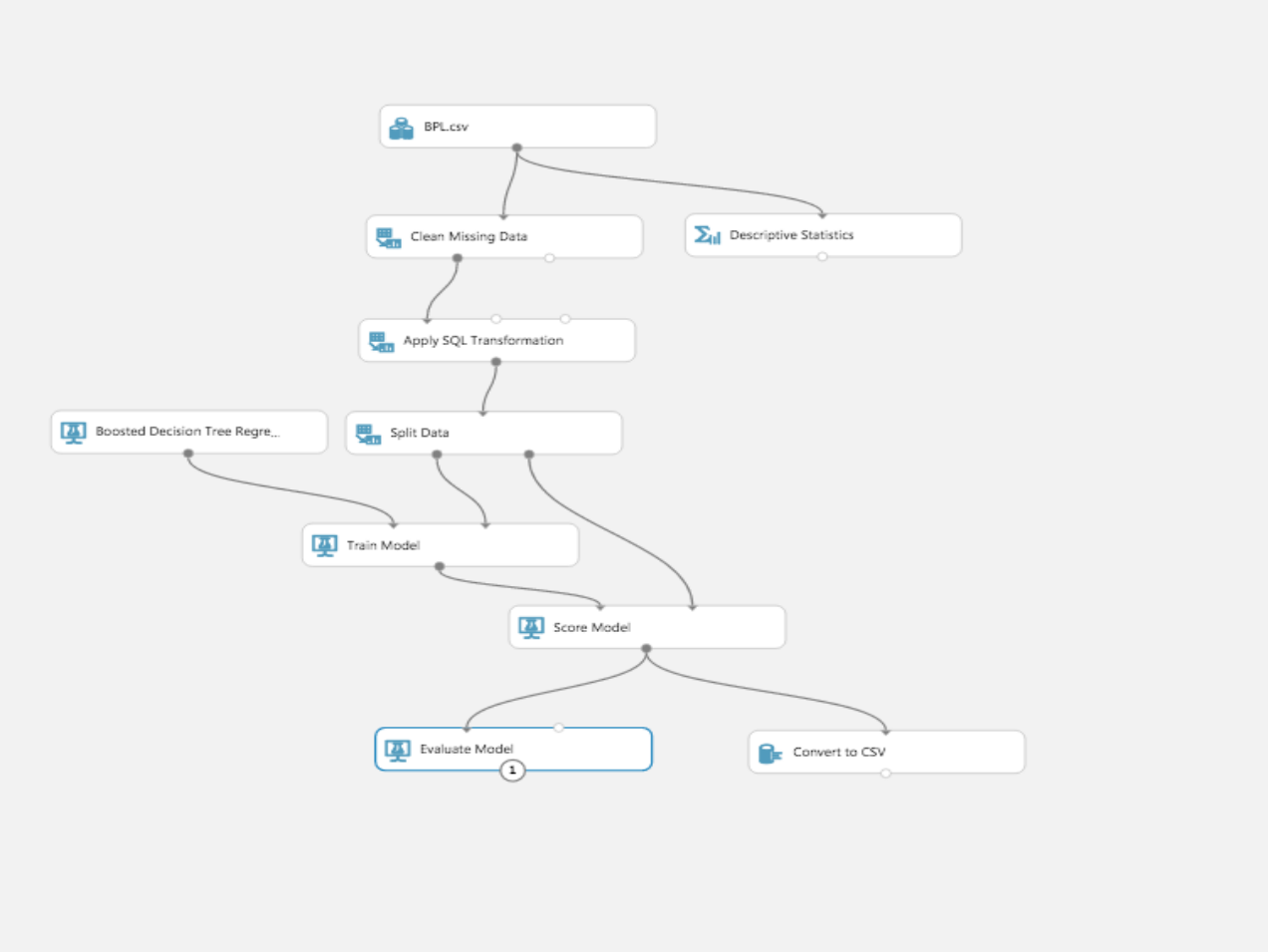
```
#read file
Agassiz1=read.csv("Agassiz Jan-june 2014.csv", header=TRUE)
Agassiz2=read.csv("Agassiz Jul-Dec 2014.csv", header=TRUE)


#merge files
Agassiz <- rbind(Agassiz1, Agassiz2)


#transpose file Agassiz
install.packages("reshape")
library(reshape)
AgassizTranspose <- melt(Agassiz, id=c("Account","Date","Channel","Units"))


#export file to Desktop as csv
write.csv(AgassizTranspose, "AgassizTranspose.csv" )
```

Building the Predictive Modeling: After preprocessing the data, all the csv files were imported to MS Azure ML Studio to build the predictive models for each area.

The following steps were taken for building the predictive model:

1. The missing values were replaced with the mean values using the clean missing data module.
2. We applied the SQL transformation module to extract the records with kWh values.
3. The data was spit into a 70:30 partition where 70% of the data was used for the training model and 30% was used for the scoring the model.
4. We used the boosted decision tree regression algorithm for predicting the column of kWh values.
5. We then evaluated the model based on metrics like Root mean squared Error and Coefficient of Determination.
6. The web service was deployed and the API key was obtained.

The following is the image of the predictive model we used for one of the area namely Boston Public Library.



Predictive Model Evaluation:

We used the following 3 algorithms for building our model:

1. **Boosted Decision Tree Regression**

Trail1 › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	1.085387
Root Mean Squared Error	1.722474
Relative Absolute Error	0.305147
Relative Squared Error	0.134527
Coefficient of Determination	0.865473

2. **Neural Network**

Trail1 › Evaluate Model › Evaluation results

Metrics

Mean Absolute Error	0.208832
Root Mean Squared Error	0.263972
Relative Absolute Error	0.668798
Relative Squared Error	0.632417
Coefficient of Determination	0.367583

3. **Linear Regression**

Metrics

Mean Absolute Error	2.401194
Root Mean Squared Error	3.251176
Relative Absolute Error	0.734228
Relative Squared Error	0.683047
Coefficient of Determination	0.316953

From the above metrics we concurred the following:

Boosted Decision Tree Regression:

Coefficient of Determination (R-squared value)= **0.86**

Linear Regression:

Coefficient of Determination (R-squared value)= **0.31**

Neural Network:

Coefficient of Determination (R-squared value)= **0.36**

- **Boosted Decision Tree Regression was well suited for the dataset** since the Coefficient of Determination **(R-squared value)= 0.86** so, 86% of the variance in the electricity consumption is predictable from the independent variables like Account, Channel, Date and nth minute fed to the model as compared to the other algorithms.
- Also **Boosted Decision Tree Regression is best suited for predicting values that are a non-linear function** of their independent variables.

The prediction model was deployed and the **web API key** was used in R script to confirm the predicted values.

Here is the R script and the output to confirm the prediction of electricity consumption for one of the areas namely, Boston Public Library (BPL):

```
install.packages("RCurl", dependencies = TRUE)
```

```
install.packages("rjson", dependencies = TRUE)

library("RCurl")
library("rjson")

# Accept SSL certificates issued by public Certificate Authorities
options(RCurlOptions = list(cainfo = system.file("CurlSSL", "cacert.pem", package = "RCurl")))

h = basicTextGatherer()
hdr = basicHeaderGatherer()


req = list(

  Inputs = list(


    "input1" = list(
      "ColumnNames" = list("Account", "Date", "Channel", "Units", "nth minute", "value"),
      "Values" = list( list( "26812030018", "6/19/2014", "605106493 1 kWh", "kWh", "2.05", "0.54" ), list( "26438261005", "2/8/2014 ", "605107806 1 kWh", "kWh", "2.35", "36.57" ) )
    ),
  GlobalParameters = setNames(fromJSON('{}'), character(0))
)

body = enc2utf8(toJSON(req))
api_key = "rVr/tzhw64+S+fQs1xYZFUZWCCwjFvnhH+4WCGgRmITz1DkCbvVs1/A5cOytze4j4OQNKc5JFiCiNP2gjyry/g==" # Replace this with the API key for the web service
authz_hdr = paste('Bearer', api_key, sep=' ')

h$reset()
curlPerform(url = "https://ussouthcentral.services.azureml.net/workspaces/7a42d134b6c64c51b80b0f36259de4c0/services/8799f80617664c4c86130a5f882c402b/execute?api-version=2.0&details=true",
        httpheader=c('Content-Type' = "application/json", 'Authorization' = authz_hdr),
        postfields=body,
        writefunction = h$update,
        headerfunction = hdr$update,
        verbose = TRUE
)

headers = hdr$value()
httpStatus = headers["status"]
if (httpStatus >= 400)
{
  print(paste("The request failed with status code:", httpStatus, sep=" "))

  # Print the headers - they include the requert ID and the timestamp, which are useful for debugging the failure
```

```
 print(headers)
}

print("Result:")
result = h$value()
print(fromJSON(result))
```

The output is as follows:

```
> print(fromJSON(result))
$Results
$Results$output1
$Results$output1$type
[1] "table"

$Results$output1$value
$Results$output1$value$ColumnNames
[1] "Account"     "Date"        "Channel"     "Units"       "nth minute"
[6] "value"       "Scored Labels"

$Results$output1$value$ColumnTypes
[1] "Int64"    "DateTime" "String"   "String"   "Double"   "Double"   "Double"

$Results$output1$value$Values
$Results$output1$value$Values[[1]]
[1] "26812030018"        "6/19/2014 12:00:00 AM" "605106493 1 kWh"
[4] "kWh"                "2.05"                  "0.54"
[7] "0.787800192832947"

$Results$output1$value$Values[[2]]
[1] "26438261005"        "2/8/2014 12:00:00 AM" "605107806 1 kWh"
[4] "kWh"                "2.35"                 "36.57"
[7] "35.3771095275879"
```

The highlighted electricity consumption values were compared to the 2 records from the CSV file to confirm the predictions:

	A	B	C	D	E	F	G
1	Account	Date	Channel	Units	nth minute	value	Scored Labels
2	26812030018	6/19/14	605106493 1 kWh	kWh	2.05	0.54	0.787800193
3	26438261005	2/8/14	605107806 1 kWh	kWh	2.35	36.57	35.37710953

Similarly, we executed the same process that we carried out for Boston Public Library (BPL), for all the areas namely Agassiz neighborhood, Boston Police Department (BPD),

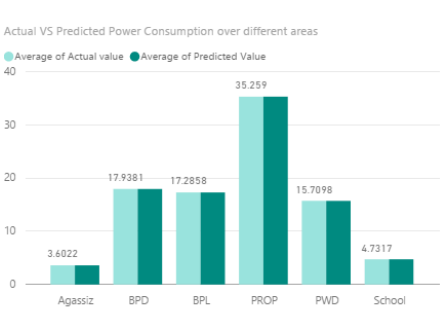
Public Work Department (PWD), Property Management (PROP), School Part I and School Part II.

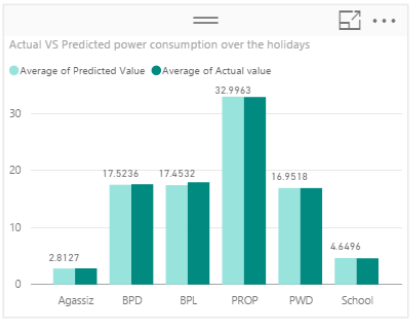
The exploratory data analysis and visualization was performed on Power BI.

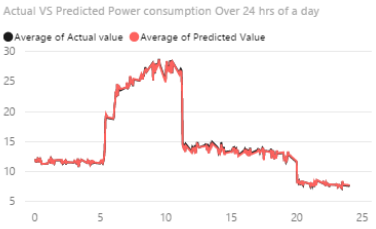
PFB the link to the Power BI Dash-Board:

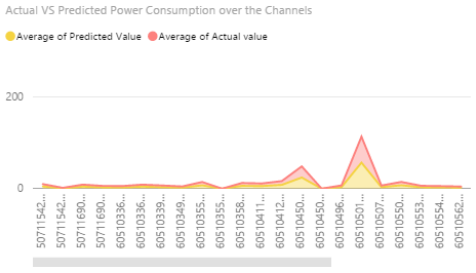
https://app.powerbi.com/view?r=eyJrIjoiZGY3MDRmNWYtNGM5ZC00ODg3LWI0NDYtNWQ5NTQwYWFiMDJkIiwidCI6IjZhYmZjNzNmLWRhNjQtNDEzNy05ZjlmLTE1ZmFhZTU2ZjY4NSIsImMiOjN9

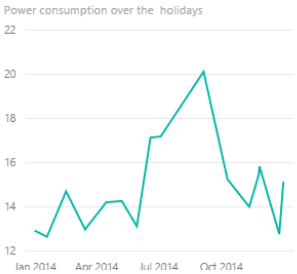
Power BI Dash Board :

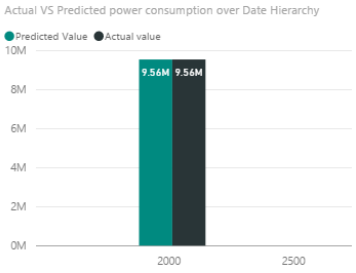






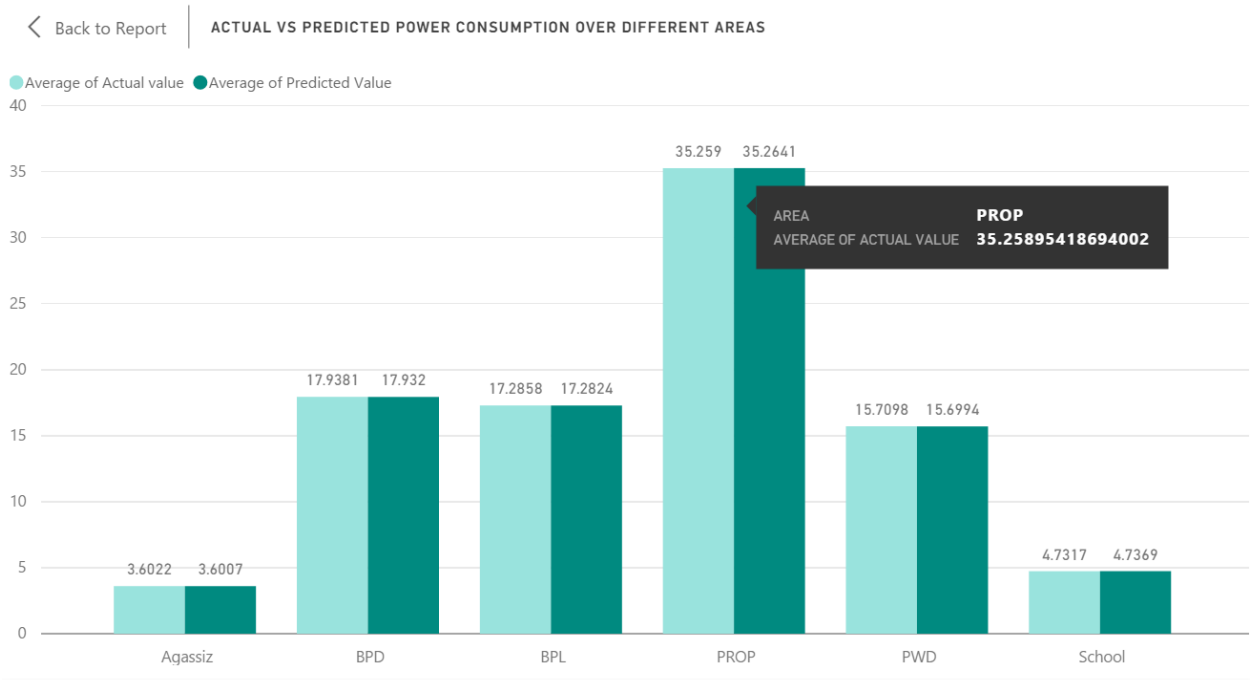






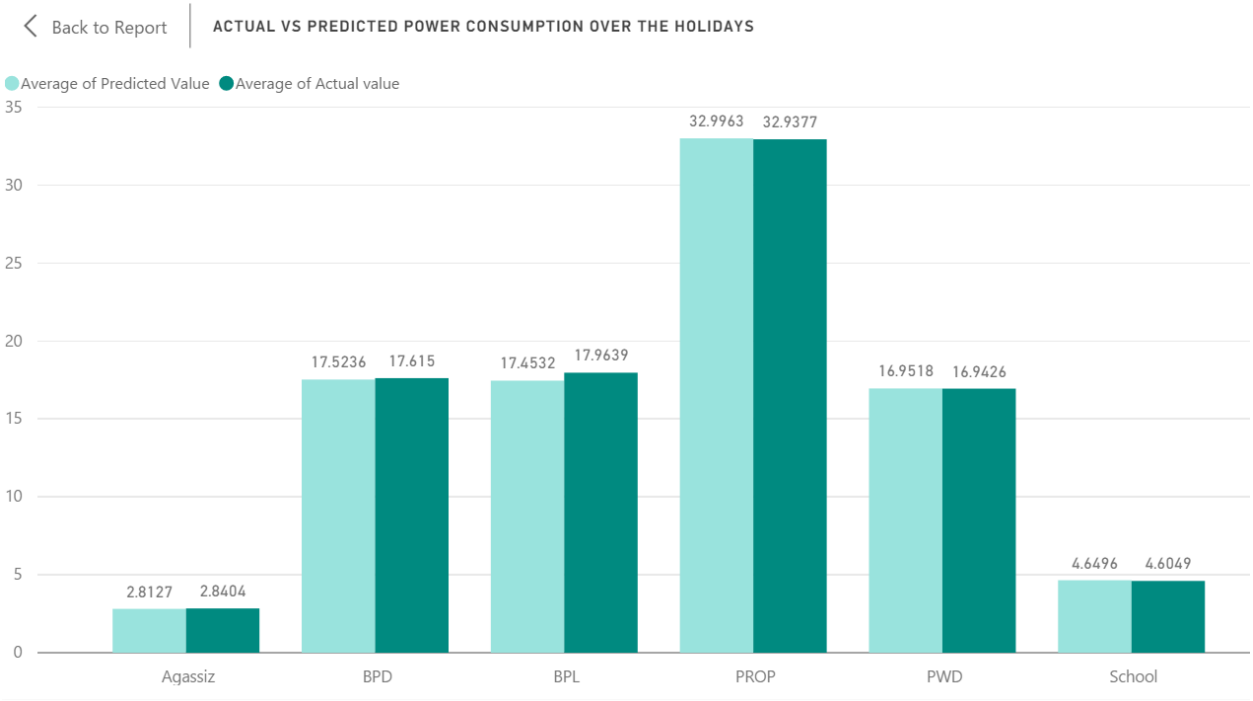
Assignment 3 - Report
Jyotirmayee Mahanandia | Aditya Shinde | Rachitha Dhanraj
INFO 7390: Advance Data Sci/Architecture

Chart 1: Actual VS Predicted Power Consumption Over Different Areas



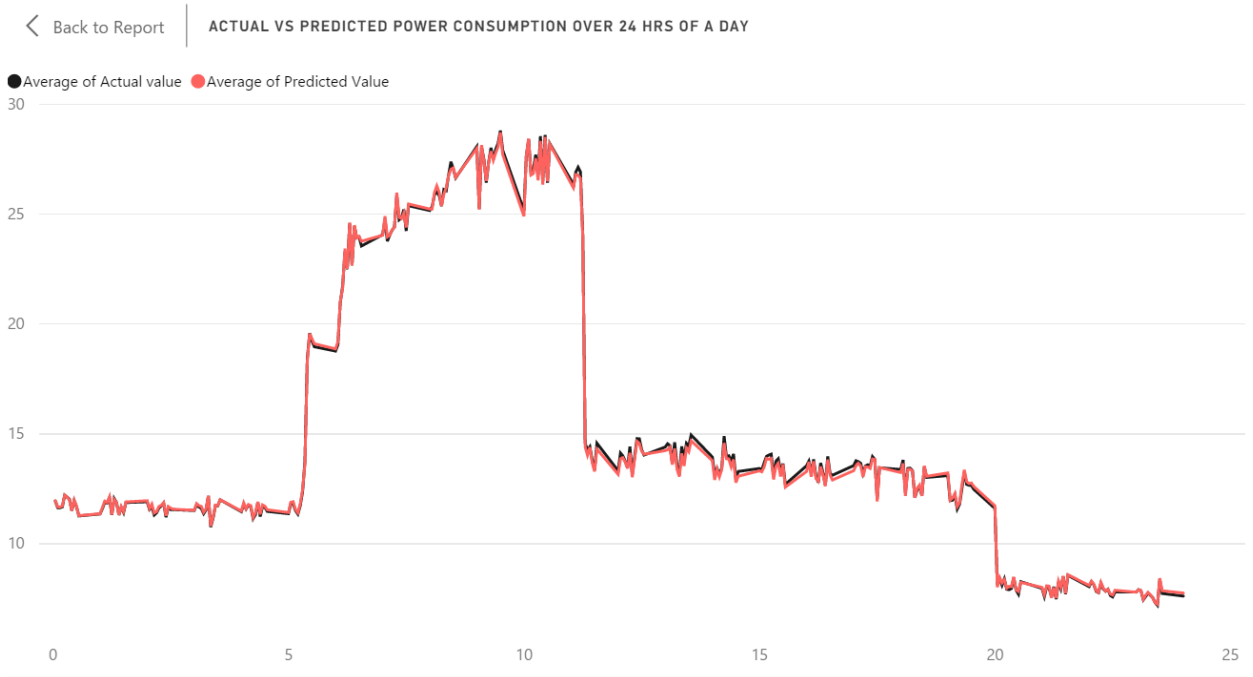
The above chart provides a pictorial representation of the power consumed in the year 2014 and expected power consumption for the next year. The bars in light green color represents the average power consumption in different areas and the bars in dark green represents the predicted power consumption by Microsoft Azure Boosted Tree Algorithm for the year 2015. From the above chart it's concluded that Property Management Area (PROP) had the maximum usage of power in the year 2014 and Agassiz had the minimum power usage.

Chart 2: Actual VS Predicted Power Consumption over the holidays.



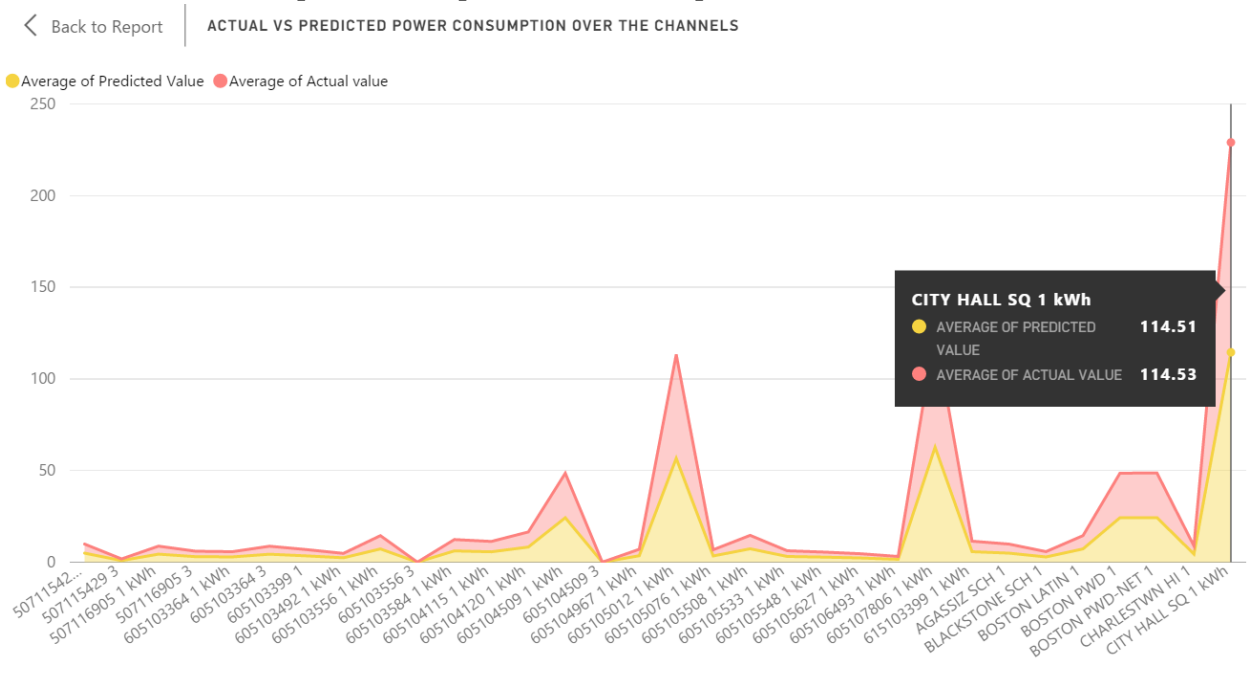
This chart provides the details of the actual vs predicted power consumption in different areas.

Chart 3: Actual VS Predicted Power Consumption over 24 Hrs of the day:



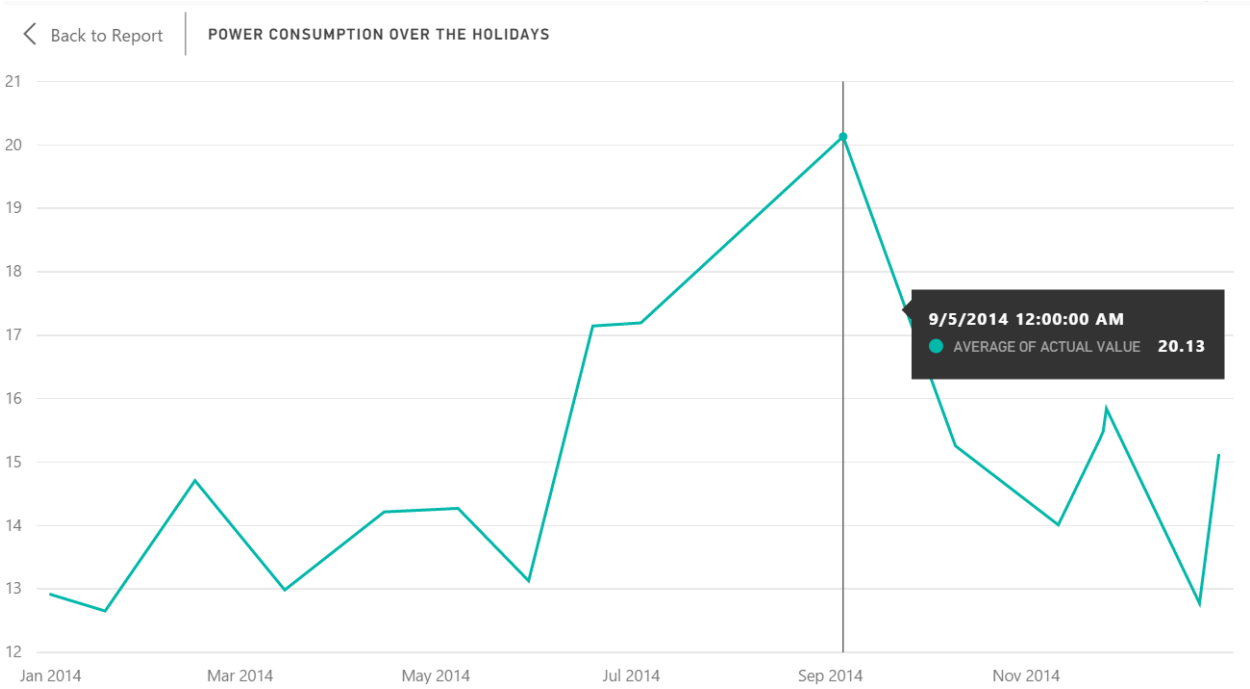
This graph is a pictorial representation of the actual vs predicted power consumption over 24 hours of a day. This shows the behavior of the electricity consumed in a day. You can see there is a sudden rise in the usage of electricity between 5 AM to 6 Am. From the graph it's clear that the peak hours of the consumption is usually from about 9 AM to 11 AM in the morning and its quiet steady from 11 AM to 7 PM in the evening and there is a sudden drop after that.

Chart 4: Actual VS predicted power consumption over different Channels:



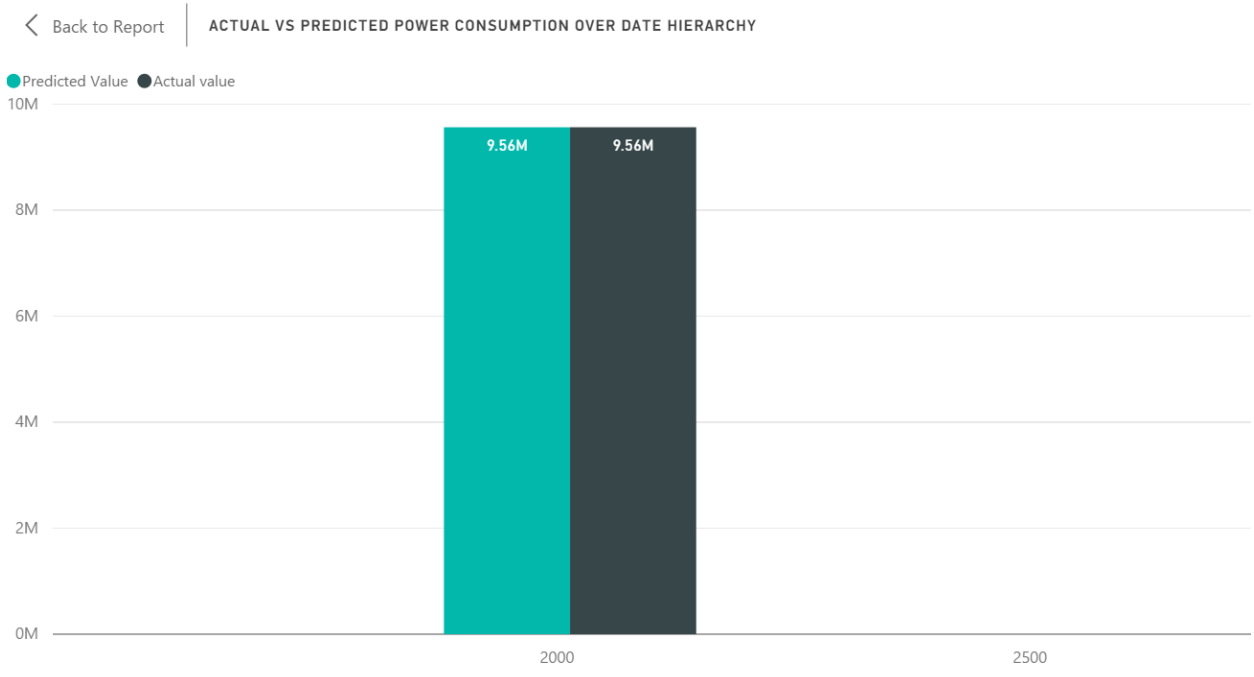
This graph shows the behavior of actual vs predicted power consumption over different channels. And from the graph it's clear that the channel City Hall got the maximum power consumption.

Chart 5: Power Consumption Over the Holiday



This chart shows the power consumption in Boston over the Holidays. And from the analysis it looks like 5th of September had the maximum power usage.

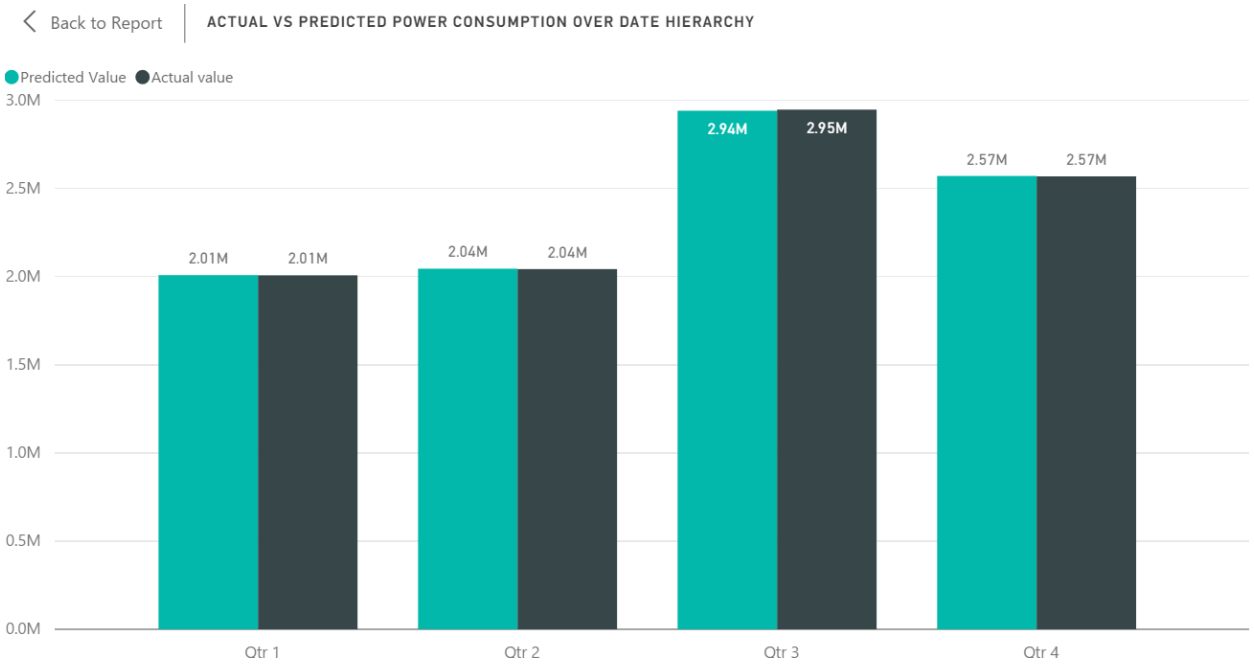
Chart 6: Actual VS Predicted power consumption over the data hierarchy



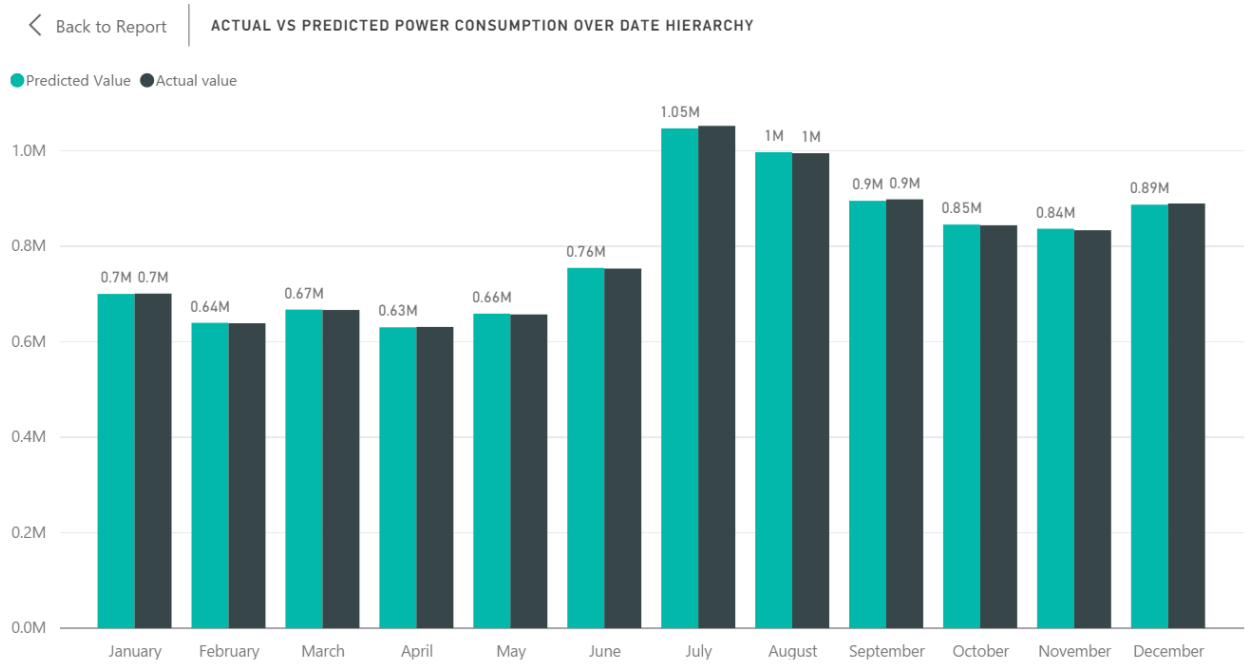
Assignment 3 - Report
Jyotirmayee Mahanandia | Aditya Shinde | Rachitha Dhanraj
INFO 7390: Advance Data Sci/Architecture

This graph shows the total amount of power consumed in year 2014 and the predicted value given by Microsoft Azure Machine Learning Algorithm. From the graph it's observed that the actual and predicted values are pretty close. This graph can be further drilled down to different quarters, months and days of the year. PFB the drilled down graphs.

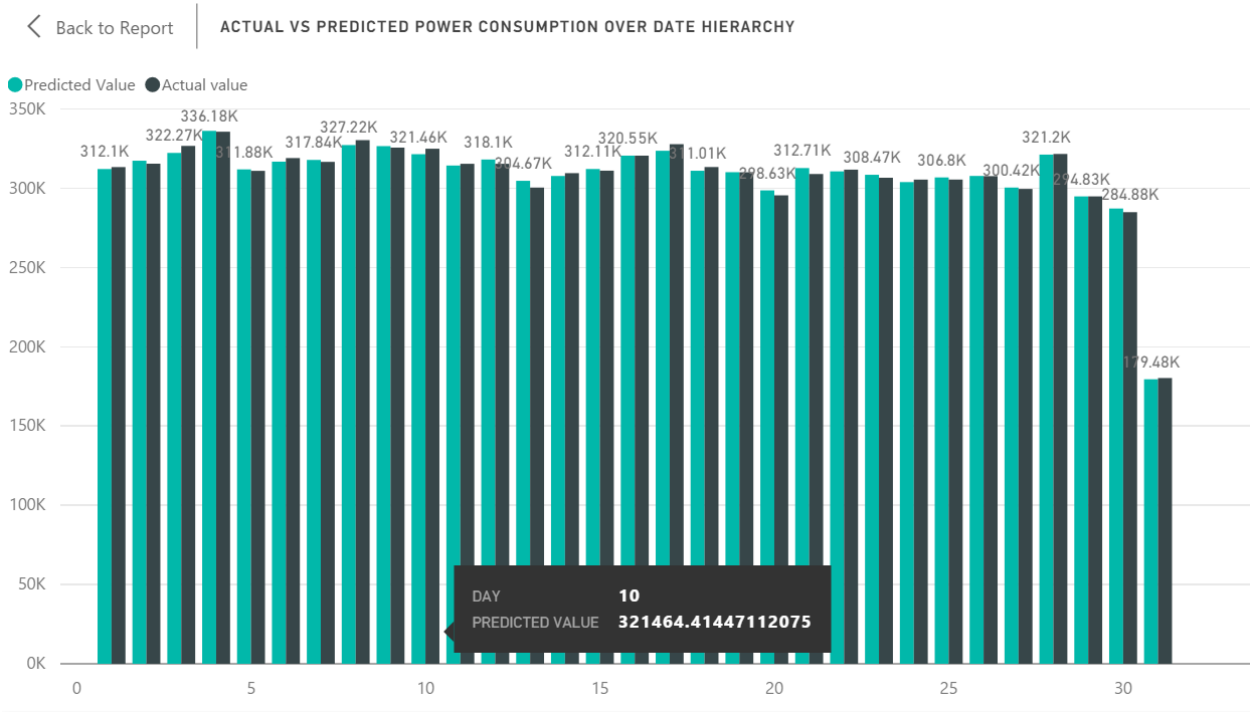
Actual VS Predicted Power Consumption Over different quarters:



From the above graph it's observed that Quarter 3 has the highest power consumption. This can be further drilled down to different months. PFB the graph for different months.



It's seen from the graph that the month of July had the maximum power consumption. This can be further drilled down to days. PFB the graph for power consumption for different days.



The graphs in the dashboard changes depending on the selection of the data in any of the charts for example if we select the property management data in the 1st chart it will show the respective data in the other charts. PFB the snap shot.

Assignment 3 - Report

Jyotirmayee Mahanandia | Aditya Shinde | Rachitha Dhanraj

INFO 7390: Advance Data Sci/Architecture