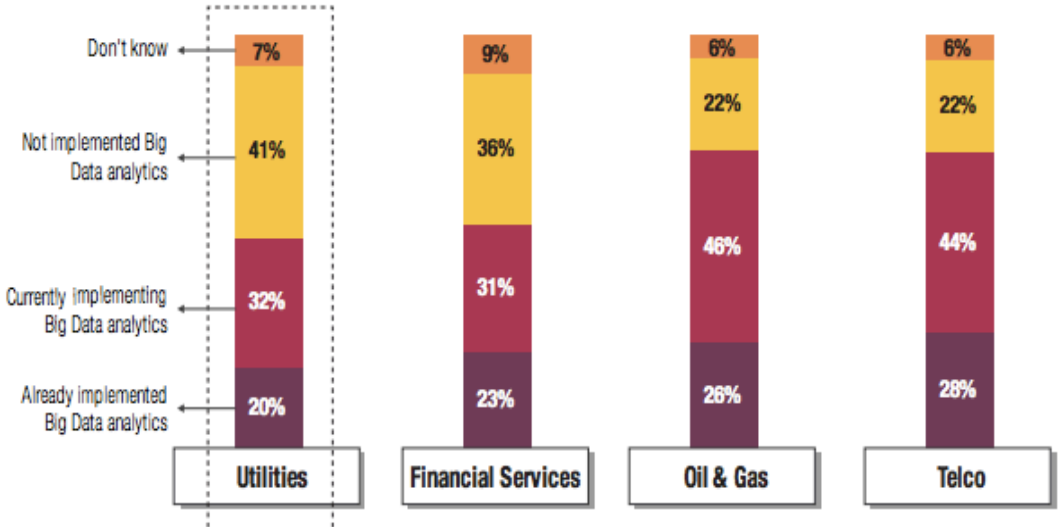
Final Project Proposal
Group4

Boston City Utility Data Analysis

Project Presented by (Team 4):
Ananya Chakravarty
Rohit Nigam
Rohit Nikam

Background

Suppliers of electricity, gas and water to U.S homes and businesses are finding ways to analyze the vast volumes of data in order to gain insights in customer trends and operational efficiencies. According to a survey conducted by Capgemini, utility spending is expected to grow 29% each year, totaling over $2 billion by 2017. Our analysis provides an insight on how utility vendors or customer can predict the total cost of utility based on several factors like location, zip and time factor.



Premise

Boston Government has many departments, which are placed, at different geographical locations in the Greater Boston Area. The people belonging to these departments work in buildings constructed as per their needs. Every building needs utilities such as Water, Gas, Steam, Fuel, Electricity, and Oil etc. The costs incurred on these utilities are massive and is taken into consideration during the planning phase of one such infrastructure construction. The valuable data that will come in from smart grids will enable utilities to:

- Streamline operational efficiency.
- Easily recognize areas of leakage or thefts.
- Enhance pricing and improve customer relationships.
- Forecast trends and demand.

Understanding the dataset

Considering the pain points mentioned above we decided to select the Boston Government Utility dataset. This dataset had the following attributes: –

- SiteName
- Department
- SiteZip
- Month
- Year
- TotalArea
- UtilityType
- UtilityVendor
- TotalCost
- TotalUsage
- WeatherNormalization
- CO2Emissions

Dataset1 - This dataset give us the Total Cost incurred and the Total Usage of different utilities such as Gas, Water, Electricity, Steam and Fuel from multiple vendors (such as VEOLIA, HESS and EVERSOURCE for electricity). The data was collected on the basis of billing month which had Day, Month, Year of Utilization and also the Zip, Area of the infrastructure.

Dataset2 - We found another dataset, which gave us the weather-normalized data for different utilities on each day for various locations in and around Boston.

Dataset3 - The third dataset contained data about CO2 Emissions in all the departments in Boston based on the utility consumption.

All datasets can be found on data.cityofboston.gov.

The utilities in the dataset used the following units –
Water in Gallons, Gas in MCF, Electricity in KWH, Fuel in Gallons and Steam in KJ.

Preprocessing:

Our first challenge was to clean the null values, incorrect values, outliers, NA and missing values in all 3 datasets. We used R for preprocessing. Below is the screenshot of the R code we used.

```
#To read the datasets
data1 <- read.csv(file.choose(), header=T, sep=",")
data2 <- read.csv(file.choose(), header=T, sep=",")
data3 <- read.csv(file.choose(), header=T, sep=",")

#To split the datetime column into Day,Month and Year
datetxt <- as.Date(data1$ServiceMonth, "%m/%d/%Y")
df <- data.frame(Month = as.numeric(format(datetxt, format = "%m")),
                 Day = as.numeric(format(datetxt, format = "%d")),
                 Year = as.numeric(format(datetxt, format = "%Y")))

#delete columns
data <- bpl[c(-3,-4, -8,-9,-10,-13,-14,-15,-19, -20,-21,-22,-23,-24,-25,-26,-27,-28,-29,-30,-31,-32,-33,-34,-35,-36, -

#delete Null values
data <- na.omit(data)

#delete NA values
data[data == 0] <- NA

#Merge date column for exploratory analysis.
data$FullDate <- paste(data$Month, data$Year, sep="/")

#Merge all 3 datasets
Temp <- merge(data1,data2)
FinalData <- merge(Temp,data3)


#Export data to excel
write.csv(FinalData, "c:/Users/RohitNikam/Desktop/finalWithDate.csv")
```

Step1 – Remove nulls and NA
Step2 – Remove unnecessary attributes
Step3 – Merge datasets
Step4 – Split date column into day, month and year.
Step5 – Align location and address of the infrastructure as per ZIP
Step6 – Split location to get latitude and longitude for exploratory analysis
Step7 – Remove Outliers
Step8 – Sort as per the month of usage and group by department

Processed Dataset:

1	SiteName	ZipCode	DivisionNameCat	DivisionName	VendorName	ServiceNa	ServiceName	Month	Day	Year	SiteAreaInSq.Ft.	Total.Cost	Total.Cost	Total.Usage	Total.Usage.Sq.Ft.
2	Clougherty Pool	2129	1	Boston Centers for Youth & Families	Hess Corporation	1	Electric Power	2	1	2011	600	444.775	0.741291	6520	10.866666
3	Clougherty Pool	2129	1	Boston Centers for Youth & Families	EverSource (NSTAR - Boston Edison)	1	Electric Power	4	1	2011	600	295.27	0.492116	4000	6.666666
4	Clougherty Pool	2129	1	Boston Centers for Youth & Families	Hess Corporation	1	Electric Power	4	1	2011	600	264.94	0.441566	4000	6.666666
5	Clougherty Pool	2129	1	Boston Centers for Youth & Families	Hess Corporation	1	Electric Power	5	1	2011	600	258.72	0.4312	3880	6.466666
6	Clougherty Pool	2129	1	Boston Centers for Youth & Families	Boston Water	5	Water	5	1	2011	600	36.41	0.060683	890	1.483333
7	Clougherty Pool	2129	1	Boston Centers for Youth & Families	EverSource (NSTAR - Boston Edison)	1	Electric Power	6	1	2011	600	413.52	0.6892	480	0.8
8	Guild	2128	6	Boston Public Schools	Hess Corporation	1	Electric Power	1	1	2011	36628	934.24	0.025506	10560	0.288304
9	Clougherty Pool	2129	1	Boston Centers for Youth & Families	Boston Water	5	Water	7	1	2011	600	9111.42	15.1857	182620	304.366666
10	Clougherty Pool	2129	1	Boston Centers for Youth & Families	Hess Corporation	1	Electric Power	7	1	2011	600	919.95	1.53325	13000	21.666666
11	Marshall	2124	6	Boston Public Schools	Direct Energy	1	Electric Power	6	1	2014	141091	1323.02	0.009377	22560	0.159896
12	Ohrenberger School	2132	6	Boston Public Schools	Boston Water	5	Water	3	1	2011	111592	605.54	0.005426	13030	0.116764
13	Clougherty Pool	2129	1	Boston Centers for Youth & Families	EverSource (NSTAR - Boston Edison)	1	Electric Power	10	1	2011	600	57.36	0.0956	960	1.6
14	Clougherty Pool	2129	1	Boston Centers for Youth & Families	EverSource (NSTAR - Boston Edison)	1	Electric Power	12	1	2011	600	54.2	0.090333	920	1.533333
15	Clougherty Pool	2129	1	Boston Centers for Youth & Families	Hess Corporation	1	Electric Power	1	1	2012	600	59.17	0.098616	960	1.6
16	Clougherty Pool	2129	1	Boston Centers for Youth & Families	EverSource (NSTAR - Boston Edison)	1	Electric Power	1	1	2012	600	55.69	0.092816	960	1.6

Approach:

Our approach towards analysis of Boston Utility Consumption is to categorize the data depending on various departments like Boston Police Department, Boston Fire Department, Boston Public Library, etc. and publish a website to predict the Total Usage and Total Cost of each utility based on different departments located in different locations.

Web Application features:

- Dashboard for easy online access and visualization of utility consumption cost for different departments in Boston
- Enabling user to predict utility consumption value and cost by selecting the Month, Year, Zip, Area, Department and Type of Utility
- Enabling user to choose between multiple departments, timeframe and unit to visualize the prediction data

Technologies:

- Data Preprocessing – R
- Prediction – Azure Machine Learning Studio
- Visualization – Tableau
- Front-End – HTML, CSS, Bootstrap, JavaScript

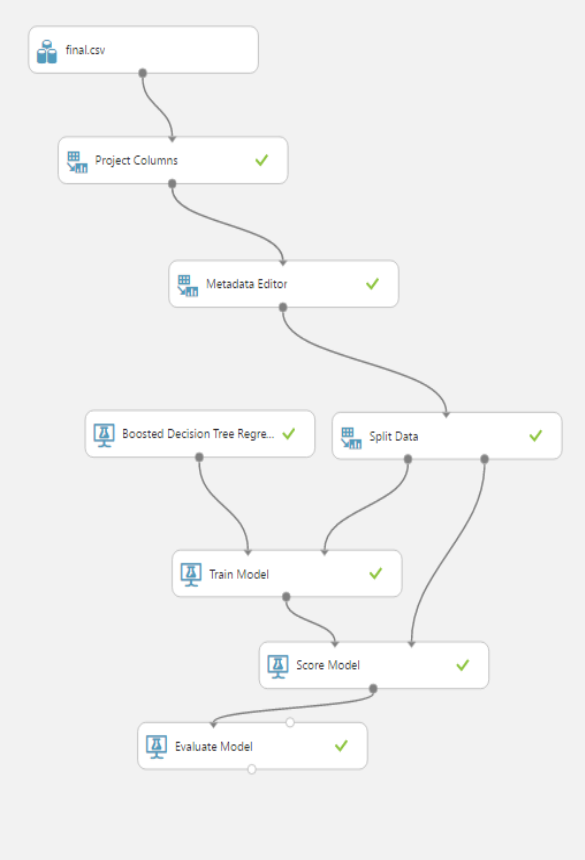
Prediction Model:

For predicting the utility consumption cost, we used **Boosted Decision Tree Regression model** in azure machine learning studio. It is a supervised learning method to create ensemble of regression trees.

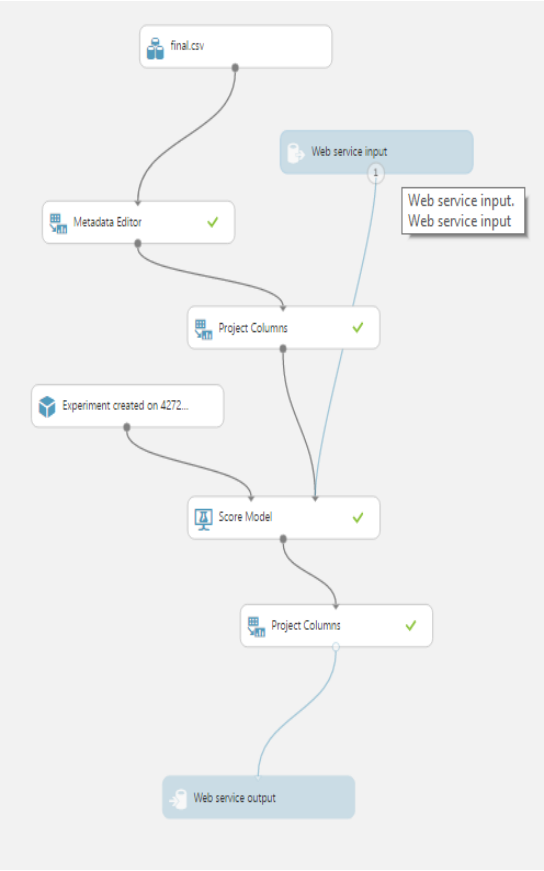
Properties:

- Used Projected Column to filter out the columns, which did not contribute to our prediction.
- Used Metadata editor to send the department and utility as categorical values.
- Split the data as Train and Score (70% traning)
- Used Boosted Decision Tree Regression to predict total cost values
- Consider utility prediction as our output/predicted value

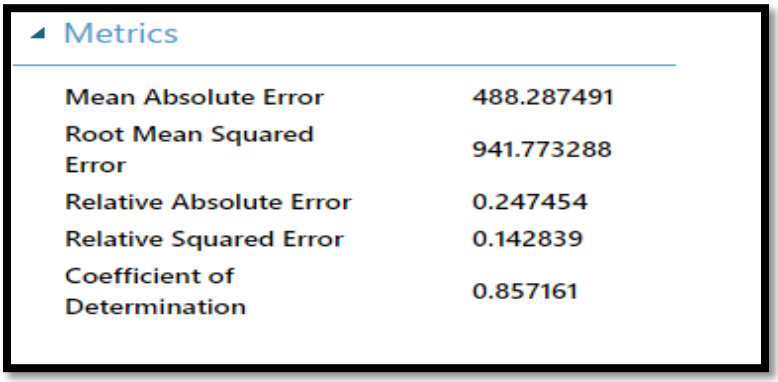
Training Experiment



Predictive Experiment



The above figure shows the Boosted Decision Tree Model (Training and Predictive Model) for Boston Utility Cost Prediction.



The figure on the right shows the metrics of the prediction model. The coefficient of Determination for the model is evaluated as 0.86, which shows that the dependent variable can be predicted from the independent variable with less error.

After evaluating the model we used the generated API and URL in Microsoft Azure Web App to generate dynamic platform for predicting the utility consumption and the cost associated.

Web Service Deployment:

After generating the API and web URL for the prediction model we used Microsoft Azure Marketplace platform so that any user can input the explanatory values/determining factors to get the predicted Utility Usage and Cost data.

The figure on the left shows the generated web URL of the Boston Utility Model using the web API and URL of the predicting model generated in Azure Studio.

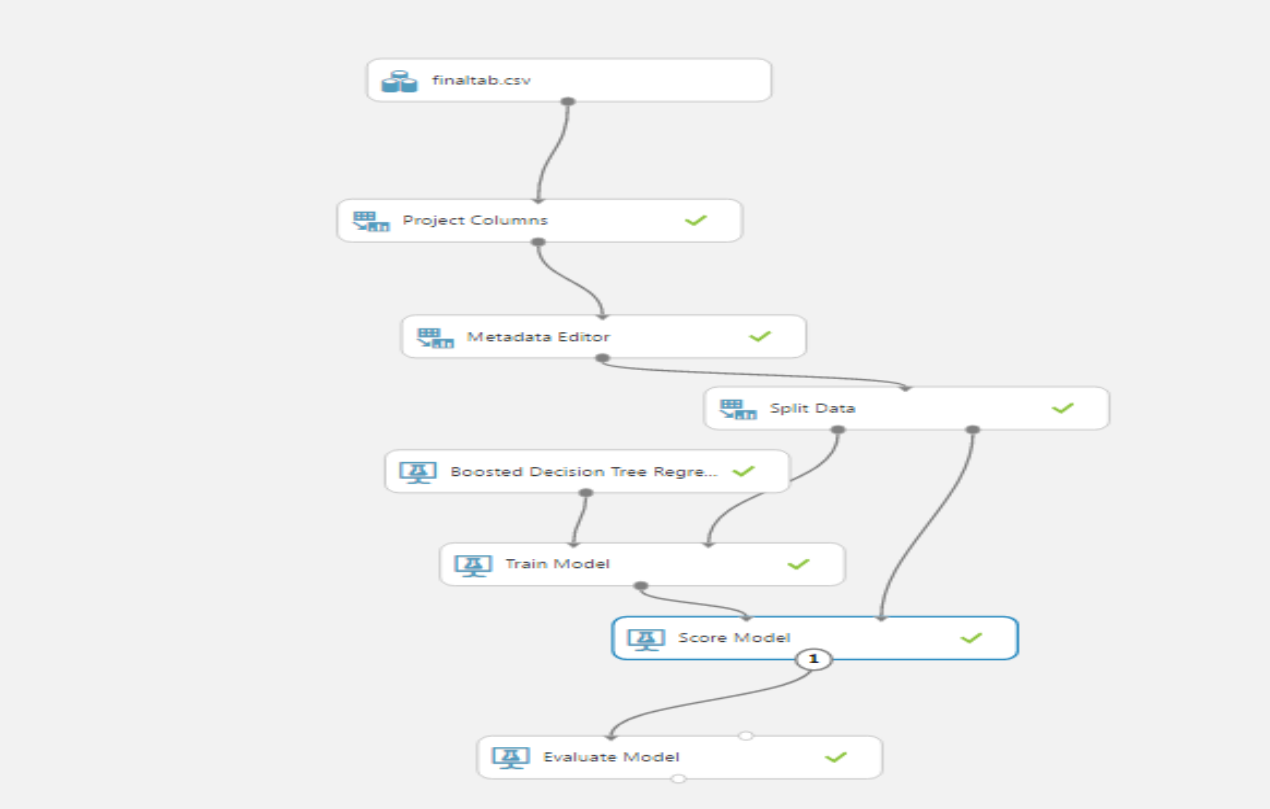
<u>Predicting CO2 Emission:</u>

This model gives the CO2 emissions due to the different utility consumption. For predicting the CO2 emission, we used **Boosted Decision Tree Regression model** in azure machine learning studio. It is a supervised learning method to create ensemble of regression trees.

Properties:

- Used Projected Column to filter out the columns, which did not contribute to our prediction.
- Used Metadata editor to send the ZIP code,department and utility as categorical values.
- Split the data as Train and Score (70% traning)
- Used Boosted Decision Tree Regression to predict CO2 Emissions

TRAINING MODEL



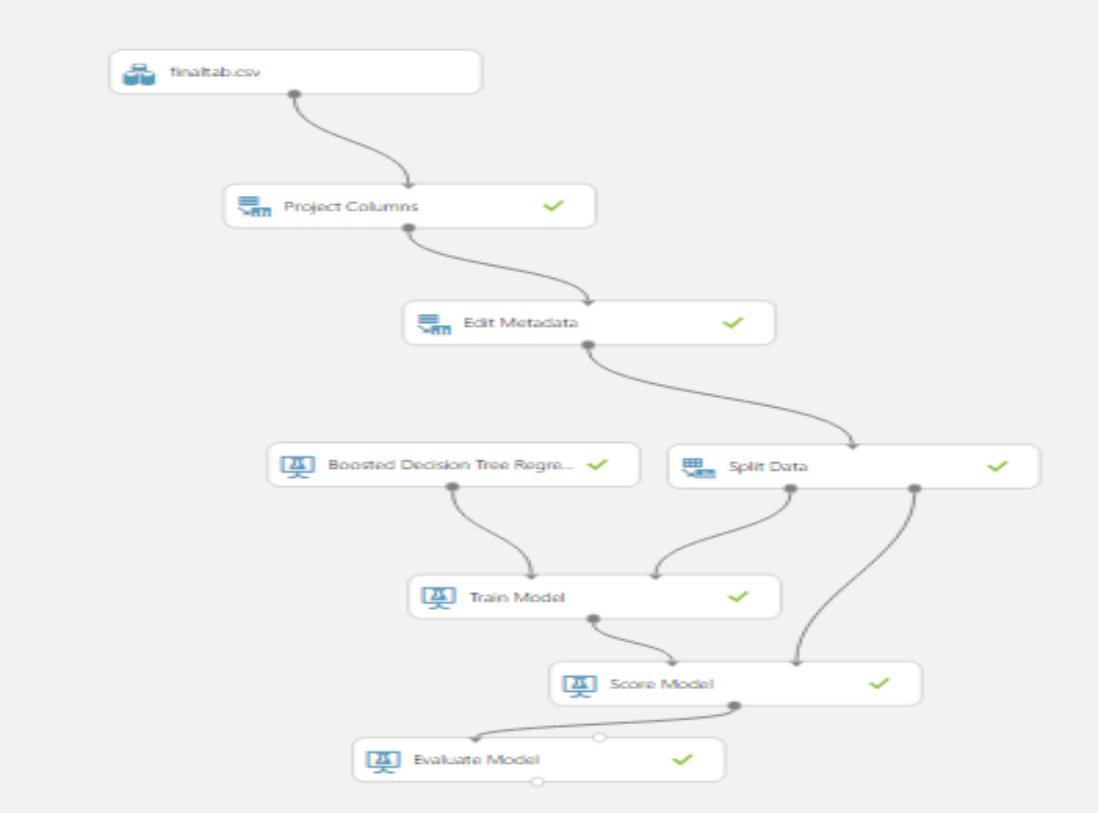
Predicting Weather Normalization Data:

This data gives the increase in the utility consumption due to changes in the weather condition. For predicting this, we used **Boosted Decision Tree Regression model** in azure machine learning studio. It is a supervised learning method to create ensemble of regression trees.

Properties:

- Used Projected Column to filter out the columns, which did not contribute to our prediction.
- Used Metadata editor to send the ZIP code, department and utility as categorical values.
- Split the data as Train and Score (70% traning)
- Used Boosted Decision Tree Regression to predict Weather Normalized data

TRAINING MODEL



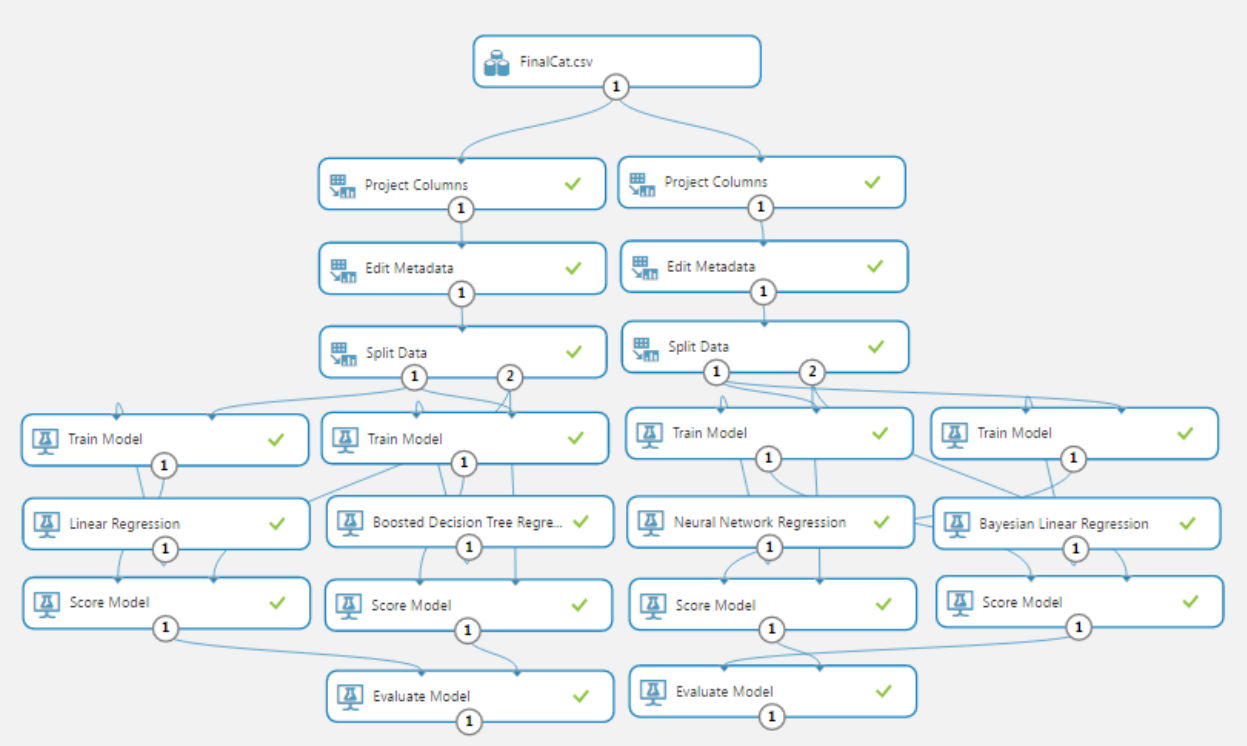
Why Boosted Decision Tree Regression?

The explanatory attributes in our dataset are mostly numeric and if they're not numeric they're categorical. We have converted all categorical data into numbers and decided to use a regression algorithm.

Boosting is one of several classic methods for creating ensemble models, along with bagging, random forests, and so forth. In Azure Machine Learning Studio, boosted decision trees use an efficient implementation of the MART gradient boosting algorithm. Gradient boosting is a machine learning technique for regression problems. It builds each regression tree in a step-wise fashion, using a predefined loss function to measure the error in each step and correct for it in the next. Thus the prediction model is actually an ensemble of weaker prediction models.

We compared 4 different regression algorithms Linear Regression, Neural Network Regression, Bayesian Linear Regression and Boosted Decision Tree Regression algorithms to figure out which one to use for our model and which one accurately predicts the Total Utility cost for each department. Below is the model that we created to compare the accuracy of the algorithm.

Algorithm Comparison Model



Linear Regression

Boosted Decision Tree Regression

Metrics

Mean Absolute Error	595.134842
Root Mean Squared Error	1275.943609
Relative Absolute Error	0.634631
Relative Squared Error	0.56469
Coefficient of Determination	0.43531

Metrics

Mean Absolute Error	346.321471
Root Mean Squared Error	90.390156
Relative Absolute Error	0.369305
Relative Squared Error	0.28182
Coefficient of Determination	0.93818

Neural Network Regression

Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
813.238523	1732.549298	0.86721	1.041161	0.549298

Bayesian Linear Regression

Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error	Relative Squared Error	Coefficient of Determination
594.760211	1276.154478	0.634232	0.564876	0.435124

Visualization

Analytics can help transform a range of operational processes. For instance, analytics helps match power generation with expected demand, manage peak load through variable pricing, and optimize the integration of decentralized energy generation. The use of predictive analytics also helps increase asset life and performance, while driving down overall asset maintenance costs. In order to provide benefit to the vendor as well as customer, we analyzed trend in our dataset to provide cost advantage, study trend in utility usage, etc.

From Customer Points of View:

In our dataset, customers refer to various departments in city of Boston like Property and Construction, Boston Parks Department, Boston Public Works Department, Boston Parks Department, Boston Public Schools, etc.

In order to provide advantage to the customer, we analyzed which department has the highest utility usage and then explored why the utility cost was higher for that department on the basis of location and vendor used by the customer.

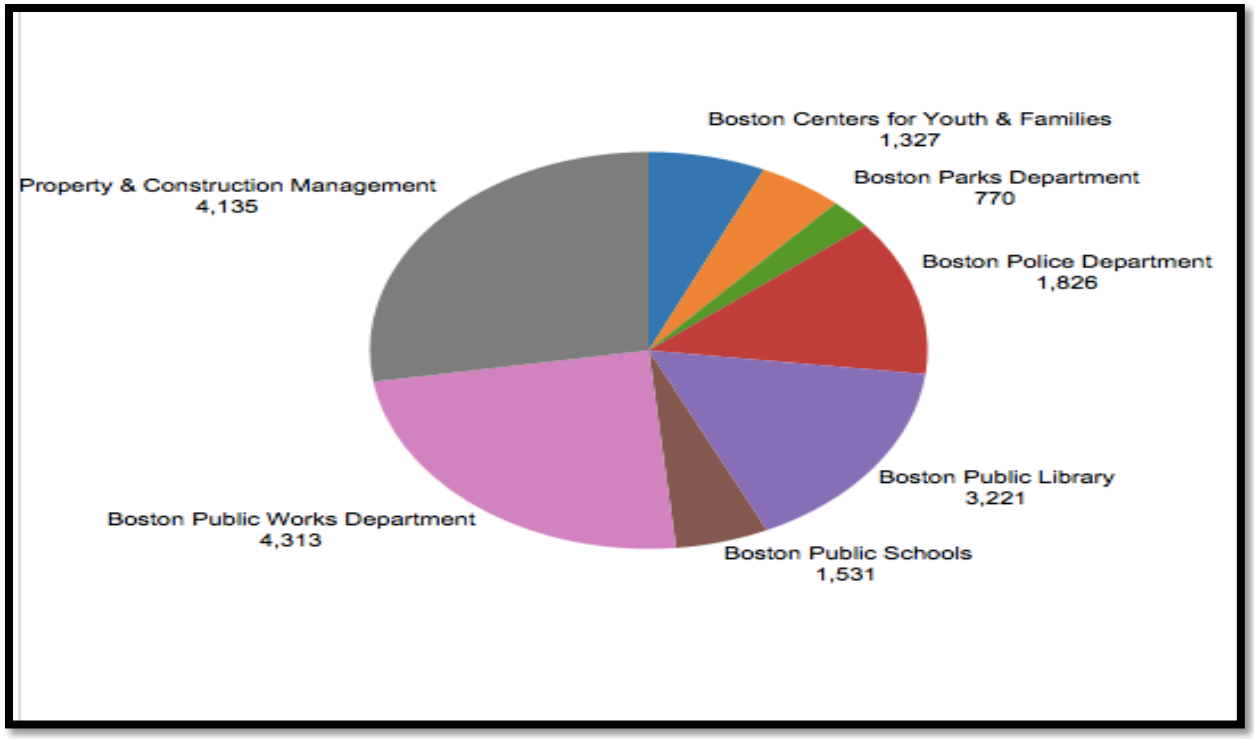


Figure: Which department has highest utility usage?

Now, we studied further the utility usage trend in Property and Construction Management (PCM) and analyzed which year showed the highest utility usage and particularly which month, which vendor and which utility source saw the highest usage (that whether electricity, water, gas or steam).

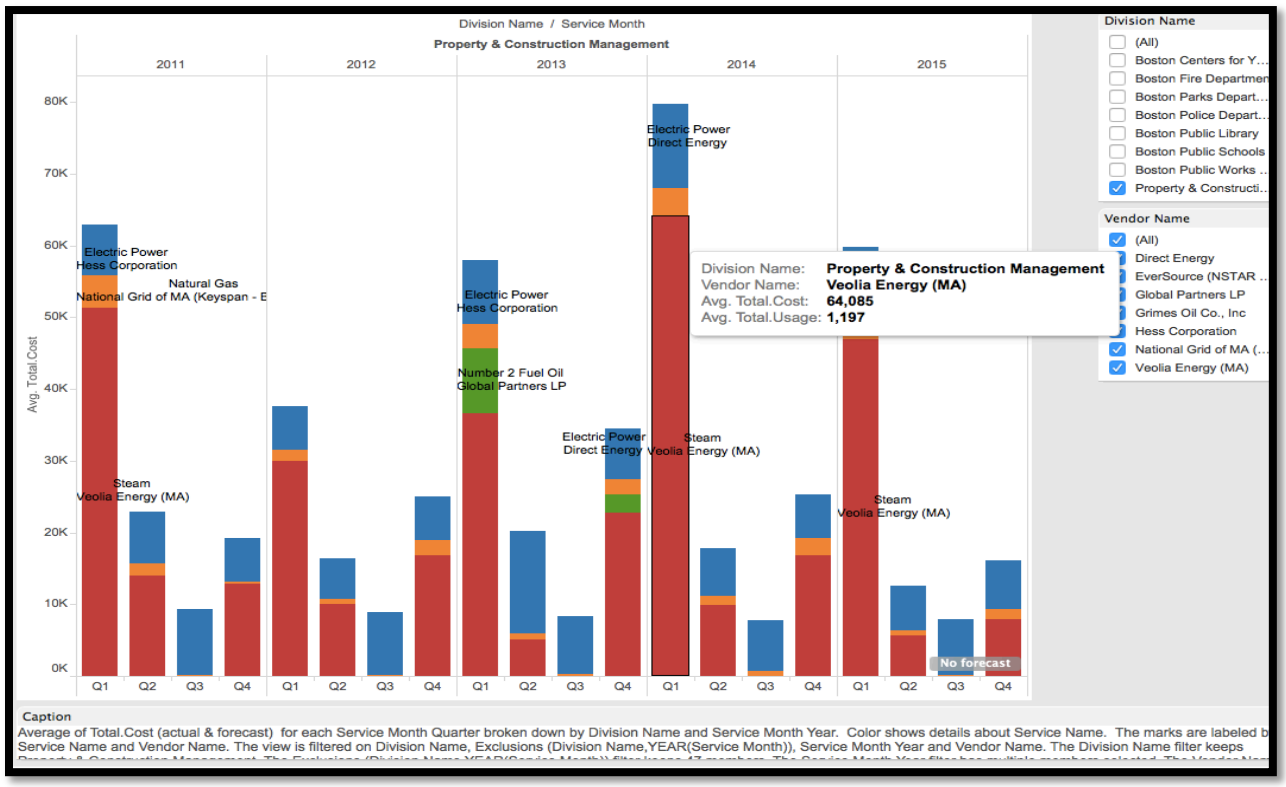


Figure: Which year shows the highest utility usage for PCM?

The above figure shows that in first quarter of 2014 PCM showed the highest utility usage and their vendor at 2014 was Veolia Energy.

Since our major focus is to reduce cost, we further analyzed how much high is the total cost of utility usage for PCM in different location and does it effect if PCM choose different vendors for the energy utility?

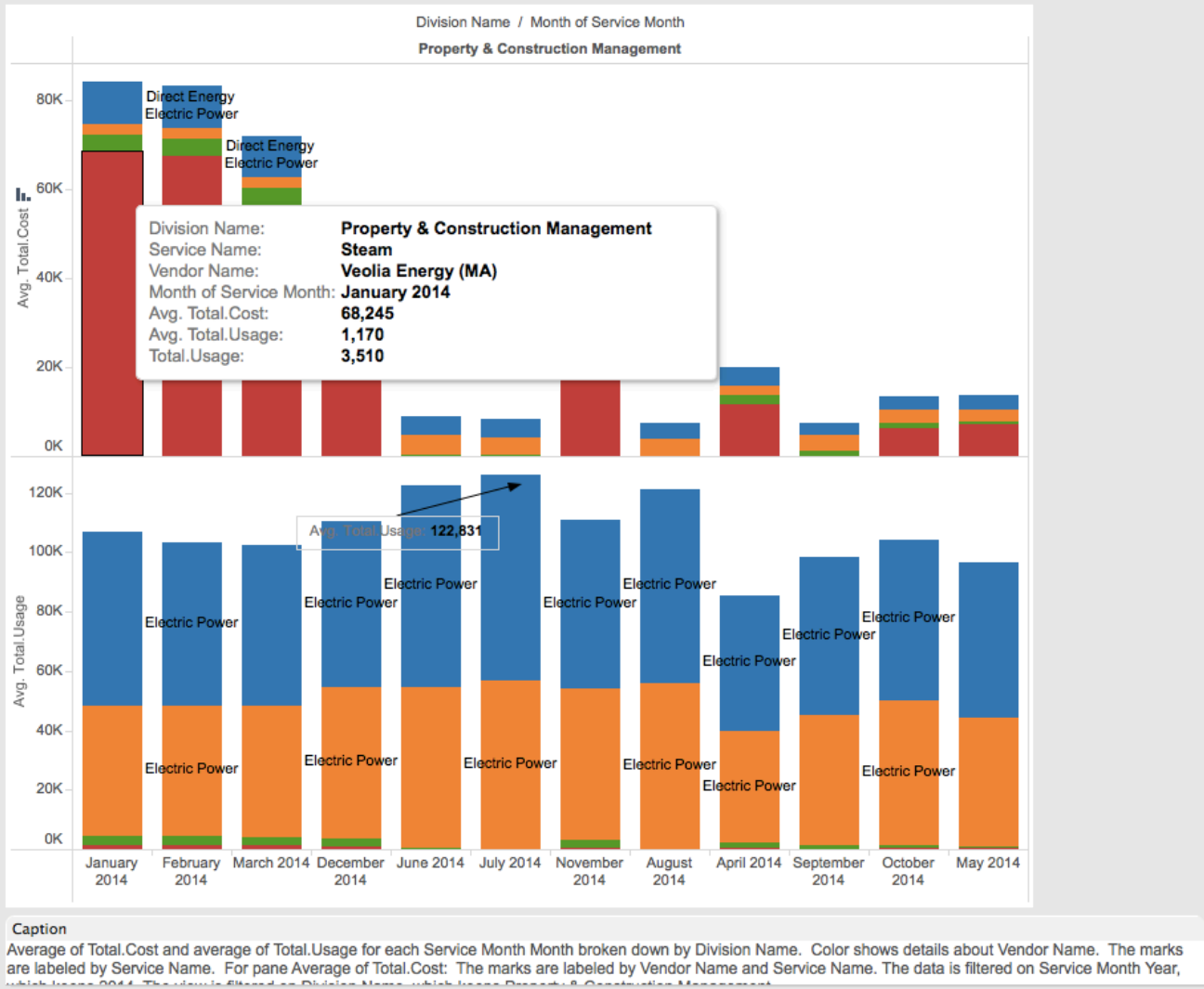


Figure: Which month shows the highest utility usage for PCM?

The above figure shows that highest cost for utility usage for Property and Construction Management vs Total usage. From this analysis, we can analyze that steam is one of the costliest source of energy since its usage is minimum and total cost is maximum.

Further, we analyze which utility vendor proved to be profitable for the PCM. So for this purpose we compared the utility usage and total cost of utility for same department but considering different location based on street and the zip code.

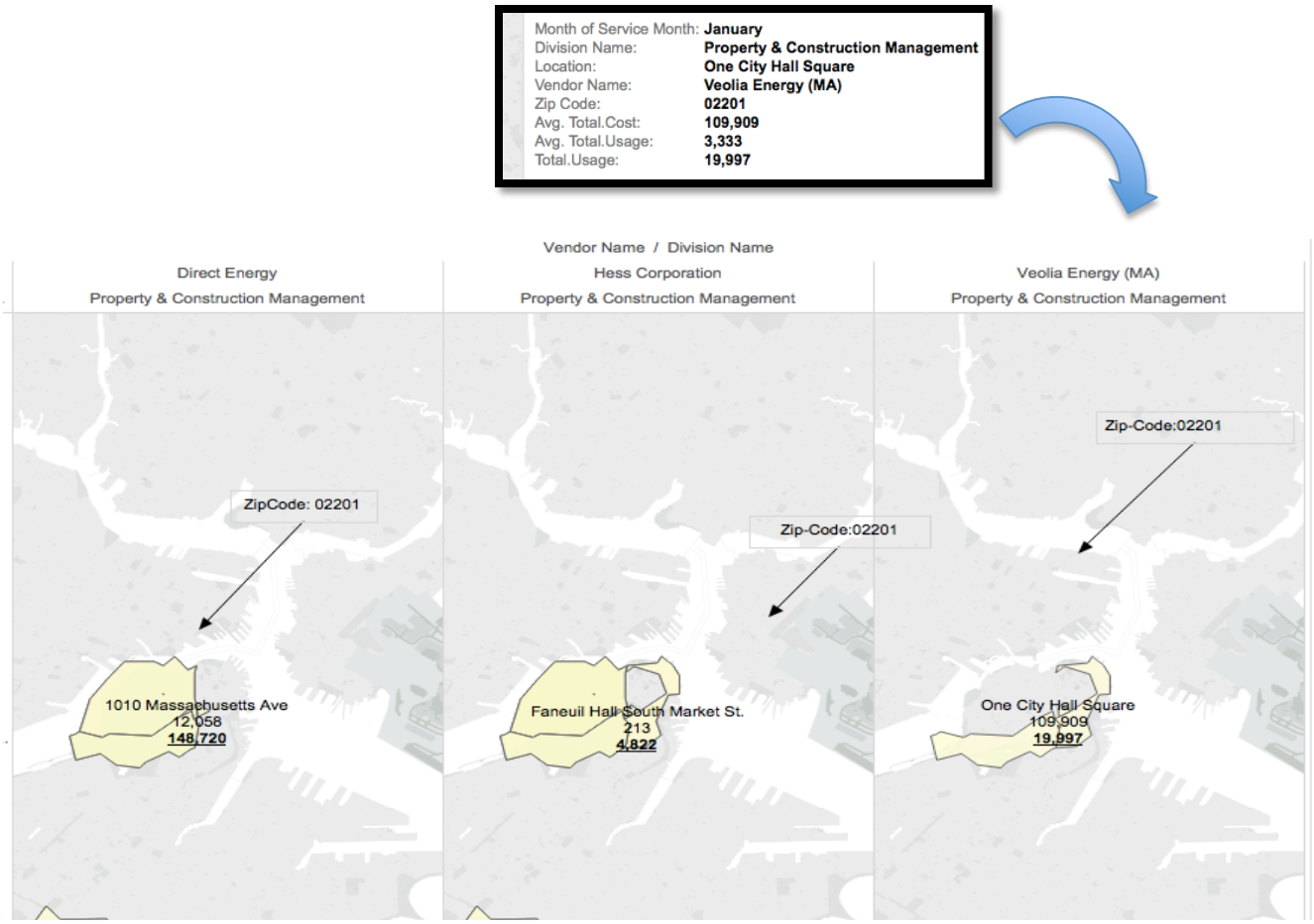


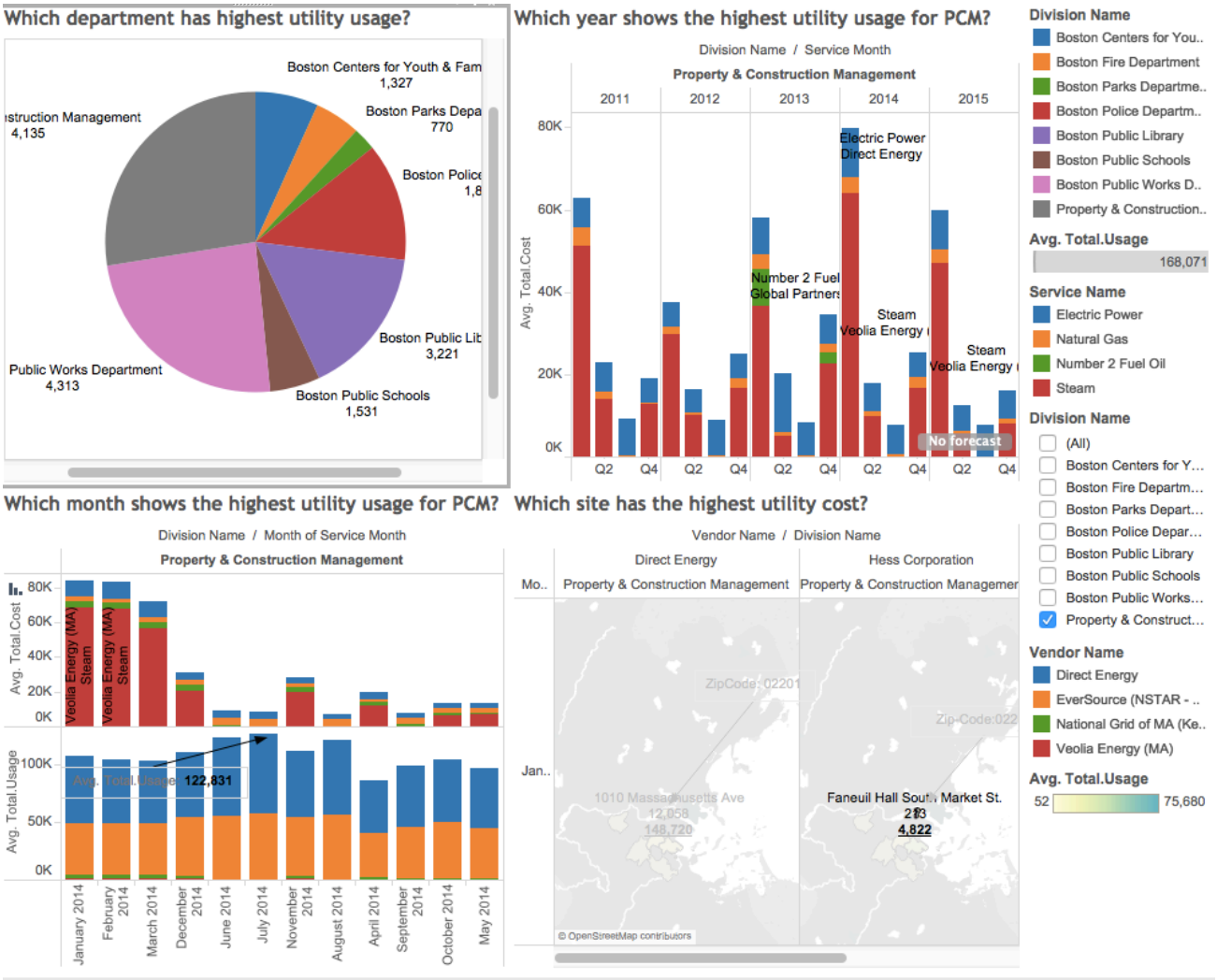
Figure: Which Vendor proved to be beneficial for PCM?

The above figure shows that for PCM, Veolia Energy proved to be the costliest vendor as compared to the other three vendors for energy usage.

This analysis answers 4 questions from customer point of view:

- Which department has highest utility usage?
- Which year shows the highest utility usage for that department (PCM)?
- Which month shows the highest utility usage for PCM and what was the cost associated with the usage of that particular source of energy?
- Which vendor should the customer choose from money saving point of view?

Here is the Dashboard answering the above four questions:



From Vendor Point of View:

Data analytics can prove beneficial to the utility vendors by improving their marketing strategy. The need for robust analysis for utility industry will provide numerous benefits to the utility vendors in multiple areas:

- Provide usage patterns to the customer.
- Identifying trends and forecasting demands.
- Using predictive analysis to improve reliability
- Establishing better customer relationship

In our analysis, we studied that which customer shows the highest utility usage and through which customer is generating maximum revenue.

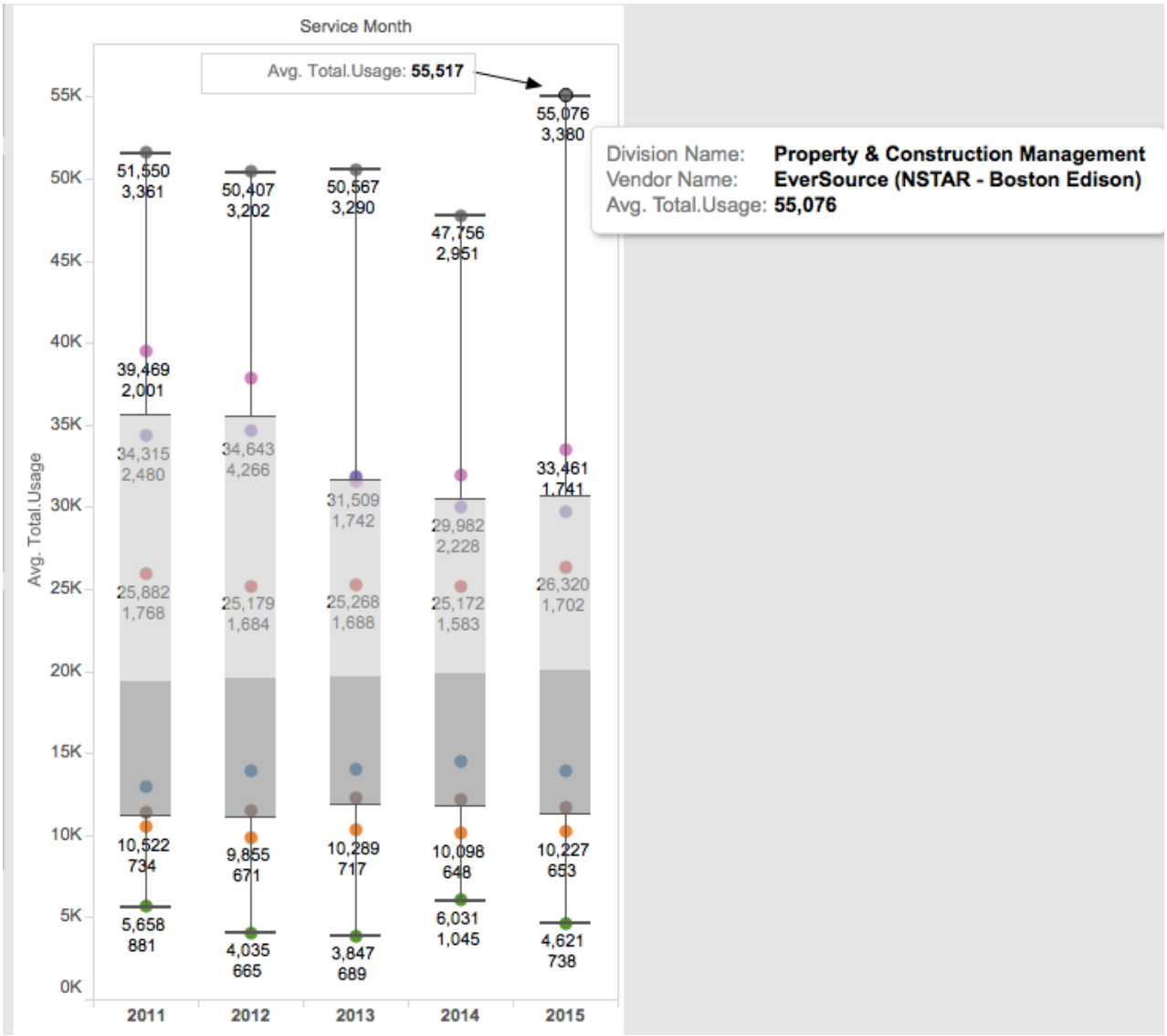


Figure: Which customer has highest utility usage?

The above figure shows the analysis from Vendors point of view (in this case we considered EverRSource). Now if EverSource wants to analyze which is their best customer or which customer is using the maximum service then the vendor can study the above graph and analyze that the best customer for EVERSource is Property and Construction Management (PCM) since the total utility usage is maximum as compared to the other departments.

Now what if the vendor wants to analyze the usage trend for PCM so that they can provide better customer service to their customer?

To answer this question we analyzed the usage trend of PCM for monthly basis and presented which month shows the highest electricity usage(NSTAR is a vendor which also provides electricity in city of Boston) and which in which month they had the highest billing cost.

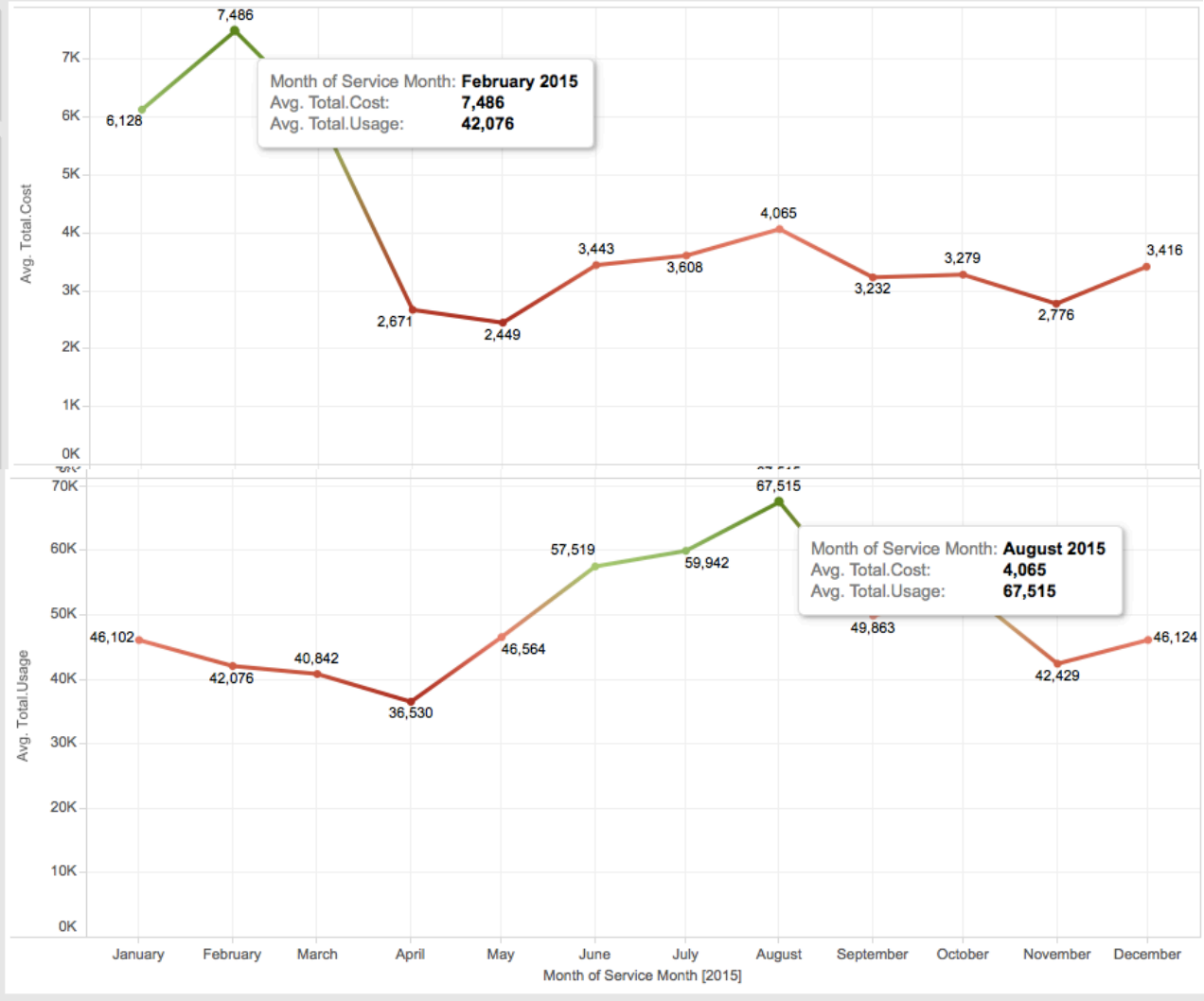


Figure: Cost and usage Analysis for PCM

The above figure shows that for PCM, the highest utility usage was in the month of February and the costliest electricity expense was generated on the month of August.

Now to further provide benefit to the user, the vendor can study the daily data and suggest him multiple ways to improve their usage pattern and decrease their expense.

Web Application

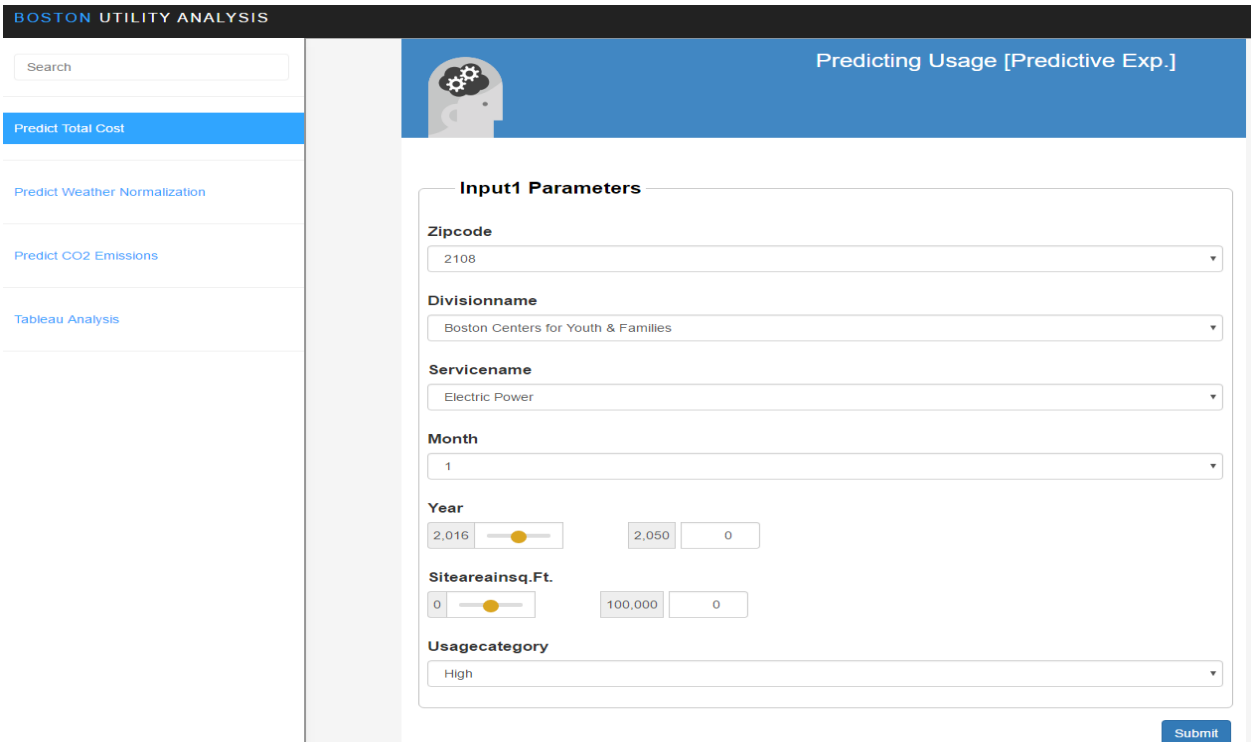


Figure: Total Cost Prediction Page

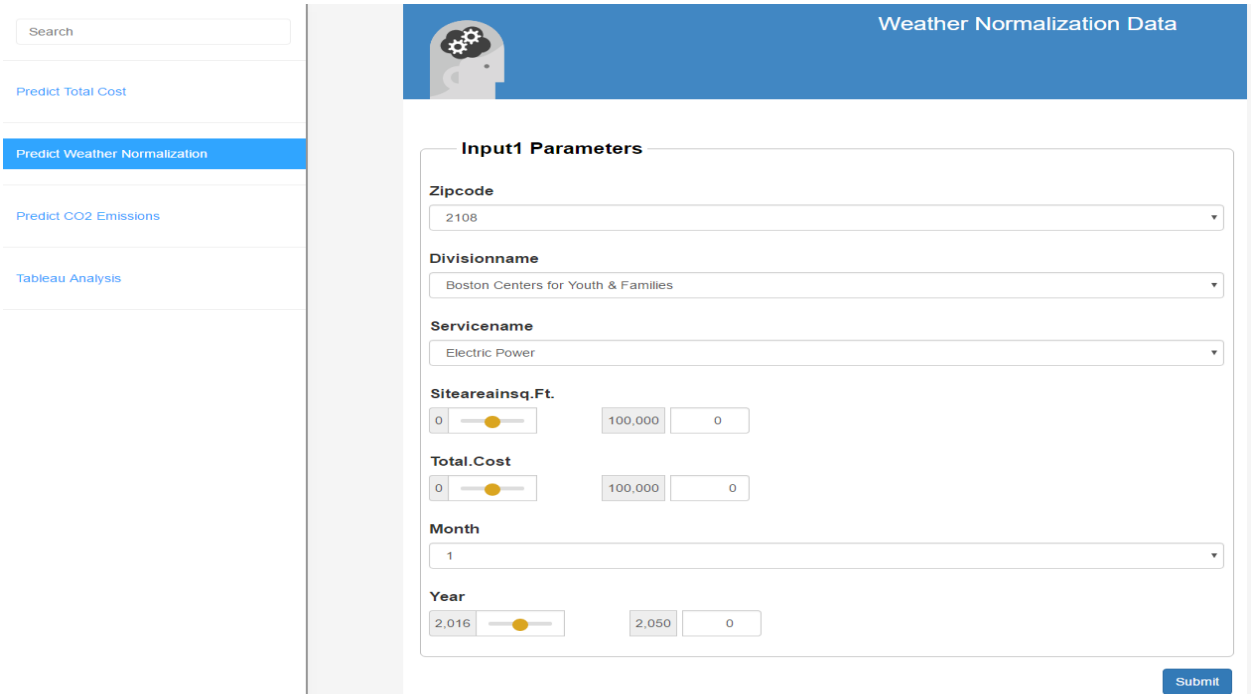


Figure: Forecast Weather Normalization

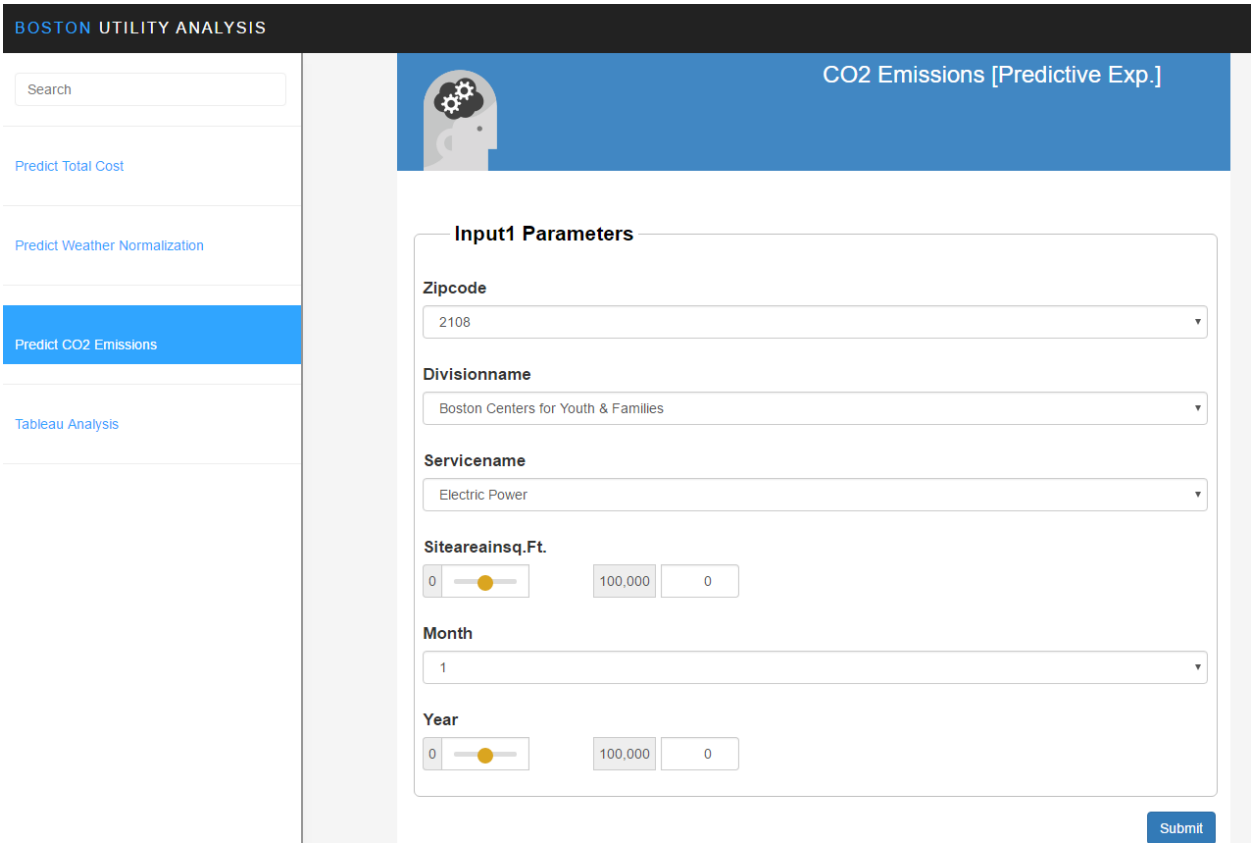


Figure: Co2 emission prediction

References

https://azure.microsoft.com/en-us/overview/what-is-azure/

https://studio.azureml.net/

https://public.tableau.com/s/