Hacking Climate Change 2020

Machine Learning for Efficient Energy Management

(Toronto, CA; photo by @nxsyed on Unplash)

Wency Go

Tinson Wang

8 February 2020

Introduction

In an increasingly connected world, the facilities management (FM) industry plays a critical role in providing a range of services to businesses such as building operations, maintenance, automation, and energy solutions. Businesses are able to streamline their workflow by leaving such services for FM experts to handle. As our global population continues to grow, so too does our energy consumption in its various forms such as electricity, natural gas, and fossil fuels. Being at the forefront of facility management, IFMA and BGIS are seeking solutions to mitigate the effects of climate change through modern ideas and technology.

Problem Statements

To generate valuable and focused ideas, hackathons are designed around a certain theme or problem for its participants to solve. The problem statements for this year are stated as follows:

#1: Define how our energy supply chain may be changed to address resiliency needs and availability for development requirements.

#2: Describe how technology may continue to evolve over the next 10, 20, 50 years and how advancements and applications of tech can contribute to address sustainability and environmental changes.

Our Solution

Our proposed solution seeks to address both problems by leveraging artificial intelligence (AI) technologies, specifically Machine Learning (ML), to predict future energy consumption which can be used to plan for resiliency needs, future developments, and energy management. More specifically, machine learning can provide a basis for which more proactive measures can be taken to better prepare for the future.

Our current energy supply chain consists of using some form of renewable or non-renewable resource to generate energy and distribute it through a network before it arrives to the end user. One way to use Machine Learning is to create a model to continuously analyze energy consumption data and geographical location to create a short- and long-term predictive model. Authorized users such as facility managers can see trends and typical usage for their facilities and make decisions accordingly. For example, decreasing non-critical services when they are underused, leading to more energy available for services that require it. With the presence of geographical data, the model can also be used to determine future development. For example, if a neighbourhood within a city is growing, as seen through their energy consumption, it would be interesting to observe its growth and perhaps plan future infrastructure near it. Machine Learning is still a developing technology; its rapid growth and implementation will continue, leading to better models and integration into our daily lives.

In this report, we develop a model that examines the consumption of natural gas resources over the years and makes a prediction on the level of consumption in following years. The US consumption of natural gases (1960-2017), US natural gas prices (1970-2017), and US natural gas expenditures (1970-2017) datasets used was supplied by the US Energy Information Administration (EIA), State Energy Data System (SEDS).

We choose to investigate natural gas as our energy source of focus due to its stronger pollution footprint on the environment compared to more renewable energy sources like solar or hydro energy. We hope to point out an increasing consumption of natural gases as something that facility managers should be wary of and may thus promote proactivity in switching to other energy sources that are less harmful to the environment.

Machine Learning and Energy Consumption in the U.S.

Background

Our goal is to demonstrate the use of Machine Learning to predict energy consumption by area for more efficient energy management. We decided to create a simple Multiple Linear Regression model using the Python scikit-learn package. The EIA datasets contain yearly data on the consumption of natural gas for each state, the price of natural gas for each state, and the amount spent on natural gases for each state. As previously noted, these datasets extend from 1960 to 2017. We believe this would give us enough data points to create an accurate model and extrapolate data for years to come.

Application

Cleaning the Dataset

The first step was to clean and format the dataset so it can be put through our model. The dataset used state abbreviations to map the energy consumption to each state. State abbreviations themselves would not lend to any useful relationships in terms of location so we mapped the states using x-y coordinates so that the ML model can see geographical proximity of the states.

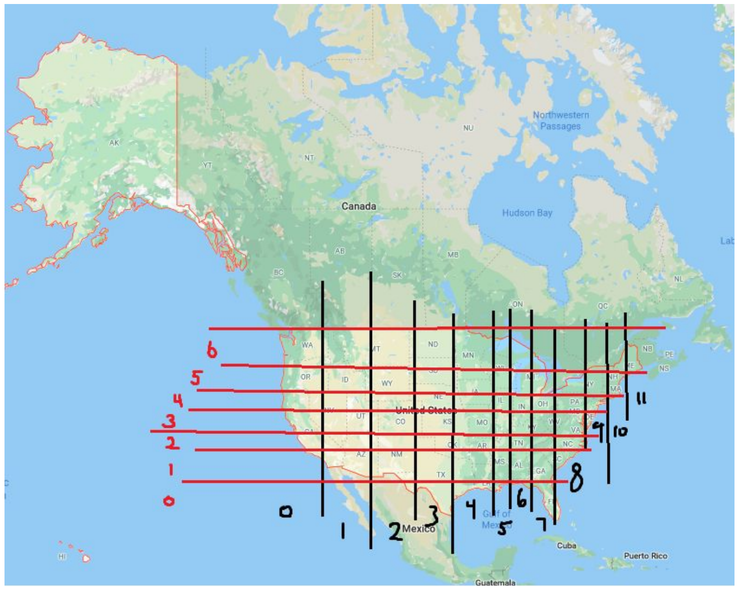

Figure 1. Map of USA overlaid with an X-Y grid; Map data by Google Maps 2020

The states of Alaska and Hawaii were removed from the dataset because of their farther geographic locations from the rest of the United States.

Creating the Machine Learning Model

The dataset was broken into two sets of data: the (x,y) coordinates of the state with respect to our map and the year that we wanted to examine the consumption of natural gas were split into one, and the actual numbers of natural gas consumption for a state in a year in the other.

Since the range of our values for natural gas consumption varies greatly, we made sure to first normalize them. Following suit, both sets of data were further split randomly into a training and testing set. We decided to place 80% of the data points into the training set, and the remainder in the testing set.

We then fit the model to our training set and used the remaining testing set to generate a set of predicted values. These values would then be judged against a set of actual natural gas consumption measurements.

Results

When the model was run using the complete dataset, we observed predicted consumption values that showed drastic differences compared to actual consumption values. Further analysis revealed that this was due to highly polar consumption values of certain states which played a critical role in adjusting the global trend. As a result, states that did not traditionally depend on natural gas as much would be predicted to consume more than that state did for that year due to the elevated global trend applied by the model. Inversely, some states had lower consumption trends. This affected the global trend too and would thus pull down the predicted consumption values of higher-usage states.

Subsequent analyses on the metrics of our model provide reasoning as to the inefficiency of our current model (see section "Evaluating the Machine Learning Model").

Understanding the limited capacity of our model given the dataset, we made the decision to limit our application of it to single states at a time. Known for its advanced tech industry, we here chose to apply the model to predict the consumption of California as well as its neighbouring states.

```
Machine Learning Model Results:
      Year Actual Consumption (Billion BTU) Predicted Consumption (Billion BTU)
5   1962.0                        [1483423]               [1734896.8709487803]
10  1964.0                        [1817241]               [1760886.4862314016]
11  1970.0                        [2241295]                [1838855.332079269]
4   1971.0                        [2265324]               [1851850.1397205777]
8   1982.0                        [1765171]               [1994793.0237750001]
3   1987.0                        [1992956]                [2059767.061981555]
6   1993.0                        [2213111]               [2137735.9078294225]
1   1994.0                        [2334777]               [2150730.7154707313]
0   1995.0                        [2109956]               [2163725.5231120437]
2   2002.0                        [2318656]                 [2254689.17660122]
7   2005.0                        [2304463]               [2293673.5995251536]
9   2008.0                        [2472612]               [2332658.0224490874]
```

Figure 2. Screenshot of our machine learning model output showing predicted vs actual natural gas consumption for California

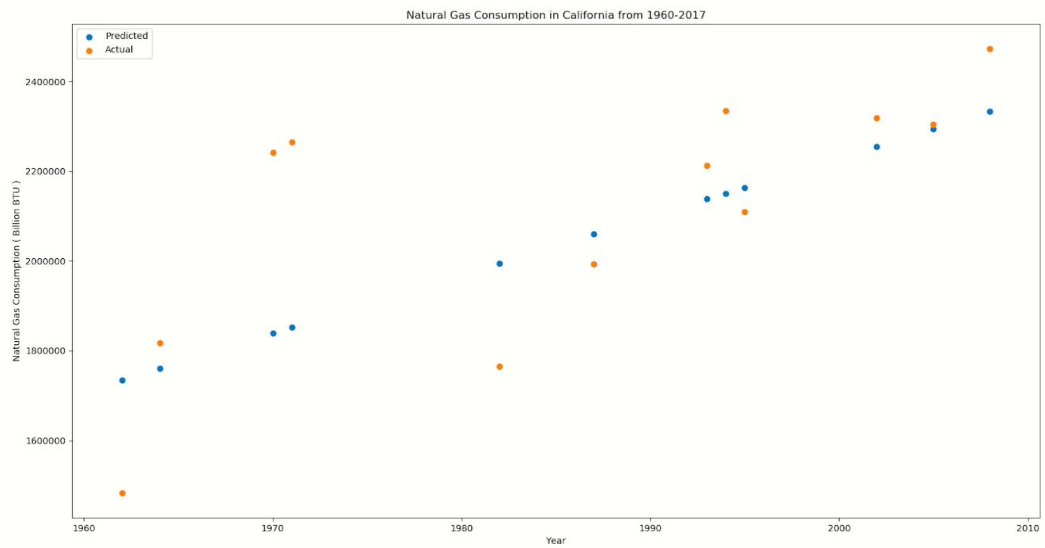

Figure 3. A graphical representation of predicted vs actual natural gas consumption of California.

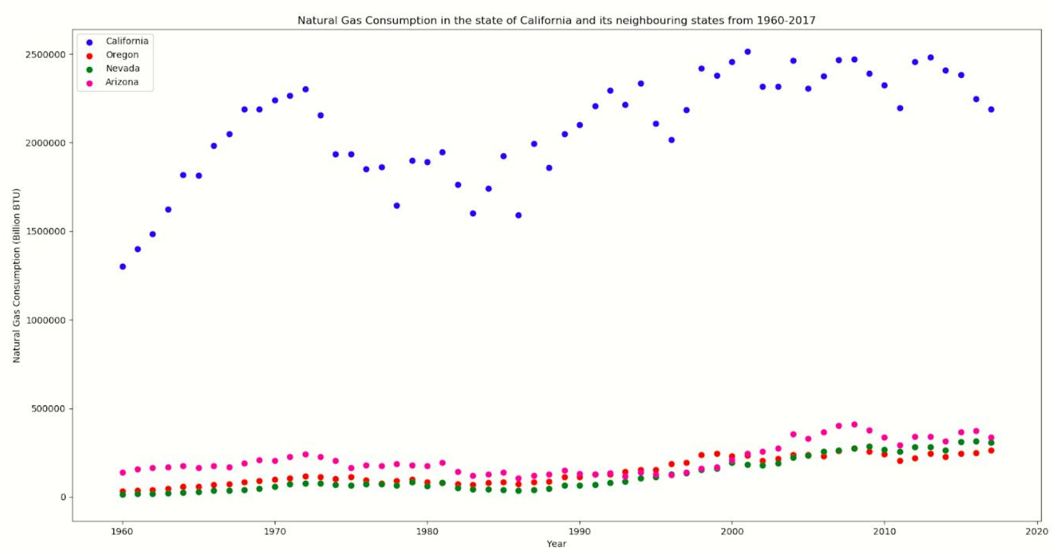

Figure 4. A graph of natural gas consumption for California and its neighbouring states

```
Machine Learning Model Results
Natural Gas Consumption in California in the future
     Year Predicted Consumption (Billion BTU)
0  [2018]                [2462606.0988621973]
1  [2019]                 [2475600.906503506]
2  [2020]                [2488595.7141448185]
3  [2030]                [2618543.7905579284]
4  [2040]                [2748491.8669710383]
5  [2050]                 [2878439.943384148]
```

Figure 5. A prediction on natural gas consumption for California in the coming years

We also chose to apply the model to forecast some more general metrics such as the combined total consumption of natural gases of the United States, the natural gas price in the United States, and the United States expenditures on natural gases for the coming years. In these applications, we remove the geographical location as an x variable and look only at the y variable of consumption, price, and expenditure as a function of time.

```
Machine Learning Model Results:
    Year Actual Consumption (Billion BTU) Predicted Consumption (Billion BTU)
5   1962                       [13730841]                [16149677.953775883]
10  1964                       [15297642]                [16526429.192760348]
11  1970                       [21692710]                [17656682.909713805]
4   1971                       [22365248]                [17845058.529206038]
8   1982                       [18515242]                  [19917190.3436206]
3   1987                       [17749685]                [20859068.441081762]
6   1993                       [21376406]                 [21989322.15803516]
1   1994                       [21870095]                [22177697.777527392]
0   1995                       [22832593]                [22366073.397019625]
2   2002                       [23582195]                [23684702.733465254]
7   2005                       [22631800]                [24249829.591941953]
9   2008                       [23898175]                 [24814956.45041865]
```

Figure 6. Screenshot of output showing predicted vs actual total natural gas consumption for the United States

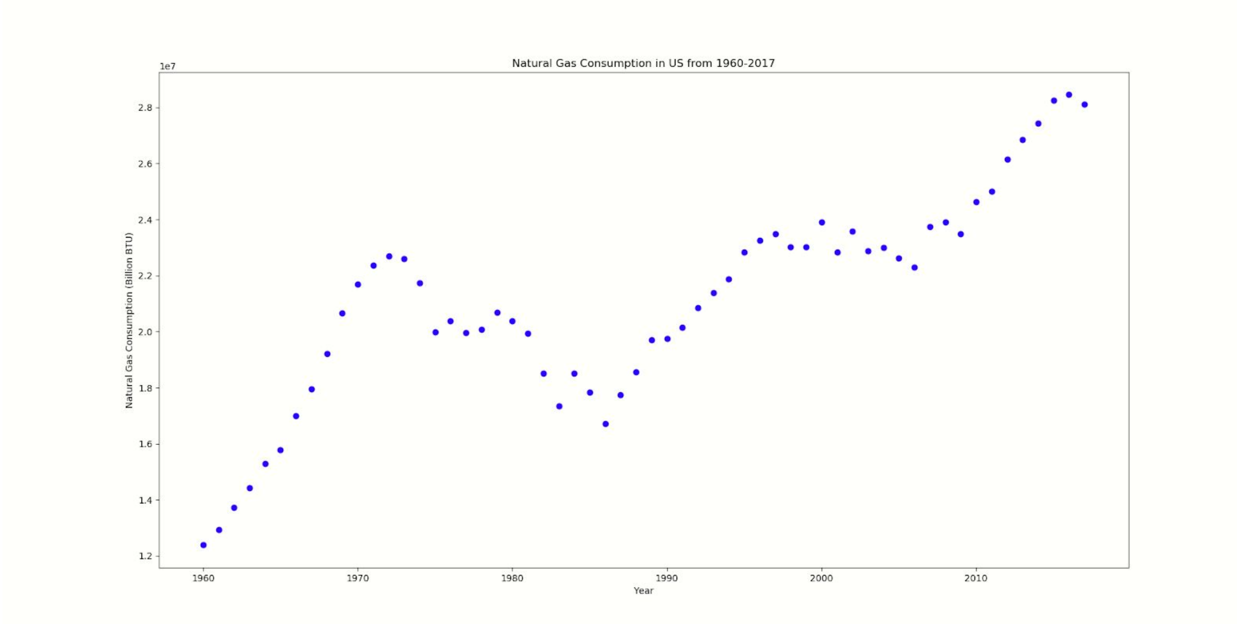

Figure 7. Graph illustrating the patterns in volume of natural gas consumption for the United States

```
Machine Learning Model Results
Natural Gas Consumption in US in the future
     Year Predicted Consumption (Billion BTU)
0  [2018]                  [26698712.64534098]
1  [2019]                 [26887088.264833212]
2  [2020]                 [27075463.884325445]
3  [2030]                 [28959220.079247832]
4  [2040]                  [30842976.27417016]
5  [2050]                  [32726732.46909249]
```

Figure 8. A prediction on natural gas consumption for the whole of United States in coming years

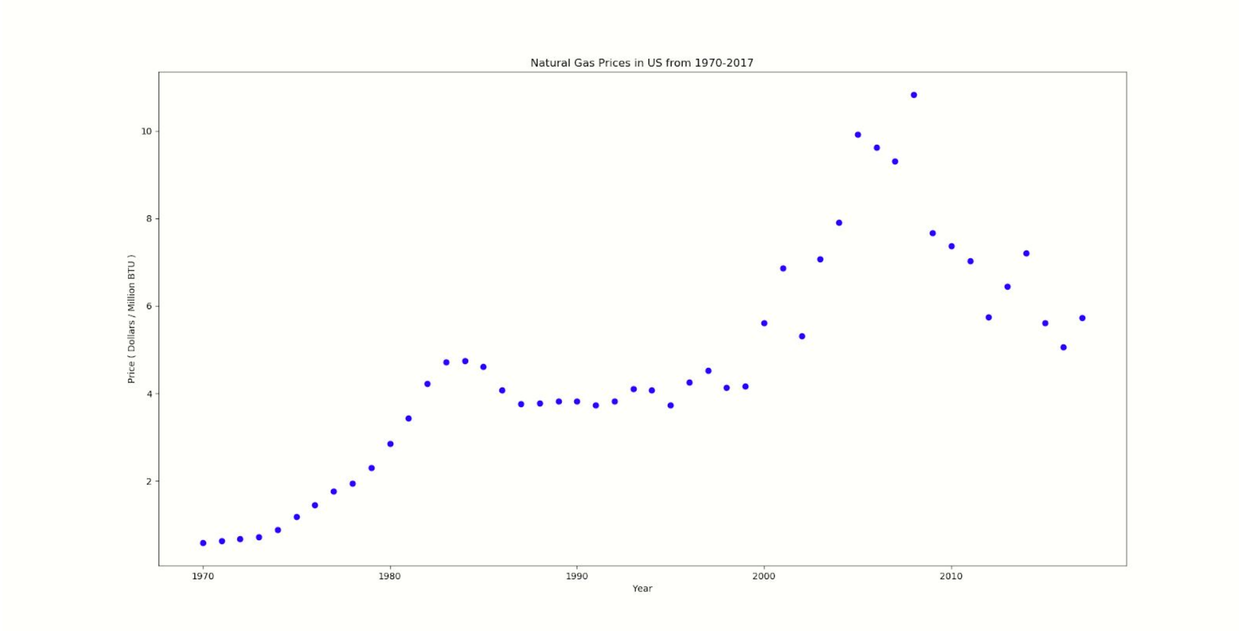

Figure 9. Graph showing the trend in natural gas prices for the United States

```
Machine Learning Model Results
Natural Gas Prices in US in the future
     Year Predicted Prices (Dollars/Million BTU)
0  [2018]                    [8.10181940668383]
1  [2019]                   [8.244040716158963]
2  [2020]                   [8.386262025634096]
3  [2030]                   [9.808475120385253]
4  [2040]                   [11.23068821513641]
5  [2050]                  [12.652901309887568]
```

Figure 10. Application of our machine learning model to project natural gas prices for the coming years

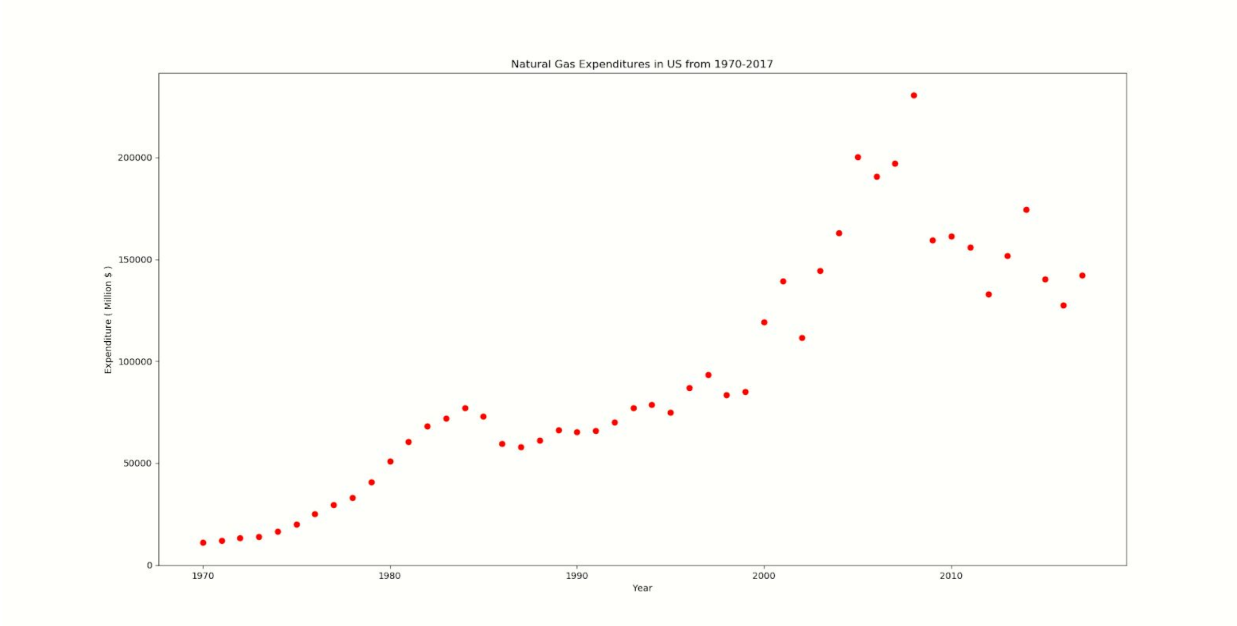

Figure 11. Graph showing the trend in total expenditures by the United States on natural gases

```
Machine Learning Model Results
Natural Gas Expenditures in US in the future
     Year Predicted Expenditures (Million $)
0  [2018]                 [179945.892158892]
1  [2019]               [183516.42348380014]
2  [2020]               [187086.95480870735]
3  [2030]               [222792.26805778593]
4  [2040]               [258497.58130686358]
5  [2050]                [294202.8945559412]
```

Figure 12. Application of our machine learning model to project natural gas expenditures for the coming years

Evaluating the Machine Learning Model

We analyzed four different metrics to evaluate the effectiveness of our model. These metrics include the Mean Absolute Error (MAE), Mean Squared Error (MSE), Root Mean Squared Error (RMSE), and the R^2 coefficient of determination.

A collection of error metrics for the appropriate figures are given below. Each of these numbers (rounded) were reported as follows:

Predicted	MAE	MSE	RMSE	R^2
CA consumption (fig. 5)	162340	4356367727	208719	0.44
US consumption (fig. 8)	1728294	4989415357537	2233700	0.51
US natural gas prices (fig. 10)	0.9152271	1.46176552	1.209035	0.78
US natural gas expenditures (fig. 12	17264	553738045	23532	0.83

Table 1. Error metrics for predicted results

Of these values, we believe that the R^2 value is worth particular notice. The R^2 value reported here is expected to be low. The R^2 value is a representation of how well the x variables together are able to explain the variations observed in the y variables, and can range from 0.0 (the x variables don't explain any y variation at all) to 1.0 (the x variables completely account for all variations in y). In our model, we look only at two x variables: that is, the geographic coordinates (of which are arbitrarily assigned by us and exists more so as ordinal variables than numerical) and the year. It is no surprise then, that these two variables alone are only able to explain a low percentage of the variations observed in the y variables.

However, the R^2 values for the US natural gas price and expenditures projections indicate that the current model, taking into account only time, can largely account for a major portion of the variations in US natural gas price and expenditures.

Limitations and Areas of Improvement

We choose to employ the model using only two variables for the sake of simplicity and as an adaptation to the restraint of time given.

The inclusion of more x variables will allow for the model to predict consumption more accurately. Factors such as population and GDP are critical and should be considered to vastly improve the power of this model. Other non-numerical factors such as prominent industries and sectors are also important. It is recommended that this model is further fleshed out and given more x variables as input to better explore and develop a more accurate prediction on natural gas consumption.

In this report, we opt to examine the consumption for the state of California and its surrounding states (Oregon, Nevada and Arizona) . This provides a better basis for a case study focused on California. Including more states may allow for a more global trend to be discovered and thus more general solutions may be adapted. However, this takes away more insight at a more state-by-state level. The ability to include more or less states may be a valuable ability that can be controlled accordingly to match management objectives.

Discussion of Results

In this report, we present projections for four different metrics:

1. The natural gas consumption for just the state of California
2. The natural gas consumption for the whole of the United States
3. The price of natural gas in the United States
4. The expenditures on natural gas by the United States

In each of these four metrics, we see an alarming upward trend in consumption, prices, and expenditures. These results when used to supplement larger studies may allow for stronger arguments to be made. For example, the results displayed in this report can provide support for larger studies examining the prominence of different energy sources and the impacts that they might have on our environment. Natural gases leave a much larger footprint on the environment compared to geothermal, solar, wind, and hydro sources of energy. The projected increases related to natural gas may come across as a point of improvement for facility managers and thus promote proactive measures to improve energy consumption by switching to more environmentally-friendly sources of energy.

In this way, we address problem topic #1 by providing critical information to allow for earlier and better planning and strategizing of making changes to energy supply chains to address resiliency needs and availability for development requirements.

Artificial intelligence is a blooming area in technology. One such application of artificial intelligence is the usage of machine learning for evaluating metrics related to energy usage and costs. This provides key information to management and alleviates the burden of those concerned with finding trends related to energy. Our machine learning model, albeit primitive, clearly demonstrates an opportunity for technology and its advancements to contribute to addressing sustainability and environmental changes.

The results presented here provide a basis for further work to be done. Under less-strict constraints of time, further efforts into better developing this machine learning model will very likely yield promising results.

Case Study: Smart Meters

Background

A smart meter is a small electronic device that tracks electricity consumption and sends the data to the electricity supplier for monitoring and billing. Traditional analog electrical meters only recorded total consumption. Smart meters are able to track this information daily or even hourly.

Advantages

One of the main benefits for customers is the ability to see their usage in real-time, allowing users to manage their energy usage and reduce their energy bill. It also gives way for more accurate energy billing to decrease discrepancies between customers and suppliers. For the electricity suppliers, smart meters can gather more data with more specific timestamps which would allow for better energy usage prediction. With the increased adoption of smart meters in a city or area, it creates a smart power grid where the electricity suppliers can monitor usage on a larger scale. According to electricity suppliers, this has led to increased efficiency in their operations and customers have generally enjoyed the benefits of live data.

Disadvantages

With new technology comes drawbacks and issues. One of the main concerns is security - because a smart meter is essentially connected to the internet in order to upload its data; it can be vulnerable to cyberattacks. Such cyberattacks may lead to stolen data or manipulated data. Despite this, many electricity suppliers are continuing its rollout while dealing with these issues in both software or hardware updates to the technology.

Integration with Machine Learning

The primary requirement for an effective application of Machine Learning is having a large enough dataset for proper training. Smart meters are able to provide hourly usage as well as their specific locations to the electricity suppliers and with a whole smart power grid, there is a massive amount of data available. If electricity suppliers implement an ML application for electricity usage, it would allow for better business modelling and analysis for trends in their customer and area usage. A continuous stream of data will allow the ML model to learn in real-time accounting for both historical and recent trends in data and advise the electricity suppliers appropriately.

Conclusion

Using machine learning, we have developed a model to predict the natural gas consumption for the state of California, the natural gas consumption for the United States of America, the price of natural gas in the United States, and the expenditures on natural gas resources by the United States for years to come. Of which, each of these metrics are expected to rise in coming years. This information may be beneficial to those in the field of FM, allowing for proactivity in addressing a variety of concerns such as energy costs and environmental pollution. Machine learning is one facet of artificial intelligence, an extremely hot topic currently in the field of technology. Another rising product of artificial intelligence is augmented reality (AR) and virtual reality (VR). As artificial intelligence technology continues to advance, attractive opportunities for solutions addressing environmental concerns will continue to present themselves. This interplay between the natural balance of our world and technology will be an increasingly interesting point of focus in coming years.

Sources of information:

1. https://www.eia.gov/ - U.S. Energy Information Administration (EIA)
2. https://www.standard.co.uk/tech/smart-meters-cyber-security-risks-a3877071.html - Here's why people aren't installing smart meters | London Evening
3. https://www.govtech.com/fs/infrastructure/Oregon-Utility-Makes-117-M-Move-to-Smart-Meters.html - Oregon Utility Makes $117M Move to Smart Meters
4. https://ecoviewhomes.com/types-of-eco-friendly-energy-sources/ - Types of Eco-Friendly Energy Sources - EcoView
5. https://www.conserve-energy-future.com/pros-and-cons-of-natural-gas.php - Pros and Cons of Natural Gas - Conserve Energy Future

Additional Figures

```r
setwd("C:/Users/tinso/PycharmProjects/BGIS-HackTheClimate2020/dataset/us_energy_census_gdp_10-14/")

library("readxl")
library("editData")

#my_data <- read.csv(file = 'Energy Census and Economic Data US 2010-2014.csv')

consumption <- read_excel("use_ng_capita.xlsx", sheet="Btu")
prodexpnd <- read_excel("pr_ex_pa_ng.xlsx")

#attach(my_data)
attach(consumption)

#coal_data <- subset(my_data, select=-c(Region:Great.Lakes, TotalP2010:BiomassC2014, ElecC2010:RNETMIG2014))

#Trimming and editing here
consumption_trimmed <- consumption[-c(1, 2),]
names(consumption_trimmed)[1] <- "States"
names(consumption_trimmed)[2:50] <- "1960":"2008"
states <- subset(consumption_trimmed, select=c("States"))
states$Position_X <- c("1":"52")
states = edit(states)
write.csv(states,'StatesXPositions.csv' )
states$Position_Y <- c("1":"52")
states <- edit(states)
write.csv(states, 'StatesXYPositions.csv')
combined = combined[,c(1, 60, 61, 2:59)]
write.csv(combined, 'TransformInExcel.csv')
```

Figure 13. Code in R to format datasets to be used in subsequent machine learning model

```python
# 9 February 2020
#
# Basic Machine Learning script using sklearn's LinearRegression() model.
#
# Datasets used was obtained from the USA Energy Information Administration (EIA)
#     and was formatted by us.

import numpy as np
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn import metrics

dataset = pd.read_csv("csv/usa_natgasc_clean.csv")

dataset_CA = dataset.loc[ (dataset['State'] == 'CA' ) ]
x = dataset_CA[['X', 'Y', 'Year']].values.reshape(-1, 3)
y = dataset_CA['NatGasC'].str.replace(",", "").astype(int).values.reshape(-1, 1)

# dataset_US = pd.read_csv("csv/usa_natgas_prices.csv")
# x = dataset_US['Year'].values.reshape(-1, 1)
# y = dataset_US['Prices'].values.reshape(-1, 1)

# dataset_US = pd.read_csv("csv/usa_natgas_expenditures.csv")
# x = dataset_US['Year'].values.reshape(-1, 1)
# y = dataset_US['Expenditures'].str.replace(",", "").values.reshape(-1, 1)

# Split our dataset into training and test sets.
x_train, x_test, y_train, y_test = train_test_split(x,
                                                    y,
                                                    test_size=0.2,
                                                    random_state=0)

# Initialize our Machine Learning model and fit the training data.
model = LinearRegression()
model.fit(x_train, y_train)

# Assign y_pred to our predicted values from x_test values.
# The y_test values will be holding the CORRECT values.
y_pred = model.predict(x_test)
```

Figure 14. Part 1 of 2 of our machine learning program in Python

```python
# Store our actual and predicted y values (Natural Gas Consumption) and store them into a Pandas DataFrame.
results = pd.DataFrame({'Year': x_test[:,2].tolist(),
                        'Actual Consumption (Billion BTU)': y_test.tolist(),
                        'Predicted Consumption (Billion BTU)': y_pred.tolist()}).sort_values(by=['Year'], ascending=True)

# Uncomment the print below to see our results.
print("Machine Learning Model Results: ")
print(results)

# plt.title("Natural Gas Consumption in California 1960-2017")
# plt.xlabel("Year")
# plt.ylabel("Natural Gas Consumption ( Billion BTU )")
# plt.scatter( x_test[:,2], y_pred, label='Predicted')
# plt.scatter( x_test[:,2], y_test, label="Actual")
# plt.legend()
# plt.show()

# Below are various error measurements printed to the screen.
# print('Mean Absolute Error: ', metrics.mean_absolute_error(y_test, y_pred) )
# print('Mean Squared Error: ', metrics.mean_squared_error(y_test, y_pred) )
# print('Root Mean Squared Error: ', np.sqrt(metrics.mean_squared_error(y_test, y_pred)) )
# print('R^2: ', model.score(x_test, y_test) )
```

Figure 15. Part 2 of 2 of our machine learning program in Python

```python
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt

# dataset = dataset.loc[ (dataset['State'] == 'CA') |
#                        (dataset['State'] == 'OR') |
#                        (dataset['State'] == 'NV') |
#                        (dataset['State'] == 'AZ') ]

# plt.xlabel("Year")
# plt.ylabel("Natural Gas Consumption (Billion BTU)")
# plt.title("Natural Gas Consumption in the state of California and its neighbouring states from 1960-2017")
#
# labels = [ 'California', 'Orgeon', 'Nevada', 'Arizona' ]
#
# dataset_CA = dataset.loc[ (dataset['State'] == 'CA' ) ]
# x = dataset_CA[['X', 'Y', 'Year']].values.reshape(-1, 3)
# y = dataset_CA['NatGasC'].str.replace(",", "").astype(int).values.reshape(-1, 1)
# plt.scatter( x[:, 2], y, c='b', label='California')
#
# dataset_OR = dataset.loc[ (dataset['State'] == 'OR') ]
# x = dataset_OR[['X', 'Y', 'Year']].values.reshape(-1, 3)
# y = dataset_OR['NatGasC'].str.replace(",", "").astype(int).values.reshape(-1, 1)
# plt.scatter( x[:, 2], y, c='r', label='Oregon')
#
# dataset_NV = dataset.loc[ (dataset['State'] == 'NV') ]
# x = dataset_NV[['X', 'Y', 'Year']].values.reshape(-1, 3)
# y = dataset_NV['NatGasC'].str.replace(",", "").astype(int).values.reshape(-1, 1)
# plt.scatter( x[:, 2], y, c='g', label='Nevada')
#
# dataset_AZ = dataset.loc[ (dataset['State'] == 'AZ') ]
# x = dataset_AZ[['X', 'Y', 'Year']].values.reshape(-1, 3)
# y = dataset_AZ['NatGasC'].str.replace(",", "").astype(int).values.reshape(-1, 1)
# plt.scatter( x[:, 2], y, c='m', label='Arizona')
#
# plt.legend()
# plt.show()
```

Figure 16. Part 1 of 2 of our code to generate graphs in Python

```python

# dataset_US = pd.read_csv("usa_natgas_prices.csv")
# x = dataset_US['Year'].values.reshape(-1, 1)
# y = dataset_US['Prices'].values.reshape(-1, 1)
# plt.title("Natural Gas Prices in US from 1970-2017")
# plt.xlabel("Year")
# plt.ylabel("Price ( Dollars / Million BTU )")
# plt.scatter( x, y, c='b', label='US')
# plt.show()

# dataset_US = pd.read_csv("usa_natgas_expenditures.csv")
# x = dataset_US['Year'].values.reshape(-1, 1)
# y = dataset_US['Expenditures'].str.replace(",", "").astype(float).values.reshape(-1, 1)
# plt.title("Natural Gas Expenditures in US from 1970-2017")
# plt.xlabel("Year")
# plt.ylabel("Expenditure ( Million $ )")
# plt.scatter( x, y, c='r', label='US')
# plt.show()
```

Figure 17. Part 2 of 2 of our code to generate graphs in Python