PROJECT REPORT

Smart Home Dataset with Weather Information

Group 2 - Storm Troopers

Vandana Chandola	:	014748604
Nikhila Churia	:	014546116
Haasitha Pidaparthi	:	012669254

INTRODUCTION

IoT brings together everything at home under one umbrella which has the potential to monitor and remote control such as air conditioning, alarm system, lighting, heating, ventilation, telephone system, tv, etc. to enhance our comfort and security with low energy consumption. The home is a specific environment, and energy management is one of the IoT use cases with which energy being sent out or consumed can be monitored. One can monitor each of the IoT appliances and how much power each of the devices is consuming, and easily switch between energy-efficient appliances across the day. The energy generation and consumption varies with weather attributes like temperature, precipitation etc. For example, during the winter, people use heaters and the use of air conditioning drops drastically and vice versa. Our project aims at predicting this change in energy consumption according to the weather based on the readings taken by a smart meter with a time span of 1 minute of 350 days of house appliances in kW from a smart meter and weather conditions of that particular region. Analysis of the data collected from the IoT devices can help in monitoring the energy consumption patterns and, in turn,control the energy consumption more efficiently.

SYSTEM DESIGN AND IMPLEMENTATION

ALGORITHMS USED:

ARIMA (AUTO REGRESSIVE INTEGRATED MOVING AVERAGE)

ARIMA is one of the easiest and effective algorithms for performing time series forecasting. It is a statistical analysis model that uses time series data to either better understand the data set or predict future trends. We have used ARIMA in our project to predict future overall energy consumption per day in a smart home.

ISOLATION FOREST

Isolation forest is a machine learning algorithm for anomaly detection. It's an unsupervised learning algorithm that identifies anomalies by isolating outliers in the data. It is based on the Decision Tree algorithm that isolates the outliers by randomly selecting a feature with the dataset and then randomly selecting a split value between max and min values of that particular feature. Using Isolation Forest not only helps us detect anomalies faster, but it also requires less memory compared to the other algorithms.

TECHNOLOGIES AND TOOLS USED

To work on this project, we collaborated on Google Colaboratory. Also, we created a google drive to add documents for Project Proposal and Project Report, PowerPoint slides for Presentation, and CSV files containing the dataset to access using the colaboratory. We also used the WhatsApp application for communicating with the project group members. We primarily used Google applications because it is easy to collaborate with everyone and view the changes the members are making instantly. Lastly, we used the WhatsApp application for communicating because all the members were comfortable with it.

SYSTEM AND SUBSETS

The dataset we worked on recorded the energy consumption of each room/appliance every minute, from January 1st 2016 (5am) to December 16th 2016 (3:29am). In order to work Smart Home Dataset, we split the data into subsets: energy_data and weather_data.

- Energy_data holds all attributes that are related to the energy consumption of rooms and appliances including, kitchen, furnace, home office, wine cellar, etc.
- Weather_data holds all attributes that are related to the weather conditions including, temperature, humidity, precipitation, dew point, visibility, etc.

We further split the energy_data into 3 subsets: energy_per_day, energy_per_week, energy_per_month.

- Energy_per_day includes the sum of energy consumption for each attribute for the entire day
- Energy_per_week includes the sum of energy consumption for each attribute for the entire week
- Energy_per_month includes the sum of energy consumption for each attribute for the entire month

EXPERIMENTS EVALUATION

DATASET USED

- The dataset we have used for this project is titled “Smart Home Dataset with Weather Information”, which has been downloaded from https://www.kaggle.com/taranvee/smart-home-dataset-with-weather-information.
- The dataset contains more than 500,000 readings with the time-span of 1 minute of the energy used (in kW) by the appliances of a smart home in the year 2016, and the weather conditions of the area at that time.

Data Preprocessing:

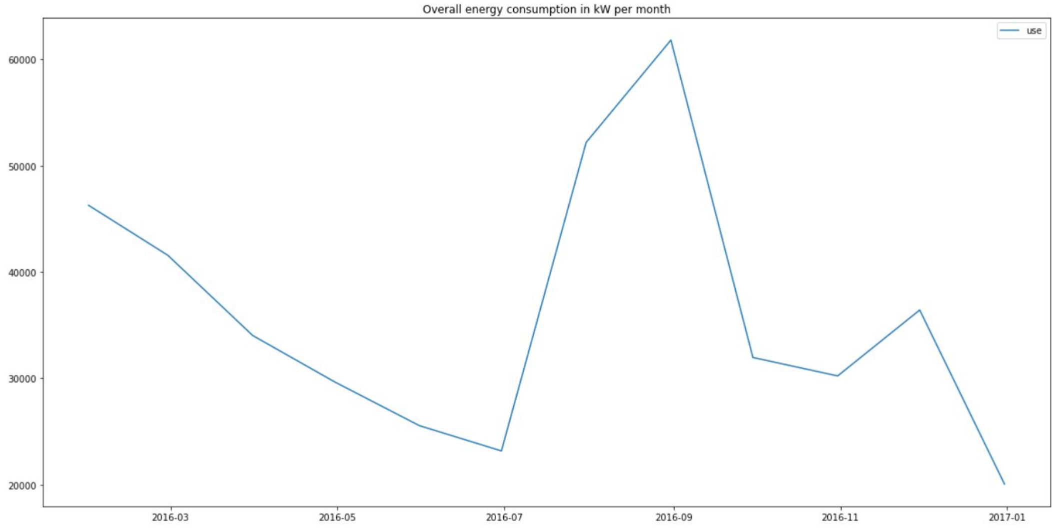
An important step in the process is to make sure that the data is complete and satisfies all the requirements for data analysis. Various pre-processing techniques have been used for the same such as removing invalid rows, changing the column names into more convenient names, performing aggregation on certain columns, converting unix timestamp to proper date format, replacing missing values in columns with the next valid observation, removing unwanted columns and duplicate columns etc. The dataset was resampled to daily and monthly datasets based on the analysis requirements.

METHODOLOGY

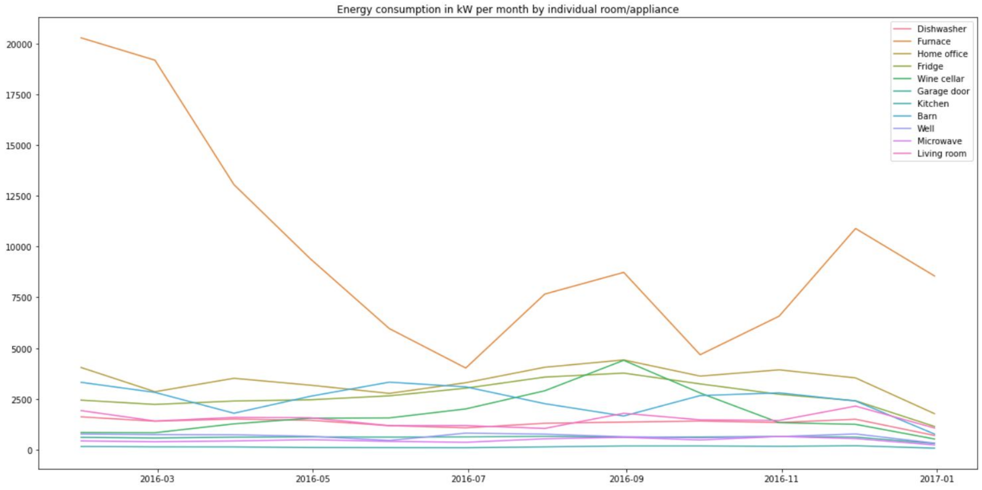
- **Visualization:**

We have visualized the time series data to find the pattern of energy consumption by the devices in a smart home. We have resampled the data accordingly and plotted the graphs to showcase the trend in a daily, weekly and monthly manner.

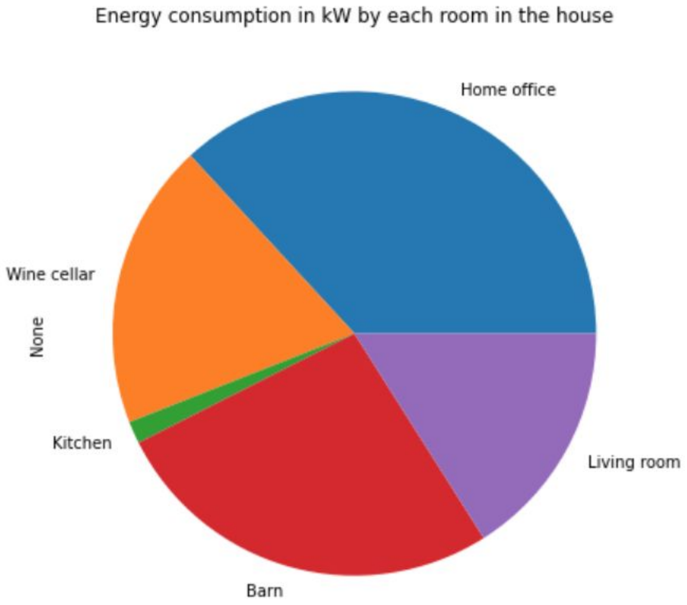
> The below plot shows that August and September are the months with the highest energy consumption.



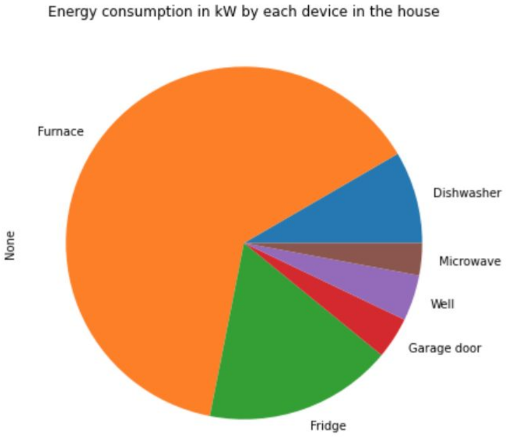
> The below plot indicates that furnace has the highest energy consumption among the rooms/devices and kitchen has the lowest in the smart home in a month.



> The below pie chart shows Home office has the highest energy consumption among all the rooms in the smart home in a month.



> The below pie chart shows Furnace has the highest energy consumption among all the devices in the smart home in a month.



- **Correlation between features:**

  Indicates the relationship between features, values ranging from -1 to 1. There are two key components of a correlation value:

  - **magnitude** : The larger the magnitude (closer to 1 or -1), the stronger the correlation
  - **sign** : Negative indicates inverse correlation. Positive indicates regular correlation.

  Here we use correlation to identify whether there is any kind of significant relationship between appliances, between weather data, and overall (i.e. interdependence between weather data and appliances) as follows:

  - Correlation between energy data
  - Correlation between weather data
  - Correlation between all data

- **Time Series Analysis:**

  3 Factors of Time-Series Analysis were used to analyze the nature of the dataset:

  1. **Autocorrelation:**

     Autocorrelation measures a set of current values against a set of past values to see if they correlate. We calculate the correlation for current time-series observations with observations of previous time steps called lags. For example, one might expect the energy usage at the 1st minute of the day to be more similar to the usage at the 2nd minute rather than at a minute during mid-day. Data that has strong autocorrelation is not random, and provides high predictability.

In this experiment, we use autocorrelation to determine the randomness and predictability of energy generation and usage.

2. **Seasonality:**

   Seasonality is the repeating short-term cycle in the series. The dataset used for this is the energy usage on a daily basis, and it showed somewhat of a cyclic behavior on a monthly basis. Therefore, used seasonal decomposition with parameters model=”additive” and freq=30.

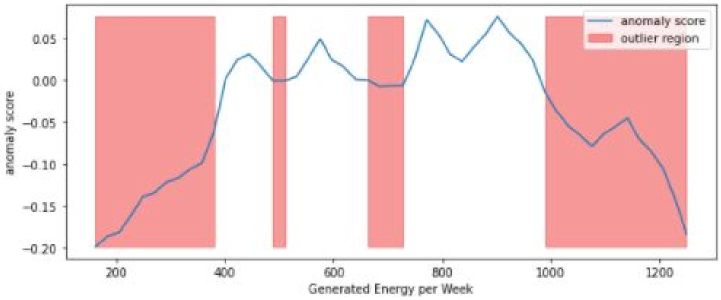
3. **Stationarity:**

   Summary statistics calculated on the time series are consistent over time i.e. the mean or the variance. To observe stationarity, we split the dataset into two contiguous halves. For each subset, we calculate mean and variance and compare the values. If they differ, and the difference is statistically significant, the time-series is non-stationary and needs to be made stationary.

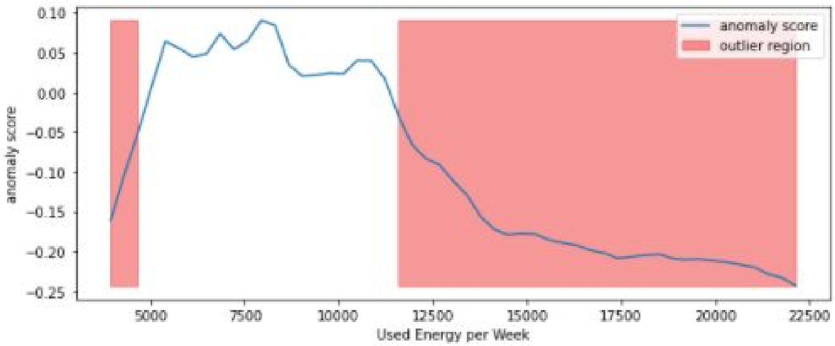
   It was observed that the values differ but are essentially in the same ballpark.

- **Anomaly Detection**
  - Isolation Forest

    Performed Isolation Forest algorithm to identify the outliers in generated and used energy attributes. The algorithm was applied to observe the outlier patterns for the subsets of energy_per_day, energy_per_week, and energy_per_month. The splits for the isolation forest were performed using the max and min values of each of the features. The anomaly scores were calculated using the decision function and the anomaly attribute using the predict feature of isolation forest. The graph below shows the isolation forest plot that has the anomaly scores of the generated energy per week and displays the regions that have outlier values in red.

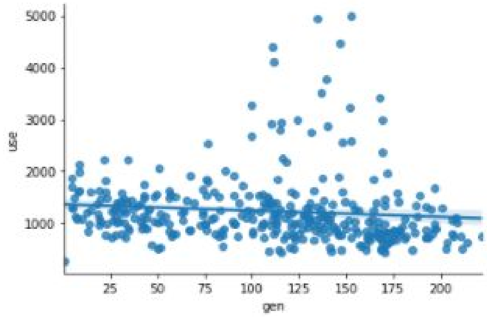


The graph below shows the isolation forest plot that has the anomaly scores of the used energy per week and displays the regions that have outlier values in red.



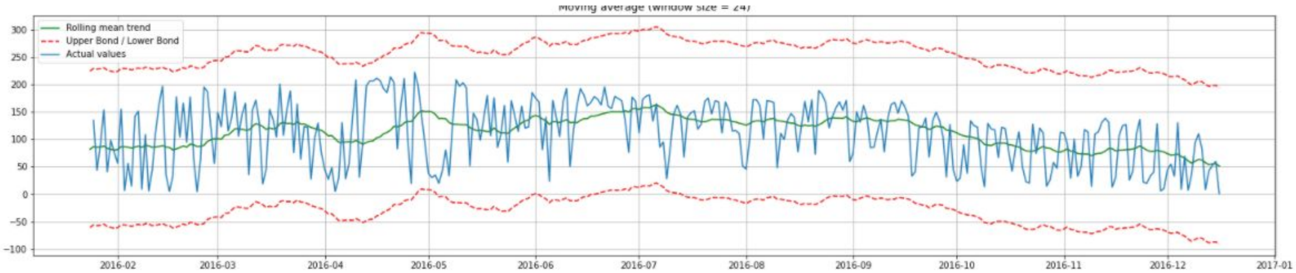
- ○ Multivariate Anomaly Detection

We used point plot with regression on generated energy and used energy to observe any patterns and anomalies found in relation to the two attributes. We plotted the graphs according to the energy_per_day, energy_per_week, and energy_per_month datasets. The graph below shows the regression plot of use vs generated energy per day. Looking at this plot,, we can notice that there is a strong relation between used and generated energy. The points that fall outside the regression line are the outliers in the data set.

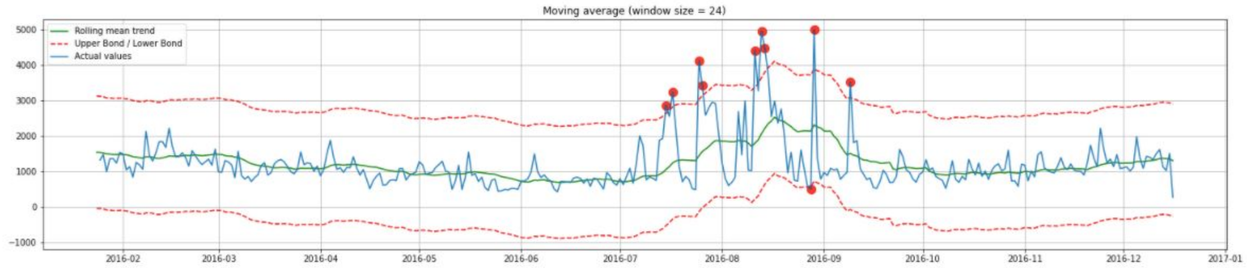


- ○ Moving Average Technique

We performed the moving average technique on both used and generated energy attributes. Moving Average takes the average of the dataset values to obtain the rolling mean trend on the dataset. For our application, we calculated the lower and upper bounds, along with the rolling average, in order to observe the outliers in the dataset. The graph below shows the moving average of generated energy using the window 24, along with lower and upper bounds that are used to detect outliers in the data.

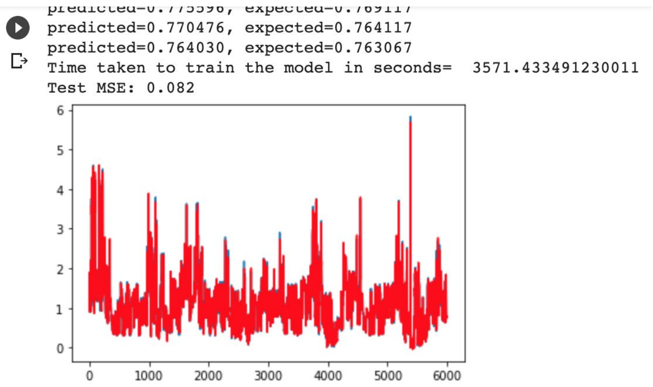


The graph below shows the moving average of used energy using the window 24, along with lower and upper bounds that are used to detect outliers in the data.

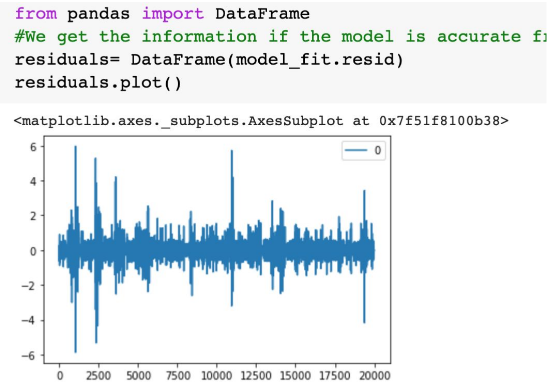


- **Training the Auto Regressive Integrated Moving Average(ARIMA) model for time series forecasting and Residual analysis:**

An autoregressive integrated moving average, or ARIMA, is a statistical analysis model that uses time series data to either better understand the data set or to predict future trends. We have used the ARIMA model in our project to predict the future values of total energy used in the smart home (column= ‘use’) by training the model on a dataset of size 20000 records. The train dataset consists of 70% of the dataset and the remaining 30% has been used as the test dataset. The p and q parameters of the model have been chosen to be 3 and based on the PACF and ACF plots respectively and since the data is stationary we could choose the d value as zero, but we have also experimented with first and second order differencing . Further details and observations have been listed below along with the screenshots.



```
predicted=0.775596, expected=0.769117
predicted=0.770476, expected=0.764117
predicted=0.764030, expected=0.763067
Time taken to train the model in seconds=  3571.433491230011
Test MSE: 0.082
```

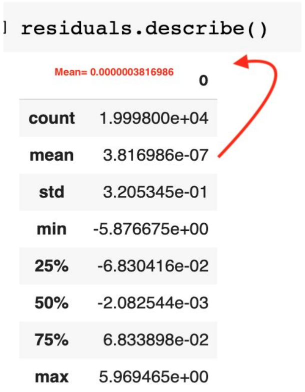


```
from pandas import DataFrame
#We get the information if the model is accurate fr
residuals= DataFrame(model_fit.resid)
residuals.plot()
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7f51f8100b38>
```

1. It took 3571.43 seconds, approximately equal to 1 hour to train the model on a dataset of the size 20000 records.
2. The prediction accuracy of the model is indicated by the MSE(Mean squared error), whose value is 0.082.
3. We have plotted the predicted values and the test values as indicated by red and blue colors respectively and from the plot it is clear that our model seems to make accurate predictions.
4. We have done residual analysis to check whether the model has adequately captured the information in the data. A good forecasting method will yield residuals with the following properties:

- The residuals are uncorrelated. If there are correlations between residuals, then there is information left in the residuals which should be used in computing forecasts. In the below plot we can see that there is no pattern or trend and does not have any correlations between residuals.
- The residuals have zero mean. If the residuals have a mean other than zero, then the forecasts are biased.



```
] residuals.describe()
```

Mean= 0.0000003816986

	0
count	1.999800e+04
mean	3.816986e-07
std	3.205345e-01
min	-5.876675e+00
25%	-6.830416e-02
50%	-2.082544e-03
75%	6.833898e-02
max	5.969465e+00

CONCLUSION

Smart Home Dataset records the energy consumption of each room/appliance in a house with an interval of one minute for the entire year of 2016. Initially while observing the dataset, there were several decisions we took in order to proceed with the data. After performing the pre-processing steps, we decided to use correlation and time series analysis to better comprehend the data and observe any patterns. Time-Series Analysis was used to understand trends and inconsistencies in time-related data. We have used it to detect abnormalities and forecast data. We added Anomaly Detection in order to observe any outliers in the attributes of generated energy and used energy. In order to perform anomaly detection, we used the Isolation Forest algorithm by using the max and min values of the attributes. Since we didn't get sufficient data, we also used Moving Average to detect any outliers that weren't observed in the Isolation Forest. Lastly, time series forecasting has been done using the ARIMA model to predict the future values of total energy used by all the devices/rooms in the smart home. One of the difficulties we faced has been the run time taken for the ARIMA model while experimenting and testing with different values for the parameter values(p,d and q) and hence we decided to test on a smaller resampled dataset and came to the conclusion that the best parameters for the model are p=5, d=1 and q=0. Since the dataset only included the energy consumption for the year 2016, the predictions we made with the given dataset might not be that accurate for future forecasting in the later years. Having data that displays the values for several years might help us make better conclusions about the energy consumption and weather conditions.

Task Distribution

1. Data Preprocessing and data preparation- Nikhila Churia, Vandana Chandola and Haasitha Pidaparthi
2. Data Visualization- Nikhila Churia
3. Correlation between features - Vandana Chandola
4. Time Series Analysis - Vandana Chandola
5. Anomaly detection- Haasitha Pidaparthi
6. Time series forecasting (using the ARIMA model) and residual analysis- Nikhila Churia