CMPE 255 - Data Mining

Smart Home Dataset with Weather Information

Project Group 4

Submitted to

Prof. Jorjeta Jetcheva

on

15/05/2021

by
Hanish Punamiya (015271971)
Rohit Sikrewal (015220634)
Sahil Bhagchandani (015264873)

GitHub Link: https://github.com/sikrewalrohit/CMPE-255-Tearm-Project.git

Table of Contents

Introduction

Motivation

Our monthly electricity bill varied over the days that we had been tracking it. We noticed a trend that the energy consumption was higher when the temperature was either very high or very low. We started wondering if the electricity consumption could be predicted based on the temperature patterns and if other weather-related information played a part in it too.
We wanted to know if we could accurately predict our energy consumption based on weather information like temperature, humidity, precipitation, etc.
During our search for such a dataset, we luckily came across the smart home dataset with weather information. This time series dataset contained extensive information about energy consumed in a household by various appliances along with the weather information for every minute. We felt this was the perfect dataset for our current needs and decided to work on it.

Objective

- The objective of our project was to determine if the amount of energy consumed and generated by a household could be predicted on the basis of weather information.
- Based on the dataset that we chose, we wanted to see if we could accurately predict the energy generation and energy consumption values of the household by supplying weather information as parameters for the prediction.
- If our predictions were reasonably accurate, we wished to predict future energy consumption and generation values for the household.

Approach

The way we approached this task was by thoroughly analyzing the data and comparing our prediction to the actual values as well as predicted values generated by not using the weather information.
To clarify, we first wanted to analyze the data to see if we could visually see any relationship between weather conditions and energy values. After which, we would predict these energy values by using two kinds of models:

- Models that would predict future energy values based on its own past values.
- Models that would take weather information into consideration while predicting future energy values.

The reason we decided to do this was to compare how well the two models compare against each other as well as to see how accurately they predict these energy values.
This would help us determine if taking weather information into consideration when predicting energy consumption and generation of a household is better than just predicting these values based on past trends in energy values.

System Design and Implementation

Algorithm considered

We decided to predict the future energy values using two different approaches:

- To predict the values based solely on its past values. For this approach we decided to use ARIMA and SARIMAX models.
    - **ARIMA** was used because our dataset was a time series dataset and we thought it would help us better understand the data and predict future values.
    - **SARIMAX** was used because during the data visualization, we noticed a certain seasonality to the energy values and thought that SARIMAX would be really helpful for future predictions in this case.
- To use models that would use weather information as parameters to predict the future energy values. We decided to use VAR, Prophet and LightGBM.
    - **VAR** was selected to model our data was because it provided us with a simple and straightforward way of plugging vital weather information as regressors into our model to predict energy values.
    - During the visualization of the time-series analysis for energy values of the dataset, we noticed a seasonal pattern in the daily, weekly and monthly energy values. We thought that we could build a time-series forecasting model with **fbprophet** to predict future energy values. By adding weather information as regressors to this model, we can take into consideration the influence of weather.
    - After some research we decided to go with **LightGBM** as it ticked all of our boxes to predict future energy values and understand the relationship between energy consumption and weather information.

Technologies, tools and libraries used

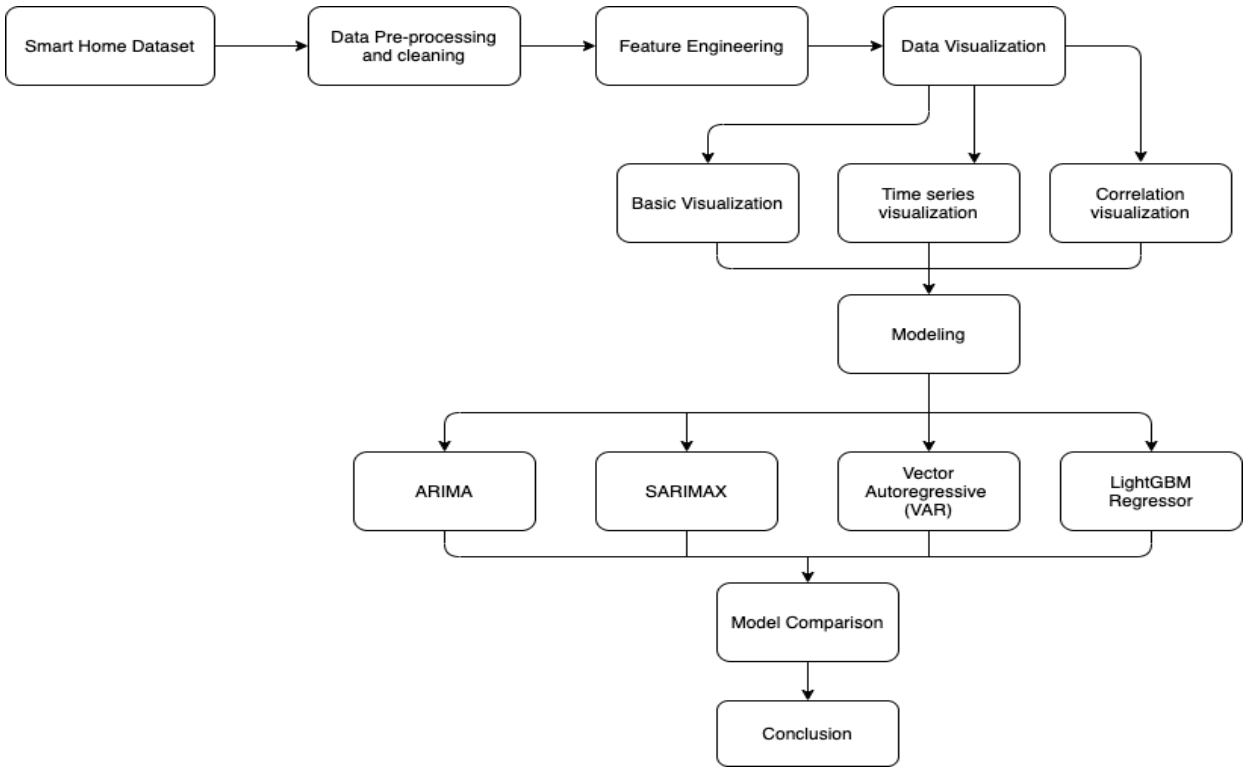
- Jupyter Notebooks
- Python
- Pandas
- Holoviews
- Matplotlib
- Seaborn
- Scipy
- Statsmodels
- Sklearn
- Tabulate
- Shap
- LightGBM

Architecture-related decisions

Our architecture decision was to build a system that had a simple straight-forward flow.

- We studied the data and performed **preprocessing**.
- We needed to visualize the data to get a better idea of how the energy values and weather information are all related to each other over time. In order to do that we needed to do some **feature engineering**.
- After which we extensively **visualized** our data. We decided to do three types of data visualization analysis on both weather information and energy values,
    - Basic analysis
    - Time-series analysis
    - Correlation analysis.
- Following visualization, we decided to apply **predictive models** to our data. Based on the MAE and MSE values that were calculated from the four models, we determined which model worked best.
- Finally, we arrived at a **conclusion** for whether future energy values could be accurately predicted based on weather information from all the work done before.

System architecture and data flow



Experiments

Dataset

- **Name:** Smart Home Dataset with Weather Information
- **Source:** Kaggle, https://www.kaggle.com/taranvee/smart-home-dataset-with-weather-information
- **Type of Data:** Time-series
- **Size of Data:** 124.89 MB
- **No of Columns:** 32
- **No of Rows:** 504K

Methodology

Date Preprocessing

- We searched for missing values and found just one row with missing values which was subsequently removed.
- We also removed the units from the column names.
- The 'cloudCover' column was of type object and had some invalid values which we replaced with the next valid value and changed the column type to float so that we could use it for evaluation.

Feature Engineering

- Since the dataset contains information for every minute, we needed to perform quite a few actions on the time column. We aggregated the time column and extracted the hourly, daily,

weekly and monthly information from it. Furthermore, we aggregated the hourly column to get time of day information by dividing it into morning, afternoon, evening and night.

- There were certain columns related to energy consumption by household appliances that held the data for the same appliances. We combined their values to get a total for energy consumption by that appliance.
- We created a correlation matrix to check for duplicate columns. We found that the pair of 'use' and 'house' and the pair of 'gen' and 'solar' had a correlation coefficient of greater than 0.95. We then dropped one column and renamed the other from both the pairs to 'use_HO' and 'gen_Sol' respectively.

Data Visualization

After the data was prepared we performed extensive data visualization analysis on it. This was done in three steps.

- Basic Analysis of
  - Total energy generated
  - Total energy consumed
  - Energy consumption by various appliances
  - Energy consumed by various rooms in the house
  - weather information checking the density of distribution of values
- As our dataset is a time series dataset, second was a time series analysis on our dataset. This time series analysis was done based on hourly, daily, weekly, monthly and time of day values on
  - Total energy consumption
  - Total energy generation
  - Energy consumption for every house appliance
  - All weather conditions
- Correlation analysis on
  - Energy consumption of all the household appliances
  - Weather information
  - Household appliances and weather information together

From all of the visualizations we could infer the weather attributes affected particular household appliances' energy consumption the most. It also helped us gain an understanding on how weather affected energy generation. This information helped us pick out the weather conditions to pass on as regressors to our models.

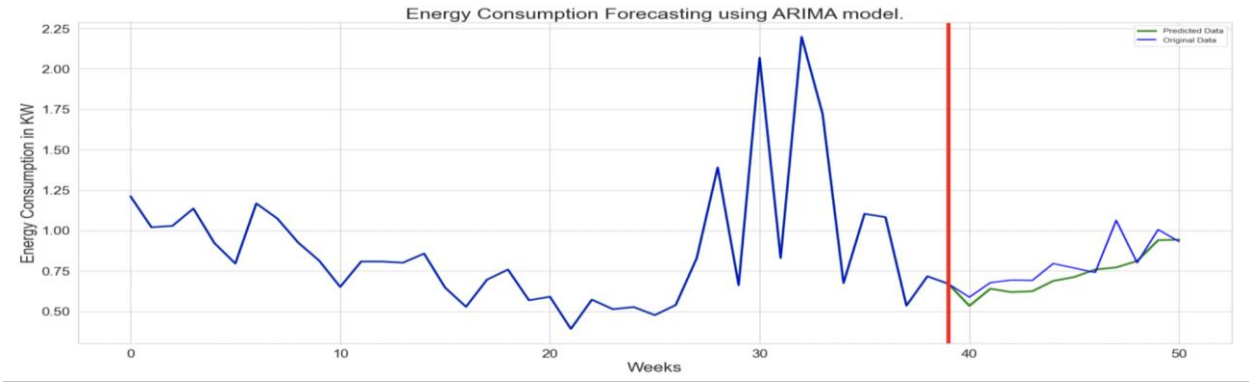
Algorithms evaluated

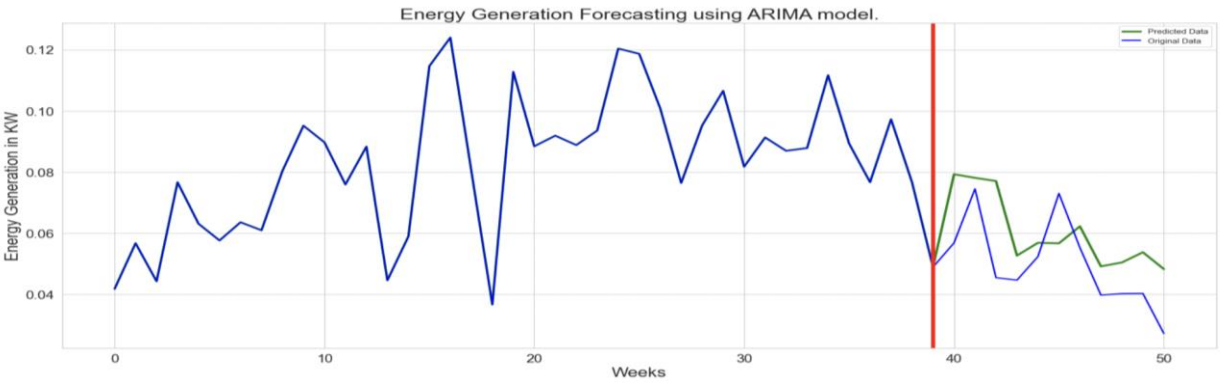
- We trained and predicted our data based on 4 models, ARIMA, SARIMAX, VAR and LightGBM.
- ARIMA and SARIMAX were used against VAR and LightGBM to compare predicting future energy values based on past trends and predicting future energy values based on weather conditions. We wanted to know how effective weather conditions would prove for future predictions against just past values.
- We used each model to predict both energy consumption and energy generation values.
- When randomly splitting the data into train and test sets, we chose a standard of 80% for our training data and 20% for our test data. This standard was maintained across all models except LightGBM.
- For each of our models we calculated the Mean Absolute Error (MAE) Value to help us compare the effectiveness of the models.

ARIMA

We used ARIMA to predict future values using past values for both energy consumption and energy generation. Parameters:

- Lag order (p)= 5
- Order of moving average(q) = 0
- Degree of differencing(d) = 1

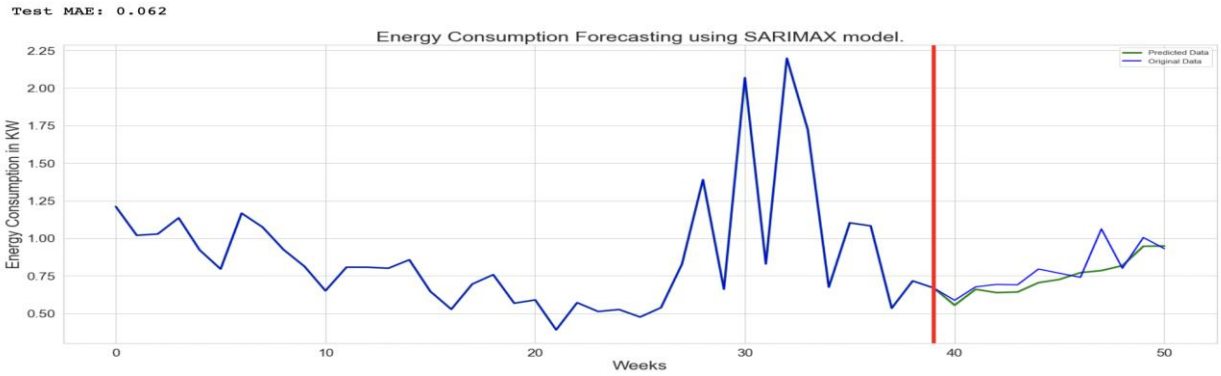


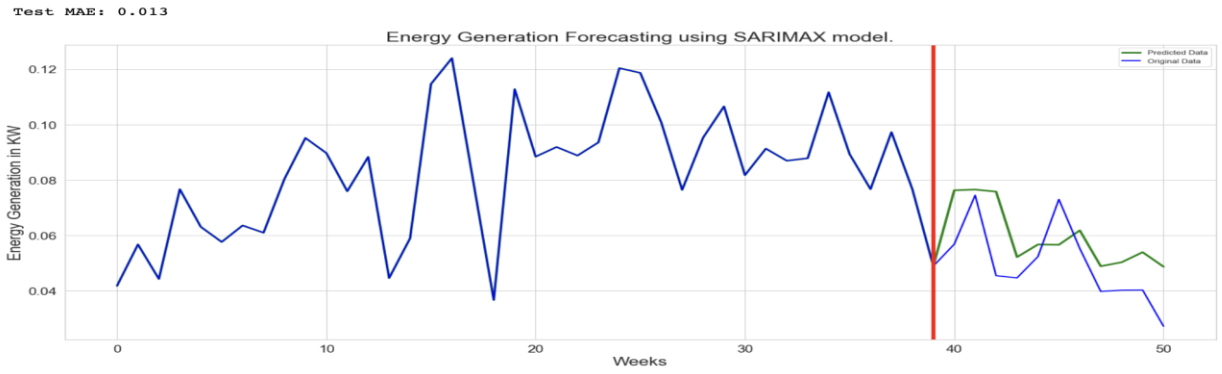


SARIMAX

We used SARIMAX to predict future values using monthly seasonality in past values for both energy consumption and energy generation. Parameters:

- Lag order (p)= 5
- Order of moving average(q) = 0
- Degree of differencing(d) = 1

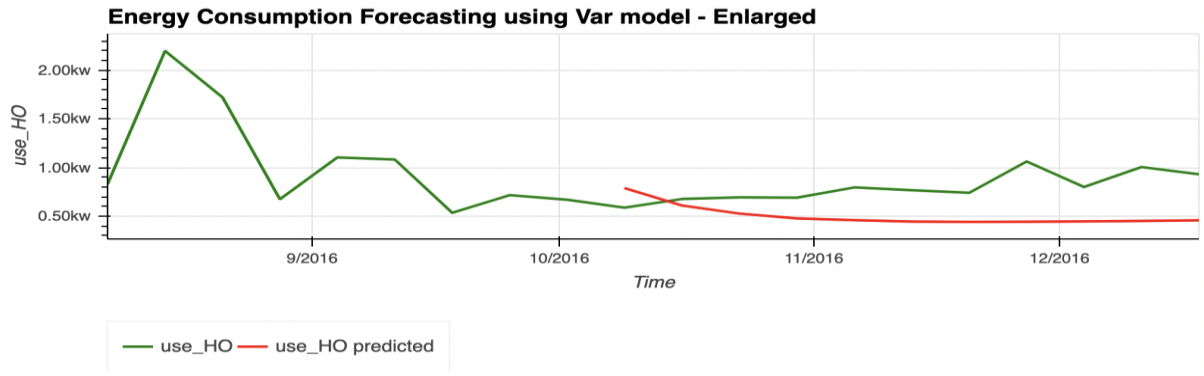




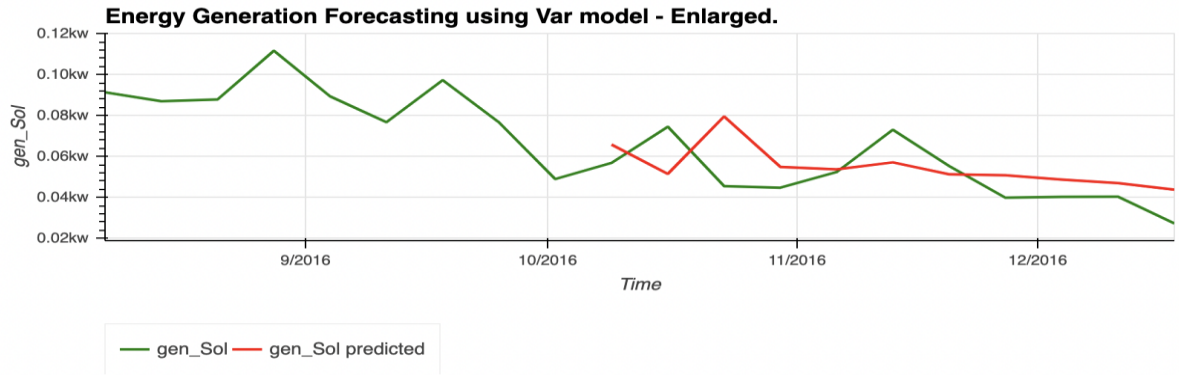
VAR

We used VAR because it allows us to use weather conditions to predict future energy consumption and generation values.

- Weather conditions used for energy consumption: 'temperature', 'humidity', 'windSpeed','cloudCover', 'windBearing', 'precipIntensity', 'dewPoint'



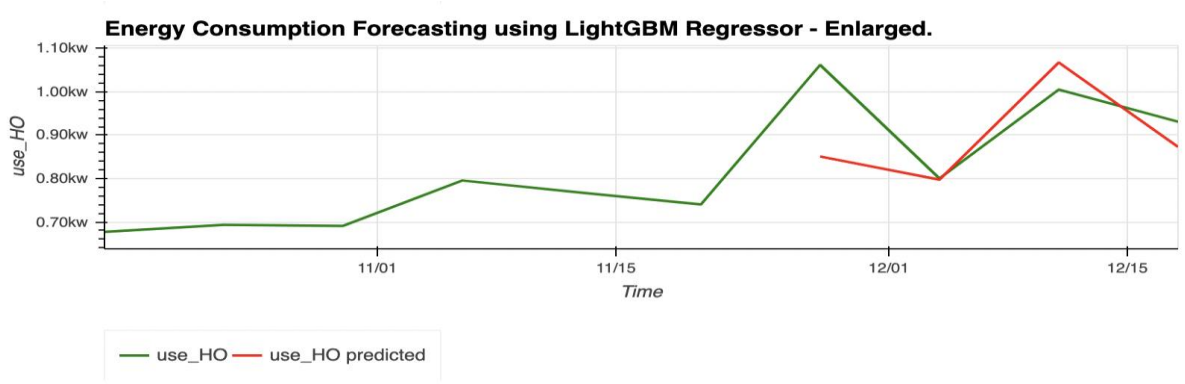
- Weather conditions used for energy generation: 'temperature', 'humidity', 'windSpeed', 'visibility', 'windBearing', 'precipIntensity', 'dewPoint'



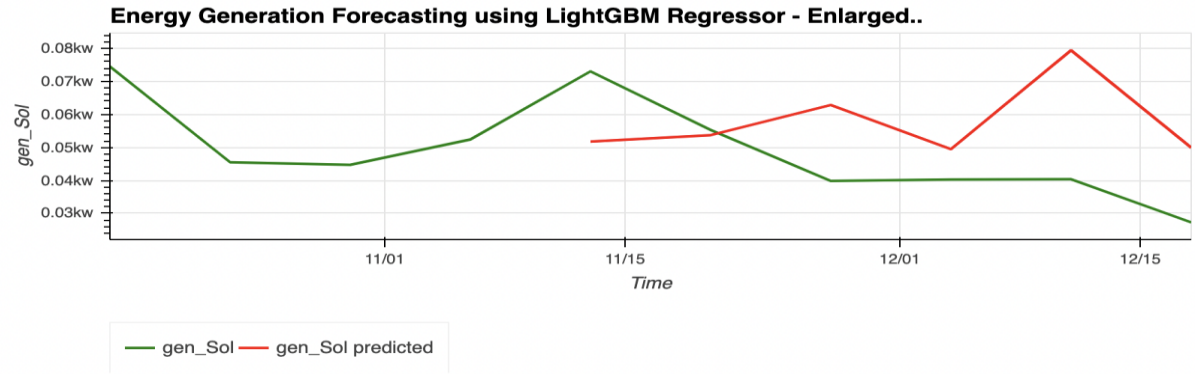
LightGBM

LGBM allowed us to use weather conditions to predict future energy consumption and generation values while understanding the relationship between energy consumption and weather information.

- Features used for energy consumption: 'temperature, ''humidity', 'windSpeed', 'pressure', 'weekday', 'hour', 'timing', 'year ', 'month', 'day'



- Features used for energy generation: 'temperature', 'apparentTemperature', 'windSpeed', 'cloudCover', 'weekday', 'hour', 'timing', 'year ', 'month', 'day'



Analysis of results

Target	Evaluation	ARIMA	SARIMAX	VAR	LightGBM
Energy Consumption	MAE	0.0720544	0.0623379	0.327561	0.0837629
	MSE	0.0107595	0.00893962	0.132992	0.013001
	Time taken (s)	0.952169	0.426018	0.0194049	0.100348
Energy Generation	MAE	0.0134282	0.0128191	0.0127327	0.0194838
	MSE	0.000248652	0.000229427	0.000241593	0.000519451
	Time taken (s)	2.45002	0.729363	0.0309839	0.142431

- We saw that SARIMAX works best followed by ARIMA, LGBM and VAR respectively.
- The prediction using done using LGBM is close enough to ARIMA and SARIMAX while VAR performs surprisingly poorly for energy consumption but the best for energy generation.
- Our analysis thus shows that future prediction of energy values is better with models that use past values and trends for prediction.
- However, our primary goal to check whether we could use weather conditions to predict energy values was met with LGBM. The accuracy we got with LGBM was close enough that it could be used to predict future energy consumption and generation values.

Discussion & Conclusions

Decisions made

- We decided upon the concept and then researched for a dataset that fit our requirements and met all the criteria necessary for project implementation.
- After we finalized the dataset, we narrowed down the tasks that were to be implemented during the project lifecycle. These tasks were then assigned between the team members.
- There were various decisions made for data preprocessing and feature engineering based upon the goal of the project.
- After thorough research, the models were decided and implemented for prediction of future values to get the best possible results.
- Finally, the team discussed and made decisions about various algorithms to be used for each task ranging from various weather conditions to be used for regressors, category of time resampling etc.

Difficulties faced

- The dataset that we selected was very big in size and needed a lot of time and effort to perform data transformation and preprocessing.
- The data visualization took the most time as we had to write functions to plot graphs for all the various time categories.
- We faced some difficulty to select the best resampling for the time category as well as the weather conditions so that we could get the best prediction results.
- We were not able to use the prophet model to predict our data because for some reason we could not properly install fbProphet on our machines and even when we did, it would not run properly.

Things that worked well

- The dataset we selected was ideal since it met all of our requirements and we also got to work and experience the complexity of handling such a huge dataset.
- The data cleaning and processing tasks worked really well and helped us easily visualize our data as well as get the best out of our algorithms.
- Although the data visualization seemed like a daunting task because we had decided to plot all combinations of values mentioned earlier, it proved to be a tremendous aid in selecting the necessary weather conditions and time categories to use in our algorithms.

Things that didn't work well

- We could not get prophet to work and that reduced the number of models we were planning to use.
- The var model we thought would have worked well gave us a very high MAE value.
- We were unable to successfully predict future energy values because based on our project objective we would have had to first predict future weather values. When we tried to predict these future weather values, we did not get a very good accuracy and this would have led to poor results for the energy values.

Future work

We came to realization that we although we can predict energy consumption and generation values using weather conditions, we would require future weather information to predict future energy values. We plan to further analyze the data for weather conditions and train a model that can accurately predict weather information values so that we can use them to predict future energy values.

Conclusion

- From the extensive analysis, we can conclude that it is indeed possible to accurately predict the energy consumption and energy generation values for the home appliances based on weather information.
- We can see that the univariate models such as ARIMA and SARIMAX perform very well by accurately predicting the trends.
- For the multivariate models, LightGBM regressor performs much better than VAR, just fallings lightly behind ARIMA and SARIMAX.
- Thus, we can use LightGBM regressor to forecast future energy consumption and generation values using the given weather information.
- Although it is possible to forecast the values using LightGBM, it is reliant on the future weather information, which needs to be predicted.

Project Plan / Task Distribution

Task	**Assigned to**	**Done by**
Dataset Selection	All	All
Data Preprocessing	Rohit	Rohit and Hanish
Feature Engineering	Rohit	Rohit and Hanish
Data Visualization	Rohit and Hanish	Rohit and Hanish
Research on Algorithms	All	All
Modelling	All	All
Analysis and Conclusion	Hanish	Rohit and Hanish
Project Report	Rohit and Hanish	Rohit and Hanish
PPT All	Hanish	Rohit and Hanish

References

- https://www.kaggle.com/taranvee/smart-home-dataset-with-weather-information
- https://machinelearningmastery.com/arima-for-time-series-forecasting-with-python/
- https://machinelearningmastery.com/sarima-for-time-series-forecasting-in-python/
- https://www.machinelearningplus.com/time-series/vector-autoregression-examples-python/
- https://machinelearningmastery.com/gradient-boosting-with-scikit-learn-xgboost-lightgbm-and-catboost/
- https://machinelearningmastery.com/time-series-forecasting-with-prophet-in-python/
- http://holoviews.org/user_guide/Plotting_with_Bokeh.html