1. Business Problem : Predicting the Energy Star Score of a building:

1.1 Description :

The NYC Benchmarking Law requires owners of large buildings to annually measure their energy and water consumption in a process called benchmarking. The law standardizes this process by requiring building owners to enter their annual energy and water use in the U.S. Environmental Protection Agency's (EPA) online tool, ENERGY STAR Portfolio Manager and use the tool to submit data to the City. This data informs building owners about a building's energy and water consumption compared to similar buildings, and tracks progress year over year to help in energy efficiency planning.

1.2 objective:

The objective that we are proposing here to achive is to use the energy data to build a model that can predict the Energy Star Score of a building(Residentail and Non Residentail spaces) and interpret the results to find the factors which influence the score or the predictions.

1.3 Source/Useful Links:

1. https://www1.nyc.gov/html/gbee/html/plan/ll84.shtml (https://www1.nyc.gov/html/gbee/html/plan/ll84.shtml)
2. http://www.nyc.gov/html/gbee/downloads/misc/nyc_benchmarking_disclosure_data_definitions_ (http://www.nyc.gov/html/gbee/downloads/misc/nyc_benchmarking_disclosure_data_definitions

1.4 Machine Learning constraints:

1. Interpretability of the results is important.
2. Feature importance of the data-point belonging to each class is needed to make the models more interpretable

2. Machine Learning Problem Formulation :

The Workflow that we will use here are as follows as guided in the assignemnt email.

1. Exploratory data analysis.
2. Data cleaning.
3. Feature engineering.
4. Feature selection.
5. Making baseline Model.
6. Use Different Machine Laerning Models to check for better performance.
7. Hyperparameter tuning the hyperparametres of different models.
8. Fiding the Best Machine learning model which best suits for the problem statement.

9. Interpret the model results and reasoning for using the model.
10. Conclusion.
11. Final report for presentation.

2.1 Type of machine learning Problem

1. As the target variable is a Continious varibale between 0-100 , so we can do it as a multiclass classification problem
2. We can also frame it as an Regression Problem.

The reason for framing it as an Regression Problem is handeling Multiclass- classification as 100 class will be very difficult to handel and will will be a very time intensive and memory intensive processes.

2.2 Error Metrix:

1. For regression task data scientist have the options of

```
1. MSE (Mean squared error)
2. MAPE (Mean absolute percentage error)
3. MAE (Mean Absolute Error)
```

2. I have used MAE (Mean Absolute Error) in this problem as MAE (Mean Absolute Error) penalises the absolute difference between the Target and the Predicted values .

3. Hypothesis for the project

1. we can build a model to predict the Energy star score of a building by using the feature that are there in the dataset

4. Context :

For sustaining the Hypothesis i performed indepth analysis of the features the work flow followed is

1. Exploratory data analysis : i explored every feature by univarient data analysis
2. Data cleaning : Performed data Cleaning and scaling as many features we lacking the feature rows.
3. Feature engineering : Performed Feature engineering on text data and Numerical data.
4. Feature selection : Performed feature selection by using colinearity and covarience as parameters .
5. Used Classical and Deep learning models for checking the genaralization error
6. Hyperparametertuned the hyperparametres of all the models used.
7. Found GBDT as the best model for this dataset

After performing all the above mentioned processes , i was able to generalise my hypothesis

Conclusion :

1. Indepth annalysis of the data provided and exploration of all the parameters i conclude that we can build a model that can predict the Energy star score of the building by using the dataset provided .

Pointing your attention to some of the important findings :

1. The important features which were very helpfull in predicting the Energy score are Site Eui , weather features ,Property id,and the type of property of use .
2. Simple machine learning models with hyperparameter tuning were also able to get good generalization score on the test data .
3. Complex Models like Random Forest and GBDT were very good at generalisation, and the error of both the models were competitively same .
4. Neural Networks(Sequence models) had less error than the classical Machine learning Models and has very low generalization error . But we cannot use those Neural netwokrs as model interpretability will be compromised.
5. Areas of imporvemnt: Given more data and time, there can be some scope of using latest machine learning models like light GBM and Catboost, did not try these algorthims as these algorithms may overfit due to lack of data, and need more time and resources for hyperparamter tuning.