**Stages of data cleaning, feature engineering and modelling**

1/ Load the latest data from the link, draft data analysis

2/ changing coma in the false object columns and transforming them to numerical

Columns to change coma and transform to numeric:

PropertyGFATotal, PropertyGFAParking, PropertyGFABuilding(s), LargestPropertyUseTypeGFA, SecondLargestPropertyUseTypeGFA, ThirdLargestPropertyUseTypeGFA, SiteEnergyUse(kBtu), SiteEnergyUseWN(kBtu), SteamUse(kBtu), Electricity(kWh), Electricity(kBtu), NaturalGas(therms), NaturalGas(kBtu), SourceEUI(kBtu/sf), SourceEUIWN(kBtu/sf)

3/ deleting Id and non-informative columns for modelling

Columns to delete:

OSEBuildingID, DataYear, PropertyName, Address, City, State, TaxParcelIdentificationNumber, YearsENERGYSTARCertified, DefaultData, Comments, Outlier

dropping the columns with electricity and gas consumption data and GHGEmissionsIntensity

'SiteEUI(kBtu/sf)','SiteEUIWN(kBtu/sf)','SourceEUI(kBtu/sf)','SourceEUIWN(kBtu/sf)','SiteEnergyUse(kBtu)','SiteEnergyUseWN(kBtu)','SteamUse(kBtu)','Electricity(kWh)','Electricity(kBtu)','NaturalGas(therms)','NaturalGas(kBtu)','GHGEmissionsIntensity'

dropping below 0 values for Y columns (TotalGHGEmissions, Electricity(kWh))

4/ splitting data into numerical and categorical

5/ replacing nans by median/most frequent

6/ implementing automatic Anova approach for treatment of the categorical columns, delete the columns which have no impact on Y (TotalGHGEmissions) (with p\_value > 0.05)

7/ selecting numerical columns with weak correlation (threshold < 0.5) to be deleted

8/ implementing OrdinalEncoder for the categorical data

9/ perturbation of the dataset (avoided due to small size of the dataset)

10/ outliers removal (therefore no need to keep the Outlier column which was dropped initially)

11/ scaling/standardizing data

14/ splitting data into X and y (emission), yel (Electricity(kWh))

15/ train\_test\_split of the dataset (check random\_state to be equal for the two splits)

16/ checking performance of a set of the different models with fit/predict (KFold/cross\_val\_score can’t be used due to a small size of a dataset)

17/ retrieving features importance

18/ hyperparameter tuning to improve the model performance

**SUMMARY OF THE RESULTS**

**PYTHON**

**Final list of X columns**: BuildingType, PrimaryPropertyType, Neighborhood, PropertyGFATotal, PropertyGFABuilding(s), ListOfAllPropertyUseTypes, LargestPropertyUseType, LargestPropertyUseTypeGFA, SecondLargestPropertyUseType, ThirdLargestPropertyUseType, NaturalGas(therms)

**Y columns**: TotalGHGEmissions, Electricity(kWh)

**TotalGHGE best model :** RandForrest: R2 = 0.89 (RMSE = 62.30)

**Electricity(kWh) best model:** XGB: R2 = 0.88 (RMSE = 950 769.28)

**GC BIGQUERY ML**

**BQ formatted columns**: BuildingType, PrimaryPropertyType, Neighborhood, PropertyGFATotal, PropertyGFABuilding\_s\_, ListOfAllPropertyUseTypes, LargestPropertyUseType, LargestPropertyUseTypeGFA, SecondLargestPropertyUseType, ThirdLargestPropertyUseType, Electricity\_kWh\_, NaturalGas\_therms\_, TotalGHGEmissions

**Label columns**: TotalGHGEmissions, Electricity(kWh)

**TotalGHGE best model :** DNN: R2 = 0.94 (RMSE = 159)

**Electricity(kWh) best model:** RF: R2 = 0.73 (RMSE = 1 090 002)

**GC VERTEX AI AUTO ML**

**BQ formatted columns**: BuildingType, PrimaryPropertyType, Neighborhood, PropertyGFATotal, PropertyGFABuilding\_s\_, ListOfAllPropertyUseTypes, LargestPropertyUseType, LargestPropertyUseTypeGFA, SecondLargestPropertyUseType, ThirdLargestPropertyUseType, Electricity\_kWh\_, NaturalGas\_therms\_, TotalGHGEmissions

**Label columns**: TotalGHGEmissions, Electricity(kWh)

**TotalGHGE best model :** Tabular Regression: R2 = 0.94 (RMSE = 142)

**Electricity(kWh) best model:** Tabular Regression: R2 = 0.84 (RMSE = 964 808.4)

The DAG failed because some tasks failed. The failed tasks are: [set-optional-inputs].

com.google.cloud.ai.platform.common.errors.AiPlatformException: code=RESOURCE\_EXHAUSTED, message=The following quota metrics exceed quota limits: [aiplatform.googleapis.com/custom\_model\_training\_cpus](https://www.google.com/url?sa=D&q=http%3A%2F%2Faiplatform.googleapis.com%2Fcustom_model_training_cpus), cause=null; Failed to create custom job for the task. Task: Project number: 733376066474, Job id: 5502089776053878784, Task id: -1411118583206903808, Task name: set-optional-inputs, Task state: DRIVER\_SUCCEEDED, Execution name: projects/733376066474/locations/us-central1/metadataStores/default/executions/14752237754040863056; Failed to create external task or refresh its state. Task:Project number: 733376066474, Job id: 5502089776053878784, Task id: -1411118583206903808, Task name: set-optional-inputs, Task state: DRIVER\_SUCCEEDED, Execution name: projects/733376066474/locations/us-central1/metadataStores/default/executions/14752237754040863056; Failed to handle the pipeline task. Task: Project number: 733376066474, Job id: 5502089776053878784, Task id: -1411118583206903808, Task name: set-optional-inputs, Task state: DRIVER\_SUCCEEDED, Execution name: projects/733376066474/locations/us-central1/metadataStores/default/executions/14752237754040863056

<https://stackoverflow.com/questions/73368320/vertax-ai-pipeline-quota>

<https://console.cloud.google.com/apis/api/aiplatform.googleapis.com/quotas?project=stage1energy>