Spain Electricity Load Shortfall

Presented by NM2 Consulting Ltd.

Table of Content

Problem Statement

The government of Spain is determined to improve the standard of living of its citizens by ensuring a steady supply of electricity. The government has identified increased investment in its renewable energy resource infrastructure as key to achieving that.

We observed that in 2020, renewable energy sources accounted for only 44% of the total electricity generated in Spain. Continued reliance on non-renewable energy sources is not sustainable.

Our team of data scientists will build a model that predicts the country's three-hourly load shortfall between the energy generated by fossil fuels and renewable energy sources. This would help the government gain insight into the trends and patterns of the country's electricity generation.

Objectives

1

Build prediction models

Our primary objective is to build models that predict the electricity load shortfall between the energy generated by the different sources.

2

Assess model accuracy

Using different test parameters, we will assess the different models to ascertain which has better accuracy.

3

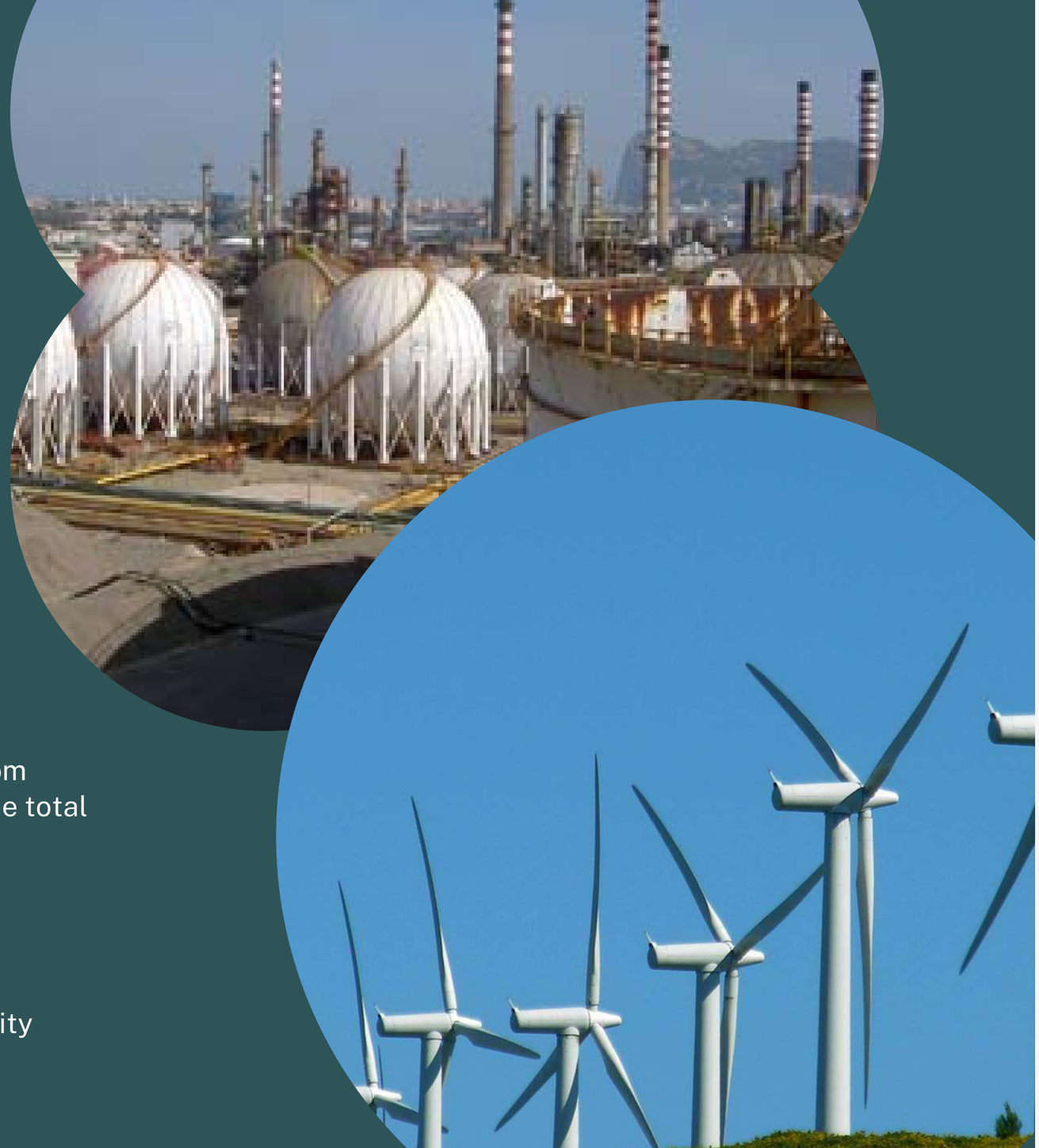
Deploy best model for prediction

Upon selecting the more accurate model, we will deploy the model on a remote server to make it accessible to the client.

Background

Electricity Generation

Spain generates electricity from non-renewable and renewable sources of energy. These sources include: nuclear, hydro, wind, solar, coal and combined cycle gas.



Renewable Energy

- In 2020, the country generated 109,269GWh of electricity from renewable sources of energy alone, representing 43.6% of the total electricity generated that year.
- Nuclear and wind were the major renewable energy sources, accounting for 22.2% and 21.7% respectively.
- Other sources of renewable energy used to generate electricity includes: combined cycle gas, hydro, solar and cogeneration.

Data Analysis Process

Data Collection

The first phase of this project entailed sourcing the dataset for our analysis.
The data was then hosted on a remote repository to enable easy accessibility and version control by all members of the team.

EDA

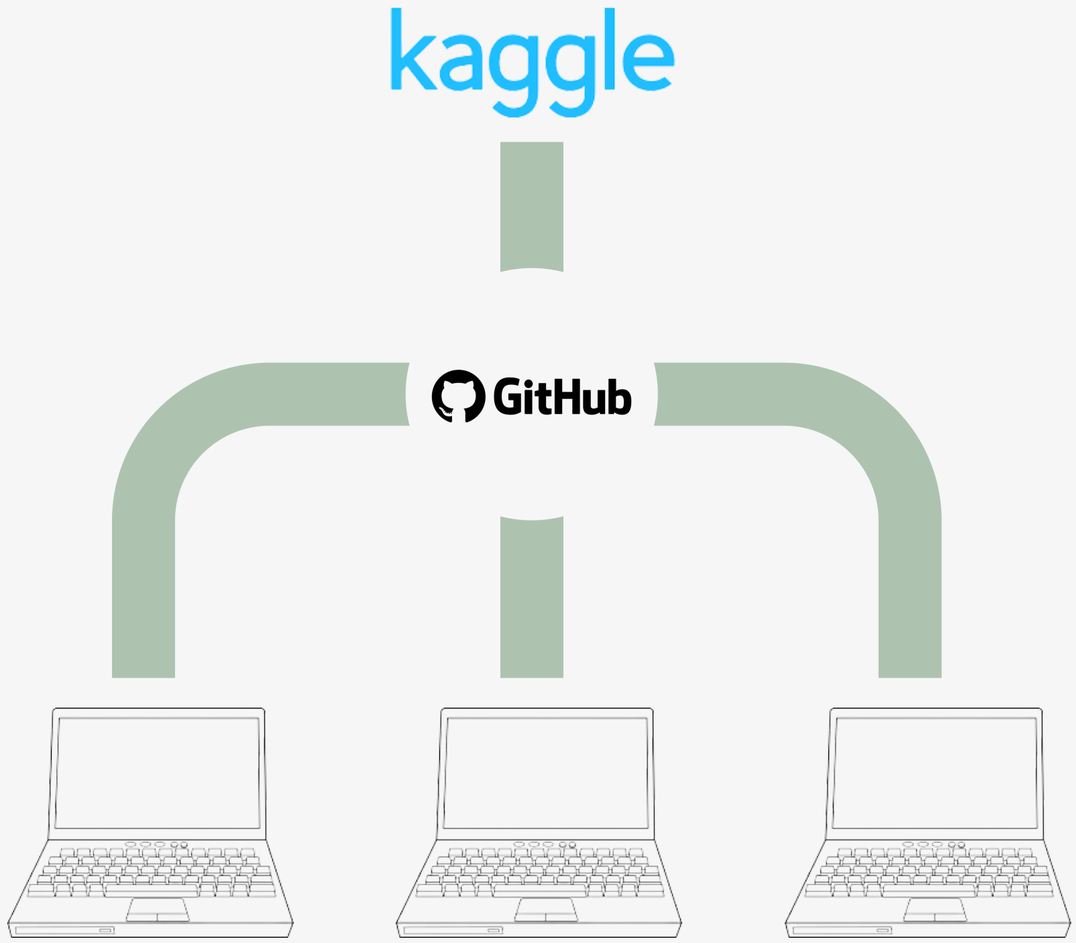
Exploratory data analysis (EDA) was performed on the two datasets provided, to have a complete overview of the different features and other characteristics that make up the dataset.

Data Engineering

Based on the observations made during the EDA phase, the datasets were cleaned, transformed, and re-engineered to include new features necessary to build the models.

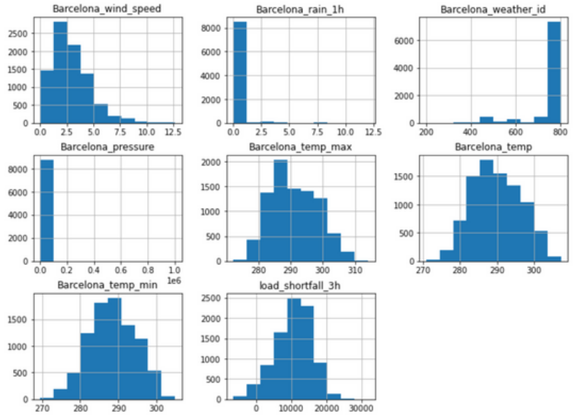
Data Collection

- Downloaded datasets from Kaggle
- Created a Github repository for the team and added all members as contributors
- Pushed the datasets and other necessary files to the repository
- Team members cloned the repository to their local PCs

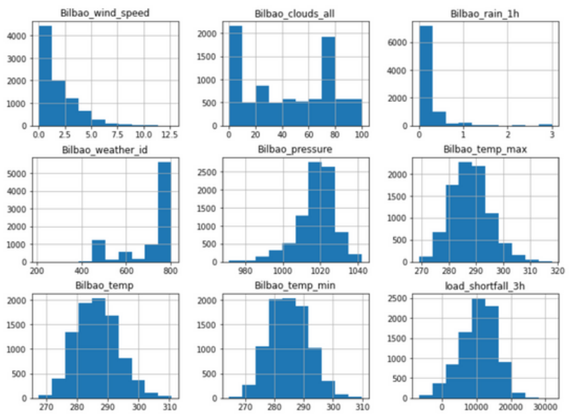


Exploratory Data Analysis

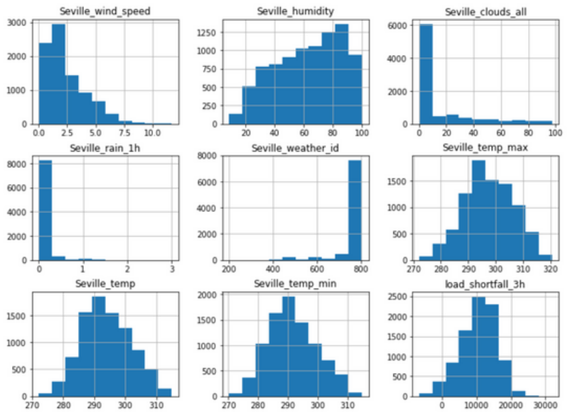
- Checked the shape, previewed the features in each of the dataset.
- Split the train dataset by city to have a view of the impact of city-specific features.
- Evaluated the correlation of the different features with the output variable.
- Checked for multicollinearity among the features.



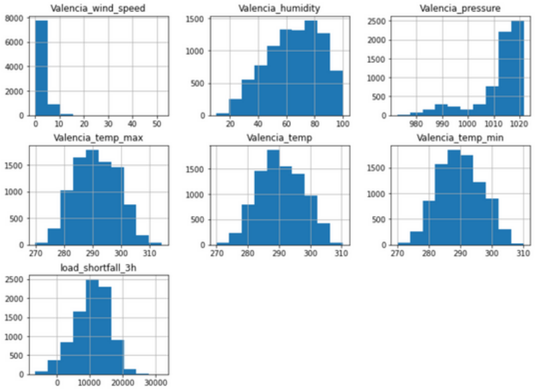
- Barcelona Features



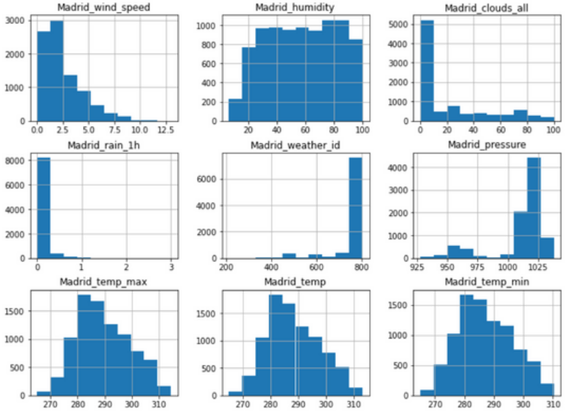
- Bilbao Features



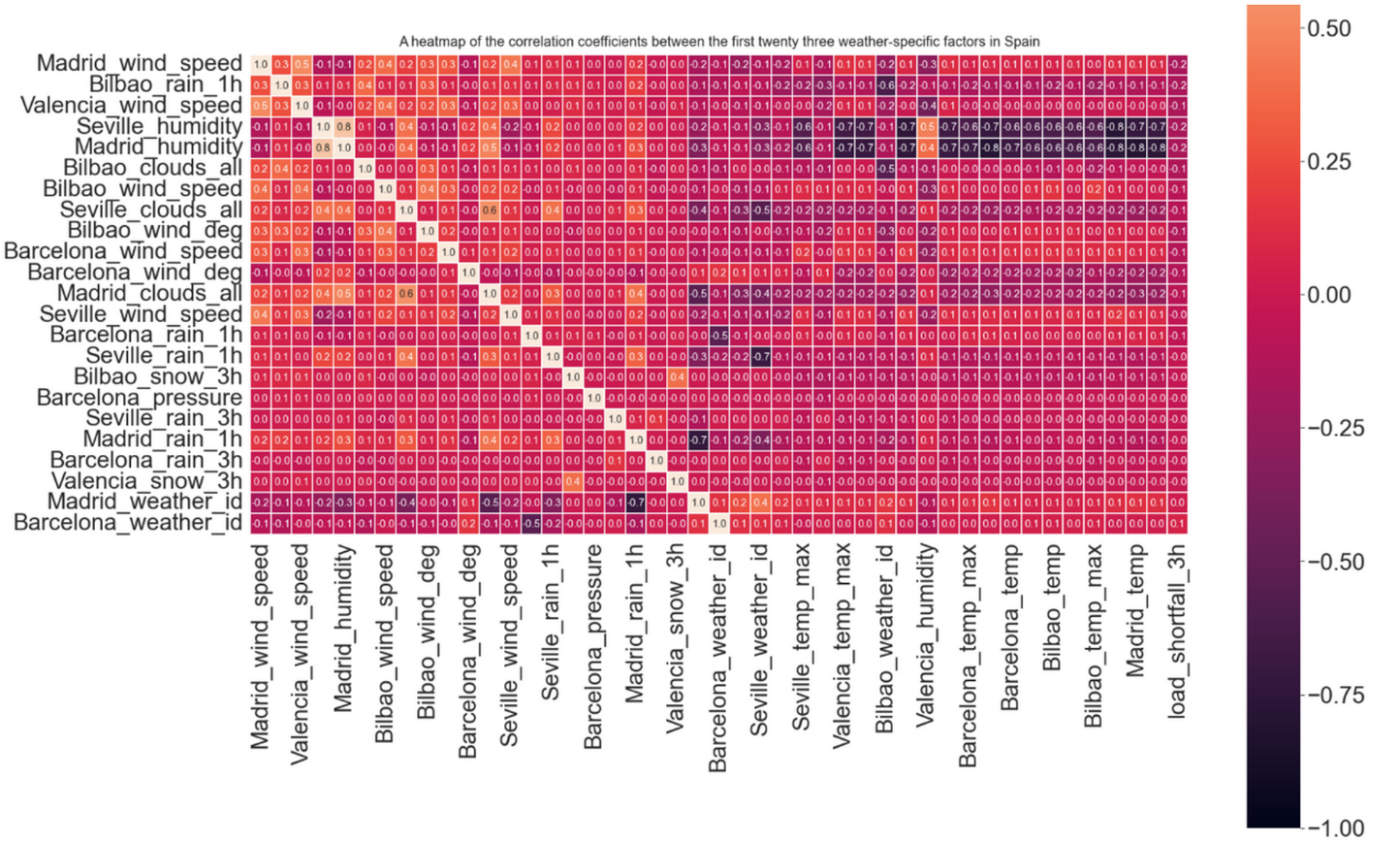
- Seville Features



- Valencia Features

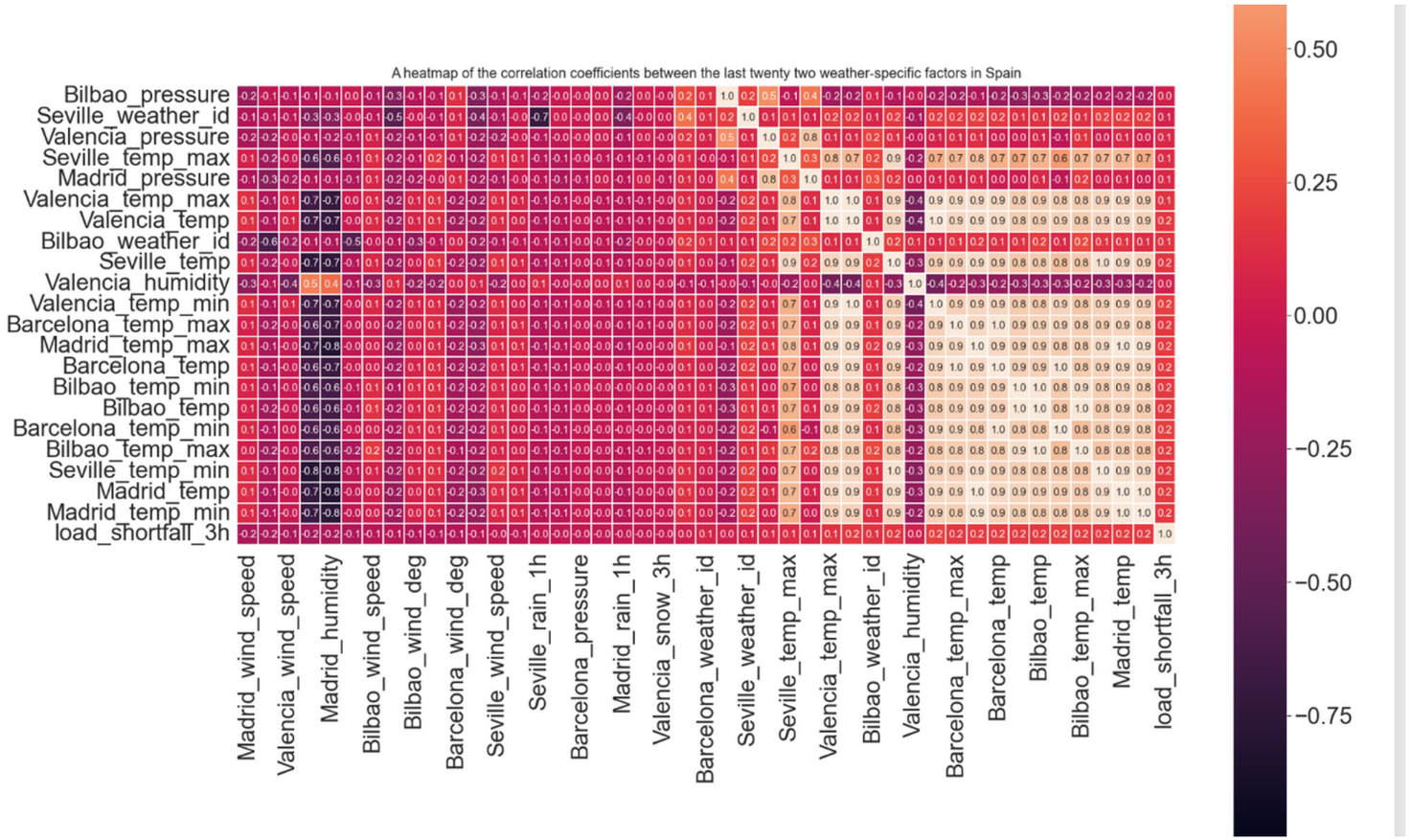


- Madrid Features



A heatmap showing the correlation among the first twenty three features.

There exists a strong multicollinearity among most of the features here



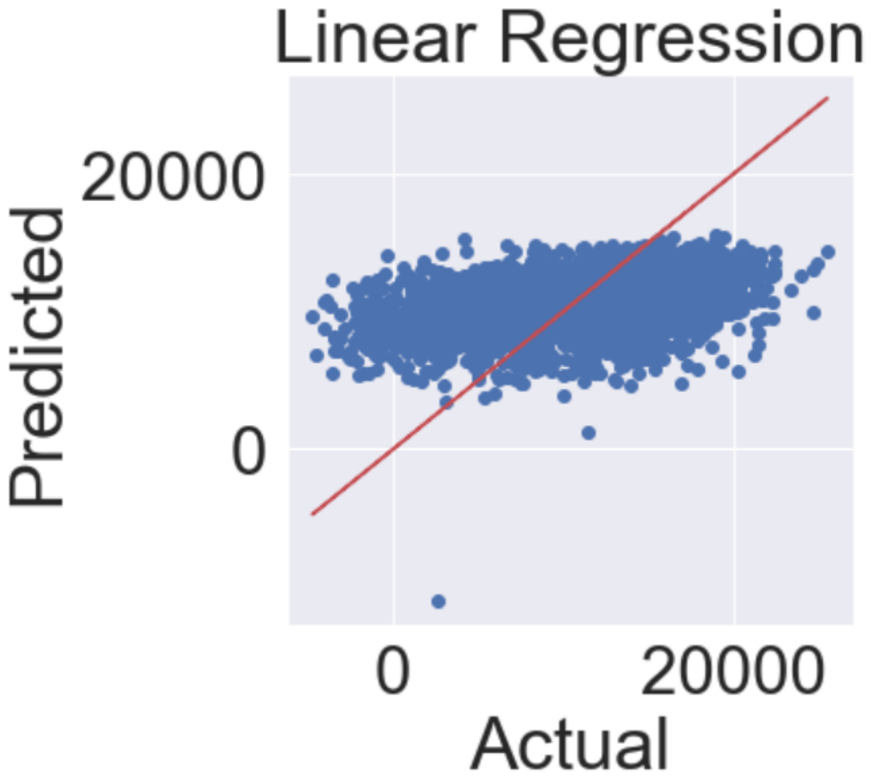
A heatmap showing the correlation among the bottom twenty two features.

Data Engineering

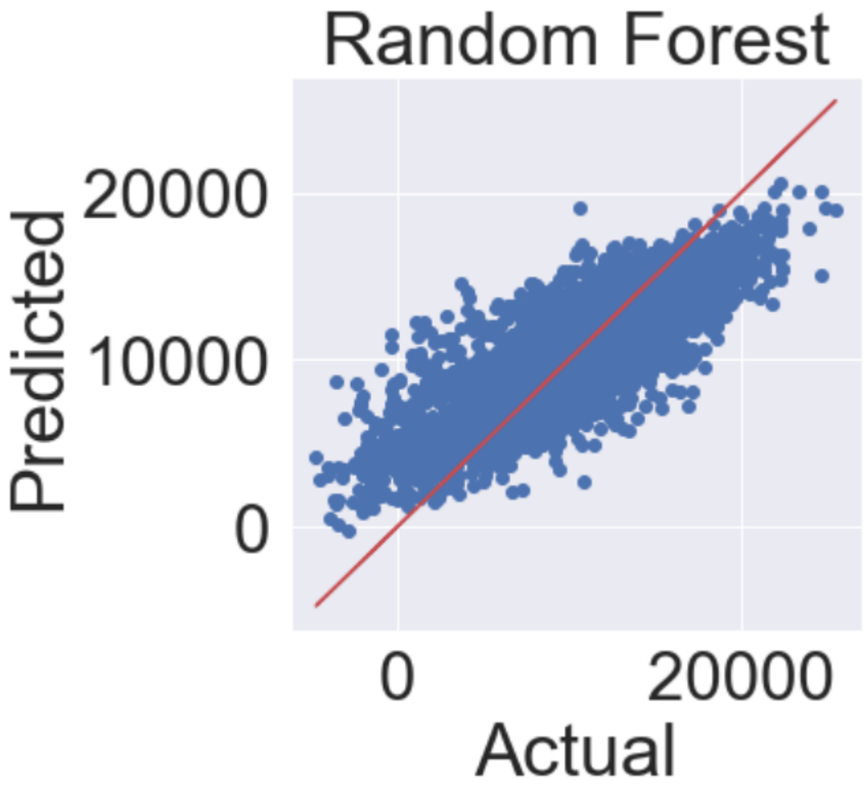
- Checked for null values in the dataset.
- Replaced the null values with the mean.
- Checked the data type of each feature, and modified it where necessary (time feature).
- Created new features to replace the time feature (seconds, minutes, hour, day, month, year)
- Dropped irrelevant features to improve model accuracy.

Modelling

- Split the data into x and y variables, and Declared train-test function, using a test size of 34%.
- Fitted a base model ie a linear regression model, and two other regression models (Random Forest & LASSO).
- Assessed the performance of all three models using two measures ie the RMSE values & R_2 score.
- Observed the Random Forest regression model had better RMSE and R_2 values, thus, making it the most accurate among the three models.



Linear Regression



Random Forest

LASSO Regression

Model Performance

Model Performance Results

```
In [50]: results_df
```

Out[50]:

	Training R-Square Score	Test R-Square Score	Training RMSE	Test RMSE
Linear Regression Model	0.14517	0.14255	4,799.09932	4,879.86636
LASSO	0.14517	0.14254	4,799.09969	4,879.90380
Random Forest	0.91704	0.62546	1,495.04609	3,225.17443

Conclusion

The predicted three-hourly load shortfall had the following summary statistics:

- Mean load shortfall of 10,545kWh
- A maximum load shortfall of 20,492kWh
- A negative minimum load shortfall of 257kWh
- The Median load shortfall is 10,891kWh

Team

Ubasinachi Eleonu

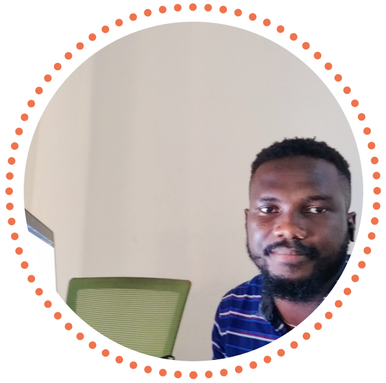
Team Leader

Elizabeth Ajabor

Member

Yinka Akindele

Member



Akinbowale Akin-Taylor

Member

Emmanuel Maisaje

Member

Tochukwu Ezeokafor

Member