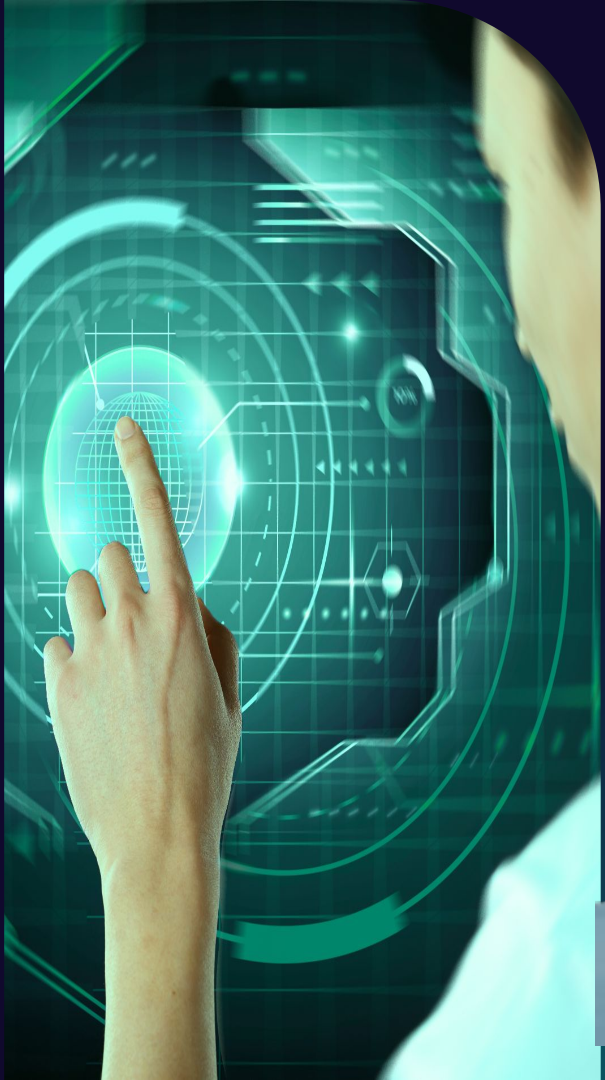


Data Worx Ltd

Spain Electricity ShortFall Project

TABLE OF CONTENTS

01. Introduction

02. Problem Statement

03. EDA & Feature Engineering

04. Model Building

05. Model Evaluation

06. Conclusion

01. INTRODUCTION

ABOUT US

Services Provided:

❏ Data collection,
❏ Data cleaning,
❏ Data analysis and
❏ Model building.

The Team

1. Ms. Mandy Rasemphe: Data Science Team lead
2. Mr. Karabo Molema: Data Analyst
3. Mr. Michael Benjamin: Machine Learning Engineer
4. Mr. John Chukwuebuka: Data Engineer

02. Problem Statement

PROBLEM STATEMENT

- ❑ Importance of electricity
- ❑ Inadequate infrastructure
- ❑ Expansion of renewable energy resource infrastructure investments.
- ❑ Model the shortfall between energy generated by means of fossil fuels & various renewable resources

03. EDA &
Feature Engineering

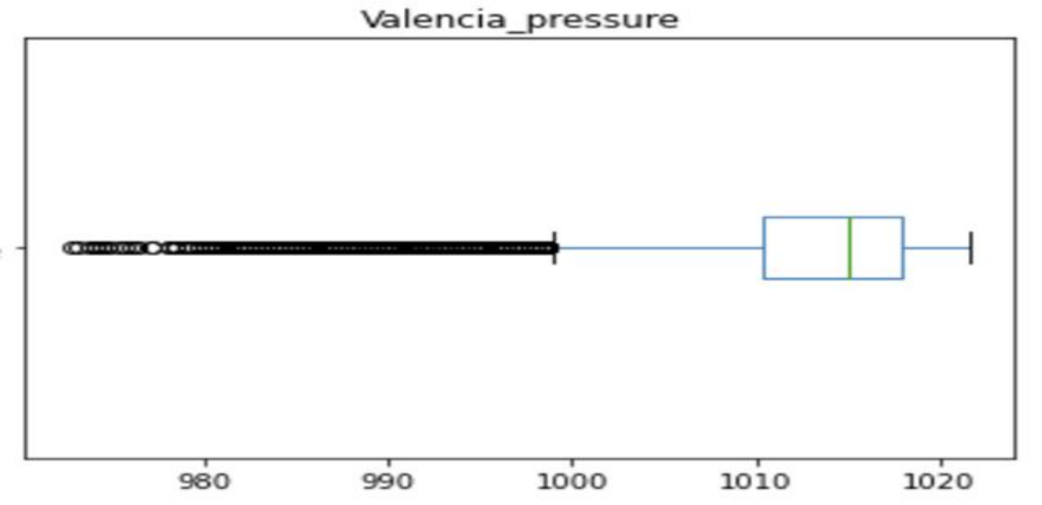
Investigating Dataset

- ❑ Provided Dataset and supporting information explored
- ❑ Contains information on weather conditions.
- ❑ Spain's 5 prominent cities were tracked.
- ❑ Duration is from 1st January 2015 to 31st December 2017

Data Issues

Outliers

- ❑ 28 Features contained outliers
- ❑ Keep Outliers
- ❑ Use outlier robust techniques.



Data Issues

Missing Values

- ❑ One feature contained Null values: Valencia_pressure
- ❑ Replaced with median

Data Issues

Incorrect Feature Data types

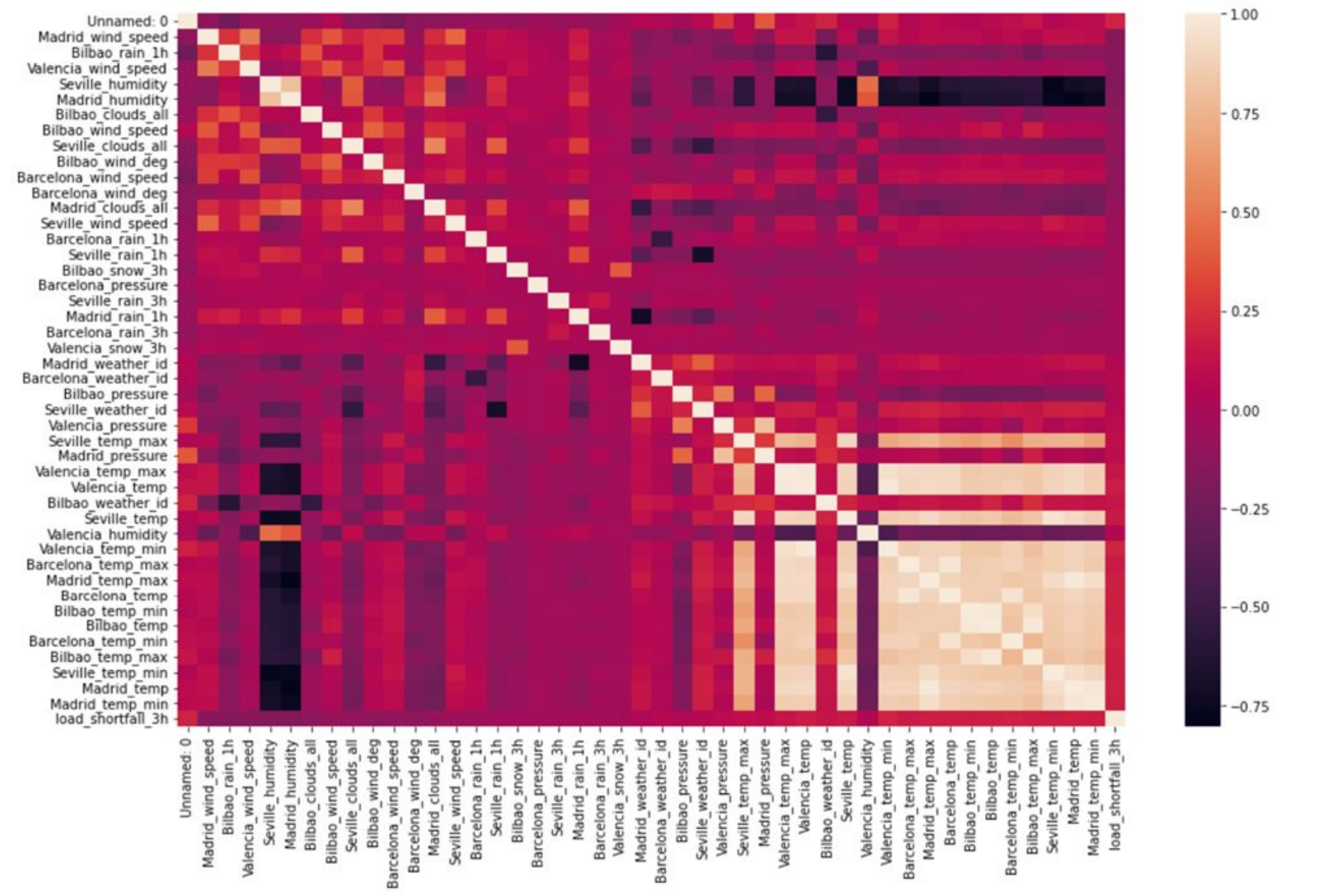
- ❑ Some features were of object data types
- ❑ Features with incorrect data types were converted

Correlation

❑ Little to no correlation between predictors and response feature

❑ Multicollinearity amongst predictor features

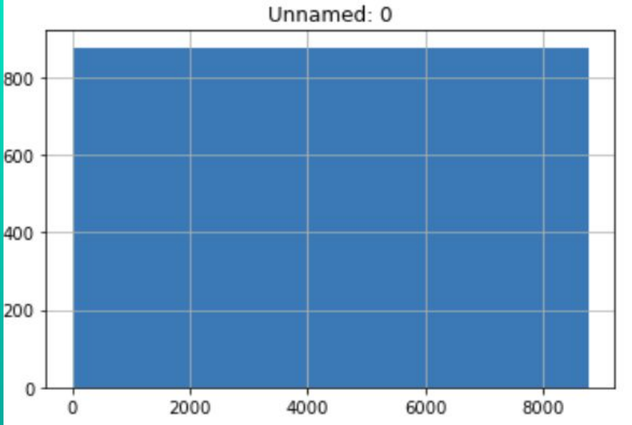
❑ Drop Multicollinear features

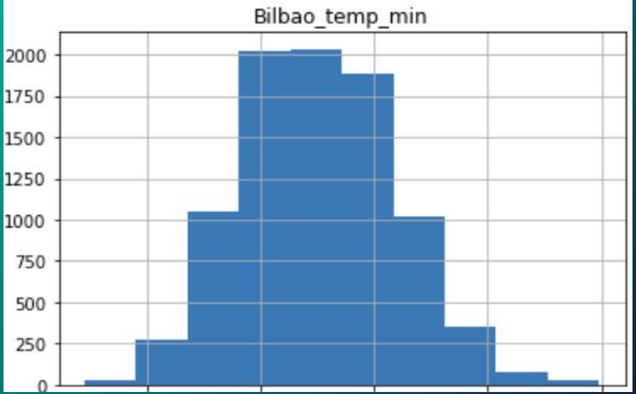

Unnamed: 0
Madrid_wind_speed
Bilbao_rain_1h
Valencia_wind_speed
Seville_humidity
Madrid_humidity
Bilbao_clouds_all
Bilbao_wind_speed
Seville_clouds_all
Bilbao_wind_deg
Barcelona_wind_speed
Barcelona_wind_deg
Madrid_clouds_all
Seville_wind_speed
Barcelona_rain_1h
Seville_rain_1h
Bilbao_snow_3h
Barcelona_pressure
Seville_rain_3h
Madrid_rain_1h
Barcelona_rain_3h
Valencia_snow_3h
Madrid_weather_id
Barcelona_weather_id
Bilbao_pressure
Seville_weather_id
Valencia_pressure
Seville_temp_max
Madrid_pressure
Valencia_temp_max
Valencia_temp
Bilbao_weather_id
Seville_temp
Valencia_humidity
Valencia_temp_min
Barcelona_temp_max
Madrid_temp_max
Barcelona_temp
Bilbao_temp_min
Bilbao_temp
Barcelona_temp_min
Bilbao_temp_max
Seville_temp_min
Madrid_temp
Madrid_temp_min
load_shortfall_3h
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

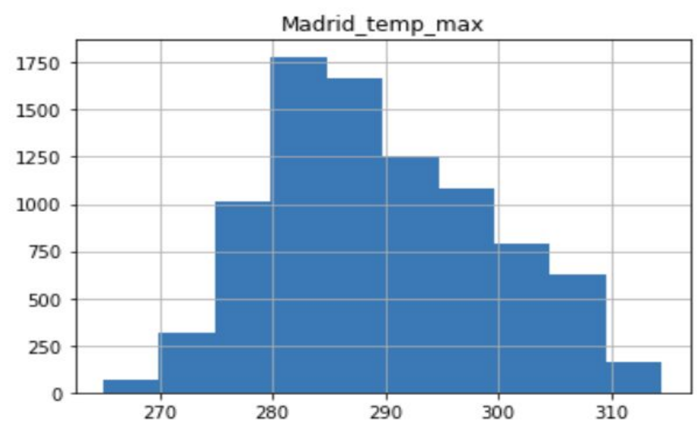
Distribution of data

- ❑ Distributions of all features
- ❑ Importance of distribution
- ❑ Too much skewness may affect prediction models

Distribution of data







Date time feature

- ❏ Split time feature

- ❏ Get different information such as Year, Month, day, time.

- ❏ Useful in capturing seasonal patterns in a dataset

04. Model Building

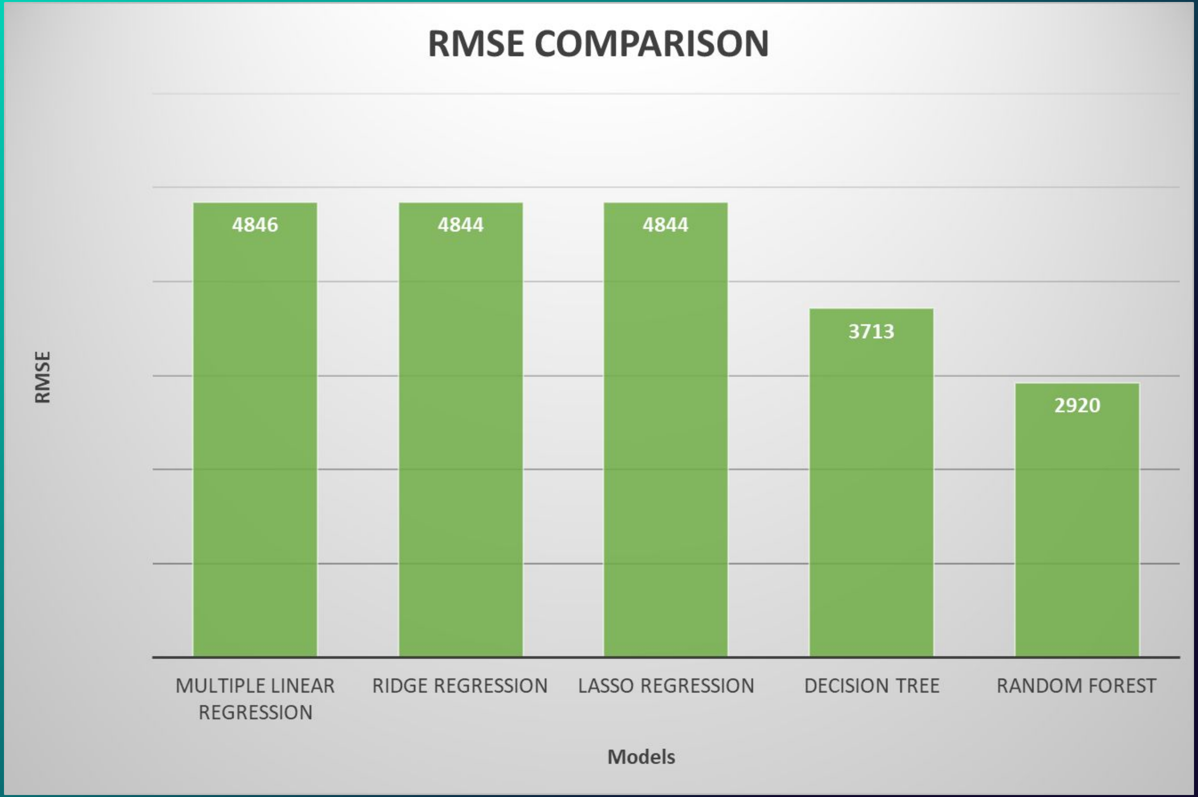
TYPES OF MODELS

- ❑ Multiple linear regression
- ❑ Ridge regression
- ❑ Lasso regression
- ❑ Decision Tree
- ❑ Random Forest

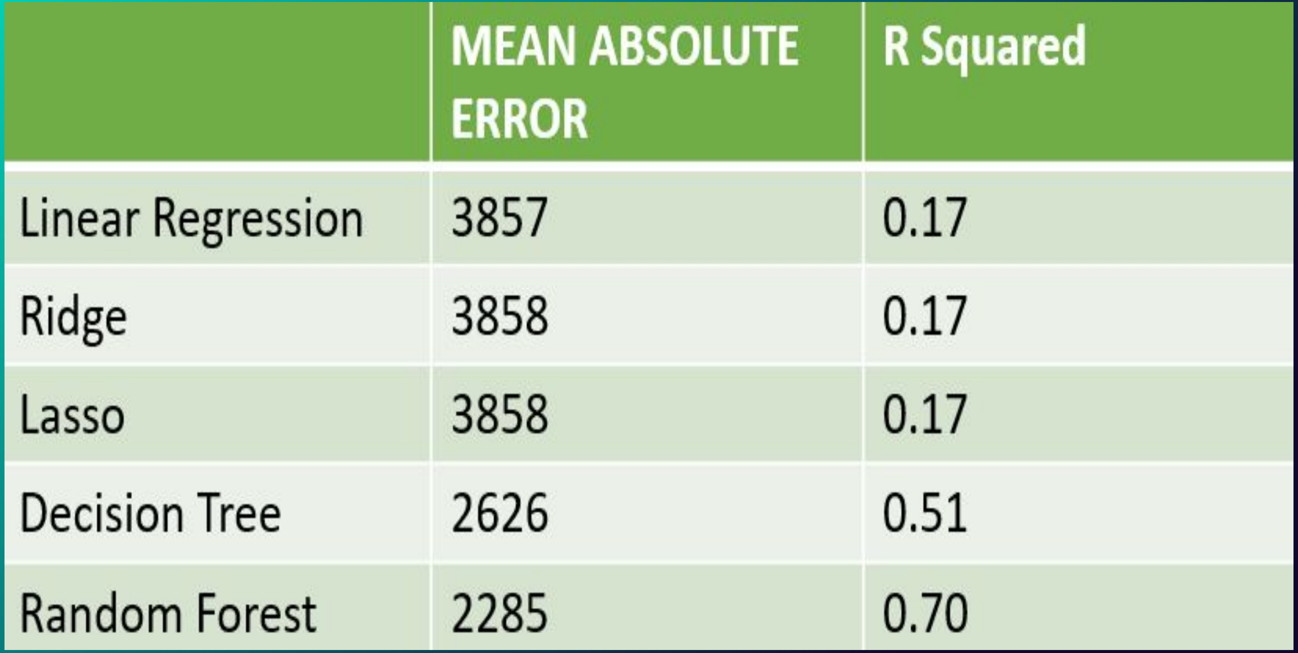
05. Model Evaluation

Evaluation Metrics

- ROOT MEAN SQUARED ERROR



MEAN ABSOLUTE ERROR
&
R squared



	MEAN ABSOLUTE ERROR	R Squared
Linear Regression	3857	0.17
Ridge	3858	0.17
Lasso	3858	0.17
Decision Tree	2626	0.51
Random Forest	2285	0.70

BUSINESS VALUE OF PROJECT

- ❑ Our robust analysis has taken most factors into account.
- ❑ Model will assist in making accurate future predictions.
- ❑ Trends and patterns identified between fossils and renewables.
- ❑ This will assist government make informed business decisions.

Conclusion

- ❑ Show the extent of the shortfall between producing electricity from fossil fuel and renewable source.

- ❑ Random Forest best modelled the data

- ❑ Limitations of project included: Outliers

- ❑ Improvement for future purposes could be further investigation into where the outliers originates.

- ❑ Findings of the project comprehensively highlights the energy deficit between the two systems.

QUESTIONS???