Project Report

Specialist Certificate in Data Analytics Essentials

GitHub URL
https://github.com/lanecolinp/UCDPA_lanecolinp

Abstract
Examination of the predictive relationship between wind speeds and Wind Poower generation in the republic of Ireland

Introduction
Renewable energy is very interesting

Datasets
This data set contains information about all locations recording meteorological data currently and historically and contains useful information such as geo coordinates, county, elevation and range of time for which recordings exist.

Station Details

This data set contains information about all weather stations in Ireland(state) recording meteorological data currently and historically and contains useful information such as geo-coordinates, county, elevation, and range of time for which recordings exist.

Source: Met Eireann

https://cli.fusio.net/cli/climate_data/webdata/StationDetails.csv

Hourly Weather Data

This dataset contains hourly recordings of various meteorological measurements. There is a file per station that records hourly data. The file is contained in a zip archive and consists of information about the station followed by comma separated header and rows. The headers are not consistent across all stations due to some measurements being taken in some stations and not others.

Source: Met Eireann

https://cli.fusio.net/cli/climate_data/webdata/hly*.zip

Wind Generation Data

Wind Power generation data measured and recorded in 15 minute intervals, data is the forecasted and actual MW of power being generated

Source: Eirgrid

https://www.smartgriddashboard.com/DashboardService.svc

Ireland Shape

https://www.geoboundaries.org/data/1_3_3/zip/shapefile/IRL/IRL_ADM1.shp.zip

Implementation Process

Wind is highly variable and unpredictable, but may have seasonal characteristics.

Get Weather Stations details available from Met Eireann :'https://cli.fusio.net/cli/climate_data/webdata/StationDetails.csv

And read into dataframe : wsdf, there are 2083 stations

The Values in 'open year' and 'close year' which describe the time range of recording for each staion are a combination of numeric text values and '(null)'

I want to analyse the wind data from Weather Stations which have data from 2013 to the present, so covering 10 years.

I converted the '(null)' value to a future year '3000' and converted it to int.

I then filtered the data based on Year which the number of stations to 429.

I removed Rathlin Island Station from the dataframe as it is not in the state.

Hourly data per station is available from from Met Eireann :'https://cli.fusio.net/cli/climate_data/webdata/StationDetails.csv

For each of the stations in the dataset, check if hourly data exists, by determining if the crafted url is valid, if not drop it from the stations dataset, otherwise download the files to the 'zipped' folder and unzip to the unzipped folder.

The files are named using the station name which is a number. For example the station name of ROCHES POINT is 1075 and the file to download is hly1075.zip. There are 25 staions remaining after this

**Show chart

I then looked at getting a distribution for the remaining stations. By finding the station's nearest neighbour and the distance between them using Balltree and Haversine method.

Distribution looks reasonably Ok, the distances are not significant

For each file in ./weatherstationdata/unzipped with the pattern hly*.csv - Use regex here

Find 'date:' and for each line until blank line create a dictionary for renaming the columns later

Find column headers starting with date, and check if there is a header called wdsp whic is Wind Speed,

If there is then write the rest of the file into the processed folder

Rename the columns as per the dictionary

Delete unused columns

23 stations remain that have Windspeed data.

Read in the processed Weather Data file.

Combine the Weather Data into allwddf dataframe

For each file read into a dataframe

Add the station name as a column

Clean the data

As we are only interested in the datetime, Windspeed and station name, we remove all other columns

We are only interested in records from 2013 to 2022 inclusive, so remove any outside that range

We are only interested in records from 2013 to 2022 inclusive, so remove any outside that range

Convert date to datetime and use the range to remove the records

This reduces the number of records from 8131546 to 2015904

Try to convert allwddf.wdsp to int

Get ValueError: invalid literal for int() with base 10: ' '

Look at unique values

As per the error message there are 1 or more values = ' '

Check for whitespace in wdsp column by using groupby with cumsum function

There are 444 records with blank data, By looking at the groups of occurences, they range from 1 or 2 hourly up to 3 whole days of recordings

A reasonable strategy to maintain the data consistency is to copy the windspeed measurement from the nearest neighbour we found earlier

This didn't work correctly and I didn't have time to fix it, so given the small number of records gibven the sample size, I set these to '0'

There was no further bad values, so I dropped the additional columns and set wdsp to int.

Removed the index and transposed allwddf to a pivoted dataset : pivwddf, such that every station is a column with the windspeed at that datetime as the value.

Next I Download WindGeneration data for Ireland(state)

Get the Actual Wind Generation data from
https://www.smartgriddashboard.com/DashboardService.svc

Last 3o days data available by using this url.

https://www.smartgriddashboard.com/DashboardService.svc/csv?area=windActual&region=ALL&datefrom=04-Feb-2023%2000:00&dateto=05-Mar-2023%2023:59

Crafting urls by manipulating the url parameters allows the possibility of downloading data for custom timeframes.

Attempting to download a full year resulted in an empty file

It was possible to download data in monthly chunks like so

https://www.smartgriddashboard.com/DashboardService.svc/csv?area=windActual&region=ROI&datefrom=01-Dec-2022%2000:00&dateto=31-Dec-2022%2023:59

Use the calendar and datetime libraries to get the first and last days of each month from 2013 to 2022 and inject these into the url to download the data

Process the files and merge into Single Dataframe

Initial clean of the Data

Remove Whitespace on ACTUAL WIND(MW) columns

Drop duplicate columns

Convert 'DATE & TIME' to datetime and make it the index

We are only interested in the ACTUAL WIND(MW) column, so drop the rest

Attempting to convert 'ACTUAL WIND(MW)' column to int, I got a ValueError: invalid literal for int() with base 10: '-'

Inspect the 'ACTUAL WIND(MW)' column for this character '-', I found there was only Only 5 months of data in 2013, so as it is at the start of the dataset dropping 2013 entirely

Also need to remove the 2013 data for windall

Rest of occurences are small, so replacing with '0'

Replace negative values with positive counterparts, seems to be a data entry error after visual inspection

Resample the data form every 15 minutes to hourly using the mean to match the wind data

Write the datset out to file

Merge the Wind Speed and Windgeneration dataframes

Modelling the data

I looked at Linear Regression and Random Forest Regression as these are suitable models for time series data.

Initially running training on the entire dataset, for the 16th December 00:00 of each year in the dataset

On this initial test, the RF regression is significantly better than the Linear regression.

Y	LR	RF	Expected
2014	569.3735968	487.65	487
2015	1063.03904	1269.61	1212
2016	582.1128675	671.08	666
2017	361.5460127	479.59	539
2018	1473.804642	1771.49	1847
2019	1183.167491	1246.72	1367
2020	2075.074323	2249.07	2514
2021	863.7137986	937.96	1016
2022	341.9531988	208.44	206

To properly evaluate the models, I ran a cross validation. As this is a time series the traditional k-fold cross-validation techniques are not appropriate.

The TimeSeries Split will be used instead with the validation derived from cross_val_score

We have 9 years of training data, so the first 8 is used for the training and validation and the last year can be used for testing proper.

I am using r2 as the score

LinearRegression: 0.702551 (0.084962)

RandomForestRegressor: 0.756548 (0.073777)

Ran in 4m 47s

The parameter ranges I tried were as follows

'n_jobs': [-1,1],

'n_estimators': [10, 20, 50],

'max_features': [1.0, 'sqrt', 'log2'],

'max_depth' : [i for i in range(5,15)]

The best hyperparameters found during the grid search are:

RandomForestRegressor(max_depth=14, n_estimators=50, n_jobs=1)

0.7529136783787332

After running the best model on the test data, I got an r2f score of 0.7438093479962249

Rerunning the model with the parameters suggested on the test data gave an

The best hyperparameters found during the grid search are:

RandomForestRegressor(max_depth=14, n_estimators=50, n_jobs=1)

r2f = 0.7529136783787332

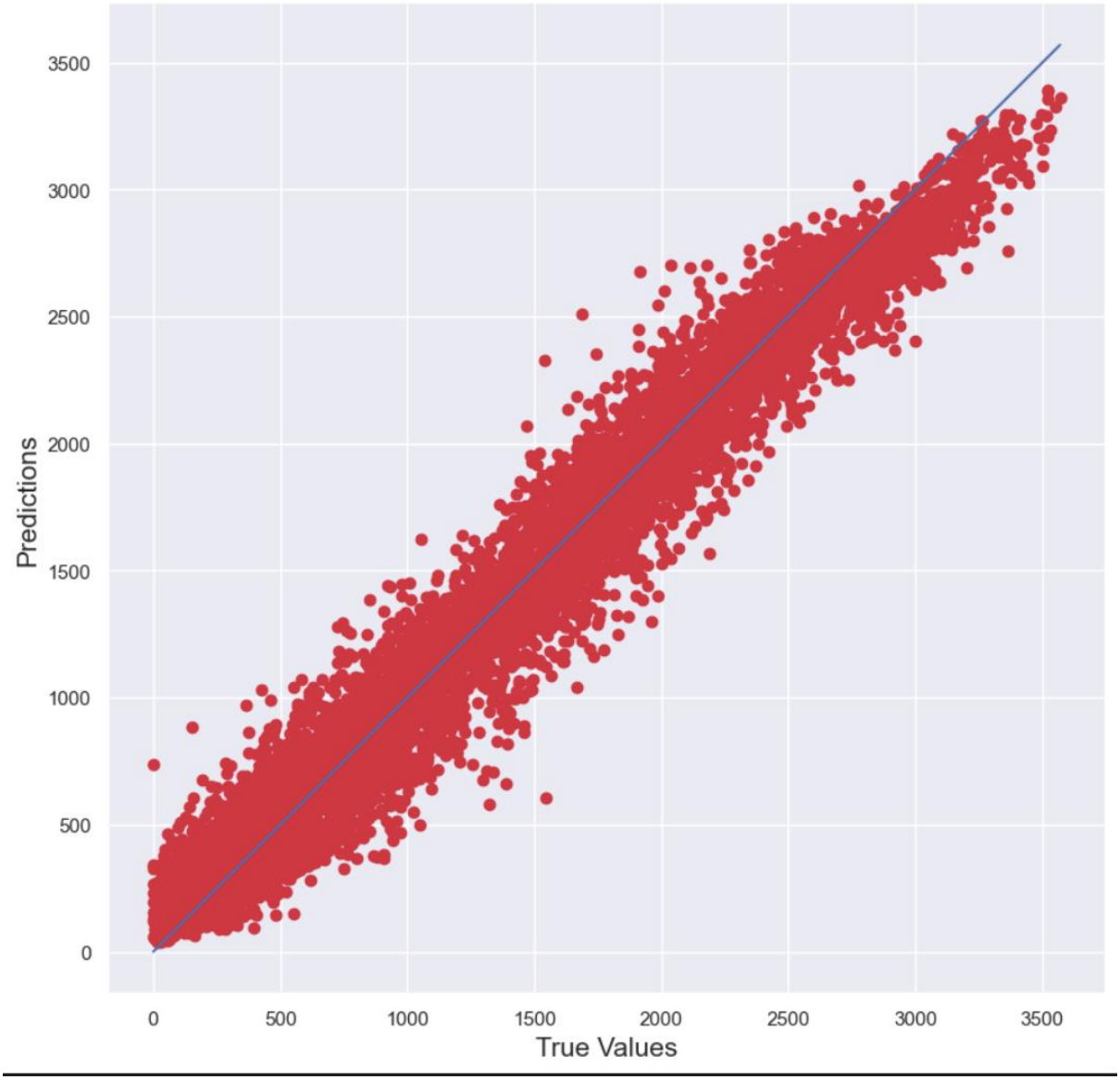
Using the test data, test the model with the best_model hyperparameters

r2f = 0.7438093479962249 in 0.1

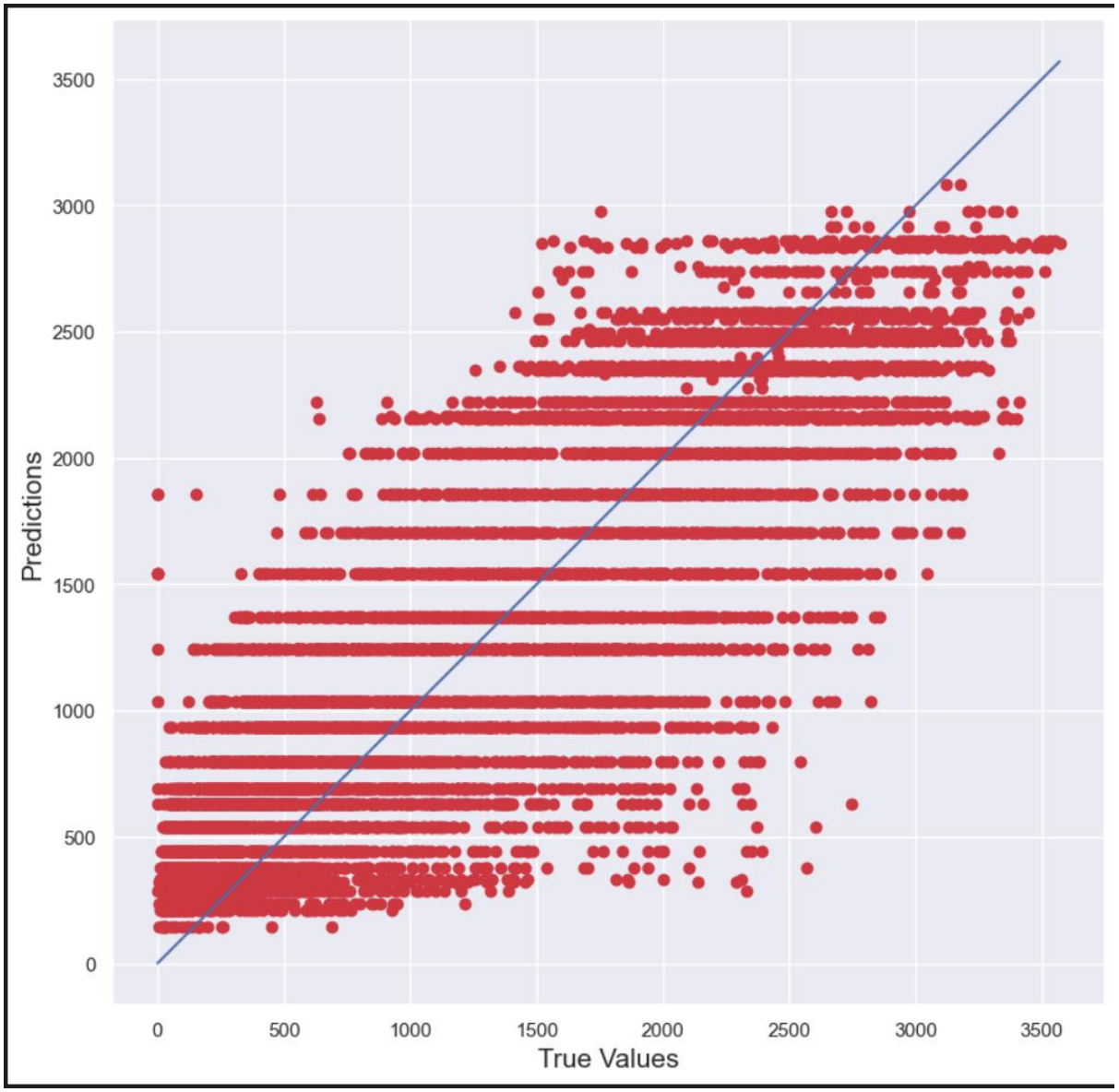
Analyse the best features

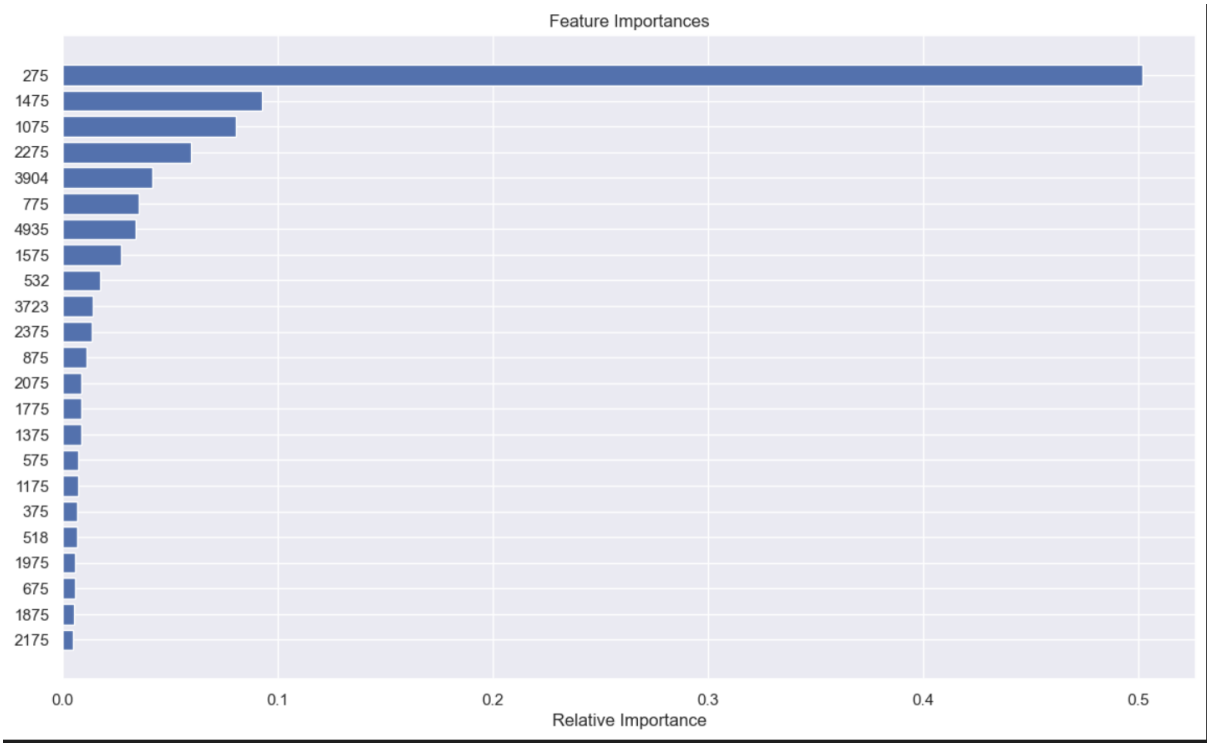
Results

This plots the predicted vs true values on the test data demonstrating a good fit for the model on the data.



This plot shows the results from predicting on the single best weather station, which is unsurprisingly on the West coast at Mace Head in Galway


Predictions
True Values
0
500
1000
1500
2000
2500
3000
3500



Feature Importance

Insights

Even in the absence of location data there is a very strong predictive relationship between wind speed and wind poer generation.

Wind Power generation, while unpredictable has a strong seasonal trend.

HyperParameter tuning is important

Data Science is hard

References

(Include any references if required