project2020

October 26, 2020

GMIT
INSTITIÚID TEICNEOLAÍOCHTA NA GAILLIMHE-MAIGH EO
GALWAY-MAYO INSTITUTE OF TECHNOLOGY

1 Higher Diploma in Science in Computing (Data Analytics)

1.1 #### Programme Module: Fundamentals of Data Analysis (COMP07084)

1.2 Project 2020

In this project you must perform and explain simple linear regression using Python on the `powerproduction` dataset available on Moodle.

The goal is to accurately predict wind turbine `power` output from wind `speed` values using the data set as a basis. Your submission must be in the form of a git repository containing, at a minimum, the following items:

1. Jupyter notebook that performs simple linear regression on the data set.
2. In that notebook, an explanation of your regression and an analysis of its accuracy.
3. Standard items in a git repository such as a README.

To enhance your submission, you might consider comparing simple linear regression to other types of regression on this data set. Rest assured, all the above concepts will be explored in lecture videos and other materials in the coming semester.

1.3 ## Wind Energy

1.3.1 Introduction

Energy demand across the world is increasing rapidly because of population and economic growth especially in emerging market economies. This has meant the search for new energy resources has intensified across the globe. The supply of energy is a key element in a countries social and economic development so much so that it can impact international relations. Conflicts in Iraq/Syria, South Sudan, the Crimea/Ukraine, and the South China Sea show the desire to control valuable oil and gas assets is fuelling long-standing historic tensions. Michael Klare argues that *"in a fossil-fuel world, control over oil and gas reserves is an essential component of national power"*. (Klare, 2014)

These factors combined with the impact of climate change mean the world will increasingly use renewable energy instead of fossil fuels in order to meet energy demand. Wind energy is seen as a positive alternative to fossil fuels but as with all renewables it has advantages as well as disadvantges. Although the manufacturer and installation of wind turbines does involve the release of some pollution into the environment, in the long run it is a source of clean energy and does not emit any greenhouse gases. It is renewable, space efficient, low cost and promotes job creation. However, it does have disadvantages, including its intermittent and unpredictable nature, noise issues, wildlife habitat dislocation as well as impacting tourism with some considering it an eye sore. It also faces location limitation issues, *"because to be economically viable, they need to be installed in a place where they will produce enough electricity, which means coastal areas, the tops of hills, and open planes - essentially anywhere with strong, reliable wind. Most of these suitable places tend to be in remote areas far outside of cities and towns, in more rural areas or offshore. Because of this distance, new infrastructure, such as power lines, have to be built in order to connect a wind farm to the power grid"*. (Lane, 2020)

1.3.2 Facts and figures

In 2019 global direct primary energy consumption stood at 158,839 TWh with renewable energy sources (RES) accounting for 19,219 TWh or 12.1% of this figure. Within (RES) wind energy contributed 6.6% or 1,270 TWh of all renewables. In 2019 Europe had 205 GW of wind energy capacity accounting for 15% of the EU-28 consumed electricity. An additional 15.4 GW of new wind power capacity was added in 2019 an increase of 27% on the previous year.

Within Europe ten member states are achieving wind power shares above 10%. *"The highest share, and a new record, was set by Denmark where 47% of the electricity demand in 2019 was met by wind energy, followed by Ireland at 32% and Portugal at 27%"* (Windeurope.org, 2020).

In Ireland wind energy is the largest contributing resource of renewable power. According to the Sustainable Energy Authority of Ireland (SEAI) in their Wind Energy Roadmap, *"wind energy has the potential to generate enough electricity to exceed domestic demand by 2030 and for its wind market to become export driven in the 2020-2030 timeframe"* (SEAI, 2011). This could generate €15 billion in economic value and lead to the creation of 20,000 jobs in the installation, operation and maintenance of wind farms. It would also present a large carbon abatement in the range of 400 to 450 metric tonnes of CO2 by 2050.

In order to achieve these ambitious goals significant investment is required and it is estimated that the wind industry would hit a peak annual investment of between €6 billion and €12 billion by 2040 (SEAI, 2011). Caution should be applied however when assessing investment in wind energy. One report has identified a payback period of 23 years for an investment in a wind energy project,

which would be a disappointing timeframe for investors. However, this does not have to be the case if a significant feasibility assessment is carried out.

According to a study carried out by TUD, if due consideration is given to *"site selection, electricity market conditions, the quality of the control system and the competencies of the design/installation/commissioning company"* (Kealy, 2015) there should be a positive outcome for investors, consumers and the environment. Their study based on a wind farm in the North East of Ireland were an average capacity factor of 34% was returned from the wind farm's turbines a payback period of 6.7 years was set. Because of the intermittent nature of wind, turbines typically produce only 20% - 40% of their maximum possible output over the course of a year which is known as the capacity factor (Miller, 2014).

1.3.3 Location

The crucial factor in the location of wind farms is calculating the annual energy production and how the energy it produces compares to alternative sources of energy. Key to this is access to and modelling of long-term data. Data should be collected from a potential site over a two to three-year timeframe. From this data the long-term annual variability needs to be calculated and can the renewable energy production output be predicted/forecast (Nelson, 2019). Modelling solutions at their most basic can be done through physical modelling via computational fluid dynamics or mathematical modelling via regular linear algebra.

Simple mathematical modelling on its own won't provide the accuracy but will suffice for common use cases like estimating maximum power production. This is the case with wind turbine manufacturers using power curve modelling techniques. The wind turbine power curve shows the relationship between wind speed and power generated *"for different wind speeds. A typical wind turbine power curve has three main characteristic speeds: 1) cut-in (Vc); 2) rated (Vr); and 3) cut-out (Vs) speeds. The turbine starts generating power when the wind speed reaches the cut-in value. The rated speed is the wind speed at which the generator is producing the machine's rated power. When the wind speed reaches the cut-out speed, the power generation is shut down to prevent defects and damages"* (Wadhvani, 2017).

Power curves can oversimplify reality though and can err by plus or minus 20% the actual power output. In actuality additional variables must be factored in (Miller, 2014). For example, *"wind speed at heights below and above the hub, wind shear, and turbulence are also strong predictors of power production"*. Any model used in predicting power output from wind turbines must also quantify uncertainty or confidence level associated with them. *"Such confidence levels are particularly of value to electric grid operators, who need both the predictions of output and the associated levels of confidence to determine an optimal schedule for turning various sources of power on and off. Quantifying output uncertainty is also crucial for siting wind farms"*(Miller, 2014). Outliers also need to be considered in any model, wind ramps being one of the most important. Because power output is proportional to the cube of the wind speed a wind ramp can result in a dramatic change in power production. *"Consequently, accurate wind ramp prediction is extremely important, leading some experts to refer to it as "the Holy Grail of wind forecasting."(Miller, 2014)*

Importing libraries Numerical Python or NumPy is the first library we require. NumPy is the fundamental package for scientific computing in Python. It is a Python library that provides a multidimensional array object, various derived objects (such as masked arrays and matrices), and an assortment of routines for fast operations on arrays, including mathematical, logical, shape ma-

nipulation, sorting, selecting, I/O, discrete Fourier transforms, basic linear algebra, basic statistical operations, random simulation and much more.

[45]:
```
import numpy as np
```

Next import the pandas library. pandas is a fast, powerful, flexible and easy to use open source data analysis and manipulation tool,built on top of the Python programming language.

[46]:
```
import pandas as pd
```

Next up import statsmodels. statsmodels is a Python module that provides classes and functions for the estimation of many different statistical models, as well as for conducting statistical tests, and statistical data exploration. An extensive list of result statistics are available for each estimator. The results are tested against existing statistical packages to ensure that they are correct.

[2]:
```
import statsmodels as sm
```

Finally we need to import Matplotlib a plotting library available for the Python programming language as a component of NumPy. Matplotlib embed plots in Python applications. Seaborn is a Python data visualization library based on matplotlib. It provides a high-level interface for drawing attractive and informative statistical graphics. The `%matplotlib inline` statement will cause our matplotlib visualizations to embed themselves directly in our Jupyter Notebook, which makes them easier to access and interpret.

[47]:
```
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

For this project we had to perform a simple linear regression on a dataset relating to wind turbines with the goal of accurately predicting wind turbine power output from the wind speed values. Regression is a statistical technique which is used to investigate the relationship between variables. When we only have one input variable, in this case wind speed, it is a simple linear regression, when there is more than one input variable then a multiple linear regression would be used.

The equation for a simple linear regression is $y = w0 + w1 * x1$. In our case we know the input variable $x1$ is wind speed and y is power output.

$w0$ and $w1$ are the two coefficients, where $w0$ is the intercept (of the y-axis), and $w1$ is the slope of the line. $w1$ shows the impact of the independent variable $x1$ on y.

For example, when $w1 = 0$, there's no impact of $x1$ on y since $(0 * x1 = 0)$. (Linear Regression in Machine Learning, 2020).

Therefore simply, linear regression is an algorithm that finds the best values of w0 and w1 to fit the training dataset. (Linear Regression in Machine Learning, 2020).

Dataset

[2]:
```
dataset = pd.read_csv('powerproduction.csv', delimiter = ',')
```

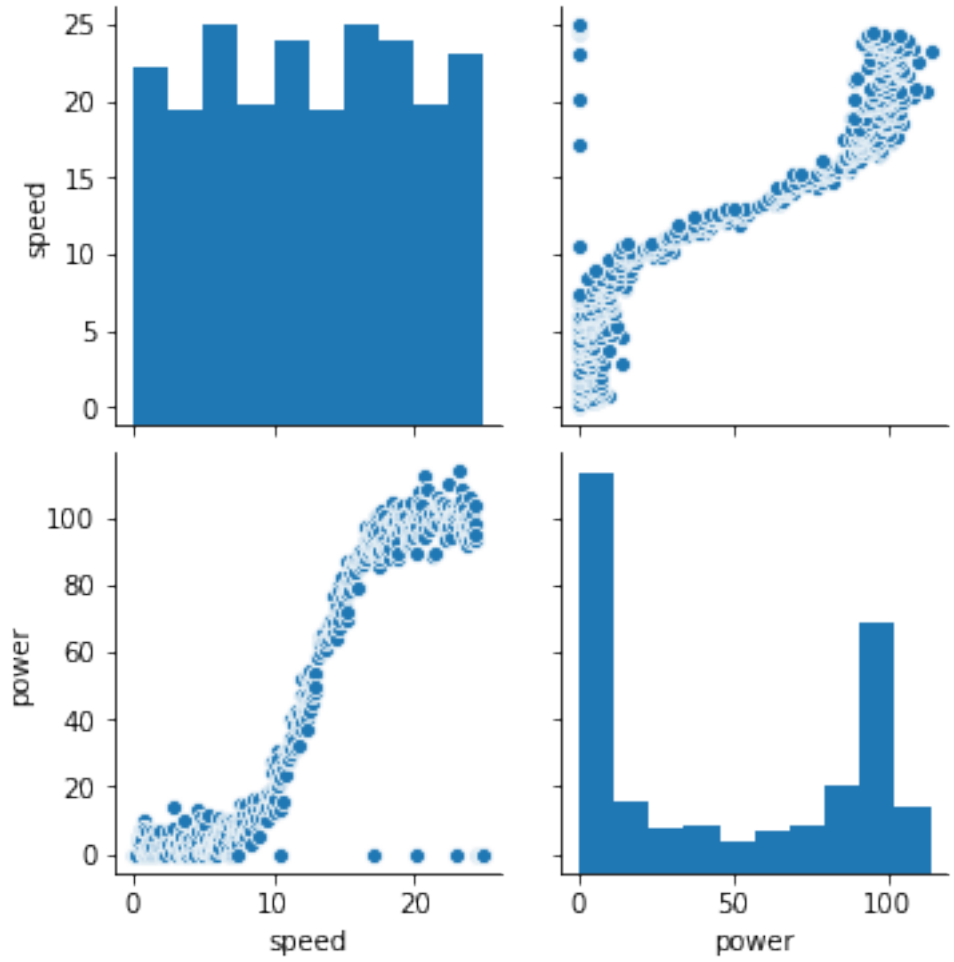
[3]:
```
dataset.shape
```

[3]:
```
(500, 2)
```

[5]:
```
dataset.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 500 entries, 0 to 499
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype  
---  ------  --------------  -----  
 0   speed   500 non-null    float64
 1   power   500 non-null    float64
dtypes: float64(2)
memory usage: 7.9 KB
```

[7]:
```
sns.pairplot(dataset);
```



[9]:
```
dataset.columns
```

[9]:
```
Index(['speed', 'power'], dtype='object')
```

[31]:
```
x = dataset.iloc[:, :-1].values
y = dataset.iloc[:, 1].values
```

[32]:
```
from sklearn.model_selection import train_test_split
```

[33]:
```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,␣
↪random_state=0)
```

[35]:
```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(x_train, y_train)
print(model.coef_)
```

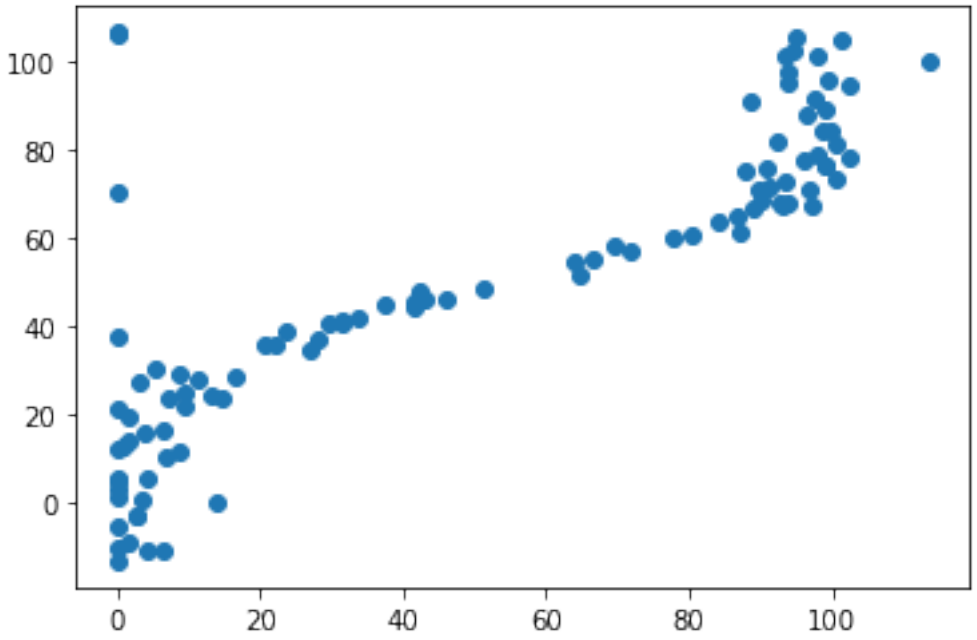
```
[4.89542079]
```

[36]:
```
print(model.intercept_)
```

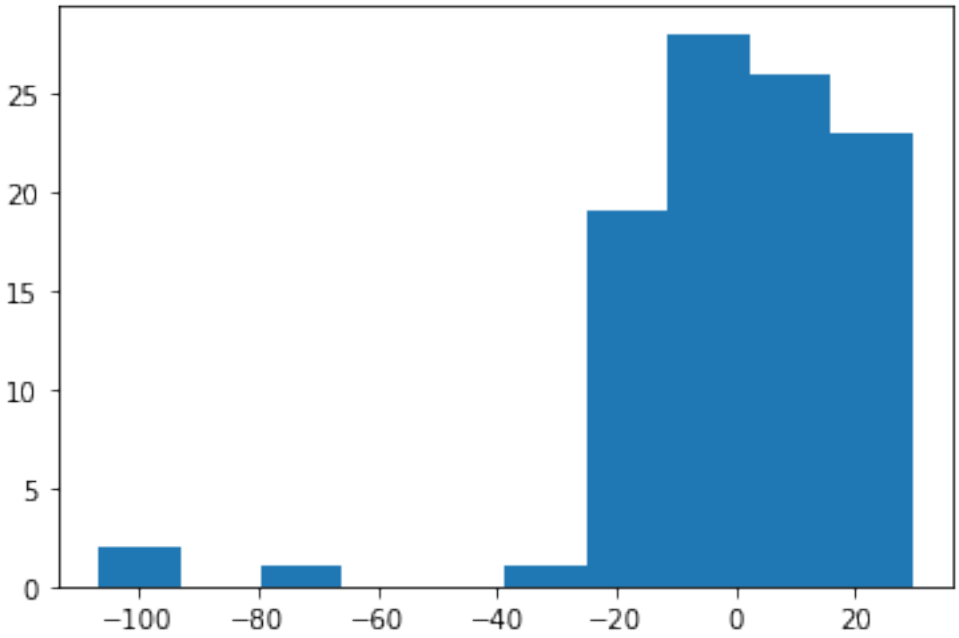
```
-13.603433993820211
```

[38]:
```
predictions=model.predict(x_test)
```

[40]:
```
plt.scatter(y_test, predictions);
```



[41]:
```
plt.hist(y_test - predictions);
```



[42]:
```
from sklearn import metrics
```

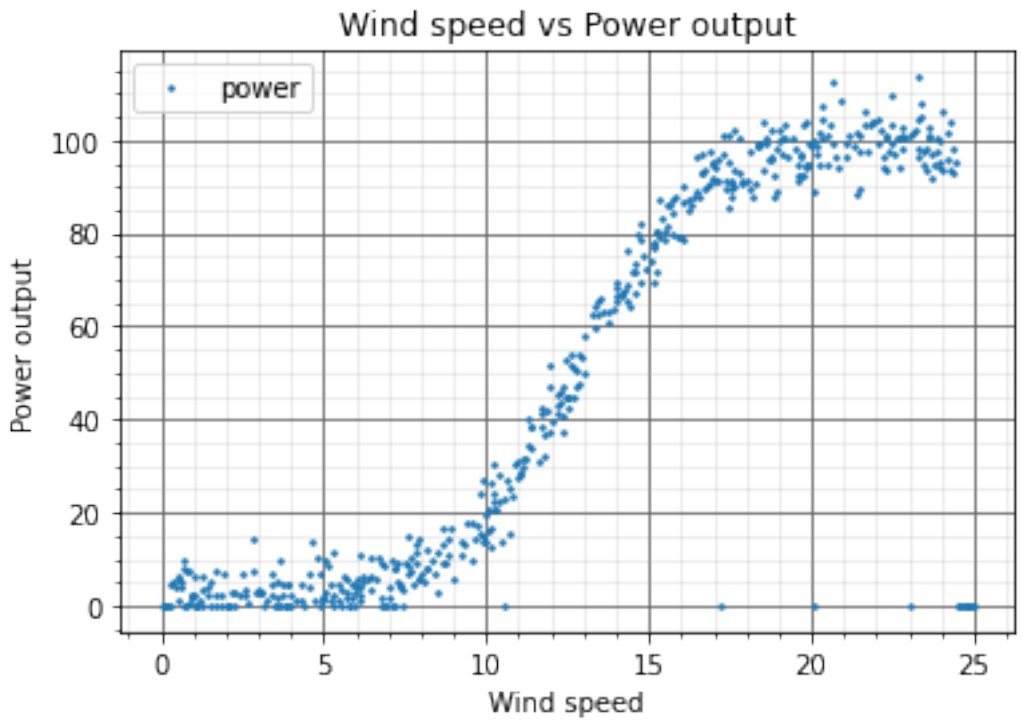
[4]:
```
dataset.describe()
```

[4]:

	speed	power
count	500.000000	500.000000
mean	12.590398	48.014584
std	7.224991	41.614572
min	0.000000	0.000000
25%	6.324750	5.288000
50%	12.550500	41.645500
75%	18.775250	93.537000
max	25.000000	113.556000

[ ]:
```
```

[5]:
```
dataset.plot(x ='speed', y='power', style=('o'), markersize=1.5)
plt.title('Wind speed vs Power output')
plt.xlabel('Wind speed')
plt.ylabel('Power output')
plt.grid(b=True, which='major', color='#666666', linestyle='-')
plt.minorticks_on()
plt.grid(b=True, which='minor', color='#999999', linestyle='-', alpha=0.2)
plt.show()
```



Preparing the data

[11]:
```
x = dataset.iloc[:, :-1].values
y = dataset.iloc[:, 1].values
```

[12]:
```
from sklearn.model_selection import train_test_split
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.2,␣
↪random_state=0)
```

Training the algorithm

[8]:
```
from sklearn.linear_model import LinearRegression
regressor = LinearRegression()
regressor.fit(x_train, y_train)
print(regressor.intercept_)
```

```
-13.603433993820211
```

[9]:
```
print(regressor.coef_)
```

```
[4.89542079]
```

Making predictions

[10]:
```
y_pred = regressor.predict(x_test)
```

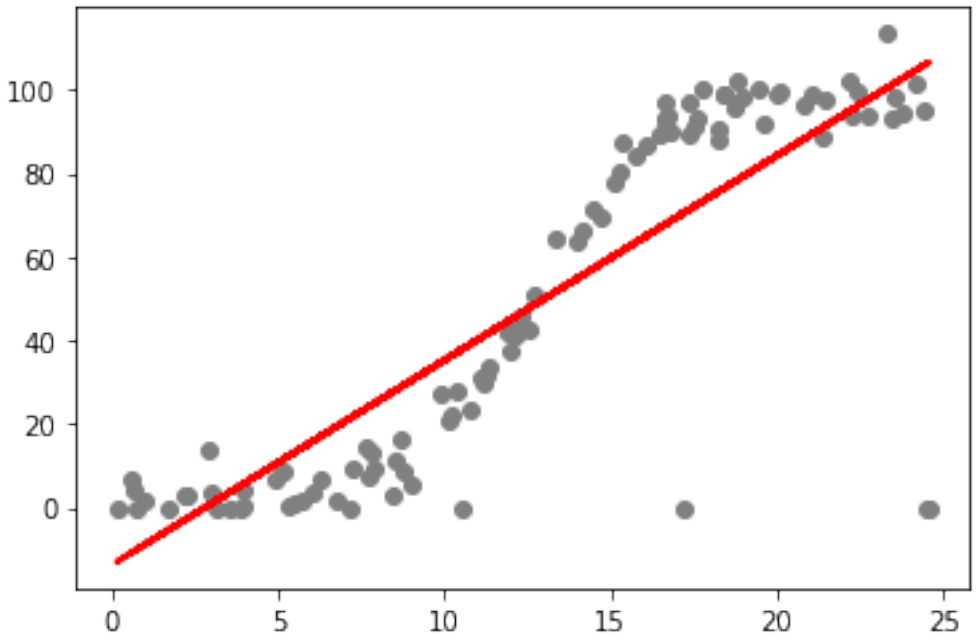
[11]:
```
df = pd.DataFrame({'Actual': y_test, 'Predicted': y_pred})
df
```

[11]:

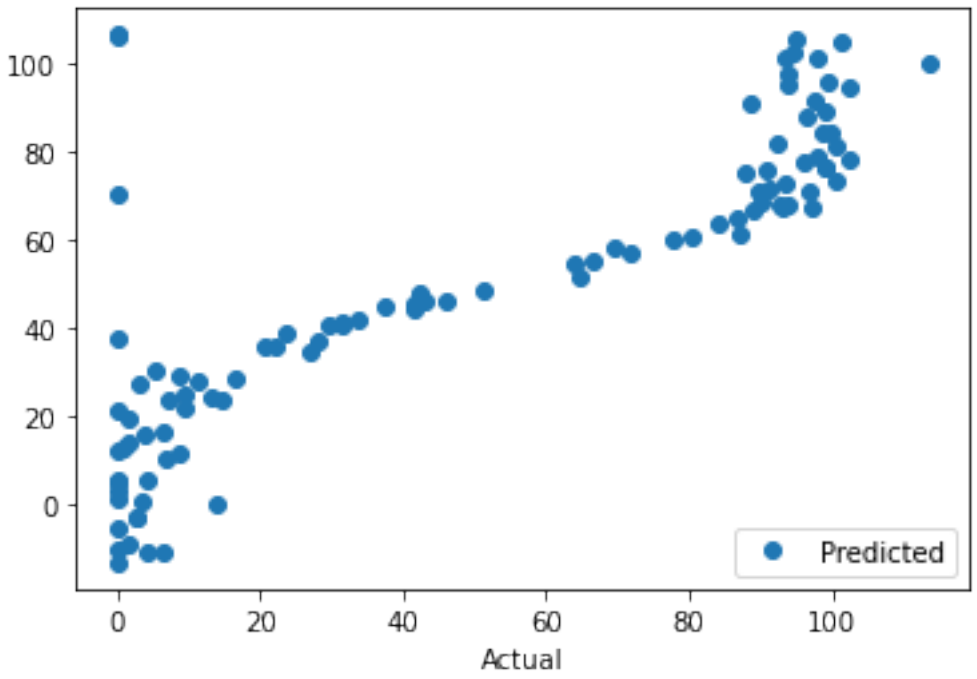
	Actual	Predicted
0	7.060	10.408605
1	51.149	48.632051
2	71.763	57.326318
3	99.357	96.161691
4	113.556	100.327694
..	…	…
95	96.058	77.911562
96	3.578	1.097515
97	93.931	95.304992
98	0.000	1.709442
99	0.000	37.852334

[100 rows x 2 columns]

[12]:
```
plt.scatter(x_test, y_test, color='grey')
plt.plot(x_test, y_pred, color='red', linewidth=2)
plt.show()
```



[13]:
```
df.plot(x='Actual', y='Predicted', style='o');
```



Evaluating the algorithm

[14]:
```
from sklearn import metrics
print('Mean Absolute Error:', metrics.mean_absolute_error(y_test, y_pred))
print('Mean Squared Error:', metrics.mean_squared_error(y_test, y_pred))
print('Root Mean Squared Error:',np.sqrt(metrics.mean_squared_error(y_test,␣
 ↪y_pred)))
```

```
Mean Absolute Error: 15.371033053882327
Mean Squared Error: 496.3930965626669
Root Mean Squared Error: 22.279880981788637
```

1.4 Bibliography

[1] Clifton, A., Kilcher, L., Lundquist, J. and Fleming, P., 2013. *Using Machine Learning To Predict Wind Turbine Power Output.* [online] ResearchGate. Available at: https://www.researchgate.net/publication/257748412_Using_machine_learning_to_predict_wind_turbine_power_output [Accessed 20 October 2020].

[2] Epaper.dk. 2020. *IEA Wind TCP - Annual Report 2019.* [online] Available at: https://www.epaper.dk/steppaper/iea/iea-wind-a-rsrapport-2019/ [Accessed 18 October 2020].

[3] Miller, W., 2014. *Predicting Wind Power With Greater Accuracy.* [online] Str.llnl.gov. Available at: https://str.llnl.gov/april-2014/miller [Accessed 18 October 2020].

[4] Windeurope.org. 2020. *Wind Energy In Europe In 2019.* [online] Available at: https://windeurope.org/wp-content/uploads/files/about-wind/statistics/WindEurope-Annual-Statistics-2019.pdf [Accessed 19 October 2020].

[5] SEAI. 2011. *Wind Energy Roadmap 2011-2050.* [online] Available at: https://www.seai.ie/publications/Wind_Energy_Roadmap_2011-2050.pdf [Accessed 18 October 2020].

[6] Sølverød, F., 2017. *Machine Learning For Wind Energy Prediction - Possible Improvements Over Traditional Methods.* [online] Duo.uio.no. Available at: https://www.duo.uio.no/bitstream/handle/10852/57735/Master_Thesis_Finn_Erik_20170525_FINAL.pdf?sequence=7&isAllowed=y [Accessed 19 October 2020].

[7] Wang, X., Guo, P. and Huang, X., 2011. *A Review Of Wind Power Forecasting Models.* [online] Elsevier. Available at: https://www.sciencedirect.com/science/article/pii/S1876610211019291 [Accessed 20 October 2020].

[8] Evans, R., 2019. *Simple Linear Regression - An Easy Introduction & Examples.* [online] Scribbr. Available at: https://www.scribbr.com/statistics/simple-linear-regression/ [Accessed 20 October 2020].

[9] Khamushkin, I., 2017. *Calculating Energy Production From Weather Forecast In Python.* [online] Medium. Available at: https://medium.com/planet-os/calculating-energy-production-from-weather-forecast-in-python-3c990047daa [Accessed 20 October 2020].

[10] Wadhvani, R., Shukla, S., Gyanchandani, M. and Rasool, A., 2017. *Analysis Of Statistical Techniques To Estimate Wind Turbine Power Generation.* [online] Paper.ijcsns.org. Available at: http://paper.ijcsns.org/07_book/201702/20170232.pdf [Accessed 21 October 2020].

[11] Kealy, T., Barrett, M. and Kearney, D., 2015. *How Profitable Are Wind Turbine Projects? An Empirical Analysis Of A 3.5 MW Wind Farm In Ireland Of A 3.5 MW Wind Farm In Ireland.* [online] Arrow.tudublin.ie. Available at: https://arrow.tudublin.ie/cgi/viewcontent.cgi?article=1101&context=engscheleart2 [Accessed 21 October 2020].

[12] Katabathun, N., Gundabathina, S. and Gummadi, D., 2020. *Prediciting Power Output Based On Weather Condition On Wind Turbines.* [online] Junikhyat.com. Available at: http://www.junikhyat.com/no_14_may_20/33.pdf?i=1 [Accessed 21 October 2020].

[13] Mester, T., 2018. *Pandas Tutorial 1: Pandas Basics (Read_Csv, Dataframe, Data Selection, Etc.).* [online] Data36. Available at: https://data36.com/pandas-tutorial-1-basics-reading-data-files-dataframes-data-selection/ [Accessed 21 October 2020].

[14] Klare, M., Dongre, S., James, M. and James, M., 2020. *Energy Wars: How Oil And Gas Are Fuelling Global Conflicts.* [online] Energy Post. Available at: https://energypost.eu/twenty-first-century-energy-wars-oil-gas-fuelling-global-conflicts/ [Accessed 24 October 2020].

[15] Robinson, S., 2020. *Linear Regression In Python With Scikit-Learn.* [online] Stack Abuse. Available at: https://stackabuse.com/linear-regression-in-python-with-scikit-learn/ [Accessed 21 October 2020].

[16] Nelson, V. and Starcher, K., 2019. *How To Select A Location For A Wind Farm.* [online] Routledge.com. Available at: <https://www.routledge.com/blog/article/how-to-select-a-location-for-a-wind farm? [Accessed 19 October 2020].

[17] Lane, C., 2020. *Wind Energy Pros And Cons.* [online] Solar Reviews. Available at: https://www.solarreviews.com/blog/wind-energy-pros-and-cons [Accessed 25 October 2020].

[18] Stojiljković, M., 2020. *Linear Regression In Python.* [online] Realpython.com. Available at: https://realpython.com/linear-regression-in-python/ [Accessed 21 October 2020].

[19] Chauhan, N., 2020. *A Beginner'S Guide To Linear Regression In Python With Scikit-Learn.* [online] KDnuggets. Available at: https://www.kdnuggets.com/2019/03/beginners-guide-linear-regression-python-scikit-learn.html [Accessed 20 October 2020].

[20] Just into Data. 2020. *Linear Regression In Machine Learning: Practical Python Tutorial.* [online] Available at: https://www.justintodata.com/linear-regression-machine-learning-python-tutorial/ [Accessed 26 October 2020].

[21] Statology. 2020. *How To Create A Scatterplot With A Regression Line In Python.* [online] Available at: https://www.statology.org/scatterplot-with-regression-line-python/ [Accessed 26 October 2020].

[22] Statology. 2020. *A Complete Guide To Linear Regression In Python.* [online] Available at: https://www.statology.org/linear-regression-python/ [Accessed 26 October 2020].

[23] McCullum, N., 2020. *Linear Regression In Python - A Step-By-Step Guide.* [online] Nickmccullum.com. Available at: https://nickmccullum.com/python-machine-learning/linear-regression-python/ [Accessed 26 October 2020].

[ ]: