**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Machine Learning for Business and Data Visualizations |
| **Assessment Title:** | Integrated CA |
| **Lecturer Name:** | James Garza, Sam Weiss |
| **Student Full Name:** | Erica Zanovello |
| **Student Number:** | Sba22179 |
| **Assessment Due Date:** | 26/11/2023 |
| **Date of Submission:** | 24/11/2023 |

**Declaration**

|  |
| --- |
| By submitting this assessment, I confirm that I have read the CCT policy on Academic Misconduct and understand the implications of submitting work that is not my own or does not appropriately reference material taken from a third party or other source. I declare it to be my own work and that all material from third parties has been appropriately referenced. I further confirm that this work has not previously been submitted for assessment by myself or someone else in CCT College Dublin or any other higher education institution. |

Contents

[1. Introduction 3](#_Toc151756481)

[2. Time Series Forecast: predicting daily energy price. 3](#_Toc151756482)

[2.1. Data cleaning and preparation for time series forecast. 3](#_Toc151756483)

[2.2. Time Series Concepts 4](#_Toc151756484)

[2.3. The methodology employed for time series forecasting. 4](#_Toc151756485)

[2.4. Results of time series forecasting. 7](#_Toc151756486)

[3. Exploratory Data Analysis and Dashboard. 9](#_Toc151756487)

[3.1. Data Cleaning and Preparation. 9](#_Toc151756488)

[3.2. Plots Created. 9](#_Toc151756489)

[3.3. Dashboard Building. 12](#_Toc151756490)

[4. Text Analysis: analysing COP26 tweets. 13](#_Toc151756491)

[4.1. Text Cleaning and Preprocessing. 13](#_Toc151756492)

[4.2. Word Cloud. 14](#_Toc151756493)

[4.3. The Latent Dirichlet Allocation. 15](#_Toc151756494)

[5. Conclusions. 15](#_Toc151756495)

[6. References 16](#_Toc151756496)

# **Introduction**

The present analysis explores the topic of energy generation, supply and sustainable energy transition. The first part of this report will focus on forecasting daily energy prices using a SARIMAX time series model. The price of energy can be extremely volatile, but a deep understanding of price trends and the ability to predict the price ahead is crucial for determining profit in energy production and for grid electricity management (Tschora et al., 2022). The data modelled consists in an extraction of energy pricing from a time series dataset on electricity generation and consumption in Spain.

The complete version of the dataset is the object if the exploratory data analysis, whose main insights regarding energy generation and consumption will be collected in a dashboard.

Lastly, Cop26 tweets are examined using text analytics and topic modelling to reveal shared ideas and subtopics about climate change's connection to energy transition.

The following are the links to the dataset employed:

<https://www.kaggle.com/datasets/dsci511g1/cop26-energy-transition-tweets>.

<https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather?select=weather_features.csv>.

GitHub link to the repository: <https://github.com/Er90Space/IntegratedCA2_ML-DV_CCT>.

# **Time Series Forecast: predicting daily energy price.**

# **Data cleaning and preparation for time series forecast.**

This analysis is performed on a subset of the complete dataset containing only the columns relatively to energy price and timestamp. Moreover, only the year 2018 is selected.

The column ‘time’ reporting the date and time of power generation is transformed from format ‘ISO8601’ to ‘datetime64’. This allows to extract the date hence effectively removing the information relative to the time so to focus only on daily energy pricing. The data is then grouped by date and the average price is determined for each day. Moreover, the frequency of the time series is set to ‘day’ (*pandas.DataFrame.asfreq — pandas 2.1.3 documentation*,2023), which will direct the execution of the forecast model towards this time unit.

The transformed time series dataset is visualized with a line plot, which shows the price trend according to the date. The dataset does not present any null value and the observations are consistently distanced by day.

A graph showing the price of a stock market

Description automatically generated with medium confidence

*Figure 1: Daily energy price for year 2018.*

**2.2. Time Series Concepts.**

The application of a time series forecast model assumes the stationarity of the data, which means that statistical properties like mean and variance should remain constant over time (ref). The Dickey-Fuller test is an instrument to assess this assumption, so that the stationarity is confirmed when the p-value < 0.05. If the data is non-stationary, consecutive differentiations are necessary until stationarity is achieved. The parameter ‘d’ in the ARIMA model controls for the degree of differentiation and corresponds to the integration element ‘I’. On the other hand, the Autoregressive element (AR), corresponding to parameter ‘p’, captures the relationship between an observation and a certain number of lagged values, while the Moving Average (MA), corresponding to the parameter ‘q’, captures variation of residual errors with lagged values, where such residuals are assumed to be normally distributed. The examination of the autocorrelation and partial autocorrelation functions, which plots the correlation between a time value and its lagged values, represents a baseline to understand the values of ‘p’ and ‘q’. In case seasonality is suspected, the model SARIMAX can decompose the seasonal element introducing the parameters ‘P’, ‘D’, and ‘Q’.

# **2.3. The methodology employed for time series forecasting.**

The data is split into train and test set with a proportion of 0.90, so to maximize the training data.

A graph showing the price of energy

Description automatically generated

*Figure 2: Train and test sets.*

The Dickey-Fuller test is performed both on the complete data (p-value= 0.40) and the train data (p-value=0.46), showing that the data is non-stationary. The one degree differentiated data results instead stationary (p-value= 6.532345087273852e-11).

The ACF and PACF plotted on both the regular and the differentiated data, show that a seasonal component of 7 days evident in the regular data decreases its significance in the differentiated data. Furthermore, the inspection of the ACF plot suggests values of 0 or 1 for the MA component after differentiation, while the PACF plot suggests values of 2 or 3 for the AR component after differentiation.

A graph with blue lines and dots

Description automatically generated

A graph with blue dots

Description automatically generated

*Figures 3 and 4: Autocorrelation and Partial Autocorrelation in the original data.*

A graph with blue dots and numbers

Description automatically generated

A graph with blue dots and numbers

Description automatically generated

*Figures 5 and 6: Autocorrelation and Partial Autocorrelation in the differentiated data.*

Since a seasonal aspect is visible, it is chosen to employ the model SARIMAX. Consequently, a grid to search for values of ‘p’, ‘d’, ‘q’, ‘P’, ‘D’, ‘Q’ is set and the results of each combination are evaluated by Akaike Information Criterion values. The objective is to select a combination of parameters that balances the necessity of minimising the AIC values while also controlling for overfitting. It is also accounted for the necessity of differentiation, so that ‘d’ must be greater or equal to 1.

After considering ACF and PACF plots, results of the grid search with corresponding AIC values and after a certain degree of experimentation, the chosen parameters are:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **p =3** | **d =1** | **q = 0** | **P = 0** | **D = 1** | **Q = 3** | **AIC Value= 1932** |

*Table 1: SARIMAX parameters.*

# **2.4. Results of time series forecasting.**

Predictions are made for the test set and a set of evaluation metrics are employed to check model adequacy. These are the model evaluations and results:

|  |  |  |  |
| --- | --- | --- | --- |
| **R^2** | **Mean Squared Error** | **Mean Absolute Error** | **Mean Root Squared Error** |
| **0.04** | **5.47** | **1.92** | **2.33** |

*Table 2: SARIMAX model evaluation metrics.*

A graph with blue and red lines

Description automatically generated

*Figure 7: Line Plot of actual price vs forecast.*

The attention is then focused only model performances on the last 10 observations, which are compared with performances on the first 10 observations. The following table resumes these results:

|  |  |  |  |
| --- | --- | --- | --- |
| **Last 10 Observations: Predictions Evaluation Metrics** | | | |
| **R^2** | **MSE** | **MAE** | **MRSE** |
| 0.29 | 4.3 | 1.67 | 2.07 |
| **First 10 test set observations: Predictions Evaluation Metrics** | | | |
| **R^2** | **MSE** | **MAE** | **MRSE** |
| 0.025 | 5.95 | 1.98 | 2.44 |

*Table 3: Comparison of evaluation metrics for the last 10 and the first 10 predictions of the test set.*

From the comparison it results, contrarily to the expectations, that the last 10 observations were predicted slightly better than the first 10 observations of the test set.

# **Exploratory Data Analysis and Dashboard.**

# **Data Cleaning and Preparation.**

The first step consists in removing completely empty columns and columns with predictions, which were part of the dataset but do not reflect the actual data. Since the dataset is in a long format, columns relatively to different types of energy generations are collected in the categorical feature ‘Energy Types’ and their values in the column ‘Energy Value’ with the use of the pandas melt function.

Moreover, similar energy types are grouped together to decrease the cardinality of the column ‘Energy Types’. Unique values are renamed for easier user understanding.

From the timestamp are extracted temporal information such as year, month, date, hour to allow the breaking down of temporal trends.

Finally, a series subsets of the original data is created by filtering out time information and aggregating values using group by according to the focus of the charts.

The libraries used for the visualizations are Altair and Plotly Express, which allow to introduce interactive elements into the graphs.

# **Plots Created.**

The analysis focuses on analysing energy generation and consumption in Spain between the years 2015-2018.

A stacked bar plot is created to investigate the share of energy produced by each energy type across all the years. This plot employers discrete colours and a legend to identify the energy types. There is a bar for each year to allow comparisons. The main energy sources are Wind, Coal, Gas, Nuclear and Hydro, with a drop in Hydro in 2017 offset by an increase of Gas. The energy generation was mostly stable with a slight increase in 2015.

A graph of energy bars

Description automatically generated with medium confidence

*Figure 8 : Stacked bars of energy generated by energy type per year.*

Consequently, the analysis focuses on the most recent year 2018 to visualize with a line plot the energy generated by energy type. Colours are consistent with the previous chart. The plot is interactive and allows to select the energy type, leaving the other grey coloured. The plot offers an insight into energy generation trends. To note is the peak in energy generated by wind and hydro in March, a peak in July for solar and more stable trends for gas, nuclear and coal.

A graph of energy generation

Description automatically generated

*Figure 9: Energy generation by energy type in year 2018.*

Follows a pie chart to more accurately visualize the shares of energy generated by type in the same year. The plot is interactive and colours consistent with the previous charts.

. A diagram of energy sources with Crust in the background

Description automatically generated

*Figure 10: Share of Energy Generated by Type.*

Furthermore, a bar chart shows difference in energy generated by renewable and non- renewable sources. The plot, realized with Plotly, presents an hover interaction aspects to visualize the values. The percentage of renewables is calculated to 40.50% of all energy produced.

A graph of energy values

Description automatically generated

*Figure 11: Energy Produced by Category.*

The analysis focuses then on the consumption side of the energy market plotting the ‘Total Load’ across all years. An area line plot is chosen to emphasize the overall magnitude of the variable. The line plot is interactive and allows to the user to zoom into specific areas. The total load trend appears mostly constant for all the time-period.

A graph of a graph

Description automatically generated with medium confidence

*Figure 12: Total Load trend years 2015-2018.*

A further area line plot highlights the daily trend focusing on a week subset. It is evident the alternance of nights, with dips in the plot, and days, with higher points. The total reaches its peaks in the central part of the week.

A graph of a load

Description automatically generated

*Figure 13: Total load trend 7 days.*

Finally, the last plot zooms into year 2018 and overlaps the line plots of energy generated and total load to highlight the gap. Complementary colours blue and red are chosen. Is evident how the energy generated is far from reaching the total load, so it is supposed that the extra energy will have to be imported in the country.

# **Dashboard Building.**

The dashboard is realized with Panel, chosen for its versatility, simplicity of use and possibility to host the output inside the Jupiter Notebook environment for a smooth development process.

The dashboard has a title and a subtitle with a description of the dataset extracted from the web page where the data is available (*Hourly energy demand generation and weather*, 2019)

The first row of the dashboard collects data per type of energy generated across all years. In the second row are the line plots of energy generated per type and indications on the share of renewables.

The third row focuses on the total load across all years, while the fourth row highlights some aspects of energy consumption: difference in energy produced and consumed in year 2018; daily energy demand in a selected week.

The wireframe is projected and realized with Draw-io. The following is its representation:

A group of rectangular objects

Description automatically generated

*Figure 14: wireframe for the dashboard.*

# **4. Text Analysis: analysing COP26 tweets.**

# **4.1. Text Cleaning and Preprocessing.**

the COP26 dataset is imported and the column of interest, ‘Tweet Content’ , is selected. Tweets are then cleaned and preprocessed following a common pipeline for text manipulation (Van Otten, 2023). This consists of:

* Removing URLs from the tweets.
* Making all words lowercase to ensure that the same words are treated in the same way.
* Remove emoji.
* Remove symbols included punctuation.
* Remove ‘stopwords’, which are uninformative part of speech.
* Remove infrequent words such as words that appear just once in the corpus.

Moreover, every row of tweets is tokenized, which means that the string text is split into individual words to facilitate further analysis.

Follows the step of lemmatization, that, unlike stemming, is a practice that reduces words to their base form, known as lemma, without simply removing prefixes or suffixes (Saumyab, 2023). In this way words can retain their original meaning while being standardized and semantic duplicates can be removed. This procedure improves the performances of topic modelling algorithms like Latent Dirichlet Allocation (Admin, 2023). Lemmatization is applied to all the parts of speech with the use of a function.

# **4.2. Word Cloud.**

The word cloud is a visual representation of text data where the size of each word in the cloud corresponds to its frequency or importance in the given text. It is commonly used to summarize the content of a document.

Upon inspection of the first word cloud, some frequent but uninformative words are detected, which it is decided to remove from the tweets content through the creation of a list of additional stopwords. A second word cloud is then generated:

A black background with colorful text

Description automatically generated

*Figure 15: Word Cloud from the tweets content.*

The inspection of the second word cloud suggests a certain preponderance of the energetic topic, considering the frequency of words such as ‘energetic transition’, ‘clean energy’ and ‘renewable energy’. It is also possible to note the existence of other subtopics such as global players countries (‘India’, ‘China’, ‘South Africa’, ‘UK’, ‘EU’), natural resources (‘oil’, ‘coal’, ‘hydrogen’), general perceptions (‘opportunity’, ‘challenge’, ‘issue’, ‘important’, ‘change’), governance indications (‘investment’, ‘government’, ‘include’, ‘expert’, ‘commitment’, ‘development’, ‘lead’, ‘support’, ‘accelerate’, ‘project’, ‘infrastructure’). These are just some examples. However, many recursive terms are also present.

# **4.3. The Latent Dirichlet Allocation.**

The LDA is a generative probabilistic model that helps uncover the underlying topics within a collection of documents. LDA aims to identify the topics and the words associated with each topic (Kapadia, 2022).

LDA is employed with the expectation of extracting subtopics for the entire corpus, which it is known to be treating the main topic of climate change. As first step, tokens are represented in numerical form as sparse matrix where each unique word in the text corpus is represented as a feature, and the frequency of each word's occurrence in a document is recorded. This representation is known as ‘bag of words’, which the LDA will use as an input.

The number of topics is set to 7, the parameter alpha, which controls the density of topics per document (tweet) is kept small, and beta, which controls the density of words per topic is kept as default (1). (Great Learning Team, 2022). The results are visualized with pyLDAvis library.

From the distance map results that that topics are well distanced, with the exception of topic 5 and 6 that overlap. However, when analysing the words appearing in the different topics it was possible to find recursive words in all the topics such as ‘clean’, ‘climate’, and ‘global’. Moreover, it appears that some words that do not have a meaning in English were present in the document, suggesting that the analysis would have benefited by further cleaning. Overall, the LDA model showed the difficulty of obtaining clear subtopic division given a common main topic.

# **Conclusions.**

The present analysis had the purpose of investigating various aspects of energy generation, consumption and pricing, while attempting at tracking the public opinion on the topic of climate change.

The SARIMAX model predicted up to 37 steps ahead, and a comparison of performance between the last 10 and first 10 showed that, against the expectation, the last observations were better predicted. Overall, the model did not achieve high R^2 (0.04), which means that it does not explain most part of the variation in the true data. This can be likely due to the high degree of fluctuations. However, the error metrics were low, meaning that the model represents a good approximation of the real trend.

The exploratory data analysis showed some key insights regarding the Spanish energy market:

* There is a good balance between renewables and non-renewables sources of power (40% of renewables).
* The main sources of power are nuclear, gas, hydro, coal and wind.
* The total load, that is the demand, is mostly constant over the period considered.
* There is a wide gap between energy generated in the country and energy demanded, which is likely filled by imports.

The text analysis gave an overall understanding of frequent words that can be an indication of tweets content on climate change. Main highlights were around topics of energy transition, perception of the phenomenon, and discussion on governance aspects. However, the LDA could not determine clear subtopics. This can be due to the nature of the documents (tweets), which is very short and doesn’t for allow much variation.

# **References**

Admin (2023) *Lemmatization to enhance Topic Modeling results*. Available at: https://www.bitext.com/blog/lemmatization-to-enhance-topic-modeling-results/.

GeeksforGeeks (2022) *Python  Lemmatization Approaches with Examples*. Available at: https://www.geeksforgeeks.org/python-lemmatization-approaches-with-examples/.

*Gensim: topic modelling for humans* (no date). Available at: https://radimrehurek.com/gensim/auto\_examples/tutorials/run\_lda.html#sphx-glr-auto-examples-tutorials-run-lda-py.

Great Learning Team (2022) *What is Latent Dirichlet Allocation (LDA)*. Available at: https://www.mygreatlearning.com/blog/understanding-latent-dirichlet-allocation/.

*Hourly energy demand generation and weather* (2019). Available at: https://www.kaggle.com/datasets/nicholasjhana/energy-consumption-generation-prices-and-weather?select=weather\_features.csv.

Kapadia, S. (2022) &apos;Topic Modeling in Python: Latent Dirichlet Allocation (LDA),&apos; *Medium*, 23 December. Available at: https://towardsdatascience.com/end-to-end-topic-modeling-in-python-latent-dirichlet-allocation-lda-35ce4ed6b3e0.

*pandas.DataFrame.asfreq — pandas 2.1.3 documentation* (no date). Available at: https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.asfreq.html.

Saumyab (2023) *Stemming vs Lemmatization in NLP: Must-Know Differences*. Available at: https://www.analyticsvidhya.com/blog/2022/06/stemming-vs-lemmatization-in-nlp-must-know-differences/.

Tschora, L. *et al.* (2022) &apos;Electricity price forecasting on the day-ahead market using machine learning,&apos; *Applied Energy*, 313, p. 118752. Available at: https://doi.org/10.1016/j.apenergy.2022.118752.

Van Otten, N. (2023) *Top 14 Steps To Build A Complete NLTK Preprocessing Pipeline In Python*. Available at: https://spotintelligence.com/2022/12/21/nltk-preprocessing-pipeline/.