**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** |  |
| **Assessment Title:** |  |
| **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

**No table of contents entries found.**

1. **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>.

1. **Time Series Forecast: predicting daily energy price.** 
   1. **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, so to focus only on investigating one year trend.

The column ‘time’ reporting the date and time of power generation is transformed from format ‘ISO8601’ to ‘datetime64’. This will allow 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.

* 1. **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’.

* 1. **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

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 x and x: 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 x and x: 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 results of any combinations are evaluated by AIC 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.

* 1. **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 X: 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.

1. **Exploratory Data Analysis and Dashboard.**
2. **Text Analysis: analysing COP26 tweets.**

**4.2. 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.

1. **Conclusions.**