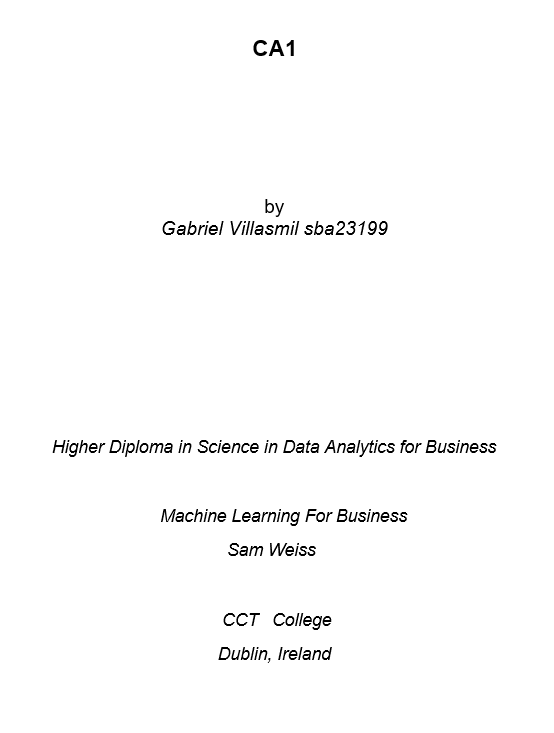
A picture containing icon

Description automatically generated

**CCT College Dublin Continuous Assessment**

|  |  |  |  |
| --- | --- | --- | --- |
| **Programme Title:** | HDip Data Analytics for Business | | |
| **Delivery Mode:** | SB+ | | |
| **Cohort Details:** | *HDIPDAB Sep24 SB+ Stage1 Semester 2]* | | |
| **Module Title(s)**: | *Machine Learning for Business* | | |
| **Assignment Type:** | *Individual* | **Weighting(s):** | *50%* |
| **Assignment Title:** | *CA1* | | |
| **Lecturer(s)**: | *Sam Weiss* | | |
| **Issue Date:** | *13th March 2025* | | |
| **Submission Deadline Date:** | *13th April 2025 @ 23:55* | | |
| **Late Submission Penalty:** | Late submissions will be accepted up to **5** calendar days after the deadline. All late submissions are subject to a penalty of **10%** of the mark awarded.  Submissions received more than 5 calendar days after the deadline above **will not** be accepted and a mark of 0% will be awarded. | | |
| **Method of Submission:** | **This assignment is submitted via Moodle.** | | |
| **Instructions for Submission:** | *A report of approx. 1250 words as a Word document (.docx)*  *Your code as a Jypyter Notebook (.ipynb)* | | |
| **Feedback Method:** | **Results posted in Moodle gradebook** | | |
| **Feedback Date:** | *Within 3 weeks of submission (including late period)* | | |



Contents

Part 1 4

Business Understanding 4

[Data Understanding 4](#_TOC_250005)

[Data Preparation 4](#_TOC_250004)

[Modeling 5](#_TOC_250009)

[Evaluation 5](#_TOC_250006)

[Deployment 5](#_TOC_250006)

[Conclusion 6](#_TOC_250005)

Part 2 7

Business Understanding 7

[Data Understanding 7](#_TOC_250005)

[Data Preparation 7](#_TOC_250004)

[One-Hot Encoding 7](#_TOC_250008)

[Transaction Conversion 7](#_TOC_250008)

[Modeling 7](#_TOC_250009)

[Apriori Algorithm 7](#_TOC_250008)

[FP-Growth Algorithm 7](#_TOC_250007)

[Evaluation 8](#_TOC_250006)

[Deployment 8](#_TOC_250006)

[Conclusion 8](#_TOC_250005)

[References 9](#_TOC_250000)

[GitHub Link 10](#_TOC_250000)

## Business Understanding

The goal is to analyse this Irish electricity data to find patterns in the electricity usage and forecast future business needs. Crucial in this new era due to the fact of increasing new sources of energies, the future for the legacy ones and the fluctuations of the market

In this report, I will:

* Understand seasonal trends in electricity generation.
* Use the Box-Jenkins methodology to fit our time series model.
* Provide a plan of action focused on data-driven solutions.

## Data Understanding

Our dataset “electricity.csv” includes energy demand from Irish people for a variety of sources in a wide range of time.

* **Year of Period and Month of Period** representing the time features.
* **Electricity generation values** are our numbers represented in GWh.

## Data Preparation

### A. Preprocessing Steps

In order to work with time series, I needed to process this data in a way that the model can read it:

### A.1. Working with Dates

* The date column was parsed into a datetime format.
* Data was indexed by the date for time series operations.

### A.2. Stationarity Check

* **Augmented Dickey-Fuller (ADF) test** was applied to check data state.
* **Data was found to be stationary.**
* **Still, I forced into Differencing because it had a better performance than not.**

### B. Data Visualization

* **Line plot** was drawn to see the individual demand for each energy source. This showed me these energy sources' demand varies in **seasons.**
* **ACF (Autocorrelation Function)** and **PACF (Partial Autocorrelation Function)** plots were used to help me set the parameters for the model later on.

## Modeling

### Box-Jenkins (ARIMA)

The Box-Jenkins methodology:

1. **Identification**: Check if **Differencing is needed to have optimal results.**
2. **Estimation**: Once I identify the parameters using the ACF and PACF plots, I fit them into the model, in this case:
3. **Diagnostic Checking**: In this, I run a summary and check the results. While I was coding, I found that if I still forced **differencing it, LogLikelihood showed a better result.**

**Evaluation**

**Log Likelihood:** Our goal is to have this value closer to 0 (Same goes for AIC, BIC, and HQIC); at least, it is not too high, but they do not tell us too much. I need to compare them with the rest with the metrics.

**Residuals:**

* Ljung-Box (L1) (Q): 0.15
* p > 0.05 = fail to reject the null; residuals look like white noise.
* JB = 5.79, p = 0.06
* p > 0.05 = fail to reject the null, residuals are close to normal

**Forecast** shows a period of Growth until 2025, then a period of Stagnation until 2026.

**ARIMA**(1, 1, 1) would perform better in this case than ARIMA(1, 0, 1) even though the data was found stationary.

There are some pikes in **Actual vs Predicted** that are not being shown but seem to be very correlated with no residuals after all, signaling a good fit.

### Next Steps

- Fitting the ARIMA model with different parameters and choose the best one

**Deployment**

For the upcoming 12 months, the ones that seem to bring **more** electricity are:

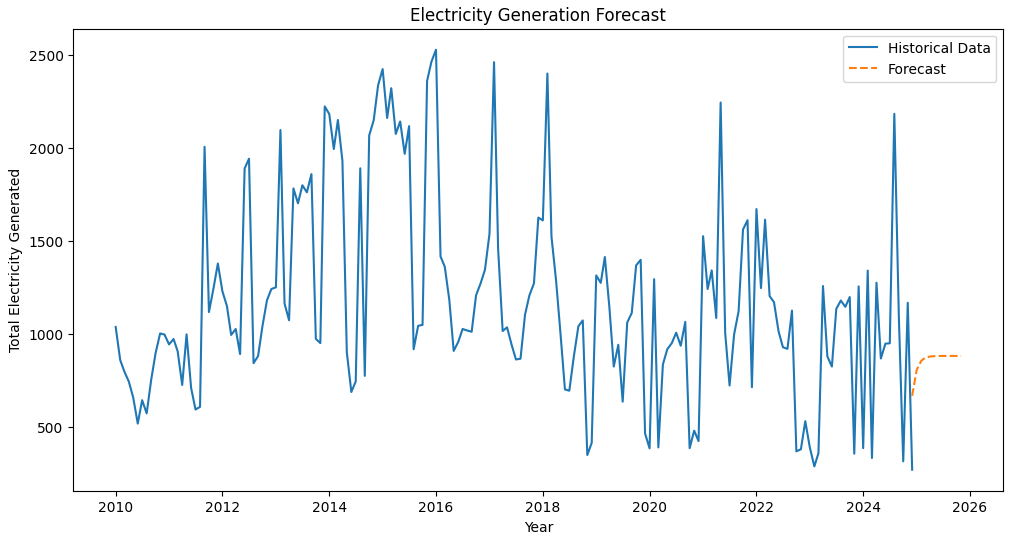
* Hydro
* Combust. Renew.
* Natural Gas
* Wastes
* Wind

If we want to focus on one with a high steady demand, I will have in mind **Combust. Renewable** energy sources.

If we want to focus on peak performance, I will take a look into **Wind** generators and **Natural Gas**.

**Conclusion**

A good model fit would have a MAPE below 15%; in this case, I have **38.9%.** It is like fitting a Random Forest and having a 0.62 in Accuracy. It is not the greatest, but it is not too bad.



*Figure 1: Final Electricity Generation Forecast*

## Business Understanding

With the use of **Market Basket Analysis (MBA)** strategies on this retail shop dataset, hidden patterns will be identified. What we aim to do is help this business make better data-driven decisions to secure profits. To do so, first of all, I am answering these questions:

* Which products are normally bought together?
* Which items can be used for promotions?

Once we have these answers, we can proceed with an action plan.

## Data Understanding

The MBA dataset contains **transactions**; each row represents items purchased in one transaction, thanks to the Boolean values (True or False) that represent whether the item was bought at that time or not.

## Data Preparation

I decided to transform the data to a more pleasurable one. Let's see the steps:

### Transaction Conversion

* Instead of each row being a transaction, now each row was forced into a list of purchased products. I used a Python loop.

### One-Hot Encoding

* The good thing about the default dataset is that it came with a binary classification. The only thing I need to do now is to pass a TransactionEnconder() to maintain its binary state. Then, I saved it into a new dataframe

## Modeling

In this step, I applied two of the most common **Market Basket Analysis algorithms** to find the most valuable itemsets and their validation rules. In parameter selections I focussed on best practices.

#### 1.1 Apriori Algorithm

* Minimum support: 0.05 is the best practice. This tells the algorithm how many times an item must appear to be considered in a set. A high number means fewer items will be considered but will have a strong relationship, and a low number will have more but less valuable patterns (Trade-off).
* Sorted by “Confidence” because I need to know how trustworthy these results were

#### 1.2 FP-Growth Algorithm

* Top five Rules matches exactly the Apriori algorithm rules.
* Performs better in larger datasets but it cost is higher (Trade-off).

## Evaluation

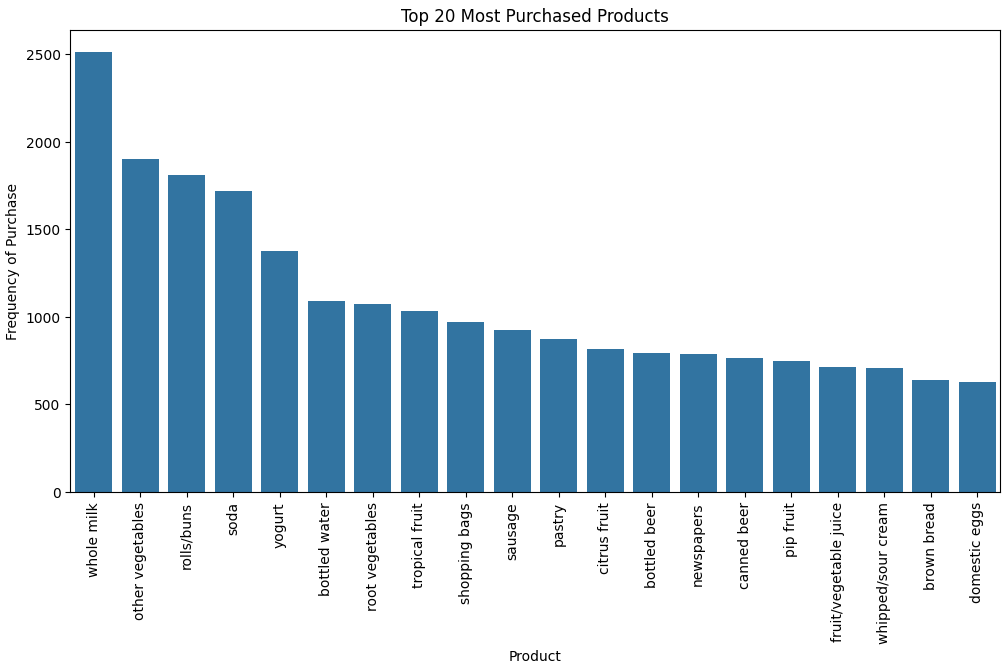
* Confirming our EDA, Whole Milk is a top product with a variety of relatively strong relationships.
* Whole Milk and Yogurt have the strongest relationship we have, with a confidence of 0.40 and the greatest lift of 1.57.
* Yogurt and Whole Milk had a confidence of 0.40. Other Vegetables and Whole Milk had a confidence of 0.38, and Rolls & Buns and Whole Milk had a confidence of 0.30. These three item sets belong to our top 3
* Whole Milk is included in the top three strongest relationships by confidence.
* Dairy, Food, and Vegetables are our strongest connection.

## Deployment

* First of all, let's put Whole Milk and Yogurt as close as possible
* Marketing Campaigns, Promos, or Discounts need to be done on Dairy Food
* The dairy, Food, and Vegetable sections need to be closed and at the very end of the store, making the customer go all the way to the end, increasing the probability that they just remembered they needed an extra item when they saw it on the shelves.
* The bread section needs to be closed as well.

## Conclusion

Dairy, Vegetables, and Bread are foods that go very fast. This store should be advertised as the freshest of the fresh first. Second, when the product is about to go off, give a massive discount on the price, to serve people with both high and low level of income.



*Figure 2: Top 20 Most Purchased Products*

## References

Ali, M. (2023). Association Rule Mining in Python Tutorial. [online] www.datacamp.com. Available at: <https://www.datacamp.com/tutorial/association-rule-mining-python>.

Brian (2018). How to elegantly one hot encode a series of lists in pandas. [online] Stack Overflow. Available at: <https://stackoverflow.com/questions/52189126/how-to-elegantly-one-hot-encode-a-series-of-lists-in-pandas>.

CodeSignal Learn. (2024). Visualizing Actual vs Predicted Prices in Regression Models. [online] Available at: https://codesignal.com/learn/courses/regression-analysis-to-model-diamond-prices/lessons/visualizing-actual-vs-predicted-prices-in-regression-models [Accessed 13 Apr. 2025].

Çoğalan, A. (2023). FP-Growth Algorithm: How to Analyze User Behavior and Outrank Your Competitors. [online] Medium. Available at: <https://medium.com/@anilcogalan/fp-growth-algorithm-how-to-analyze-user-behavior-and-outrank-your-competitors-c39af08879db>.

Kadlaskar, A. (2021). Market basket Analysis | Guide on Market Basket Analysis. [online] Analytics Vidhya. Available at: <https://www.analyticsvidhya.com/blog/2021/10/a-comprehensive-guide-on-market-basket-analysis/>.

KUMAR, S. (2022). Reading ARIMA Results. [online] Medium. Available at: <https://medium.com/@sachinnayak18/reading-arima-results-35eb1df7a565>.

LEONIE (2022). Time Series: Interpreting ACF and PACF. [online] kaggle.com. Available at: <https://www.kaggle.com/code/iamleonie/time-series-interpreting-acf-and-pacf>.

Muhammad, U.S. (2023). Evaluating forecast accuracy in time series with Python code example. [online] Medium. Available at: https://medium.com/@umarsmuhammed/evaluating-forecast-accuracy-in-time-series-with-python-code-example-7dad5af2dd8 [Accessed 13 Apr. 2025].

rajaram, abhinaya (2021). Time Series in Python. [online] Data Science & Business. Available at: <https://medium.com/data-science-business/time-series-arima-sarima-10f5b6a528d5>.

Selvaraj, N. (2022). How to Perform Market Basket Analysis. [online] 365 Data Science. Available at: <https://365datascience.com/tutorials/python-tutorials/market-basket-analysis/>.

Stack Overflow. (n.d.). python - Extracting just Month and Year separately from Pandas Datetime column. [online] Available at: <https://stackoverflow.com/questions/25146121/extracting-just-month-and-year-separately-from-pandas-datetime-column>.

support.minitab.com. (n.d.). Interpret the key results for ARIMA. [online] Available at: <https://support.minitab.com/en-us/minitab/help-and-how-to/statistical-modeling/time-series/how-to/arima/interpret-the-results/key-results/>.

Zach (2021). Augmented Dickey-Fuller Test in Python (With Example). [online] Statology. Available at: <https://www.statology.org/dickey-fuller-test-python/>.

## GitHup Link

<https://github.com/CCT-Dublin/machine-learning-for-business-ca1-Gabriel-studies>