**Capstone Project Proposal**

**Author: Eduardo Matos**

**Lecturer: Sam Weiss / David McQuaid**

**Subject: Machine Learning for Business / Data Visualisation Techniques**

**CCT College Dublin**

*Integrated and Individual CA*

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | **Machine Learning for Business**  **Data Visualisation Techniques** |
| **Assessment Title:** | *Integrated and Individual CA* |
| **Lecturer Name:** | **David McQuaid**  **Sam Weiss** |
| **Student Full Name:** | **Eduardo Matos** |
| **Student Number:** | **Sba23298 - Eduardo Matos** |
| **Assessment Due Date:** | **28/07/2024** |
| **Date of Submission:** | **28/07/2024** |

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.

Index

[Question 1 1](#_Toc173086303)

[Question 3 for Question 1 7](#_Toc173086304)

[Question 4 for Question 1 7](#_Toc173086305)

[Question 2 8](#_Toc173086306)

[Objectives (246 words) 8](#_Toc173086307)

[Problem definition (165 words) 8](#_Toc173086308)

[Scope (73 words) 8](#_Toc173086309)

[Methods (2033 words) 8](#_Toc173086310)

[Data Source (110 words) 8](#_Toc173086311)

[Ethical Considerations (269 words) 8](#_Toc173086312)

[Acronyms (28 words) 8](#_Toc173086313)

[References 8](#_Toc173086314)

# Question 1

* Briefly describe the importance of time series forecasting models for business applications that can be used in a particular scenario and explain this by providing at least two examples.
* Your goal is to forecast based on the use of the experimental data provided at the above-mentioned link  to create time series machine learning models for forecasting the appliances energy usage in a low energy building.
* Use an appropriate train/test split to develop a model, and determine its forecast errors. Evaluate the performance in the context of the dataset.

#### Briefly describe the importance of time series forecasting models for business applications that can be used in a particular scenario and explain this by providing at least two examples.

Time series forecasting models are of significant importance in business for a number of reasons.

The application of demand forecasting models enables businesses to anticipate future product demand, thereby facilitating optimal inventory management and supply chain optimization. Such models assist businesses in forecasting future product demand, thereby facilitating the management of inventory and the optimisation of supply chains.

Financial planning is also facilitated by the use of time series forecasting models. Such models assist in the prediction of revenue, expenses and cash flows, which is essential for the formulation of budgets and financial plans.

The following example scenarios illustrate the potential applications of time series forecasting in various business contexts.

The retail sector employs time series forecasting to predict daily sales, thereby optimising stock levels and reducing the likelihood of overstock or stockouts.

The energy sector can also benefit from the use of time series forecasting. An energy provider utilises electricity demand forecasting in order to efficiently manage grid loads and schedule maintenance.

#### Your goal is to forecast based on the use of the experimental data provided at the above-mentioned link  to create time series machine learning models for forecasting the appliances energy usage in a low energy building.

Exploratory data analysis

Given that the target variable in the dataset is the electricity consumption of household appliances (hereafter referred to as 'Appliances'), it is plausible that other features, such as temperature, may exert an influence upon it. However, it can be reasonably assumed that there is a direct correlation between electricity consumption and time. For instance, it is reasonable to posit that electricity usage in the evening will be higher than during the day, given the typical use of lighting and the prevalence of other household activities, such as watching television. Conversely, the target variable should be at its lowest from midnight to early morning, as most people are asleep, resulting in minimal electricity usage.

Accordingly, in order to more accurately observe the distribution of the target variable over time, the following variables have been introduced.

**Overview the entire dataset.**

Diagrama

Descripción generada automáticamente con confianza mediaImage 1. Overview the entire dataset.

It is possible to observe the shape of the distribution for each feature. For instance, certain features, such as temperature and humidity, may exhibit a normal distribution, whereas others may display a skewed or multimodal distribution.

Furthermore, it can be observed that the values RV1 and RV2 are merely random variables that have been added and could be disregarded if data other than that pertaining to 'Appliances' were to be employed.

**Energy Consumption by Hour**

Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente

Image 2. Energy Consumption by Hour.

The graph above corroborates our hypothesis, demonstrating that electricity consumption is highest between 17:00 and 20:00, and lowest between 23:00 and 06:00 the following day. This corroborates our previous hypothesis, thereby providing further evidence of the reliability of the data.

**Energy Consumption by Day of the Week**

Gráfico, Gráfico de barras

Descripción generada automáticamente

Image 3. Energy Consumption by Day of the Week

**Relationship between Date and Appliances with simple moving average**

As with the analysis of the hours within a single day, it is possible that a pattern may emerge when the data from an entire week is considered. However, when the data set is split by the day of the week, no significant features emerge in the distribution of the target variable. It is noteworthy that electricity consumption on Sundays is approximately 10% higher than on other days.

Gráfico, Gráfico de barras

Descripción generada automáticamente

Image 4. Relationship between Date and Appliances with simple moving average.

There is no clear trend in the SMA plot. Appliance usage is stable with minor fluctuations.

Seasonality

The data does not show strong seasonal patterns. There are no obvious cycles or repeating patterns by time of year, month or week.

There are no unusual spikes or drops in appliance usage. The fluctuations are normal.

**Pairwise Correlation Heatmap**

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

Image 5. Pairwise Correlation Heatmap

The objective of this visual representation is to present the correlations among different features using a heatmap. The data demonstrate a robust positive correlation between temperature and humidity levels. It is notable that RH\_6 (outdoor humidity) exhibits a strong negative correlation with temperatures, indicating that as temperatures increase, humidity levels decrease.

In contrast to the observed correlations among temperatures and humidities, the correlation between our target variable, 'Appliances', and these variables appears to be insignificant.

#### Use an appropriate train/test split to develop a model and determine its forecast errors. Evaluate the performance in the context of the dataset.

ARIMA model implementing

* ARIMA is generally insensitive to feature scaling, but LSTM models are sensitive to feature scaling because they are a type of RNN. Our dataset has no missing values, so imputation is not needed. This makes the modelling process smoother.
* 'Appliances', 'date', 'Windspeed', 'Visibility', 'Tdewpoint', 'rv1', 'rv2 are removed from the training set for the following reasons:
* Appliances: This is the target variable, so it should be separated from the features.
* Date: It is time-related.
* rv1 and rv2: These are random variables introduced for testing purposes, not actual features of interest.
* Windspeed, Visibility, Tdewpoint: These columns are not relevant for this analysis.

**Seasonal decomposition**

Imagen que contiene vara

Descripción generada automáticamente

Image 6. Seasonal decomposition

ARIMA works well with linear trends and seasonality. However, the data seems to be stationary.

**ARIMA Model Results**

Gráfico

Descripción generada automáticamente

The autoregressive (AR) model relies on past observations to predict future values, so I have carefully chosen the appropriate start and end times.

**Arima Model Scores**

Mean Absolute Error (MAE): 30.430244453539146

Root Mean Squared Error (RMSE): 62.33467043548941

Mean Absolute Percentage Error (MAPE): 31.594001370590114

Gráfico, Gráfico de rectángulos

Descripción generada automáticamente

Image 7. Arima Model Scores.

* Mean Absolute Error (MAE): 30.43

The model's predictions are, on average, 30.43 units off. This is a simple way to measure how accurate a prediction is.

* Root Mean Squared Error (RMSE): 62.33

This value suggests that the model makes larger errors sometimes.

* Mean Absolute Percentage Error (MAPE): 31.59%

This means that the prediction errors are about 31.59% of the actual values. The model is moderately accurate.

The ARIMA model provides a reasonable prediction, but there's room for improvement.

# Question 3 for Question 1

#### Create an interactive Dashboard aimed at older adults (60+) with specific features to summarise the most important aspects of the data and communicate through your visualisation based on the results obtained from Question 1 and Question 2.

The histogram shows the distribution of energy consumption levels. Histograms are simple and effective for showing the distribution of data. They help users understand the frequency of different energy consumption levels. The x-axis represents energy consumption levels, and the y-axis represents the count of occurrences. This visualization helps users understand how often certain levels of energy consumption occur. The chart updates dynamically based on the selected date range, providing relevant and timely information.

# Question 4 for Question 1

#### Explain how your dashboard is designed with the above-mentioned demographic in mind and your rationale for the types of the visualisations that you have developed in Question 3.

The interactive dashboard is designed with the needs of older adults (60+) in mind, prioritising simplicity, clarity and ease of use.

Firstly, the text and layout are tailored to ensure readability. The dashboard uses large fonts (18px or larger) in Arial, a simple and legible font. This choice helps older adults who may have visual impairments. High contrast colours, such as #FCB040 for titles, are used against a white background to improve visibility. The overall layout is straightforward and uncluttered, presenting information in a clear and organised manner. This minimises cognitive load and makes navigation easier for non-technical users.

Secondly, the interactive elements are designed to be intuitive and accessible. The date range selector allows users to filter data by selecting a range of dates. This element is large and easy to click on to accommodate users with limited dexterity. When users select a date range, the dashboard provides immediate visual feedback by updating the relevant charts. This feature keeps users engaged and helps them understand the impact of their actions.

The dashboard includes two main types of visualisations: a time series chart and a histogram.

The time series chart shows energy consumption over time. Line charts are simple and effective for showing trends over time, helping users easily understand how energy consumption changes within the selected time period. The x-axis represents the date, and the y-axis represents the energy consumed. The graph is updated based on the selected date range, allowing users to see trends over different time periods.

The histogram shows the distribution of power levels. Histograms are a simple and effective way of showing the distribution of data, helping users to understand the frequency of different energy usage levels. The x-axis represents the energy consumption levels and the y-axis represents the number of occurrences. This visualisation helps users to understand how often certain levels of energy consumption occur.

# Question 2

Perform Market Basket Analysis on the chosen dataset by using Apriori and FP growth algorithms. Can you express 3 similarities between these models? Address the following questions for both algorithms as mentioned below

●   Determine the top 10 most frequently purchased items based on the chosen dataset.

●   Use the Apriori algorithm to find frequent itemsets with a minimum support of 0.01.

●   Extract the association rules with a minimum confidence of 0.5.

●   Identify the top 5 association rules based on metric, lift or leverage.

●   Explain the meaning of these rules in the context of the dataset.

● Provide an explanation of the top 5 association rules.

Compare the machine learning results obtained based on both algorithms and show the exact time used for the evaluation of the number of rules generated in both cases. Use any dataset that has not been used in the class, tutorials and previous assignment.

#### 3 similarities between these models?

- The algorithms find frequent item sets and extract association rules.

- They can both identify strong relationships between items in the dataset.

- You need to set a minimum support threshold to filter item sets.

#### Determine the top 10 most frequently purchased items based on the chosen dataset.

**Items Distribution**

Gráfico, Gráfico de barras

Descripción generada automáticamente

Image 9. Items Distribution.

**Top 10 Most Popular Items**

Gráfico, Gráfico de barras

Descripción generada automáticamente

The top 10 items are:

- Bananas, cheese, coffee, cereal, milk, Oranges, tomatoes, sugar, potatoes, bread, and  bananas are the most popular items sold at the store.

#### Use the Apriori algorithm to find frequent item sets with a minimum support of 0.01 | Extract the association rules with a minimum confidence of 0.5 | Identify the top 5 association rules based on metric, lift or leverage.

Top 5 rules from Apriori:

antecedents consequents \

4353 (Oranges, Butter, Yogurt) (Bread, Eggs)

4360 (Bread, Eggs) (Oranges, Butter, Yogurt)

5596 (Pasta, Chicken, Milk) (Onions, Apples, Eggs)

5597 (Onions, Pasta, Apples) (Chicken, Milk, Eggs)

5598 (Pasta, Apples, Eggs) (Onions, Chicken, Milk)

antecedent support consequent support support confidence lift \

4353 0.013072 0.013072 0.013072 1.0 76.5

4360 0.013072 0.013072 0.013072 1.0 76.5

5596 0.013072 0.013072 0.013072 1.0 76.5

5597 0.013072 0.013072 0.013072 1.0 76.5

5598 0.013072 0.013072 0.013072 1.0 76.5

leverage conviction zhangs\_metric

4353 0.012901 inf 1.0

4360 0.012901 inf 1.0

5596 0.012901 inf 1.0

5597 0.012901 inf 1.0

5598 0.012901 inf 1.0

Top 5 rules from FP-Growth:

antecedents consequents \

1089 (Pasta, Eggs, Potatoes) (Onions, Apples, Milk)

1090 (Pasta, Milk, Potatoes) (Onions, Apples, Eggs)

1093 (Pasta, Milk, Eggs) (Onions, Apples, Potatoes)

1094 (Onions, Apples, Potatoes) (Pasta, Milk, Eggs)

1097 (Onions, Apples, Eggs) (Pasta, Milk, Potatoes)

antecedent support consequent support support confidence lift \

1089 0.013072 0.013072 0.013072 1.0 76.5

1090 0.013072 0.013072 0.013072 1.0 76.5

1093 0.013072 0.013072 0.013072 1.0 76.5

1094 0.013072 0.013072 0.013072 1.0 76.5

1097 0.013072 0.013072 0.013072 1.0 76.5

leverage conviction zhangs\_metric

1089 0.012901 inf 1.0

1090 0.012901 inf 1.0

1093 0.012901 inf 1.0

1094 0.012901 inf 1.0

1097 0.012901 inf 1.0

**Top 5 Rules from Apriori Algorithm**

Oranges, Butter, Yogurt => Bread, Eggs:

Customers who buy Oranges, Butter, and Yogurt also buy Bread and Eggs. This rule is always true in the data set. The high lift value of 76.5 shows these items are bought together much more often than by chance.

Bread and eggs => oranges, butter and yoghurt.

Similarly, customers who buy bread and eggs also buy oranges, butter and yoghurt. It has a confidence of 1.0 and a lift of 76.5, showing a strong association.

Pasta, chicken, milk => onions, apples, eggs:

This rule shows that customers who buy pasta, chicken, and milk also buy onions, apples, and eggs. The perfect confidence value of 1.0 and a lift of 76.5 show a strong relationship between these items.

Onions, pasta, apples => chicken, milk, eggs

Onions, pasta and apples are bought with chicken, milk and eggs. This rule also has a confidence of 1.0 and a high lift, showing a consistent pattern.

Pasta, apples, eggs => onions, chicken, milk

Customers who buy pasta, apples and eggs also buy onions, chicken and milk. This rule shows a strong association.

**Top 5 Rules from FP-Growth Algorithm**

Pasta, Eggs, Potatoes => Onions, Apples, Milk:

Pasta, eggs and potatoes customers also buy onions, apples and milk. The rule shows a strong correlation.

Pasta, Milk, Potatoes => Onions, Apples, Eggs:

This rule shows that customers who buy pasta, milk, and potatoes also buy onions, apples, and eggs. The perfect confidence value and lift show a strong association.

Pasta, milk, eggs => onions, apples, potatoes

Pasta, milk and eggs customers also buy onions, apples and potatoes. The confidence level is 1.0 and the lift is 76.5, which suggests a consistent purchasing pattern.

Onions, apples and potatoes lead to pasta, milk and eggs. The rule shows a strong relationship between these items.

Onions, apples, eggs => pasta, milk, potatoes

Customers who buy onions, apples and eggs also buy pasta, milk and potatoes. The perfect confidence value and high lift show a strong link.

# Question 3 for Question 2

#### Create an interactive Dashboard aimed at older adults (60+) with specific features to summarise the most important aspects of the data and communicate through your visualisation based on the results obtained from Question 1 and Question 2.

Gráfico de barras

Descripción generada automáticamente con confianza media

Image 8. Top 10 Most Popular Items.

# Question 4 for Question 2

#### Explain how your dashboard is designed with the above-mentioned demographic in mind and your rationale for the types of the visualisations that you have developed in Question 3.

The interactive dashboard is designed with the needs of older adults (60+) in mind, prioritising simplicity, clarity and ease of use.

Firstly, the text and layout are tailored to ensure readability. The dashboard uses large fonts (18px or larger) in Arial, a simple and legible font. This choice helps older adults who may have visual impairments. High contrast colours, such as #FCB040 for titles, are used against a white background to improve visibility. The overall layout is straightforward and uncluttered, presenting information in a clear and organised manner. This minimises cognitive load and makes navigation easier for non-technical users.

Secondly, the interactive elements are designed to be intuitive and accessible. The date range selector allows users to filter data by selecting a range of dates. This element is large and easy to click on to accommodate users with limited dexterity. When users select a date range, the dashboard provides immediate visual feedback by updating the relevant charts. This feature keeps users engaged and helps them understand the impact of their actions.

The dashboard includes two main types of visualisations: a time series chart and a histogram.

The time series chart shows energy consumption over time. Line charts are simple and effective for showing trends over time, helping users easily understand how energy consumption changes within the selected time period. The x-axis represents the date, and the y-axis represents the energy consumed. The graph is updated based on the selected date range, allowing users to see trends over different time periods.

The histogram shows the distribution of power levels. Histograms are a simple and effective way of showing the distribution of data, helping users to understand the frequency of different energy usage levels. The x-axis represents the energy consumption levels and the y-axis represents the number of occurrences. This visualisation helps users to understand how often certain levels of energy consumption occur.

## References