**CCT College Dublin**

**Assessment Cover Page**

|  |  |
| --- | --- |
| **Module Title:** | Data Visualisation Techniques / Machine Learning for Business |
| **Assessment Title:** | Integrated CA2 |
| **Lecturer Name:** | James Garza / Sam Weiss |
| **Student Full Name:** | Ivan Saravanja |
| **Student Number:** | sbs23026 |
| **Assessment Due Date:** | 26th November 2023 |
| **Date of Submission:** | 26th November 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. |

**Content**

Introduction………………………………………………………………………………………..3

Sources…………………………………………………………………………………………….3

Libraries…………………………………………………………………………………………...4

Data Preparation…………………………………………………………………………………...4

Data Visualisation assignment…………………………………………………………………….4

EDA……………………………………………………………………………………….5

Creating Interactive Dashboard……………………………………………………………6

Machine Learning for Business assignments………………………………………………………9

Time-series Analysis………………………………………………………………………………9

ADF Test…………………………………………………………………………………..9

ARIMA…………………………………………………………………………………..12

Forecast…………………………………………………………………………………..12

Model Evaluation………………………………………………………………………...13

Text Analytics……………………………………………………………………………………13

Text Preprocessing……………………………………………………………………….14

Text Categorisation/Sentiment Analysis…………………………………………………14

Topic Modelling …………………………………………………………………………15

Document Summarisation………………………………………………………………..17

Challenges………………………………………………………………………………………..17

References………………………………………………………………………………………..18

**Introduction**

Our final assignment for this course was integrated assignment between Data Visualisation Techniques and Machine Learning for Business. It asked for a dashboard creation and analysis of a dataset. For the dashboard, a plan (wireframe) was needed first, showing where plots and information sections would go. At least four visual plots had to be included. The dashboard had to have option for audience to interact with it. It also needed to have a brief explanation about the data and it’s main finding. We were suggested to use Python libraries like Plotly Dash or Altair.

In the machine Learning part, understanding time series data and using text analysis tools on social media data were key. Time series analysis focused on making forecasts and checking patterns in data (we were supposed to do one-step-ahead forecast of the last 10 observations). Text analysis dealt with things like sorting text and summarising it.

**Sources**

During this assignment, I was using 2 datasets:

* First one was for DV and Time series parts of the assignment. This was weather dataset, providing a comprehensive overview of various weather parameters for Istanbul in period from April 7th 2018 until September 27th 2023. (Missioner, 2023)
* Second dataset was used for last part of the assignments – Text Analytics. It’s dataset from the same source, Kaggle and it includes top daily tweets containing the keyword ‘Climate Change’ and it covers the period from January 1st 2022 until July 19th 2022. I wanted to stay within the same domain of climate. (Scweet, 2022)

**GitHub Repository**

I am leaving link to my GitHub here, as per assignment requirements:

<https://github.com/IvanSaravanja2306/ML_and_DV_IntegratedCA2>

I added both, James and Sam, as collaborators and invite is pending.

**Libraries**

To start with the assignment, first I had to upload all the libraries and modules I needed for data analysis, visualisation, natural language processing and machine learning tasks throughout the assignment. These libraries include Pandas for data manipulation, Plotly, Seaborn, and Matplotlib for visualization, Dash for creating interactive web-based dashboards, Statsmodels for time series analysis, NLTK for natural language processing, and scikit-learn for machine learning. My import cover a wide range of functionalities such as statistical analysis, time series modelling (SARIMAX), text analysis (TF-IDF vectorization, LDA), sentiment analysis (TextBlob), handling imbalanced data (RandomOverSampler), classification algorithms (SVC, MultinomialNB), and more.

**Data Preparation**

As I mentioned before, I was using 2 datasets throughout this assignment. I picked this type of the dataset for the first part as it has daily inputs and will be usable for Time series, but also, weather data is always good fit for visualisations. In order to better understand the dataset, I’m sharing list of columns and terminology explanation:

A white text with black text

Description automatically generated

Data preparation part was done by checking first which variables contain NaN values:

A screenshot of a computer

Description automatically generated

Since this is weather dataset, I looked for the unique values within columns that contain NaN values, to see if missing value is just matter of mistake or it indicates that value for that column was not available in the moment of recording. Later on, I used dropping columns, Interpolation and Imputation to remove or replace missing values.

A screenshot of a computer code

Description automatically generated

**EDA**

While performing EDA, I decided to put focus on few aspects of the dataset:

* Data Distribution
* Correlation Analysis
* Categorical Variables Analysis
* Relationship between Continuous Variables.
* Temperature Variation
* Time-related Analysis
* Mean data per year compared to overall mean

These were my conclusions:

A group of blue and white bars

Description automatically generated

A screenshot of a graph

Description automatically generated

A graph of different colored squares

Description automatically generated

A graph of different colored squares

Description automatically generated

Based on this information, weather conditions in this area could be described as moderately warm, with moderate wind gusts, but mostly rainy. This would imply comfortable climate with temperatures typically not too hot or cold, along with occasional breezy conditions. We can see on minimum temperature that it barely goes under 0C. Also, due to high UV index, it would be advisable to take precautions when UV index is in most common range (8-9).

As we can see, columns with higher positive correlation are temperature, dew, visibility and UV Index in combination with each other.

There are no outliers.

**Interactive Dashboard**

As per assignment, we were supposed to create a interactive dashboard which will include 3 rows and 2 columns, making total of 6 sections. When it comes to visualisation components, in my workbook, I used Plotly Dash and plots wise, I used Heatmap, Line Chart, Histograms, Bar Chart, Stacked Bar Chart, Boxplot and Scatter plots to visualise my dataset. Also, I used Interactive features like dropdowns and interactive plots.

* “Overview" is a text section, it's giving all the relevant information on the dataset I picked, as well as link to its source. Also, sums up all the conclusions I made by performing the EDA.
* "Distribution of Weather Parameters" is histogram, created to give insights on distribution of the weather parameters.
* "Mean Temperature per Month for Each Year" is a heatmap type of plot and it displays mean temperature variations across on month level, for 5 years.
* "Correlation Heatmap of Continuous Variables" is a Heatmap (Correlation Matrix) type of plot. I picked this type of plot as it shows relationship between continuous variables in simple and understandable way
* "Weather Trends over Time" is time-series plot, Line Chart type. Purpose of this plot is to show trends or patterns over a continuous time period.
* "Type of Weather for Each Day" is categorical time series plot. I picked this plot to give audience more insights on what kind of weather was recorded each day

These plots together offer a comprehensive view of the weather dataset, covering distribution, trends over time, relationships between variables and correlations among continuous variables, ensuring a holistic understanding for audience. The selection of plots still effectively covers various aspects of the weather dataset, providing audience with multiple perspectives to explore and understand the data's nuances.

Choice of colours: I used blue colour for bars, dots and lines as it's perceived as calm colour, it's easy on eyes and can create sense of comfort for viewers, making it pleasant to look at for extended periods. In cases I had more options to pick at same time, I had to bring in other colours as well, for better visibility, but blue was main colour in dashboard plots. Also, I was adding Grid Lines, as per instructions from previous assignment again, for better visibility, as well as using white background.

**Machine Learning for Business Assignment**

**Time Series Analysis**

Times series analysis involves analysing data points collected or recorded at specific time intervals. It is used for forecasting, understanding underlying patterns and making predictions based on the time-dependent data. In machine learning, various models like ARMA, ARIMA, SARIMA etc. are applied to time series for prediction.

Time series used on my specific dataset:

Temporal Patterns and Trends: Time series helps in uncovering patterns, trends, and seasonality within the data. It allows us to understand how weather attributes like temperature, humidity, wind characteristics, etc. change over time.

Forecasting and Predictions: Time series models are used to forecast future values based on historical patterns. For instance, predicting future temperature trends or the likelihood of certain weather conditions occurring based on past data.

Detecting Anomalies or Outliers: Analysing time series data helps in identifying unusual or anomalous events. Sudden spikes in temperature, unexpected weather patterns, or irregularities can be detected through time series analysis.

**ADF Test**

The ADF test is used to determine whether a given time series is stationary or not. Stationarity is a crucial assumption in time series analysis, and the ADF test helps in confirming or rejecting the presence of a unit root in the data. A unit root suggests non-stationarity.

A white screen with red and blue text

Description automatically generated

With the ADF test indicating that the 'windspeed' time series is stationary, it suggests that there might not be a need for differencing or other transformations to make the data stationary before proceeding with time series modelling or analysis. This stationary characteristic simplifies the modelling process and makes it more suitable for various time series modelling techniques without requiring additional adjustments for stationarity.

A screen shot of a graph

Description automatically generated

I decided to go with ARIMA based on observed characteristics of the windspeed time series data, particularly from ACF and PACF plot, along with some previous assumptions:

* ACF and PACF Patterns: The significant spikes observed in the PACF plot at the first two lags and the ACF plot indicating a slow decay in autocorrelation suggest a potential autoregressive (AR) behaviour at these lags. This pattern aligns with the characteristics that ARIMA models aim to capture.
* Stationarity: As per the earlier analysis using the Augmented Dickey-Fuller (ADF) test, the 'windspeed' series was found to be stationary. ARIMA models are applicable to stationary time series data, and stationary series tend to exhibit consistent behaviour over time, making them suitable for ARIMA modelling.
* Time Dependency: ARIMA models are designed to capture time dependencies in data, especially those with observed autocorrelation patterns (as seen in ACF and PACF plots).
* Forecasting: Since assignment requires one-step-ahead forecasting, ARIMA is used in short forecasting.

After deciding which model to do, I was looking for Best parameters. As dataset is stationary, I didn’t need to find it for “d”, as it’s 0. The best result was [4,0,4] and I used [1, 0, 4] as well.

A screenshot of a computer code

Description automatically generated

Once I had best parameters, I split dataset in 80/20 ratio, did ADF test and ACF/PACF on train data as well.  
  
A screenshot of a computer

Description automatically generated

**ARIMA**

I fitted ARIMA model and did further investigations into dataset. Then, I forecasted following value and following 10 values as well, but also, forecasted errors as well but my model wasn’t showing good scores.

A blue line on a white background

Description automatically generated

A graph with a line and a red dot

Description automatically generated

A screenshot of a computer error

Description automatically generated

The Mean Absolute Error (2.62) and Root Mean Squared Error (3.22) values suggest the model's predictive accuracy in terms of absolute and squared errors, but it still show that the model's performance might have room for improvement.

Lower values for MAE and RMSE are better, indicating smaller prediction errors. For R², a value closer to 1 is better, as it suggests the model explains a larger portion of the variability in windspeed. However, in my case, the negative R² indicates that my model might not be a great fit for this specific data, possibly due to the inadequacy of the model or the complexity of the underlying patterns in the data.

I tried to do model with different parameters, but scores were even lower.

**Text Analytics**

“Text analysis (TA) is a machine learning technique used to automatically extract valuable insights from unstructured text data. Companies use text analysis tools to quickly digest online data and documents, and transform them into actionable insights.

You can use text analysis to extract specific information, like keywords, names, or company information from thousands of emails, or categorize survey responses by sentiment and topic.” (Anon., 2023)

**Text Pre-processing**

As mentioned above, I was using different dataset for this assignment. As usual, first I had to pre-process it. As a part of pre-processing data, this is what I did:

* Converting text to lowercase
* Removing special characters
* Tokenising text
* Removing stopwords

**Text Categorisation / Sentiment Analysis**

After I prepared dataset, I could proceed with Text Categorisation / Sentiment Analysis. First, I needed to do data labelling. In order not to do it manually or import external labelled datasets, I used TextBlob to calculate sentiment polarity. Newly created column will contain the polarity scores ranging from -1 (negative) to 1 (positive), indicating the sentiment polarity of the text.

A screenshot of a computer

Description automatically generated

I also created discrete categorical labels directly from the sentiment polarity scores.

After that, I run Multinomial Naive Bayes classifier but it had low accuracy (59.61%). I tired to improve it by using Oversampling technique, using Fine- tuning Hyperparameters and using SVM model but scores improved only to 64.5%.

**Topic Modelling**

When I proceeded to Topic modelling, I used Latent Dirichlet Allocation (LDA) on a document-term matrix. I used it to identify five topics within a collection of text documents represented in the document-term matrix. It then displays the top words associated with each topic to provide insights into the main themes or topics present in the documents. Once I got results, I did visualisation on those results:

A close-up of words

Description automatically generated  
A close-up of words

Description automatically generated

A close-up of words

Description automatically generated

Later on, I also performed Topic-based Analysis to get count of tweets associated with each identified dominant topic, providing an understanding of the distribution of topics across the dataset.

A graph with blue rectangles

Description automatically generated

I performed, as well, Sentiment Analysis on topics and got results that Topics 4 and 5 predominantly exhibited a higher volume of positive sentiments compared to negative sentiments, with Topic 4 displaying the highest positive sentiment count. Conversely, Topic 4 also had a notable count of negative sentiments, showcasing a more balanced sentiment distribution.

After that, I wanted to check Temporal Trends and could notice that there was a huge spike in tweets on July 12th, 2022.

A graph showing the time of a climate change

Description automatically generated with medium confidence

**Document Summarisation**

Document summarization involves condensing a document's main points or content into a shorter representation while preserving its essence. I tried to use TextRank Algorithm, but I was facing many issues with gensim.summarization module and couldn't run it, even though I installed multiple needed libraries so I performed TF-IDF scoring, to get top 10 words for each tweet.

A screenshot of a computer screen

Description automatically generated

**Challenges**

Challenges I faced throughout this assignment were mostly related to creating dashboard and time series analysis. Due to limited word count we can use, I was leaving comments in my notebook and there you can find more detailed information and explanations about logic I followed for any decision and issues I encountered.

Word count: 2195

# **References**

Anon., 2023. *Monkey Learn.* [Online]   
Available at: https://monkeylearn.com/text-analysis/  
[Accessed 25 November 2023].

Missioner, R., 2023. *Kaggle.* [Online]   
Available at: https://www.kaggle.com/datasets/rukenmissionnier/weather-in-istanbul/data  
[Accessed 16 November 2023].

Scweet, A., 2022. *Kaggle.* [Online]   
Available at: https://www.kaggle.com/datasets/die9origephit/climate-change-tweets  
[Accessed 16 November 2023].