**CCT College Dublin**

**Assessment Cover Page**

*To be provided separately as a word doc for students to include with every submission*

|  |  |
| --- | --- |
| **Module Title:** | Programming for Data Analysis, Machine Learning for Data Analysis, Statistics for Data Analytics, Data Preparation & Visualisation |
| **Assessment Title:** | CA2 50% Integrated Assessment |
| **Lecturer Name:** | *Marina Iantorno*  *Sam Weiss*  *Muhammad Iqbal*  *David McQuaid* |
| **Student Full Name:** | Patrick O’ Connor |
| **Student Number:** | sbs23060 |
| **Assessment Due Date:** | 26/05/2023 |
| **Date of Submission:** | 26/05/2023 |

**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. |

**Github Repository:** repo:sbs23060/POC-CA2

**Word Count:** 2,770

**Project Management Framework**

The CRISP-DM project management framework was chosen for implementing this project due to its focus on the business question at hand, its flexibility and the level of supporting material available. These three factors are considered central considerations when implementing any data science project compared with alternative approaches such as SEMMA or KDD (Dåderman & Rosander 2018).

**Raw Data Gathering**

The datasets selected for this research project are chosen with the below requirements in mind:

|  |  |
| --- | --- |
| **Specific Requirements** | **Module** |
| Gather and process data that has been stored in at least two distinct formats | Programming |
| Document and evaluate a testing and optimisation strategy for your analysis     Plan and document how you ensured your code is doing what it is meant to, as well as ensuring that the code is making good use of your resources (eg computing, time etc). | Programming |
| Inferential stats for possible population values | Statistics |
| 5 different inferential statistics tests Ireland + 1 other country to compare  Parametric and non-parametric testing | Statistics |
| At least 2 models to compare and contrast | ML (Machine learning) |
| At least 2 approaches (prediction, classification etc.) with GridSearchCV | ML |
| Collect and develop a dataset based on the construction topic related to Ireland as well as other parts of the world. Perform a sentimental analysis for an appropriate construction topic (e.g., house price, availability of labour etc…) for producers and consumers point of view in Ireland. | ML |

The above requirements dictate the kind of data and data sources that need to be considered when choosing a dataset for this study. One example of this includes the requirement to gather and process data that has been stored in at least two distinct formats. To fulfil this requirement, data needs to be chosen from two separate formats and collated together in one format in order to be analysed, or a dataset can be split into two with one of the split halves transformed to an alternative format. The timeline of Github commits demonstrates the project planning associated with this research project.

Sourcing the data involved searching reputable data providers such as that of governments, banks and research agencies who provide free licences access to databases. One of the constraints in sourcing data is the format in which the data is retrievable in addition to the to the structure of the data. Searches where made by going to the reliable provider such as the Central Statistics Office (CSO) and searching/filtering within sites according to the desired topic. This approach was taken to ensure the validity of the data based on the authority and reputation of the provider of this data. Later searches took the reverse approach when seeking data of a specific country, searching for the topic and country first before finding a reliable provider. This approach was taken as the researcher had less knowledge of the bodies providing open access to info for the specific country in question. The country in question was Denmark which was chosen due to its similarity to Ireland in size, which supports comparative analysis without the scale of population size needing to be considered. The eventual providers of the datasets chosen for this research include the CSO (Anon n.d.) for Irish specific data, Eurostat (Anon n.d.) for data on the European area and the Federal Reserve Bank of St. Louis (Organization for Economic Co-operation and Development 1995) for data on the Danish construction industry.

**Preparation for Statistical Analysis**

The aim for the dataset first read into the notebook was to prepare it for statistical analysis. One of the requirements of the research was to compare a country’s data related to the construction industry with another country/ This requirement necessitated focusing on data that could be directly compared with data from another country. This involves sharing the same format, measuring the same metric, utilising the same scale of measurement as well as using the same time period in the case of time series data. For these reasons the column measuring the ‘seasonally adjusted’ volume of construction production was taken which was directly comparable to the data sourced for the construction industry of Denmark. Cleaning the data is an important activity to support the trustworthiness of the analysis which stems from the data (Xiaoou Ding, Hongzhi Wang, Genglong Li, *et al.* 2022). Rows containing null values were removed. This is a form of listwise deletion called complete case analysis (CCA) which is when a row is removed from a dataset if it contains any missing values (Yan, Lang n.d.). This contrasts with variable deletion (DV) which removes a column if more than x% of values are missing from that column. On this occasion it wasn’t required as this method usually requires a threshold of 30% of values to be missing from a column before it’s removed, which wasn’t present in this dataset.

**Descriptive Analysis**

The data for Ireland appeared to skew to the left, with a min of 76, median of 135 and max of 347. The average of 173 was skewed due to outliers in the data. In general the data was skewed to the left which can be seen in the boxplot outlined below:

A picture containing text, line, screenshot, diagram

Description automatically generated

A boxplot was chosen as a visualisation method due to its compatibility with skewed data. The initial analysis performed using the ‘describe’ function provided the initial insight that the data could be skewed.

Denmark’s production volume data contained less variability, which was demonstrated through the proximity between its mean (102) and median (98) values. The standard deviation highlighted this difference further with a value of 12 for Denmark vs. a value of 87 for Ireland. A boxplot was used to highlight the data distribution for Denmark also to compare its distribution with Ireland. It’s evident from the boxplot that Denmark’s distribution also skews to the left but to a lesser degree than the Ireland data:

A picture containing text, line, diagram, screenshot

Description automatically generated

A line plot was then devised using Matplotlib to compare Ireland and Denmark production data over a period of time. This method was chosen due to its suitability to time series data, displaying any trends which may occur in the data. It was clear from graph that both countries follow a similar trend, however Ireland experiences excessive variability compared with Denmark, which is also displayed in the scale on the left of the graph representing Ireland values. Generally, there is a one or two year gap between the peaks and troughs of either country.

**Inferential Statistics**

Multiple tests were carried out in comparing Ireland and Denmark. The first test consisted of a Sign test which was performed on Demark and Ireland individually before being applied together. The null hypothesis for the individual Sign tests stated the production volume index did not change each year for the countries overall measuring declines and rises year over year. The null hypothesis test rejected for Ireland and accepted for Denmark with a p-value of 0.28 and 0.016 respectively. The final Sign test was a comparative of Ireland and Denmark, with the null hypothesis stating both countries share the same index. This was done to assess the similarity or dissimilarity of the countries index scores over time. The null hypothesis was rejected with a p-value of 0.016. Z-testing was performed on Ireland and Denmark individually to estimate population mean ranges (despite the mean already being provided) as well as T-testing comparing the means of two countries.

**Outcome**

The analysis displays that despite the fact Ireland experiences higher variability in production volume, it is Denmark is growing at a faster rate in most recent years. This is evidenced through the Sign tests showing Denmark’s volume is changing year on year as well as the line plot which demonstrates the recent growth rate of Denmark. Ireland’s high activity early in the time series before the economic recession of 2007/2009 was not followed by the same rise in activity which was experienced by Demark which may explain the housing supply constraints in the Irish market. This research could be expanded further by assessing construction costs over the same time period between the two countries to compare their performance in the construction industry more accurately.

**Machine Learning Model Selection**

Due to the size of the datasets used in the previous section, it was deemed appropriate to seek a larger dataset for the machine learning in order to improve its performance (Maciej Śliwowski, Matthieu Martin, Antoine Souloumiac, *et al.* 2023). House price data for the European area was selected due to its fertility in comparing construction data across multiple countries in the same region of the world, who are likely to be equally impacted by global supply chain shocks and other factors which affect construction industry performance. Data was cleaned removing null values and placing the countries as labels to allow for country comparisons. The exploratory data analysis showed Turkey (Türkiye) displaying disproportionate variability through its standard deviation. This precipitated the isolation of Turkey in the analysis to come and assess the effect the change in house prices of the other countries in the dataset have on the house prices in Turkey. Firstly a linear regression was performed, to assess the numerical properties of the house prices. It was deemed a prediction approach would be most suitable here to determine the relationship between house prices in other European countries and the house prices in Turkey. This analysis was explored further through the use of ridge regression, in order to generalise the data further and maximise test performance due to the small nature of the data.

For a complete view a backwards facing approach was also employed through classification, which sought to assess whether a quarter over quarter rise in prices in other countries results in a rise in Turkey also. Despite the lack of categorical data, it was possible to generate this in a binary form through the addition of new columns per country. A random forest classifier was selected as the classification model to pursue as its ability to group multiple decision tree outcomes into an overall outcome strengthens its reliability as well as the facility of adjusting the number of estimators involved in the random forest model as required.

**Machine Learning Results**

The linear regression test produced high accuracy results with a training accuracy score of 1.0 and testing accuracy of 0.97. These results indicate a strong relationship between on the predictability of house prices in Turkey based on prices in the rest of Europe. Although the high nature of the results provokes suspicion also that the dataset may be too small in nature to provide reliable results. Ridge regression was employed in an effort to generalise the model, however accuracy measures remained high at 0.99 for training and 0.97 for testing. A change did occur in the error measurements for both root means squared error and mean absolute percentage error. The hyperparameters were tuned further using GridSearchCV, however on the next iteration the model indicators did not change to any substantial degree. An iteration was then made with a smaller hyperparameter value entered which sabotaged the accuracy scores, while having a positive effect on the error measures. Scatter plots were selected to display the results of these linear models, due to the suitability of scatter plots in displaying correlation and proximity to a defined reference such as a residual line. One example of this plot is displayed below:

**Ridge Regression Accuracy**

A picture containing text, screenshot, plot, line

Description automatically generated

The outcome of the random forest classification displayed an accuracy of 1.0 in predicting whether Turkey experiences a quarter over quarter increase in house prices based off the same data for the other countries in the region. This can explain the variability of Turkey’s index when there’s a strong dependence on other countries markets in determining Turkey’s market prices. Similar to the linear and ridge regression, the accuracy score introduces suspicion as to the reliability of the model, which may benefit from a larger amount of data, which can be fulfilled through introducing quarters from earlier years into the data, going back as far as the year 2000 for example. Feature importance was sought to assess which countries could be impacting Turkey’s house prices. A barplot was chosen to display these results due to its capability in displaying comparisons in categories of data. Although Lithuania was originally observed as the highest contributor to Turkey house prices, the model appears to produce a different output each time it’s run. The example below shows Iceland as the most impactful country on Turkey:

**Most Impactful Countries on Turkey House Prices**

A picture containing text, screenshot, diagram, line

Description automatically generated

**Interactive Visualisation**

The deemed machine learning results weren’t deemed suitable to be displayed in an interactive visualisation due to the nature of the data values. To display feature importance graphically was deemed not to take advantage of the geographic visual capabilities available, as more informative information can be shared to provide more relevant information to the reader. The regression results did not provide a data individually for each country which also limited its use for interactive representation. It was deemed suitable to utilise the index values for European countries in a given quarter in order to isolate a column and give a geographically rounded view for that quarter. This is beneficial over covering a number of quarters but narrowing the spread of countries included when showing a comparative analysis.

A choropleth using Plotly express was selected as the means of visualisation due to its ability to capture data for a geographic area in an intuitive manner. A specific area can be focused on using the geo\_scope option which provides an appropriate scale to the visualisation. This approach also allows many data points to be displayed in a simple and clear manner which is one of the core staples of Tuft’s principles (Tufte, 2007). A country code file was merged together with the European house price data in order for the choropleth to be able to represent the location of each country accurately. The data appeared to not display in the visual despite ensuring the correct data values match the expected format and range, verifying the ‘fig.show’ call and ensuring the correct column names were included as parameters.

**Data Formats**

For the sentiment analysis element of the research it was necessary to source textual data. This was sourced via a Reddit API which provides opinion on chosen topics in the subreddits. The Python Reddit API Wrapper (PRAW) was utilised in order to simplify the interaction with the API and avoid inefficient code which can be created through raw API requests when not constructed optimally. For the sake of this research a subreddit was selected containing opinions on the construction industry in Ireland as well as another subreddit containing opinion on construction in general, including those not based in Ireland. Environmental variables were used to encrypt the credentials required for the authentication of the Reddit API. Variables were saved containing the credential values which were processed through a ‘.env’ file, without being explicitly displayed in the notebook. This is to aid security of codebases as part of best practices in companies using version control tools such as Github, which can be accessed and exploited in the hands of bad actors. The information from the API was stored as in a JSON file initially before being read in as a dataframe. The initial data was appended to a list to allow it to be serialised and readable in JSON. This exercise was performed for the data that was accessed from both APIs that were called. The two dataframes were then concatenated together for the sentiment analysis. Through the project unit testing was performed to ensure code both performed as expected both upstream and downstream in the notebook. Some code which was commented out of the execution has been left in the notebook due to its support in identifying issues and resolving them in the code.

**Sentiment Analysis**

TextBlob library was used for the sentiment analysis due to its ability to assign sentiment scores without data which is already assigned sentiment labels. A ‘get\_sentiment\_function’ was created to assign the sentiment labels to the data and was reusable across the second column examined which created efficiency in the code. The results were displayed in a stacked bar plot due to its ability to demonstrate the values for two different columns in a single bar across positive, neutral and negative points on the X axis. This is consistent with created a simplifying display of the information compared with multiple bars being used on the chart to cover each column’s data individually. The visual display highlights the dominance of neutral opinions in the textual data analysed. For further research more specific textual data can be examined and processed to focus on relevant comments which may provide a more diverse spread of sentiment versus a discussion on broad topics which may have been the case with the subreddit selections.

**Library Choices**

Environmental variables were set in dotenv when other choices were available such as setting the variables in the system environment . This approach can be straightforward but involves managing the variables outside of your codebase. Creating a configuration file is another option which involves handling the variables in a separate file. It provides more flexibility, however requires additional handling to read and parse the file. Pandas was used throughout the project for manipulating and aggregating data. This is a powerful library due to its ability to create dataframes which form the basis of preparing data for statistical analysis which can be performed by Numpy. An alternative to Pandas may be Dask, which has scalable parallel computing capabilities and seamless integration with existing Pandas code. However the overhead associated with parallelisation may not always provide justified benefit on smaller datasets, which makes it more suitable for big data.

**Bibliography**

Anon (n.d.) *BEA04 - Indices of Total Production in Building and Construction Sector (Base 2015=100)*. [Online]. Available at: https://data.cso.ie/ (Accessed: 26 May 2023a).

Anon (n.d.) *Statistics | Eurostat*. [Online]. Available at: https://ec.europa.eu/eurostat/databrowser/view/TEICP270/default/table (Accessed: 26 May 2023b).

Dåderman, A. and Rosander, S. (2018) *Evaluating Frameworks for Implementing Machine Learning in Signal Processing : A Comparative Study of CRISP-DM, SEMMA and KDD*. [Online]. Available at: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-235408 (Accessed: 7 April 2023).

Maciej Śliwowski, Matthieu Martin, Antoine Souloumiac, Pierre Blanchart, et al. (2023) Impact of dataset size and long-term ECoG-based BCI usage on deep learning decoders performance. *Frontiers in Human Neuroscience*. [Online] 17. Available at: doi:10.3389/fnhum.2023.1111645.

Organization for Economic Co-operation and Development (1995) *Production: Construction: Total Construction: Total for Denmark*. [Online]. 1 January 1995. FRED, Federal Reserve Bank of St. Louis. Available at: https://fred.stlouisfed.org/series/DNKPROCONAISMEI (Accessed: 26 May 2023).

Tufte, E.R. (2007) *The Visual Display of Quantitative Information*

Xiaoou Ding, Hongzhi Wang, Genglong Li, Haoxuan Li, et al. (2022) IoT data cleaning techniques: A survey. *Intelligent and Converged Networks*. [Online] 3 (4), 325–339. Available at: doi:10.23919/ICN.2022.0026.

Yan, Lang (n.d.) Use Case and Performance Analyses for Missing Data Imputation Methods in Big Data Analytics. *California State Polytechnic University*.