CCT College Dublin

Assessment Cover Page

Module Title:	Data Preparation & Visualisation, Machine Learning for Data Analytics, Statistics for Data Analytics, Programming for Data Analytics
Assessment Title:	Construction Industry in Ireland and Europe
Lecturer Name:	David McQuaid, Dr. Muhammad Iqbal, Marina Iantorno, Sam Weiss
Student Full Name:	Paul Ryan
Student Number:	sbs23013
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.

Abstract

A comparison of the Irish construction industry with wider Europe including predicting future trends using time series analysis.

Keywords:

Introduction

This is an analysis of the Irish construction industry, looking at different indicators, measured as index values, across a range of European countries. It involves loading a dataset from the Eurostat website, controlled by the European Union. After a small amount of preparation, the dataset was explored and visualised. As it was found to have quarterly values assigned to different countries and measurements, it was then tested using both parametric and non parametric statistical tests to compare the values of different countries, with a specific focus on Ireland.

Machine learning models will be created to attempt to predict future values using both time series and support vector regression models. These models will have their hyperparameters tuned, to deliver the best accuracy. The results of the two models will be compared. In addition, a sentiment analysis will be performed on text pulled from the r/Ireland and r/Europe subreddits from the Reddit discussion platform.

As part of the analysis, an interactive dashboard will be created to allow the user to see the Hours Worked Index for a selected Country, with a dynamic line chart showing both the actual past values as well as future predicted values.

There is also a discussion at the end of the report on testing and optimisation of the code and a comparison of two different data manipulation libraries in Python.

Materials

Overview

A .tsv file is downloaded from the Eurostat website, via API. The file is compressed using gzip so after loading it needs to be extracted using the gzip package. It is then loaded into a data frame for further modification.
In the format it is downloaded, the quarterly values are stored in separate columns. The melt function is used to combine the quarters and respective values into two columns.
There are records in the "Value" column with no value, but they are represented with a ':'. To enable calculations on this column, they are set to NaN, and the 'Value' column is then set to a data type of float, which will allow for results using decimal points such as 1.5.
As the data contains information around countries, it will likely be visualised via a map or choropleth plot. To allow for this there needs to be compatible values representing each country. The ISO3 country codes are one such format. The 'pycountry' package allows for the conversion of the existing two character country codes to the ISO3 three character country codes. These are mapped and assigned to a new column named 'Country_Codes'.
Finally two new columns are created; 'pvalue' and 'normal'. These are indicators based on the possible groupings of the values, with a p-value corresponding to each, indicating if that particular segment is normally distributed. If the values are normally distributed the 'normal' column will have 'True', if not, it will have 'False'.

Methods

Exploratory Data Analysis

Data Exploration

For the overall data set, the first 5 rows are printed using the '.head' function. Columns and the standard statistical information is then shown using the '.describe(include='all')' function. The same descriptive statistics are shown at a group wide level using the '.groupby' function in conjunction with the '.describe' function.
The rows relating to Ireland are then separated into a different dataframe named 'IE'. The same techniques are applied to this dataset, including listing the columns and observations, statistical values overall, and statistical values for grouped values within the dataset.

Visualisations

First a choropleth plot is created showing the Numbers of Persons Employed Index for each country over time. This is an interactive plot that allows you to filter by quarter.
After this the IE data frame is focused on. Histograms of the 'Value' column for each indicator and seasonal adjustment option are created, with differentiation for whether the values are part of a normal distribution or not. They are displayed using the 'FacetGrid' function which is part of the 'seaborn' visualisation library.
Then the data is divided further and instances where the data is normally distributed are visualised using histograms, split on the indicator type and seasonal adjustment option. This data is also shown using boxplots spread across indicator types, split on the seasonal adjustment option.

Statistics

T-Test

A T-Test was performed to determine if the number of persons employed index for Ireland, differed significantly from the Numbers of Persons Employed Index for Europe overall. This test was chosen as the distribution for both groups was found to be normal. As the variances between groups are unequal, this will be a Welch's T-Test. The assumptions are:

- The two samples are independent.
- The populations from which the samples are drawn follow approximately normal distributions.
- The variances of the two populations may not be equal.

The Hypothesis test is as follows:

Null Hypothesis (H_0): The mean of the Numbers of Persons Employed Index in Ireland is equal to the mean of the Numbers of Persons Employed Index in Europe.

Alternative Hypothesis (H_1): The mean of the Numbers of Persons Employed Index in Ireland differs significantly from the mean of the Numbers of Persons Employed Index in Europe.

With an alpha (Significance level) $\alpha : \ 0.05$

One - Way Anova

A One-Way Anova test was performed to understand if there were significant differences in the mean of the Numbers of Persons Employed Index for a range of European countries. The assumptions are:

- The samples are independent.
- The populations from which the samples are drawn follow approximately normal distributions.
- The populations have equal variances.

The Hypothesis test is as follows:

Null Hypothesis (H_0): The mean of the Numbers of Persons Employed Index in Ireland is equal to the mean of the Hours Worked Index in other European countries.

Alternative Hypothesis (H_1): The mean of the Numbers of Persons Employed Index in Ireland differs significantly from the mean of the Hours Worked Index in other European countries.

With an alpha (Significance level) $\alpha : \ 0.05$

Post Hoc tests were performed to determine which groups are significantly different from each other; this was done with a Tukey's Range Test.

Two - Way Anova

A Two-Way Anova was performed to understand if there were significant differences in the mean of the Numbers of Persons Employed Index for a range of European countries, and if there were differences between the seasonally and non-seasonally adjusted data. The assumptions were:

- The samples are independent.
- The populations from which the samples are drawn follow approximately normal distributions.
- The populations have equal variances.
- The observations are randomly and independently assigned to the groups.

The Hypothesis test is as follows:

Null Hypothesis (H_0): The mean of the Numbers of Persons Employed Index in Ireland is equal to the mean of the Hours Worked Index in other European countries, and there is no difference between seasonally adjusted and non-seasonally adjusted values.

Alternative Hypothesis (H_1): At least one of the means of the Numbers of Persons Employed Index in Ireland and other European countries is significantly different, and there is a difference between seasonally adjusted and non-seasonally adjusted values.

With an alpha (Significance level) $\alpha : \ 0.05$

Post-Hoc tests were performed to determine any interaction and which groups are significantly different from each other.

Wilcoxon Signed-Rank Test

A Wilcoxon Signed-Rank Test was performed to understand if there was a significant difference in the distribution of the Hours Worked Index for Ireland, when compared to Europe overall. The Wilcoxon Signed-Rank Test is suitable as the data is not normally distributed, therefore necessitating a non-parametric test, and this "is a nonparametric test for the one-sample location problem" (Lovric, 2011, p.1658). The assumptions were:

- The paired observations are independent of each other.
- The population distributions of the paired observations are identical.
- The data is not normally distributed

The Hypothesis test is as follows:

Null Hypothesis (H_0): The distribution of the Hours Worked Index values in Ireland is the same as the distribution of the Hours Worked Index values in Europe.

Alternative Hypothesis (H_1): The distribution of the Hours Worked Index values in Ireland differs significantly from the distribution of the Hours Worked Index values in Europe.

With an alpha (Significance level) $\alpha : \ 0.05$

Kruskall Wallis

A Kruskall Wallis test was performed to understand if there were significant differences in the distribution of the Hours Worked Index for a range of European countries. The assumptions are:

- The samples from each country are independent of each other.
- The observations within each country are independent and identically distributed.

The Hypothesis Test is as follows:

Null Hypothesis (H_0): The distributions of the Hours Worked Index values are the same across European countries.

Alternative Hypothesis (H_1): The distributions of the Hours Worked Index values differ significantly across European countries.

With an alpha (Significance level) $\alpha : \ 0.05$

Post Hoc tests were also performed to further examine pairwise comparisons between the European countries, this was accomplished using Dunn's test.

Machine Learning - Sentiment Analysis

Model Overview / Data Processing

Sentiment Analysis is a machine learning algorithm that allows for the scoring of words or groups of words based on whether the meaning is positive or negative. This is achieved through "processing text, also known as natural language processing (NLP)" (Müller and Guido, 2017, p.355).
The data is gathered from Reddit, a social messaging platform where users can post news and opinions and have discussions in comments sections. Using the PRAW API wrapper and an access point set up via a Reddit profile, a for loop is used to collect posts from the 'Ireland' community subreddit, and the 'Europe' community subreddit, using 'house prices' as the search query. These are stored as a list and then converted to a dataframe.

Sentiment Analysis

Within the dataframe a new column named 'polarity' is created. A lambda function is used to apply the .sentiment.polarity function from the TextBlob package to the text and store the results in the 'polarity' column. The results are then printed and compared to see which area is more or less positive in regards to house prices.

Machine Learning - Time Series Analysis

Model Overview / Data Processing

A Time Series analysis is a problem where "observations are collected at regular time intervals and there are correlations among successive observations" (University of Cambridge, p. iii). The data concerning the construction industry is spread across multiple quarters and can be approached as a time series problem. The data is first divided to focus on the Hours Worked Index in Ireland. Multiplicative and Additive Decomposition are performed and plotted and an Augmented Dickey-Fuller test is performed to check if the data is stationary or not.

Time Series Analysis

The data is split into training and test sets and a SARIMAX model is fit on the training set. The Mean Absolute error (MAE) and Root Mean Square error (RMSE) are calculated from the test set and printed as well as the predictions being plotted against the actual values.

Hyperparameter Tuning

To attempt to improve the accuracy of the model, the hyperparameters are tuned. For time series analysis these are the p, d and q parameters, as well as the seasonal order parameters. The different possible options are stored in lists and then using a for loop they are all tested

by fitting each combination on the training data, with the best results being stored and then printed. These are then applied to a SARIMAX model and the MAE and RSME are printed with the predicted values plotted against the actual values.

Machine Learning - Support Vector Regression

Model Overview / Data Processing

Support Vector Regression (SVR) is a regression technique that uses similar principles to Support Vector Machines. It trains using “a symmetrical loss function which equally penalizes high and low misestimates” (Awad and Rahul Khanna, 2015, p.67). To prepare the data for the SVR, the values are scaled using the StandardScaler function. The feature and target columns are selected, and one-hot encoding is used to transform the categorical columns to numerical representations.

Support Vector Regression

The data is split into training and testing sets. The SVR model is created and fitted on the training set. The model is then used to make predictions on the test set. The predicted results are compared against the actual results and the MSE and R-squared values are printed.

Hyperparameter Tuning

The hyperparameters are tuned in an effort to improve the accuracy of the model. This is done using the GridSearchCV package to run through a grid of possible parameters, testing each combination on the model to return the best parameters for the models accuracy. These hyperparameters are then used on a new SVR model, and the MSE and R-Squared score are again printed. The actual and predicted values are then plotted against a diagonal line to display the accuracy.

Interactive Dashboard

Overview / Data Processing

To allow users to easily access the results of the time series analysis, specifically the projections of future performance, an interactive dashboard is created using the ‘Dash’ and ‘plotly.express’ packages. These allow for different plots to be created and placed on a dashboard which is accessed via a web browser. The predicted data is created using a for loop where a dataframe containing the actual and predicted values for each country are stored. Data is also added to give better functionality, this includes full country names, mapped to the country codes, and the maximum and minimum values and their respective quarters for each country. In addition, longitude and latitude information is loaded via the ‘geopy’ package for the corresponding country codes.

Dashboard

A line chart showing the actual values for the past three years and the predicted values for the next year is created, as well as a choropleth map plot. The Dash package allows for the layout of the dashboard, including the title, a filter using radio style selection buttons, and then the map and line plots. An ‘update_charts’ function is created to update the title and plots based on which country is selected, as well as what information is displayed. Finally the command to run the dashboard allows it to be created and displayed.

Results

Exploratory Data Analysis

Data Exploration

The data frame contained 11 columns with features representing time (‘Quarter’), value (‘Value’), country (‘geo\\TIME_PERIOD’) and whether or not the data was seasonally adjusted (‘s_adj’), as well as a ‘indic’ column which denoted the index being measured (‘Hours Worked’, ‘ Number of Persons Employed’, etc.).
The separated IE dataframe contained the same columns, 1730 rows of data with 804 null values. Examining the data, the majority of these null values occurred pre 1999. Any analysis taken should focus on a timeframe after this period.

Visualisation

A choropleth map with values animated by quarter shows that for the majority of countries, 2000 or later was the period when values were recorded, again this would lead to focusing analysis on this period.

Histograms of the IE dataframe faceted by the Index type show that the ‘Number of Persons Employed Index’ is normally distributed, while the other four indexes are non-normally distributed. When comparing Ireland to other countries, this should be taken into account when deciding whether to use parametric or non-parametric tests.

Box plots show that the ‘Building Permits Index’ (IS-PEI) has the largest range of values, and that the distribution of seasonally and non-seasonally adjusted data is similar for all indexes.

Statistics

T-Test

Ireland had a sample mean of 129.13 and a standard deviation of 38.59.
Europe had a sample mean of 111.118 and a standard deviation of 8.26.
The t-statistic was calculated by comparing the means and came to -3.96.
The degrees of freedom were calculated using the Welch-Satterwaite and were 81.86
The p-value associated with the t-statistic was 0.0002. As this is less than the alpha of 0.5, we reject the null hypothesis.
The difference between the means is statistically significant.

One - Way Anova

Ireland had a sample mean of 129.52 with a standard deviation of 34.91

Denmark had a sample mean of 118.43 with a standard deviation of 17.15
Europe had a sample mean of 114.7 with a standard deviation of 12.52

The F-statistic calculated from comparing the means came to 11.20, with degrees of freedom of 2.
The p-value associated with the F-statistic was found to be 0.000. Since the p-value is less than the alpha of 0.05, we reject the null hypothesis.
The differences between the means is statistically significant.

Tukeys Range post hoc test showed a p-value of 0.49 between Denmark and Europe, the difference in means is not statistically significant. Ireland is the group with a different mean, with a p-value of less than 0.05 when compared to either group.

Two - Way Anova

The F-statistic was calculated by comparing the means of the different groups and factors.

F-statistic for the main effect of seasonality was 0.00
F-statistic for the main effect of country was 86.81
F-statistic for the interaction between seasonality and country was 0.00

The degrees of freedom for the main effect of seasonality was 1
The degrees of freedom for the main effect of country was 2
The degrees of freedom for the interaction effect was 1

The p-value for the main effect of seasonality is 0.99, and as it is larger than the alpha of 0.05, we fail to reject the null hypothesis that there is no statistically significant difference in means.
The p-value for the main effect of countries is 0.00, and as it is smaller than the alpha of 0.05, we reject the null hypothesis and conclude that there is a statistically significant difference in means.
The p-value for the interaction effect between seasonality and country is 1, and as it is larger than the alpha of 0.05, we fail to reject the null hypothesis that there is no interaction effect between factors.

Performing the post hoc test shows that of the country groups, Ireland is the one that has a statistically significant difference in means when compared to the other two countries, with Poland and Slovakia having a p-value of 0.139, which, being larger than the alpha of 0.05, leads to failing to reject the null hypothesis that there is no statistically significant difference in their means.

From the two anova tests performed, Ireland was the group causing the rejection of the null hypothesis with a difference in its mean from the other groups tested.

Wilcoxon Signed-Rank Test

The test statistic for this test was calculated by looking at the difference between the two groups. In this case the t-statistic came to 1300.50.
The p-value associated with the t-statistic was 0.0062, which is smaller than the alpha of 0.05, therefore leading to the rejection of the null hypothesis, which was that there was no statistically significant difference between the distributions.

Kruskall Wallis

The test statistic for this test was calculated by looking at the differences between the distributions of the groups. In this case the t-statistic came to 8.93.
The p-value associated with the t-statistic was 0.0115, which is smaller than the alpha of 0.05, therefore leading to the rejection of the null hypothesis, which was that there was no statistically significant difference between the distributions.

The Dunn's post hoc tests showed that there were significant differences in the distribution of Finland and Ireland.

Machine Learning - Sentiment Analysis

Sentiment Analysis Results

The sentiment analysis showed a polarity of 0.0481 for the Ireland subreddit and 0.0214 for Europe. Both areas were positive overall about house prices, however only by a small measure, they were close to neutral. Ireland was more positive than Europe, but again only by a small measure. Ireland did have the most positive post with a polarity of 0.9, but it also contained the most negative sentiment with a polarity of -0.3125.

Machine Learning - Time Series Analysis

Initial Model Results

The SARIMAX model returned a result of a MAE of 20.51 and an RMSE of 24.31. Looking at the plot of the predicted versus the actual values, it is clear that the accuracy is weak.

Hyperparameter Tuning Results

The hyperparameter tuning results showed that a pdq value of 0, 1, 0 respectively was the best selection, with seasonal parameters of 0, 1, 0, 4.
When applied to the SARIMAX model the results are a MAE of 17.13 and an RMSE 20.85. The new plot shows the predicted values corresponding more closely to the actual values, although there are still variances.

Future predictions are made and plotted for the next four quarters. This same process is applied to each country in the dataframe for the ‘Hours Worked Index’

Machine Learning - Support Vector Regression

Initial Model Results

The initial MSE is 0.72 with an R-squared value of 0.29. When plotted it can be seen that there is a block of correctly predicted data, but this is skewed by larger, unpredictable actual value outliers.

Hyperparameter Tuning Results

After a long tuning process, the best hyperparameters were chosen as a ‘C’ of 100, ‘gamma’ of 0.1 and a linear ‘kernal’. When applied to a new SARIMAX model it resulted in a MSE of 0.67 and an R-squared value of 0.33. Both were improved from the previous model but only slightly. The plotted values shows a better range of predicted values, which align more with the actual values.

Even after hyperparameter tuning the results were not optimal, in subsequent analysis a different model might be used.

Interactive Dashboard

The resulting dashboard shows a map side by side with a line plot which has two different styles of line, one for actual values and one for predicted values. Above the plots there are single select buttons with full country names.
When a country is selected the map highlights and focuses on the country, and the line plot shows the actual and predicted values for the Hours Worked Index of the selected country. Hovering over the country gives the name as well as the minimum and maximum values with the corresponding quarters. Hovering over the line gives the corresponding Quarter and Value.

Discussion

Testing & Optimisation

Data Library Comparison

Processing

T

Aggregation

T

Conclusion

References