**CCT College Dublin**

**Assessment Cover Page**

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

|  |  |
| --- | --- |
| **Module Title:** | Data Preparation & Visualisation - Machine Learning – Programming – Statistics for Data Analytics |
| **Assessment Title:** | Construction industry in Ireland and Comparison with Other Countries |
| **Lecturer Name:** | David McQuaid, Dr. Muhammad Iqbal, Sam Weiss, John O'Sullivan |
| **Student Full Name:** | Burak Gulecyuz |
| **Student Number:** | 2023069 |
| **Assessment Due Date:** |  |
| **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. |

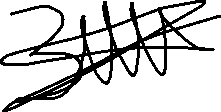


Table of Contents

Summary............................................................................3

Introduction........................................................................4

Github Link………………................................................5

Data Access and Licensing................................................5

Data Processing..................................................................6

Visualization and Statistical Analysis................................6

Hypothesis Tests.................................................................8

One-Sample t-test...............................................................10

One-Way ANOVA.............................................................11

Wilcoxon Signed-Rank Tests.............................................11

Chi-Square Test..................................................................11

The difficulties I encountered(statistic)..............................11

Machine Learning...............................................................12

Regression Analysis............................................................13

Cross-validation..................................................................13

PCA.....................................................................................14

Random Forest Regressor...................................................15

Comparison of Results of Model........................................15

Sentiment Analysis.............................................................17

Global Sentiment Analysis.................................................17

Ireland Sentiment Analysis.................................................17

Summary of Machine Learning..........................................18

Alternative libraries and techniques...................................18

Summarize and Conclusion................................................19

Construction industry in Ireland and Comparison with Other Countries

Summary

The assigned task focused on examining the contribution of the construction industry to the economies of different countries. In addition to comparing the data, two separate variables were incorporated: workforce and unemployment rate, to analyze their impact on the construction sector. To enhance the analysis of our dataset, we performed data cleaning and manipulation, followed by visualizations to gain a better understanding and interpretation of the data. Furthermore, statistical analysis methods and hypothesis testing, as specified in the tasks, were applied, and the results were obtained. In the machine learning section, appropriate models were selected for our dataset, along with an explanation of why they were chosen. Finally, the desired tasks in the field of machine learning were implemented.

Introduction

Abstract: The construction sector plays a pivotal role in driving economic growth and development. This article explores the impact of the construction industry on national economies, focusing on workforce dynamics and the unemployment rate. A skilled workforce is crucial for efficient construction operations, while the unemployment rate influences labor competition and costs. Understanding these dynamics is essential for optimizing the construction sector's potential for economic progress.

Introduction: The construction sector significantly contributes to national economies through employment generation, income redistribution, and infrastructure development. Workforce dynamics and the unemployment rate are critical factors influencing the sector's performance. Analysing these factors provides valuable insights into the impact of the construction sector on economies.

Workforce Dynamics: A skilled workforce is vital for the construction industry's efficient operation. Sufficient skilled labor ensures timely project completion and drives economic growth. Countries with well-trained construction-related trades have a competitive advantage. Conversely, labor shortages can lead to delays and increased costs.

Unemployment Rate and the Construction Sector: The unemployment rate affects the construction industry by influencing labor competition and costs. High unemployment expands the labor pool, intensifying job competition and potentially affecting wages. Government initiatives targeting unemployment stimulate the sector by increasing demand for construction services.

Economic Contribution: The construction sector contributes to economic growth through employment generation, income growth, and infrastructure development. It reduces unemployment rates and stimulates economic activity. Investments in construction projects benefit workers, suppliers, and service providers, driving consumer spending. Infrastructure development attracts investments and enhances productivity.

Conclusion: The construction sector has a significant impact on national economies. Analyzing workforce dynamics and the unemployment rate helps optimize its potential. Policymakers and industry stakeholders can use these insights to devise strategies for sustainable economic growth and development. Maximizing the construction sector's contributions enhances employment, income distribution, and infrastructure development, driving economic progress.

Data Access and Licensing

The World Bank Group makes data publicly available according to [open data standards](http://opendefinition.org/) and licenses datasets under the [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0) (CC-BY 4.0). Many datasets are available under other licenses. They are labeled accordingly, and when they are accessed by users, users agree to comply with all of the terms of the respective licenses, as explained below.

Creative Commons Attribution 4.0 (CC-BY 4.0)

The [Creative Commons Attribution 4.0 International license](https://creativecommons.org/licenses/by/4.0) allows users to copy, modify and distribute data in any format for any purpose, including commercial use. Users are only obligated to give appropriate credit (attribution) and indicate if they have made any changes, including translations. CC-BY 4.0, with the additional terms below, is the default license for all Datasets produced by the World Bank itself and distributed as [open data](http://opendefinition.org/).

Open Database License (ODbL)

The [Open Database License](https://opendatacommons.org/licenses/odbl/summary) allows users to copy, modify and distribute data, including for commercial purposes. Users are obligated to give appropriate credit (attribution), and to provide any redistributions of data under the same license and in open formats. The World Bank distributes data under this license when required to do so by the original data provider or by the conditions of a data partnership.

Microdata Research License

Microdata are unit-level data obtained from sample surveys, censuses, and administrative systems, and are primarily managed through the [World Bank’s Microdata Library](http://microdata.worldbank.org/). While a small percentage of microdata datasets are available under a Creative Commons license, and some are available under other licenses, most are distributed under the [World Bank Microdata Research License](http://microdata.worldbank.org/index.php/terms-of-use), which is designed to protect the privacy of individuals that have provided data, and ensure that use conforms to the terms under which the data were originally collected.

Reddit License

Reddit grants you, subject to your complete and ongoing compliance with the Data API Terms, a non-exclusive, non-transferable, non-sublicensable, and revocable license solely to access and use the Data APIs as expressly permitted under and in accordance with the Data API Terms. All rights not expressly granted to you are reserved by Reddit.\

Github link:

<https://github.com/burakgulecyuz/assignment/tree/87e3fd48770c5cb1e18dd443c3b712bb2107168a>

Data Processing

Using the datasets "Industry\_including\_construction\_value\_added\_ of\_GDP.csv," "Labor\_force\_total.csv," and "Unemployment\_total \_of\_total\_labor\_force.csv," we aim to examine the influence of the labor force and unemployment rates, as well as the contribution of the construction sector, on the annual income of countries.

Once we gained a general understanding of the data by using the. head and .info commands to process our imported datasets, we proceeded to transform the data into a desirable format for examination, specifically into a tabular structure, by employing relevant functions and necessary operations.

Visualization and Statistical Analysis.

We have data on the "Contribution of the construction sector to the Gross Domestic Product (GDP) %," "Total labor force," and "Unemployment, total (% of total labor force)." In order to analyze the "Contribution of the construction sector to the GDP %" in countries, it is necessary to consider the other values of the countries and examine them logically. Here, we will focus on the total labor force of countries and select two countries that are similar to Ireland. We will then display them on a graph.

Our topic is countries, and since we have data on over 60 countries, it becomes quite challenging to examine them on a line graph or bar graph. Therefore, we will use a Choropleth map to visualize the labor force data of countries over the years.

We specifically selected the "matter" color\_continuous\_scale parameter to effectively visualize the differences between countries. Additionally, we set the range\_color parameter to filter the data between 1 million and 10 million to easily demonstrate the proximity of other countries' total labor force to Ireland. Currently, it is not essential for us to compare Ireland with a country that has a labor force of more than 10 million.  
metin, harita içeren bir resim

Açıklama otomatik olarak oluşturuldu

[..\OneDrive\Desktop\assignment2\Labor\_Force.html](file:///C:\Users\burak\OneDrive\Desktop\assignment2\Labor_Force.html)(Click display to interactive graph)

Now, we clearly observed two countries, Finland and Croatia, which closely resemble the labor force changes in Ireland over the years. To demonstrate this effectively, we will proceed with this step.

When we run the code to identify the countries that are closest to Ireland in terms of the labor force, we obtain the following result: ['Croatia', 'New Zealand', 'Lithuania', 'Costa Rica', 'Singapore', 'Norway', 'Albania', 'Finland'].

As we can see, Finland and Croatia are the two countries that resemble the labor force changes in Ireland. Now, let's focus on these two countries by displaying them on a map of Europe, highlighting only Finland and Croatia.

[metin, ekran görüntüsü, harita içeren bir resim

Açıklama otomatik olarak oluşturuldu](file:///C:\Users\burak\OneDrive\Desktop\assignment2\Labor_Force_by_cont.html)

[..\OneDrive\Desktop\assignment2\Labor\_Force\_by\_cont.html](file:///C:\Users\burak\OneDrive\Desktop\assignment2\Labor_Force_by_cont.html)(Click display to interactive graph)

Now, to have a clearer view of the values for the "Contribution of the construction sector to the GDP %," "Total labor force," and "Unemployment, total (% of total labor force)" for these three countries (Ireland, Finland, and Croatia), we will combine their data into the same table. By doing so, we can observe their trends on a line plot.

metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, çizgi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram, çizgi içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, çizgi, öykü gelişim çizgisi; kumpas; grafiğini çıkarma, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

[..\OneDrive\Desktop\assignment2\comparison\_1\_countries.html](file:///C:\Users\burak\OneDrive\Desktop\assignment2\comparison_1_countries.html)

[..\OneDrive\Desktop\assignment2\comparison\_2\_countries.html](file:///C:\Users\burak\OneDrive\Desktop\assignment2\comparison_2_countries.html)

[..\OneDrive\Desktop\assignment2\comparison\_3\_countries.html](file:///C:\Users\burak\OneDrive\Desktop\assignment2\comparison_3_countries.html)

"Based on the provided graphs, we can observe certain trends and draw comparisons between Ireland, Finland, and Croatia during the period of 2014 to 2015. Notably, the contribution of the industry sector to the income of Ireland experienced a significant increase compared to Finland and Croatia. Additionally, it is important to analyse why certain trends have shown growth while others have declined, and how labor force and unemployment rates can potentially influence these patterns. It is also crucial to acknowledge that there may be other factors at play which we currently lack data for but could have influenced these observed trends."

metin, ekran görüntüsü, paralel, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

These Results delve into the complex analysis of the correlation matrix, focusing primarily on the interactions among the percentage contribution of the construction sector to the Gross Domestic Product (GDP), total workforce, and total unemployment rate.

In conclusion, in this correlation heatmap and matrix, we observed the relationships among the percentage contribution of the construction sector to the GDP, the total workforce, and the total unemployment rate. We noticed a positive relationship between the percentage contribution of the construction sector to the GDP and the total workforce, meaning that as the construction sector grows, the workforce generally tends to increase as well.

Hypothesis Tests

Most of the time, with the help of the information obtained from the sample, the population trying to reach a decision about the parameters. For example, a new education systemit can be decided whether it is different from the old one. The important point here is that whether the differences are due to sampling errors that are the result of random selection or is to determine that there is indeed a change. Some tests are carried out to determine whether these differences are statistically significant (significant).decision is made as a result. Populations are defined by the probability distributions they show. These dispersions are known Otherwise, the decisions to be made about the populations become final. But the population their distribution is generally unknown. Therefore, it is difficult to make such decisions. Another way to learn about populations is by sampling. a suitable with the help of chance samples selected in this way, parameters are estimated. Based on certain assumptions, this populations in response to a certain risk with the help of estimates or Various decisions are made about the distributions. While making these decisions, a guess is made. Or a certain assumption is made about the subject. come true or Such assumptions, which are claimed not to be realized, are called HYPOTHESES.A hypothesis is an attempt to test its accuracy with a research or experiment. Called predictions. Hypothesis tests were drawn from this sample with a sample mean. Whether the difference around the mean we think is significant (i.e.It is the tests that allow us to investigate whether there is a significant difference. If we are interested in the difference between the averages of the two populations; of these by making hypothesis tests of the difference between the means of the samples drawn, We can find out if it's true. Statistical hypotheses put forward regarding population parameters and special propositions whose validity can be investigated according to the laws of probability. The difference between statistical hypotheses and other hypotheses is that the hypothesis is divided by a frequency. That it is related. For example, “the average life of batteries carrying a certain brand is 2.5 hours. We propose a hypothesis when we assert that This means a normal the arithmetic mean of the distribution is equal to 2.5 hours. A hypothesis is either true or false. Randomly from the population to investigate this the units in a selected sample are examined and based on this sample, the hypothesis is determined. A decision is made as to whether it is valid or not. From sample statistics the process of demonstrating whether a hypothesis is valid or not It is called statistical hypothesis testing or hypothesis testing. (CENGİZ, 2018)

Confidence Interval

Confidence interval is a kind of interval estimation for a population parameter in statistics and is an inferential statistical solution tool. Instead of estimating a population parameter value with a single number, there is a range of two (lower and upper limit) numbers that can cover this parameter value. Thus, confidence intervals indicate how reliable a prediction is. (Anon., 2018)

Confidence Interval for Contribution of the construction sector to the GBD %:

25.20930104101185 - 27.77251714080634

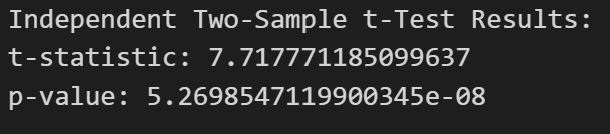
Confidence Interval for Labor Force, total:

2165916.279814146 - 2343784.629276763

Contribution of the construction sector to GDP: The contribution of the construction sector to the economy is between 25.21% and 27.77% and

Total workforce: Total workforce values ​​are also between 2,165,916.28 and 2,343,784.63. We can say that these estimates are correct within the 95% confidence interval.

Independent Two-Sample t-test:



The Independent Two-Sample t-test is a test used to see if there is a significant difference between the two groups. Here, we reject H0 by stating that there is a significant difference between the contribution of the construction sector to GDP between Ireland and Croatia, and that the p value is less than 0.5, and we can say that Ireland's contribution is higher than that of Croatia, since the t statistic is positive.

One-Sample t-test:

metin, yazı tipi, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu

P value is greater than 0.5, so the result is not statistically significant. H0 hypothesis is accepted, showing that there is no significant difference with the expected 30% average of the Irish construction industry's contribution to GDP.

One-Way ANOVA:

metin, yazı tipi, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Since our P value is well below the 0.5 value, it is said that there is a significant difference between the 3 countries, further analysis is required to tell which groups there is a significant difference between Ireland, Croatia and Finland.

Wilcoxon Signed-Rank Tests:

metin, yazı tipi, ekran görüntüsü, tipografi içeren bir resim

Açıklama otomatik olarak oluşturuldu

Since our P value is well below 0.5, H0 is rejected and we say that there is a significant difference between Ireland and Croatia according to this test, but since the Wilcoxon test does not give information about the size of the difference between the two paired groups, we can only say that there is a significant difference.

Chi-Square Test:

The chi-square test is commonly used to test independence or association between two categorical variables. In this case, the data focuses on the continuous variable "contribution of the construction sector to the GDP (%)." Therefore, we cannot apply the chi-square test directly in here. However, if we categorize our data, we can apply the chi-square test.

In order to apply the chi-square test, we will create 3 categories according to the contribution of the countries to the annual income of the construction sector, separate them as low, medium and high contribution levels, and then do the chi-square test between Ireland and Croatia.

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

The chi-square value of 30.0 shows a strong relationship between the parameters and the p value is very small, indicating that the relationship we observed is unlikely to be due to chance alone. As a result, we can say that the contribution categories are not independent from each other, that is, there are important differences between Ireland and Croatia regarding the contribution level of the construction sector to annual revenues.

The difficulties I encountered in this process were that while applying the chi-square test, I had to categorize my data because my data set was not suitable. Apart from that, it was quite challenging to find the right data, especially in the construction sector, and the data I found regularly and in detail was always paid.

Additionally, I would like to add that since our data did not follow a normal distribution, it would have been more logical to apply distribution-free tests such as the Mann-Whitney U test or Kruskal-Wallis test. However, in this case, we had to assume that our data was normally distributed.

Machine Learning

With the digitalizing world, internet of things (IoT), cloud computing, big data (Big Data), cyber-Concepts such as security, artificial intelligence and machine learning have begun to be used frequently. From the concept of big data to be mentioned, the records that could be kept in notebooks before computers started to be used are today’s It does not seem possible to preserve it in the same way in the world. Information is growing exponentially every day. The growth of the data has created the complexity of finding the necessary information from such a large amount of data. Data This is where mining mining emerges, extracting meaningful relationships from large amounts of data, is the analysis of relationships (Savaş, 2012)). In addition to extracting useful information from the data, future events can be predicted with this information. Forecasting is very important. People are looking forward to the future in line with their experiences and the training they have received. Can make predictions. However, while making decisions, people make decisions rationally with the influence of their emotions. It can be difficult to give information about the details that people overlook when there is too much data. Can lead to wrong decisions. But trained machines can evaluate all situations and can make rational decisions in a very short time by examining the data very quickly. In machine learning It is very important to be able to model the human mindset and decision-making ability (İnik, 2017)Because an evolving branch of computational algorithms for machine learning designed to mimic human intelligence (El Naqa, 2015)it can be said that a problem is modelled with an algorithm suitable for data (Atalay, 2017)Big data brings big problems due to diversity, speed and volume.

(Suthaharan, 2014). Machine learning algorithms can appropriately classify big data and solve these problems precedes it.

There are different machine learning algorithms that should be used according to the type of data. Algorithms have the features of clustering, classification, estimation and estimation according to their feature. These algorithms are supervised, unsupervised and Machines are taught with 3 strategies on reinforcement. In supervised learning, both input and output are valued, in unsupervised learning only input is valued to the machine, whereas reinforcement learning is input with the finding of the most appropriate values of the coefficients during the most appropriate output corresponding to the values provided (Ülker, 2002). Implement which machine learning algorithm should be used determined by the programmer. There are so many algorithms that can be used in machine learning, that It is important to choose the most suitable algorithm that can process the data among the algorithms. Algorithms functions grouping in terms of.

The most common machine learning algorithms k-ennear neighbour, Bayes classifier, decision trees, logistic regression, support vector machines and artificial neural networks. Big data is used in computers and internet such as education, health, commerce, banking, shopping, logistics. Appears in many fields. Each sector extracts data from their own data structures with data mining methods. He tries to improve himself by using the information. Diagnosis made by doctors in the health sector and to obtain more accurate results in treatments, to prevent human-induced errors and to help the doctor's decision. Machine learning-based decision support systems are used in order to2014). The fact that the data in the field of econometrics is very large in volume also necessitates the analysis of these data. Machine learning algorithms are used for data analysis (Akay, 2018).Twitter In social media environments such as social media, the shares of individuals are classified by machine learning algorithm and profile extraction studies are carried out (Tataroğlu, 2019).The electricity created by the emotions felt in the human brain machine learning is used in emotion analysis by examining signals (Er, 2019). from psychology while machine learning is used in every field where big data exists, from sociology, economy to politics, education use in the field is inevitable.

Machine Learning: It is a computer science branch that enables machines to produce a solution or gradually reach a solution in the face of a problem, using various data and algorithms.

Machine learning is based on artificial intelligence. According to the way artificial intelligence works, machine learning can take place in different ways. However, machine learning basically takes place in the following 3 steps.

Decision Process: The algorithm used in machine learning makes a decision by making a certain classification or prediction.

Error Function: In order to evaluate whether the prediction is correct or not, it is passed through various processes and a model is created.

Model Optimization Process: The accuracy of the model is checked. If the model is not correct or compatible, the process starts from the beginning and the process continues automatically until a correct model is reached.

Regression Analysis

Our goal is to analyse and predict the impact of independent variables on continuous dependent variables in our dataset. The dependent variables include "Labor force, total" and "Unemployment, total (% of total labor force)", while the independent variable is "Contribution of the construction sector to the GBD %”.

Regression analysis is a widely used machine learning model that helps us understand relationships between variables, make predictions, and assess their effects. It is particularly useful when dealing with continuous dependent variables, such as estimating the influence of income on housing prices or evaluating the impact of advertising campaigns on sales.

Regression analysis employs statistical techniques to mathematically express the relationships between variables in the dataset. By calculating parameters, it enables us to analyse these relationships and make predictions about future values.

Choosing regression analysis provides a strong statistical framework for analysing relationships, making predictions, and assessing the effects between variables in your dataset. It is crucial to select the appropriate regression method based on the specific requirements and objectives of your analysis.

I couldn't achieve the desired performance with linear ridge and lasso regression models, so through my research, I came across an open-source machine learning library called XGBoost and applied it. XGBoost operates based on the boosting algorithm and is particularly known for delivering high performance in regression and classification problems.

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

The first model results indicate a low MSE value and a high R^2 score in the training set, while achieving slightly higher MSE value and lower R^2 score in the test set. On the other hand, the second model results, obtained using Grid SearchCV with the best parameters, exhibited slightly worse performance. It can be said that the second model has room for improvement.

Cross-validation

Cross-validation is a method used to objectively evaluate the performance of machine learning models and assess their generalization abilities. By splitting the dataset into multiple partitions for training and testing, it enables us to make more reliable performance assessments. It also helps mitigate the risk of overfitting, provides a basis for parameter tuning and model selection, and allows us to work more effectively with limited datasets.

metin, yazı tipi, ekran görüntüsü içeren bir resim

Açıklama otomatik olarak oluşturuldu

According to the cross-validation results, the model calculated the mean squared error (MSE) value as 51.45 with a standard deviation of 13.09. Lower MSE and standard deviation values indicate better model performance. Based on these results, we can aim to reduce the MSE and improve the consistency of the performance to enhance our model.

PCA

We can use PCA to understand the relationships between variables in the dataset, reduce the dimensionality of the data, and leverage the latent structures present in the dataset. Additionally, PCA allows us to observe how it performs in a different model.

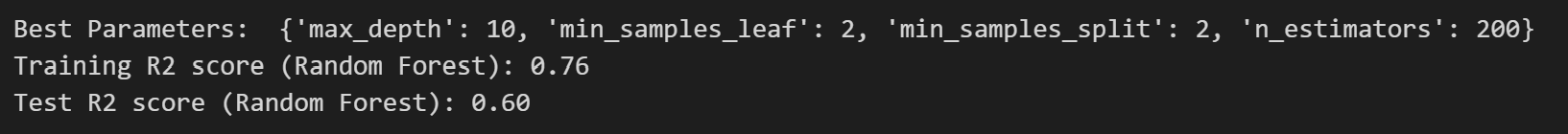
metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

According to the XGBoost regression results obtained with PCA, it can be observed that the model performs better. Lower MSE values and higher R^2 scores indicate that the model fits better and makes better predictions. The use of PCA has improved the model's generalization ability.

Random Forest Regressor

Random Forest is a preferred machine learning algorithm for various reasons. It offers high performance, the ability to select important features, good generalization capabilities, and adaptability to different types of data. Moreover, it is efficient in handling large datasets thanks to its fast-training feature. Even though our dataset may not be large, using Random Forest can still provide valuable insights and demonstrate how the data performs in a different model.



The Random Forest model has shown a good fit on the training set with the best parameters (R^2 = 0.76), but it has exhibited lower performance on the test set (R^2 = 0.60). The model's generalization ability may be somewhat weak.

Comparison of Results of Model

metin, ekran görüntüsü, diyagram, renklilik içeren bir resim

Açıklama otomatik olarak oluşturuldu

According to the results, the XGBoost model outperforms other models in terms of performance. Initially, when trained on the original dataset, the XGBoost model achieves a high R^2 score on the training set but slightly lower on the test set, indicating overfitting and limited generalization ability. However, after applying PCA, the XGBoost model shows improved performance. By reducing dimensionality, PCA helps the model generalize better and make more accurate predictions. The model obtains higher R^2 scores on both the training and test sets, indicating enhanced performance and better utilization of underlying dataset structures. In contrast, the Random Forest model demonstrates good fit on the training set but performs less effectively on the test set, suggesting weaker generalization ability compared to the XGBoost model. In conclusion, the XGBoost model consistently performs better, both on the original dataset and after applying PCA, making it the preferred choice over other models.

According to the graph, while the XGBoost regression model showed a tendency to overfit in the original data set, this tendency may have decreased when PCA was applied. Random Forest regression model, on the other hand, shows an underfitting tendency.

Sentiment Analysis

I have extracted my data from the Reddit API in JSON file format, which includes comments made by individuals regarding global and Ireland-specific topics such as new building construction costs, the state of the construction sector, and unemployment rates. We will apply sentiment analysis to these contexts using a pre-trained supervised model and evaluate the results.

Global Sentiment Analysis

metin, ekran görüntüsü, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu

After we applied our trained model on text our conclude is in below.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

Ireland Sentiment Analysis

metin, ekran görüntüsü, yazı tipi, menü içeren bir resim

Açıklama otomatik olarak oluşturuldu

After we applied our trained model on text our conclude is in below.

metin, ekran görüntüsü, diyagram, dikdörtgen içeren bir resim

Açıklama otomatik olarak oluşturuldu

As observed in the results, there is a higher occurrence of negative comments globally, whereas in Ireland, there are more neutral comments. However, in both analyses, the number of positive comments is significantly low. Overall, when we take a broad look, it can be concluded that Reddit users' opinions regarding the construction sector and its influencing factors tend to be neutral or negative.

metin, ekran görüntüsü, diyagram, yazı tipi içeren bir resim

Açıklama otomatik olarak oluşturuldu metin, ekran görüntüsü, yazı tipi, diyagram içeren bir resim

Açıklama otomatik olarak oluşturuldu

Summary of Machine Learning

In summary, we applied supervised learning models, specifically XGBoost regression and Random Forest Regressor, to assess the impact of "Contribution of the construction sector to the GBD (%)" on the variables "Labor force, total" and "Unemployment, total (% of total labor force)". Additionally, we applied the unsupervised learning model, PCA, for analysis. It's worth noting that we treated sentiment analysis, which is typically an unsupervised task, as a supervised model since we used a pre-trained model.

Why did we choose the libraries and techniques we chose for each of the different data sources?

Pandas vs. csvkit:

Pandas is preferred for processing CSV files because:It offers advanced functionality for data manipulation and analysis. It has a wide ecosystem and an active community. It provides high performance and can perform operations quickly. It has extensive functionality for customization and visualization. CSVKit, on the other hand, is a lighter and more specialized library, but generally, Pandas is preferred because it is more comprehensive and easier to use. (Anon., n.d.) (Anon., n.d.)

plotly.Express & Plotly.Graph\_Objects vs. matplotlib & seaborn:

We chose Plotly.Express and Plotly.Graph\_Objects over other visualization libraries like Matplotlib and Seaborn due to their ability to create interactive graphics, ease of sharing, customization options, and integration capabilities. Additionally, we couldn't use Matplotlib and Seaborn for visualizing country-specific data over the years on a world map in an interactive manner. Also, we could use bokeh library. (Anon., n.d.) (Agrawal, 2021)

sklearn.model\_selection vs statsmodels.api:

The statsmodels.api library is a tool that simplifies working with statistical analysis and models. This library provides functions for creating, analysing, and presenting statistical models such as linear regression analysis, time series analysis, logistic regression, ANOVA, and more. It offers a range of functions for model creation, analysis, and result presentation in statistical analysis tasks.

On the other hand, the sklearn.model\_selection library simplifies basic operations in machine learning models such as model selection, validation, and hyperparameter tuning. It offers functionalities like splitting data sets into training and test sets, performing cross-validation, conducting hyperparameter optimization, and evaluating model performance.

In fact, we could have used the statsmodels.api library for statistical analysis and for regression analysis. However, since we performed hyperparameter optimization and evaluated model performance in our machine learning models, we preferred to use the sklearn.model\_selection module. (Anon., n.d.)

Summarize and Conclusion.

"This study was conducted with the aim of analysing the impact of countries' labor forces and unemployment rates on the contribution of the construction sector to national annual income. Three distinct datasets were compiled and processed in CSV format. Through data manipulation and cleaning, the data was brought into a form suitable for analysis.

An extensive visualization process was applied in the study. Colors and charts were meticulously selected in accordance with data visualization rules, and an interactive world map was used to compare countries' labor force data. Emphasis was put on the data from Croatia and Finland, which closely match Ireland's labor force.

A thorough statistical analysis was performed on the contribution of countries' construction sector to annual revenues, and trends were identified based on this data. Our commentary focused on the contributions of the construction sector to the annual revenues of Ireland, Croatia, and Finland.

Subsequently, machine learning models such as XGBRegression analysis, PCA, and Random Forest were employed to train and test the dataset. Sentiment analysis was applied to users' comments about the construction sector, labor force, and unemployment rates using the Reddit API. This analysis was part of our effort to understand public opinion.

We arrived at the findings of our study and meticulously discussed our interpretations. We grappled with a dataset that was more restricted than we would have liked; we did stumble upon a more comprehensive dataset but, alas, it was a paid resource which we couldn't make use of. However, every cloud has a silver lining - throughout this journey, we delved into and learned a multitude of things about machine learning models, visualization strategies, statistical analysis methods, and various coding techniques. So, even though we were dealing with a less than ideal dataset, the wealth of insights and practical experience we gathered from this investigative process was genuinely gratifying and rewarding."