# Feb 2023 – SB+ - MSc in Data Analytics

Author: Sajjan Bhattarai

e-mail: [sbs23010@student.cct.ie](mailto:sbs23010@student.cct.ie)

Student ID: sbs23010

Word count: ~3240 excluding Abstract, Tables and References

# Abstract

This report presents a study on the impact of construction material prices on the production in construction. The objective of this study was to analyse the historical price indices for 40 different construction materials, and the indices of production in construction during the same timeline. The research methodology comprised of the descriptive and inferential statistics, regression and classification machine learning models, and sentiment analysis.

The findings indicate that the prices of construction materials are correlated to the production in construction sector. Drastic increase and decrease in certain key materials can impact the production of a construction sector which heavily utilizes such materials. This makes it challenging to predict the value and volume of production in construction for a specific sector when the prices of key materials become volatile.

Through the application of machine learning models, this study provides actionable insights into how the materials prices are related to which construction type. Based on this, one could identify the potential choke-points in construction in regards to acquiring the necessary materials. Then, to plan the sourcing, supply and storage of such materials as per the goal for production in construction. The aim of this study has been to lay some ground-works so that further research can be performed to deepen the understanding in this area.

Table of Contents

[Feb 2023 – SB+ - MSc in Data Analytics 1](#_Toc136006156)

[Abstract 2](#_Toc136006157)

[Introduction 3](#_Toc136006158)

[Materials and Methods 3](#_Toc136006159)

[Selection of Programming Language and Libraries 3](#_Toc136006160)

[Project Management Framework 3](#_Toc136006161)

[Research Understanding 4](#_Toc136006162)

[Data Understanding 4](#_Toc136006163)

[Modeling 11](#_Toc136006164)

[Evolution 17](#_Toc136006165)

[Deployment 17](#_Toc136006166)

[Results 17](#_Toc136006167)

[Data Summary 17](#_Toc136006168)

[Model Development 18](#_Toc136006169)

[Sentiment Analysis 18](#_Toc136006170)

[Discussion/Conclusions 18](#_Toc136006171)

[References 19](#_Toc136006172)

# 

# Introduction

Construction sector is one of the key drivers of the economy globally. In the context of European Union, it accounts for 18 million jobs and almost 9% of GDP. It is estimated that a full-scale digitalization can contribute up to 20% annual global savings in the design and construction phases, and up to 15% in the operational phases (Arthur, 2022). Hence, applying the concepts of Data Analytics on to Construction industry could help further our understanding of the relationship between different components, and to potentially optimize the sector.

# Materials and Methods

## Selection of Programming Language and Libraries

For this study, two programming languages Python and R were considered for their simplicity and availability of wide range of libraries for Data Analytics and Machine Learning. Ultimately, Python was chosen because it provides more human readable syntax, and versatility in terms of object-oriented and modular programming. Additionally, it has a larger ecosystem of libraries and frameworks, that offer more choices and easy integration with other technologies. Therefore, for the purposes of performing statistical analysis, machine learning and visualization on dashboard, Python was the most effective language within the scope of this study.

Similarly for data manipulation, three Python libraries Pandas, PySpark and NumPy were considered. This study mainly utilized Pandas, along with NumPy in several situations. PySpark was also explored and tested, however it was deemed infeasible for the scale of data being analyzed in this study. Additionally, the differences and limitations in PySpark’s syntax made it more challenging to perform small or basic tasks. For example, lack of *apply()* method and *lambda* function in PySpark DataFrame required the use of *User Defined Functions* (UDF). Such added complexities would have been justified if the concerned dataset was huge and required distributed processing (SparkByExamples, 2023). Nevertheless, it was kept as an option during this study in case a much larger dataset needed to be processed.

## Project Management Framework

Two commonly used frameworks CRISP-DM and SEMMA were considered for this study. CRISP-DM stands for Cross Industry Standard Process, and was developed by a consortium of numerous data-mining companies (Kotu and Deshpande, 2014). It comprises of iterative phases that include business/research understanding, data understanding, modelling, evolution and deployment. On the other hand, SEMMA stands for Sample, Explore, Modify and Assess, was developed by Statistical Analysis System (SAS).

Although both frameworks follow similar model of iterative progress and feedback loop, CRISP-DM was chosen for this study because it offers a more comprehensive framework from understanding of research requirements to final deployment of the model.

With this, the project deliverables were planned as below:

1. Understanding the problem domain and exploring the available datasets during first week
2. Performing descriptive and inferential statistical analysis on the relevant datasets, and working with Machine Learning models during second week
3. Further Machine Learning testing and optimization, along with Sentiment Analysis on Reddit data during third week
4. Improving code quality with reusable functions and unit tests, as well as preparing the dashboard visualization in the fourth week

### Research Understanding

This research sought to analyse and establish relationship between construction material prices and production in construction. Although it was primarily focused on the context of Ireland’s construction sector, it also made a high-level comparison with other countries. Additionally, it aimed to perform sentiment analysis from Reddit posts and comments to further support the outlined relationship.

### Data Understanding

The datasets used in this study were mostly published by Central Statistics Office of Ireland under Creative Commons Attribute 4.0 license (Government of Ireland, 2023). Other datasets were published by Eurostat under Public access type (European Commission, 2023) and Reddit API data (Reddit, 2023). All datasets that were explored and considered during this study can be found in attached Jupyter Notebook (data-exploration.ipynb).

After searching for relevant datasets, the Production in Construction Indices and Wholesale Price Indices of Construction Materials datasets from Ireland were selected for further analysis. Additionally, the Production in Construction Indices from Eurostat, and Building Construction Costs from Netherlands were selected for high-level comparison with Ireland. Lastly, the Reddit APIs were used to search for relevant topic and conversations from concerned sub-reddits to perform public sentiment analysis. All these analysis and results are available in the attached Jupyter Notebook (construction-materials-and-production.ipynb).

#### Descriptive Statistics

The first dataset represented the monthly wholesale price indices and percentage change over 1 and 12 months for 40 major construction materials from January 2015 to April 2023. After filtering the required statistic label and pivoting the table to put the materials as columns, there were 100 rows in total. Out of these records, 2 materials had null values for the last 3 months, and no duplicates were present. Upon performing descriptive statistical analysis on this dataset, 10 materials had standard deviation value greater than 20, which implies higher volatility in their prices. Their mean/average price indices were also higher than the median, thus implying positive skew or presence of outliers or extreme values in the right tail.

Table 1: Statistical Description of Some Materials

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Type of Material** | Other structural steel | Other treated timber | Plaster | Structural steel | Structural steel and reinforcing metal | Structural steel fabricated metal |
| **count** | 100.0000 | 100.0000 | 100.00000 | 100.00000 | 100.00000 | 100.00000 |
| **mean** | 119.3100 | 122.3140 | 122.01200 | 127.65400 | 124.53800 | 128.67800 |
| **std** | 32.48594 | 44.93026 | 26.717203 | 31.898473 | 31.227089 | 32.694406 |
| **min** | 95.70000 | 95.30000 | 98.400000 | 97.600000 | 98.800000 | 97.300000 |
| **25%** | 99.60000 | 95.30000 | 104.12500 | 110.35000 | 106.32500 | 111.75000 |
| **50%** | 104.6500 | 102.8000 | 112.80000 | 120.45000 | 113.60000 | 122.70000 |
| **75%** | 121.1750 | 104.7000 | 125.62500 | 124.12500 | 120.57500 | 124.50000 |
| **max** | 210.5000 | 216.6000 | 206.30000 | 222.40000 | 205.70000 | 229.90000 |

The presence of outliers or extreme price indices for some construction materials can also be demonstrated in below BoxPlot.

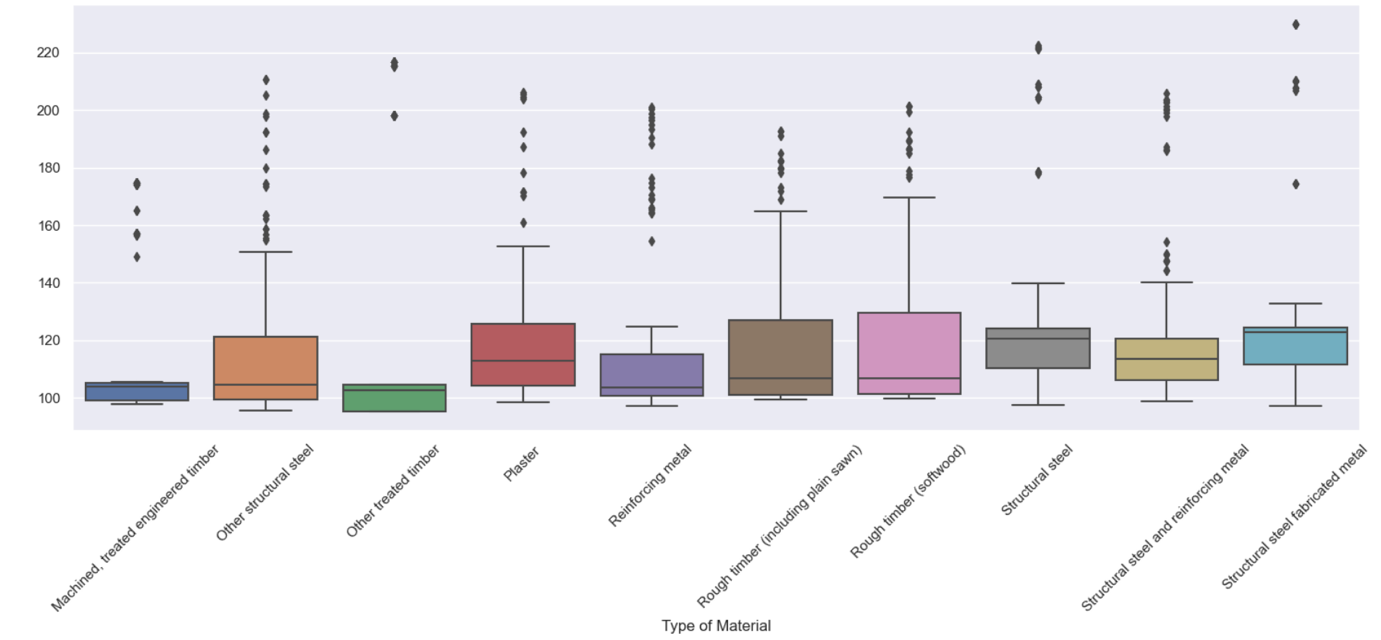


Figure 1: BoxPlot of Construction Materials with High Variation

It is preferable to address such outliers when applicable, using methods like transformation, removal, Winsorization, etc. BoxCox Transformation was applied on these features, which helped to reduce the outliers. However, with such approaches, it is important to consider aspects such as interoperability, impact on relationships, and reversibility. Therefore, based on the use case, it was decided to not transform the data.

Similarly, the second dataset represents the Production in Construction Indices with same base of 100 for 2015. It is also mainly positive skewed, with Residential Buildings standing out. This construction type had highest maximum of 789.9 and lowest medium of 55. It also had highest mean at 268.5 with standard deviation of 219.3.

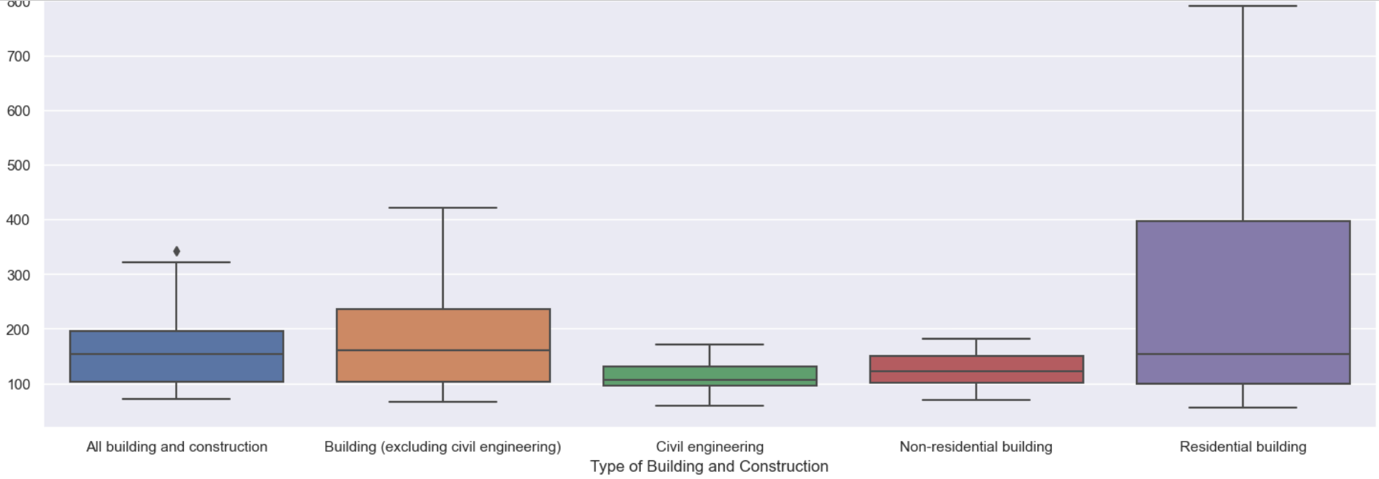


Figure 2: BoxPlot of Production in Construction by Types

In overall, only Non-residential Building type had relatively balanced distribution.

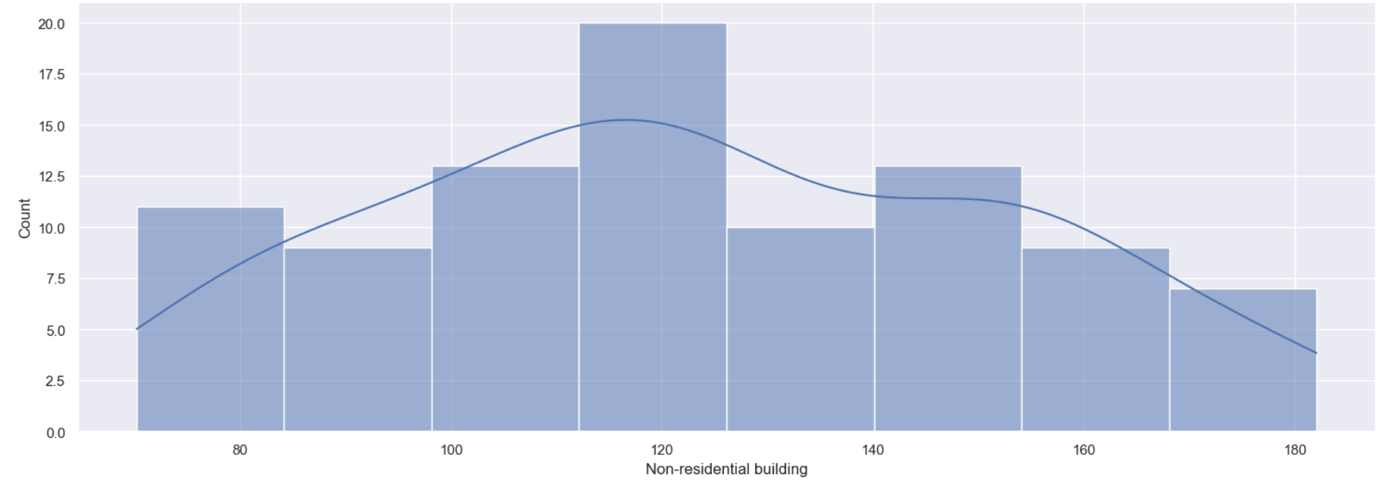


Figure 3: HistPlot of Production in Construction Indices for Non-residential Building

#### Inferential Statistics

Below inferential statistical analysis were performed to gain more insight and to make inferences about a population based on available sample data.

##### Pearson Correlation Coefficient Test

10 construction materials were shortlisted based on their Pearson correlation coefficient and target feature i.e. Production in Construction index for Non-residential buildings (Sheposh, 2022). Below test calculates the p-value and confidence interval for each of these materials and the target variable.

Stating the Null Hypothesis:

* **Null Hypothesis (H0):** There is no correlation between each of the 10 selected construction materials and the value of production in Non-residential buildings.
* **Alternative Hypothesis (H1):** There is a correlation between each of the 10 selected construction materials and the value of production in Non-residential buildings (the population correlation coefficient is not zero for at least one construction material).
* **Significance Level (alpha):** 0.05

Table 2: Pearson Correlation Coefficient Test results

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Material | Coefficient | p-value | Accept H0? | Confidence interval |
| All other metal fittings | 0.814911 | 5.419483e-24 | False | 0.733-0.873 |
| Cement | 0.750128 | 1.421157e-18 | False | 0.645-0.827 |
| HVAC (heating and ventilation equipment) | 0.742812 | 4.573935e-18 | False | 0.636-0.822 |
| Other steel products | 0.735394 | 1.438152e-17 | False | 0.626-0.817 |
| PVC pipes and fittings | 0.698355 | 2.589154e-15 | False | 0.577-0.789 |
| Paints, oils and varnishes | 0.797549 | 2.381846e-22 | False | 0.709-0.861 |
| Plaster | 0.700247 | 2.024512e-15 | False | 0.58-0.791 |
| Precast concrete | 0.729288 | 3.589822e-17 | False | 0.618-0.812 |
| Rough timber (hardwood) | 0.687658 | 1.004452e-14 | False | 0.563-0.782 |
| Wooden windows and doors | 0.687138 | 1.071252e-14 | False | 0.563-0.781 |

For all selected materials, the calculated p-values are lower than the significance level 0.05, so the null hypothesis is rejected. Hence, there is statistically significant correlation between the materials price and production in construction. Additionally, the confidence interval doesn't include 0, so the correlation coefficient is statistically significant and shouldn’t have occurred by coincidence alone.

##### T-test to Calculate Confidence Intervals for Population

Assuming the population mean as 110 for each selected materials, below test tried to infer whether there is significant difference between sample mean and population mean.

* **H0:** There is no difference between the sample mean and population mean.
* **H1:** The sample mean and population mean are different.
* **Alpha:** 0.05

Table 3: T-test for Sample Mean and Population Mean

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Material | T-statistic | p-value | Accept H0 | confidence interval |
| All other metal fittings | -3.502 | 0.00071 | False | (107.016, 109.175) |
| Cement | 3.190 | 0.00193 | False | (112.034, 118.739) |
| HVAC (heating and ventilation equipment) | -12.498 | 0.00000 | False | (103.421, 105.225) |
| Other steel products | -11.535 | 0.00000 | False | (102.627, 104.792) |
| PVC pipes and fittings | 2.591 | 0.01108 | False | (111.064, 118.044) |
| Paints, oils and varnishes | 3.144 | 0.00222 | False | (111.614, 117.146) |
| Plaster | 3.989 | 0.00013 | False | (114.295, 122.805) |
| Precast concrete | 1.987 | 0.04980 | False | (110.002, 115.548) |
| Rough timber (hardwood) | -3.649 | 0.00043 | False | (104.665, 108.425) |
| Wooden windows and doors | -2.006 | 0.04765 | False | (105.667, 109.977) |

##### Independent Samples T-test

In this independent samples t-test (Gio and Rosmaini, 2018), the yearly average cost for construction materials were compared between Ireland and Netherlands. Since the Dutch dataset contained yearly records for all materials combined, matched the Irish dataset by grouping the data on year and calculating the mean of all materials.

* **H0:** There is no difference between the mean cost indices.
* **H1:** There is a difference.
* **Alpha:** 0.05

It returned the t-statistic as -2.28 and p-value as 0.029, which suggest a significant difference between the mean building material cost indices of these two countries.

##### Analysis of Variance with Levene’s Test

It is helpful to compare the production in construction across Ireland and other European countries. Therefore, an equivalent dataset from Eurostat was used for this analysis, and aggregated to quarterly from monthly basis. While comparing more than two groups, Levene’s Test can be used to assess the equality of variances.

* **H0:** Population variances are equal.
* **H1:** Population variances aren't equal.
* **Alpha:** 0.05

The returned p-value was less than 0.05 in both cases of using mean and median as center. Therefore the Null hypothesis is rejected, which indicates difference in the variances.

##### Welch’s ANOVA Test

Based on results of above Levene’s Test, the variances between chosen countries aren’t same. Therefore, the use case doesn’t meet one of the criteria for using ANOVA (Holt, 2023). Therefore, a more flexible Welch’s ANOVA Test was performed to compare the means across more than two groups (Delacre, 2019).

* **H0:** There is no difference in means of production in construction across different countries.
* **H1:** There is a difference between means across countries.
* **Alpha:** 0.05

From above test results, the f-value of 355.27 suggests there might be some differences in means of the countries. And the p-value is less than the chosen significance level of 0.05, so the Null hypothesis is rejected. Therefore, there is sufficient evidence to suggest statistically significant differences across the means of these countries.

##### Mann-Whitney U Test

Similar comparison was also made using non-parametric test called Mann-Whitney U test.

* **H0:** There is no significant difference between the production in construction indices between Ireland and Netherlands.
* **H1:** There is difference between the two countries
* **Alpha:** 0.05

The resulting p-value was 0.028, which is lower than the chosen significance value. Therefore, the Null Hypothesis is rejected.

##### Wilcoxon Signed-rank Test

Another non-parametric test was used to compare two groups of production in construction indices for Netherlands.

* **H0:** There is no significant difference between the production in construction indices in Netherlands before and after 3rd quarter of 2022.
* **H1:** There is difference between the group of production in construction indices.
* **alpha:** 0.05

It resulted in p-value of 0.50, so the Null Hypothesis is accepted.

#### Data Preparation

After shortlisting the relevant datasets and performing statistical analysis, they were pivoted to put the independent and target features as columns and joined together. This resulting data contained price indices for 40 construction materials, and production in construction indices for 5 types. This was done to understand the relationship between the materials and certain construction types. At the end of the study, the objective was to identify key materials affecting a construction type, and to perform prediction and/or classification. Such information could help to better plan the supply of proper materials as per the needs of each construction type and their demands. Below line plots represent the trends in Production in Construction and Construction Material Prices. The different colored lines indicate different construction types or materials.

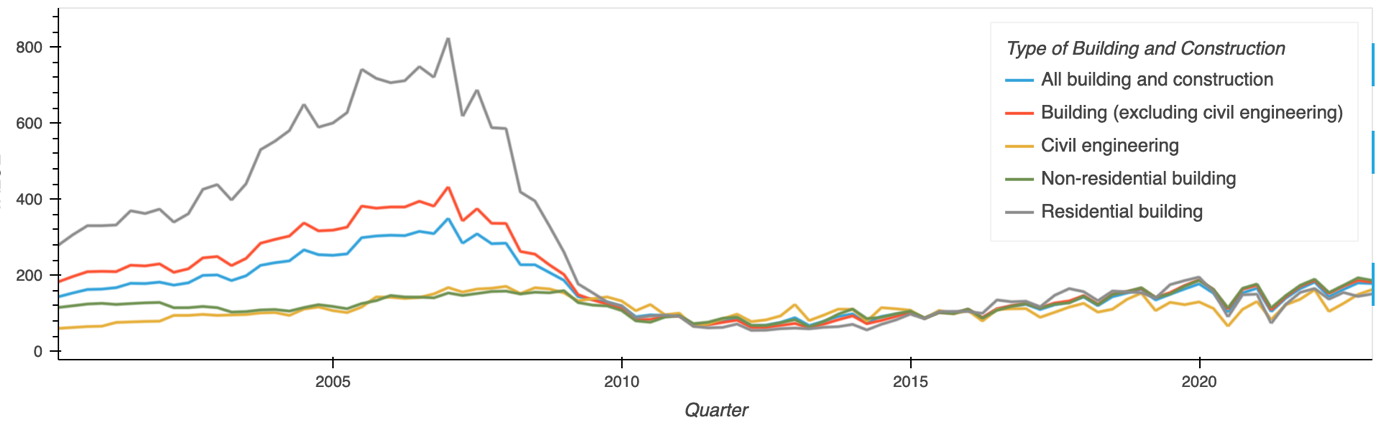


Figure 4: Production in Construction Indices in Ireland

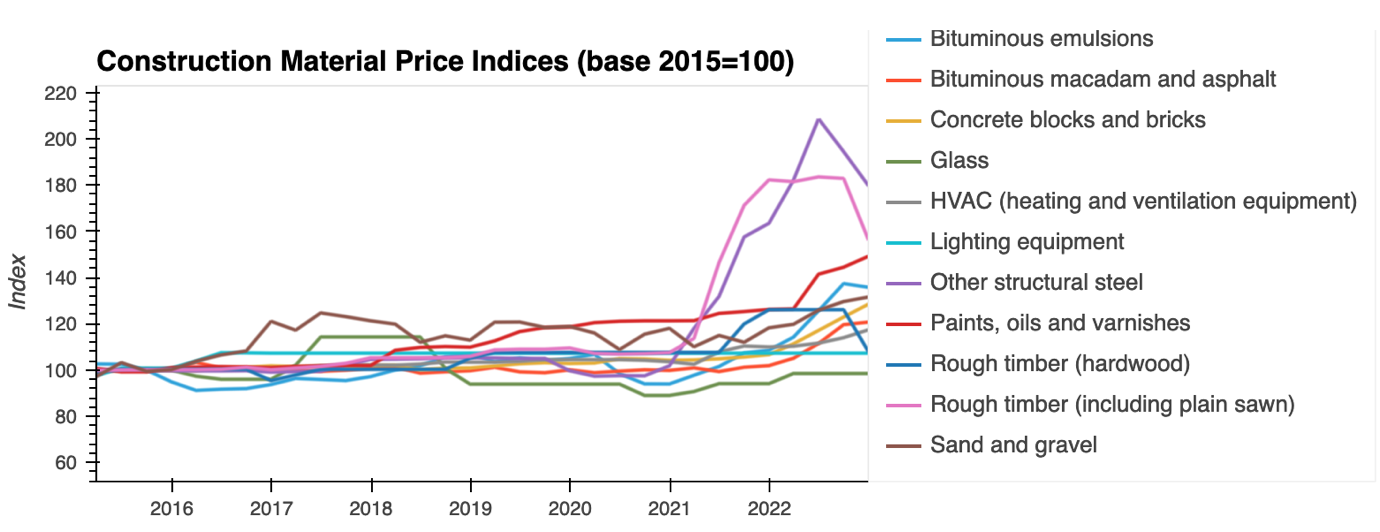


Figure 5: Construction Materials Price Indices in Ireland (Top 10)

Before proceeding with Machine Learning, a new target feature was created to label whether the production in construction was increasing quarter-over-quarter. This label was then used to perform Classification models.

##### Feature Selection

Out of all available independent features or construction materials, it was important to identify the most significant ones for more effective modelling. In addition to Correlation Coefficient Test during Inferential Statistics, further analysis was done using SelectKBest and Pearson Correlation heatmap. As part of machine learning, Lasso was also used to identify the features with highest magnitude (Jomthanachai, Wong and Khaw, 2022).

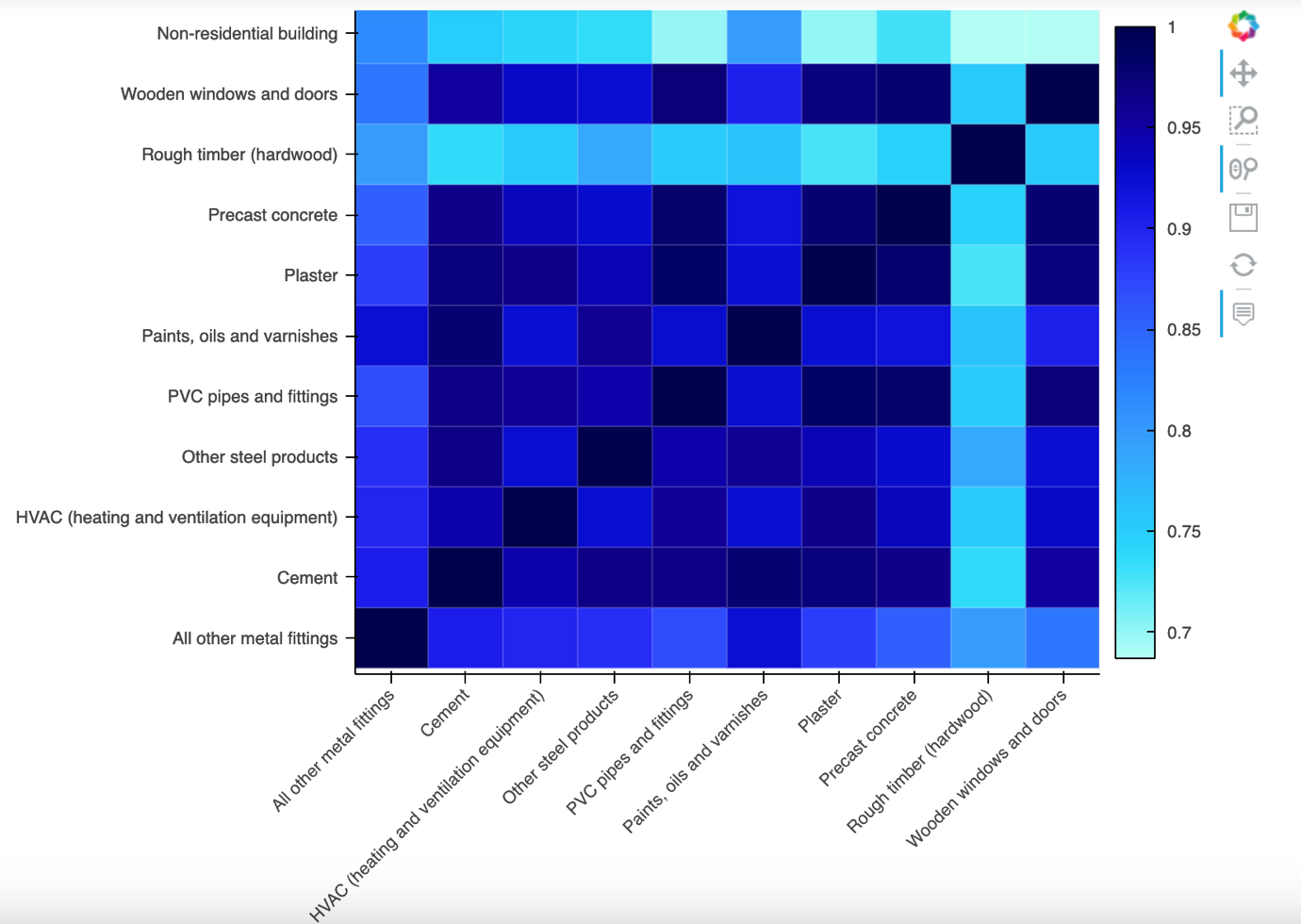


Figure 6: Heatmap of Correlation Coefficients

With this, “All other metal fittings” was identified as the most important feature, along with 9 other materials.

##### Data Scaling

All the features in the finalized dataset represent index values taking base of 100 on 2015. Since these are on same scale, it might not be necessary to scale the data using methods like MinMaxScaler and/or StandardScaler. Nevertheless, it was important to note the presence of outliers for some of the features, therefore some Regularization models like Ridge or Lasso could potentially benefit from data scaling. Hence, both of these scaling methods were used during the modelling phase.

### Modeling

All features including target variable in the prepared dataset are continuous or numeric, therefore Regression models were the ideal match. Additionally, a new target variable was created to classify whether the construction was increasing or not, thus allowing the use of Classification Models. Provided that the data was labelled, the models belonged to Supervised learning.

#### Regression Models

As part of regression, 3 different models were used: Linear Regression, Ridge and Lasso. These models provided insight into how changes in material prices affected the seasonally adjusted value of production in residential and non-residential buildings. Below table represents the accuracy scores for each model:

Table 4: Regression Model Results with All Features

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Target | Training Score | Testing Score |
| LinearRegression | Non-residential | 0.94 | 0.49 |
| Residential | 0.82 | 0.29 |
| Ridge | Non-residential | 0.93 | 0.54 |
| Residential | 0.79 | 0.47 |
| Lasso | Non-residential | 0.92 | 0.78 |
| Residential | 0.77 | 0.64 |

Above results showed much higher training score than testing score, thus implying overfitted model. To address this, feature reduction was attempted based on the feature coefficients.

Table 5: Regression Models Results with Top 25 Features

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Target | Training Score | Testing Score |
| LinearRegression | Non-residential | 0.92 | 0.80 |
| Residential | 0.73 | 0.42 |
| Ridge | Non-residential | 0.91 | 0.79 |
| Residential | 0.72 | 0.44 |
| Lasso | Non-residential | 0.90 | 0.80 |
| Residential | 0.68 | 0.48 |

Limiting the number of features helped to improve the testing scores. Next, also tested the models after scaling the data with MinMaxScaler and StandardScaler.

Table 6: Regression Models Results with Scaled Data

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Target | Scaler | Training Score | Testing Score |
| LinearRegression | Non-residential | StandardScaler | 0.85 | 0.16 |
| Residential | 0.71 | -2.15 |
| Non-residential | MinMaxScaler | 0.85 | -3.67 |
| Residential | 0.71 | -13.87 |
| Ridge | Non-residential | StandardScaler | 0.83 | 0.66 |
| Residential | 0.62 | 0.52 |
| Non-residential | MinMaxScaler | 0.75 | 0.73 |
| Residential | 0.49 | 0.50 |
| Lasso | Non-residential | StandardScaler | 0.75 | 0.64 |
| Residential | 0.50 | 0.43 |
| Non-residential | MinMaxScaler | 0.64 | 0.66 |
| Residential | 0.31 | 0.39 |

In overall, scaling the data didn’t improve the model performance. In fact, it degraded the test scores for LinearRegression. Lastly, below table shows the results after performing hyper-parameter tuning with GridSearchCV (Lewinson, 2020).

Table 7: Regression Models Results with Best Estimator from GridSearchCV

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Target | Training Score | Testing Score |
| LinearRegression | Non-residential | 0.85 | 0.75 |
| Residential | 0.68 | 0.64 |
| Ridge | Non-residential | 0.83 | 0.78 |
| Residential | 0.63 | 0.66 |
| Lasso | Non-residential | 0.84 | 0.76 |
| Residential | 0.65 | 0.61 |

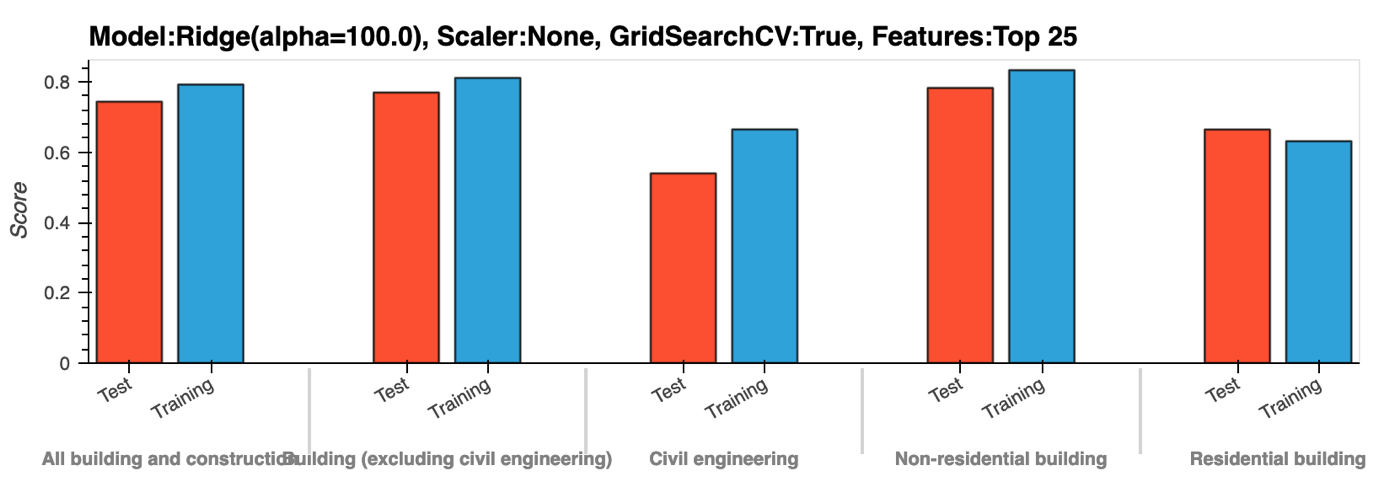


Figure 7: Ridge Model's Training and Testing Scores for Each Construction Type

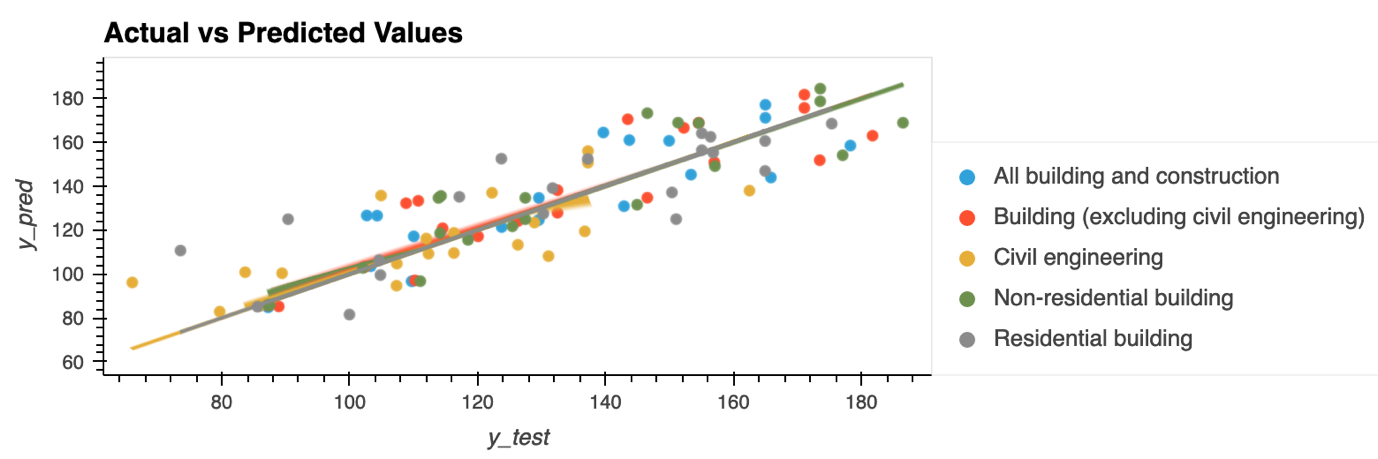


Figure 8: Scatter Plot of Ridge Model's Predicted Values vs. Actual Values

#### Classification Models

For classification, 4 different models were used: Logistic Regression, Decision Tree, Random Forest and Support Vector Machine. Similar to regression, these models attempted to classify whether the production was increasing for each construction type.

Table 8: Classification Model Results for All Features

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Target | Training Score | Testing Score |
| LogisticRegression | Non-residential | 0.88 | 0.9 |
| Residential | 0.88 | 0.9 |
| DecisionTree | Non-residential | 1.0 | 0.80 |
| Residential | 1.0 | 0.80 |
| RandomForest | Non-residential | 1.0 | 0.9 |
| Residential | 1.0 | 0.9 |
| SVC | Non-residential | 0.66 | 0.65 |
| Residential | 0.66 | 0.65 |

The scores were already better than the regression models, with Decision Tree and Random Forest Classifiers performing the best.

Table 9: Classification Models Results with Top 25 Features

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Target | Training Score | Testing Score |
| LogisticRegression | Non-residential | 0.84 | 0.8 |
| Residential | 0.86 | 0.8 |
| DecisionTree | Non-residential | 1.0 | 0.65 |
| Residential | 1.0 | 0.80 |
| RandomForest | Non-residential | 1.0 | 0.90 |
| Residential | 1.0 | 0.90 |
| SVC | Non-residential | 0.66 | 0.65 |
| Residential | 0.66 | 0.65 |

Upon reducing to 25 features, only Random Forest performed better, while others didn’t improve. Next, the models were run with scaled data.

Table 10: Classification Model Results with Data Scaling for Residential Buildings

|  |  |  |  |
| --- | --- | --- | --- |
| Model | Scaler | Training Score | Testing Score |
| LogisticRegression | StandardScaler | 0.75 | 0.75 |
| MinMaxScaler | 0.67 | 0.7 |
| DecisionTree | StandardScaler | 1.0 | 0.75 |
| MinMaxScaler | 1.0 | 0.75 |
| RandomForest | StandardScaler | 1.0 | 0.90 |
| MinMaxScaler | 1.0 | 0.85 |
| SVC | StandardScaler | 0.68 | 0.75 |
| MinMaxScaler | 0.67 | 0.7 |

Performance of SVC improved with scaled data, whereas other models degraded. Lastly, the models were run with GridSearchCV, and below results was obtained for the best estimator:

Table 11: Classification Model Results with Best Estimator from GridSearchCV

|  |  |  |
| --- | --- | --- |
| **Model** | Training Score | Testing Score |
| LogisticRegression(C=0.1,solver=’liblinear’) | 0.84 | 0.85 |
| DecisionTreeClassifier(criterion='entropy', max\_depth=5) | 0.95 | 0.75 |
| DecisionTreeClassifier(max\_features='sqrt',…) | 1.0 | 0.90 |
| SVC(C=10, kernel='linear') | 0.89 | 0.75 |

In overall, Random Forest Classifier had the best performance for this use case, followed by Decision Tree and SVC.

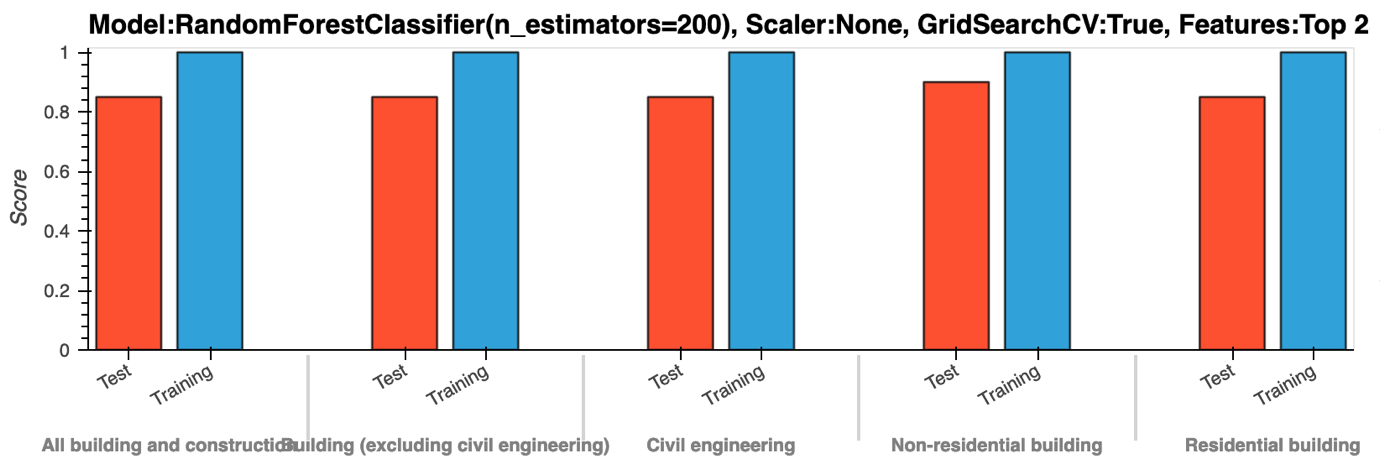


Figure 9: Random Forest's Training and Testing Scores

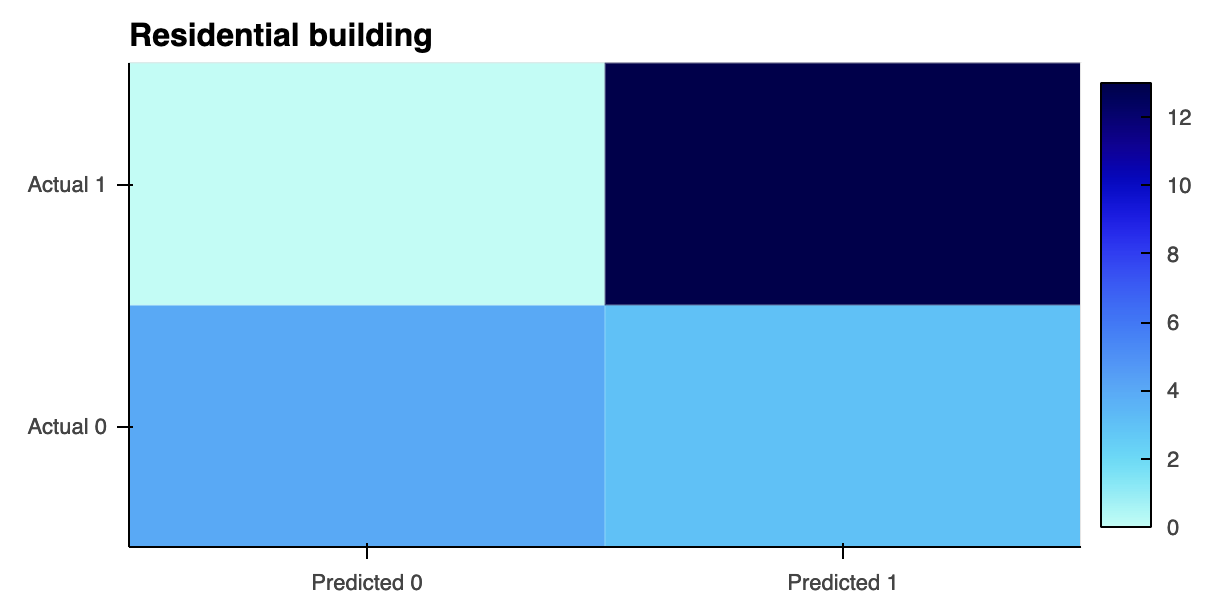


Figure 10: Confusion Matrix of Predicted Values by Random Forest

The dark coloured box represents higher number of matches between actual and predicted classification, while the light blue boxes indicate lower matches.

#### Sentiment Analysis on Reddit

Reddit API was used to gather search results and comments on the topic of construction material prices and production in construction across Ireland, Netherlands and UK. The texts were then analysed with SentimentIntensityAnalyzer using a pre-trained model called VADER, and Textblob.

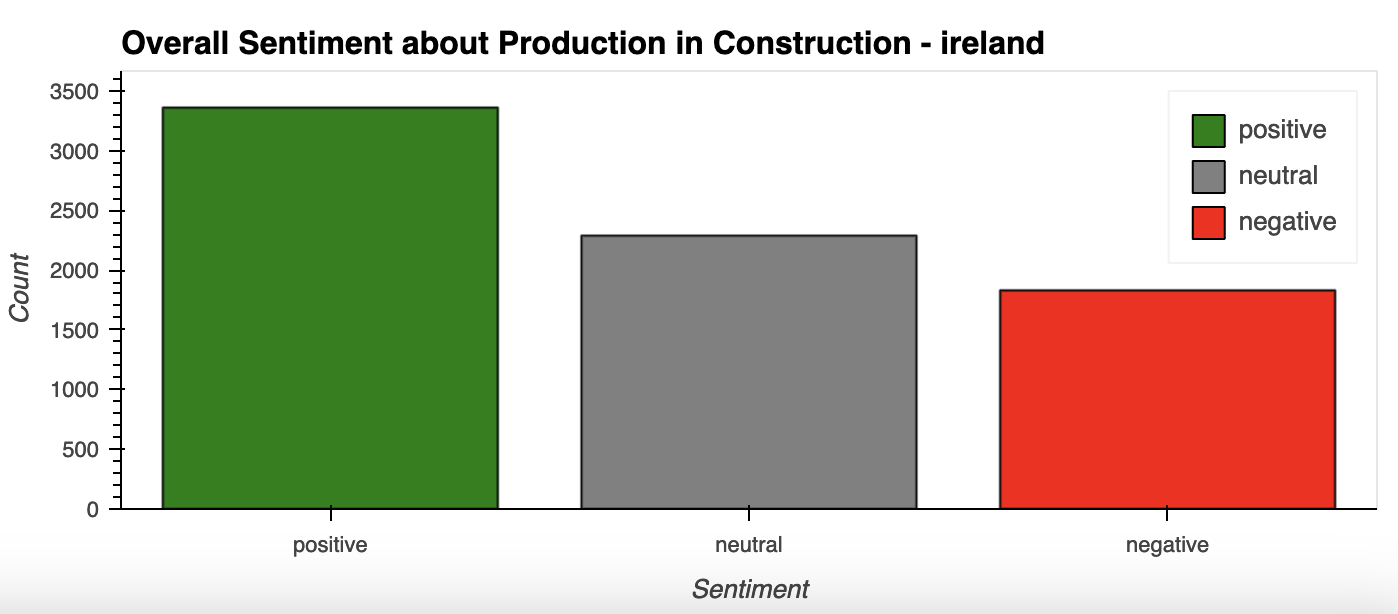


Figure 11: VADER Prediction

Above barplot indicates total number of texts categorized as positive (green), neutral (grey) and negative (red). Whereas TextBlob classified slightly less posts as positive, and more as neutral.

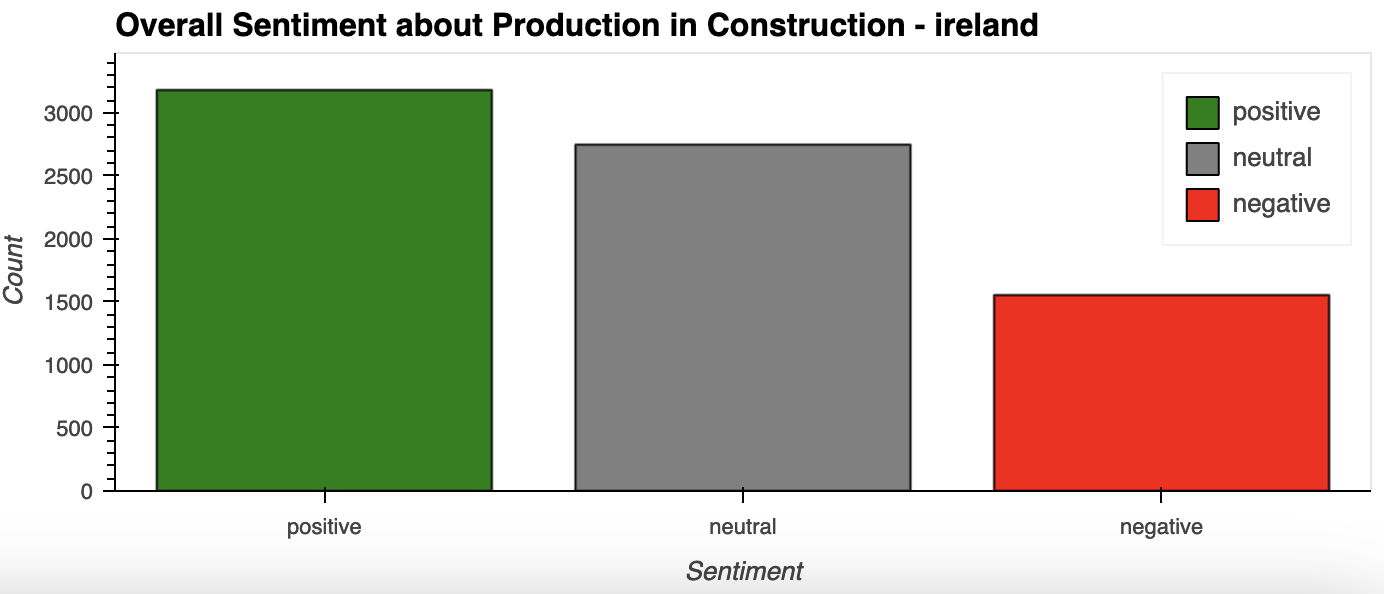


Figure 12: TextBlob Sentiment Prediction

Both VADER and TextBlob had common labels for majority of the sentiments. For texts with differing sentiments, it was challenging to decide which one was correct. Therefore, it was decided to only use the common sentiments. Similarly, below lineplot represents the trend in the posts and sentiment.

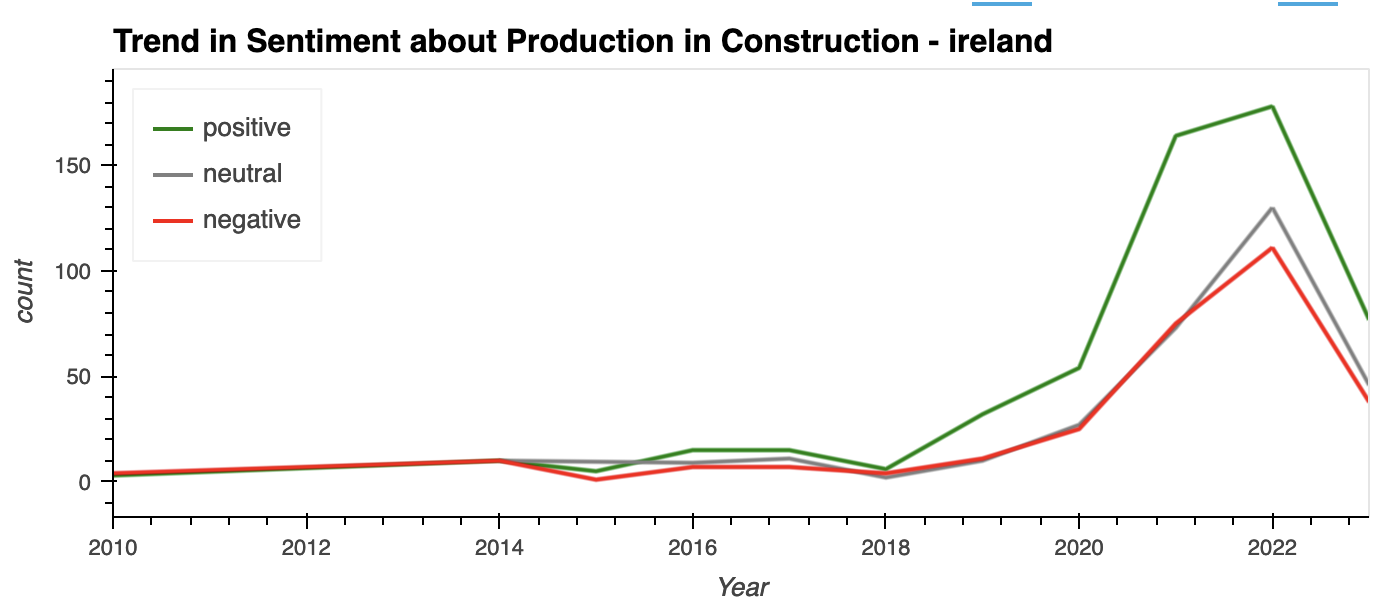


Figure 13: VADER Prediction over years

For comparison, below plot shows trend in Netherlands:

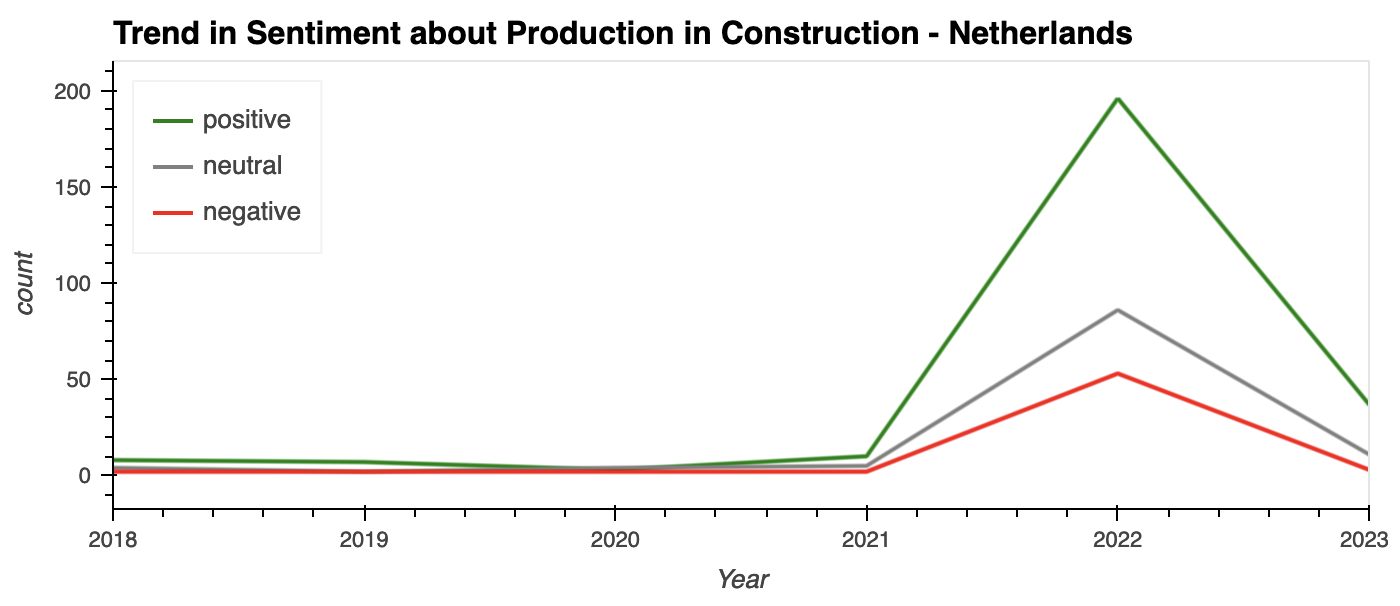


Figure 14: Sentiment Trend in Netherlands

Overall, the sentiment across Ireland and Netherlands were positive in Reddit. The Irish posts saw a rise since 2018 and rising. Whereas Dutch posts only started to grow in 2021. For 2023, the numbers were expected to be lower since the year was still in progress.

### Evolution

During evolution phase, the previously created models were expanded by creating an interactive visualization with *Panel* library. This allowed user to choose the parameters like statistics label, model, scaler, number of features, and GridSearchCV. As an output, it displayed the visualization of the data and modeling results. This made the Machine Learning more effective, and easier to understand the relationship between different features and parameters. Similar interactive visualization was created for sentiment analysis as well.

Additionally, for all the Python functions created to achieve above tasks, unit-testing was performed to ensure their quality and correctness.

### Deployment

Finally, all the models and sentiment analysis were deployed along with a common interactive dashboard. For sentiment analysis, all the reddit data with sentiment labels were persisted to MongoDB collection and local JSON file. Reddit post/comments can range from few words to long paragraphs, and can also contain unexpected characters. Therefore, MongoDB was chosen over relational database like MySQL (Bierer, 2020), and JSON format was preferred over structured formats like CSV.

# Results

### Data Summary

The finalized dataset contained 96 observations with 40 independent features (materials) and 5 dependent features (construction types). And for 4 different statistic labels, 4 different datasets were present. Therefore, there were in total 20 different modelling for each Model configuration. From the datasets, materials like timber and steel appeared to have more extreme values. Whereas for construction types, residential building had bigger fluctuations.

### Model Development

The datasets were trained on 3 regression models. After adding a feature to classify if the production was growing, trained the resulting dataset on 4 classification models. All the models were also cross-validated and tuned using GridSearchCV with 5 folds. Depending on which construction type was used as target variable, applied models returned varying results. In overall, Ridge for regression and Random Forest for classification had better training and testing scores.

Above test results supported that there indeed existed relationships between prices of materials and production in construction sectors.

### Sentiment Analysis

Based on the conversations from Reddit, the number of interactions have been growing in recent years. This can partly be contributed to the rising popularity of Reddit, and partly to the rising issues in the housing and construction sector. The trend was similar across all observed countries: Ireland, Netherlands and United Kingdom.

# Discussion/Conclusions

From this study, five conclusions can be drawn:

1. Research understanding is key to get familiar with the problem domain.
2. Gathering the relevant datasets, and performing descriptive and inferential statistics provide additional insight for the research.
3. Implementing appropriate visualization can help to deepen the understanding of data and the trend it represents.
4. During modeling, it was found to perform better after applying feature reduction and hyper-parameter tuning.
5. Reddit posts and comments allow us to peek into general opinion on the topic, which can be used to perform sentiment analysis on the construction related topics.

In conclusion, the relationship between prices of construction materials and production in construction was further validated by this study. Nevertheless, it is important to consider that not all listed materials are required for the mentioned construction types. For instance, rough timber price increased by about 40% between 2020 and 2021, and then decreased by approximately 30% between January 2022 and May 2022 (Department of Public Expenditure and Reform, 2022). Such drastic fluctuation in price affected construction types that relied more on rough timber. Additionally, certain construction might have been impacted by other factors as well, such as Covid19 restrictions, Russia-Ukraine war, etc. Therefore, the material prices might have had varying degree of impact on different construction types. Hence, further research would be required to identify as many contributing factors as possible to get more accurate results.

# References

* SparkByExamples (2023). *Pandas vs PySpark DataFrame With Examples*. Available at: <https://sparkbyexamples.com/pyspark/pandas-vs-pyspark-dataframe-with-examples/> (Accessed: 14 April 2023).
* Kotu v. and Deshpande B. (2014) Predictive Analytics and Data Mining : Concepts and Practice with RapidMiner. Amsterdam: Morgan Kaufmann. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=e250xww&AN=919334&site=eds-live> (Accessed: 12 April 2023).
* Arthur I. Tsarev (2022) ‘Implementation of BIM technologies in the construction sector: the European experience’, Stroitel’stvo: Nauka i Obrazovanie, 12(3), pp. 125–136. doi:10.22227/2305-5502.2022.3.8.
* Government of Ireland (2023). Datasets, Publisher: Central Statistics Office. Available at: https://data.gov.ie/dataset/beq04-onstruction-sector-base-2015100-by-type-of-building-and-construction-year-and-statistic-fb86?package\_type=dataset (Accessed: 27 April 2023).
* European Commission (2023). Datasets, Publisher: Eurostat. Available at: <https://data.europa.eu/data/datasets/sb1sa3jwwgiu0buqjyq5g?locale=en> (Accessed: 28 April 2023).
* Reddit Inc. (2023). Data API Terms. Available at: <https://www.redditinc.com/policies/data-api-terms> (Accessed at: 23 May 2023).
* Sheposh, R. (2022) ‘Pearson correlation coefficient (PCC)’, Salem Press Encyclopedia of Science [Preprint]. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=155876503&site=eds-live> (Accessed: 23 May 2023).
* Prana Ugiana Gio and Elly Rosmaini (2018) ‘The Robustness of Two Independent Samples t Test Using Monte Carlo Simulation with R’, IOP Conference Series: Materials Science and Engineering, 300, p. 012030. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=edsair&AN=edsair.doi...........7fc4a5ce2ebec8a7cf2573c0cef7776a&site=eds-live> (Accessed: 23 May 2023).
* Holt, M.P.. Ms. (2023) ‘Analysis of variance (ANOVA)’, Salem Press Encyclopedia of Science [Preprint]. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=ers&AN=113931283&site=eds-live> (Accessed: 23 May 2023).
* Marie Delacre et al. (2019) ‘Taking Parametric Assumptions Seriously: Arguments for the Use of Welch’s ’F’-test instead of the Classical ’F’-test in One-Way ANOVA’, International Review of Social Psychology, 32(1). doi:10.5334/irsp.198.
* Jomthanachai, S., Wong, W.P. and Khaw, K.W. (2022) ‘An application of machine learning regression to feature selection: a study of logistics performance and economic attribute’, Neural Computing & Applications, 34(18), pp. 15781–15805. doi:10.1007/s00521-022-07266-6.
* Eryk Lewinson (2020) Python for Finance Cookbook : Over 50 Recipes for Applying Modern Python Libraries to Financial Data Analysis. Birmingham, UK: Packt Publishing. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=e250xww&AN=2366460&site=eds-live> (Accessed: 26 May 2023). Doug Bierer (2020) Learn MongoDB 4.x : A Guide to Understanding MongoDB Development and Administration for NoSQL Developers. Birmingham: Packt Publishing. Available at: <https://search.ebscohost.com/login.aspx?direct=true&db=e250xww&AN=2621660&site=eds-live> (Accessed: 26 May 2023).
* Department of Public Expenditure and Reform (2022). *Build 2022: Construction Sector Performance and Capacity*. Available at: <https://www.gov.ie/pdf/?file=https://assets.gov.ie/229941/676ef592-29b8-4dd0-84b3-bfb0aa375ec7.pdf#page=null> (Accessed: 26 May 2023).