**Introduction**

As part of Project Ireland 2040 the National Development Plan sets out the Government’s over-arching investment strategy and budget for the period 2021-2030. It is an ambitious plan that balances the significant demand for public investment across all sectors and regions of Ireland with a major focus on improving the delivery of infrastructure projects to ensure speed of delivery and value for money (NDP Paper Ref). An essential component to ensure the delivery of the increased ambition set out in the National Development Plan 2021 – 2030 is an efficient, productive and sustainable construction sector. A build report was published in 2020 and it detailed a number of encouraging trends in the sector, such as increasing construction apprenticeships (BUILD 2022 Paper). Trends relating to the construction sector in Ireland will be analysed in this report along with international statistics to act as a comparison. Specifically, construction costs, labour input in the construction sector and construction sector productivity levels will be analysed. Other key economic indicators are also referenced for the purpose of enhancing the development of prediction models for the aforementioned construction sector statistics.

The structure of the paper is as follows; Data Extraction & Manipulation, Statistical Insights, Machine Learning Methods, Final Recommendations & Conclusions, and Limitations.

**Data Extraction & Manipulation**

Background

Raw data was extracted from multiple sources for the purpose of this research paper. The main source for the construction sector information was Eurostat. Key statistics related to each individual country were sourced from the relevant country’s statistical database. For the purpose of drawing international comparisons, the other countries included in this report are; United Kingdom, Germany, France, the Netherlands, the Eurozone (27 countries) and the Euro Area (20 countries). These four countries were selected for inclusion for comparative purposes as:

1. They have broadly comparable climate conditions to Ireland as mid/northern European countries (notwithstanding the south of France) and their building construction methodologies are broadly comparable also.
2. They are economically comparable on a GDP/GNP per capita basis to Ireland.
3. Their construction labour costs are broadly comparable to Ireland.

(Constr Costs Paper)

The Eurozone and Euro Area countries serve as an aggregate measure across the continent to compare against.

*Eurostat*

Eurostat acted as the primary source for the construction information. Eurostat is the statistical office of the European Union with the declared mission to provide high-quality statistics and data on Europe (Eurostat website). The datasets obtained from Eurostat are defined in the below table.

|  |  |  |
| --- | --- | --- |
| **Title** | **Definition** | **Link** |
| Production in construction - monthly data | Volume index of production (Index = 2015) | [Link here](https://ec.europa.eu/eurostat/databrowser/view/STS_COPR_M__custom_6124978/default/table?lang=en) |
| Construction producer prices or costs, new residential buildings - quarterly data | Output price index in construction - in national currency (Index = 2015) | [Link here](https://ec.europa.eu/eurostat/databrowser/view/STS_COPI_Q__custom_6125040/default/table?lang=en) |
| Labour input in construction - quarterly data | Employment (number of persons employed) (Index = 2015) | [Link here](https://ec.europa.eu/eurostat/databrowser/view/STS_COLB_Q__custom_6124890/default/table?lang=en) |

*Ireland*

The key economic indicators relating to Ireland came from a number of sources. The main source used was the online database for the Central Statistics Office (CSO). The CSO is Ireland’s national statistical office with the purpose to impartially collect, analyse and make available statistics about Ireland’s people, society and economy (CSO website). The Banking & Payments Federation Ireland (BPFI) statistics on mortgage drawdowns was also used. The datasets obtained relating to key Irish economic indicators are defined in the below table.

(**INSERT TABLE)**

An API data query on JSON-stat formatted data was also submitted on the ‘Indices of Total Production in Building and Construction (Base 2015=100)’ dataset. This was done as there was no information available for Ireland’s construction production field in the dataset obtained from Eurostat.

*United Kingdom*

*Germany*

*France*

*Netherlands*

*Eurozone (27 countries)*

*Euro Area (20 countries)*

Methodology

*Data Manipulation*

The datasets obtained and used in this paper required in depth manipulation and cleaning in order to form a robust analysis. Some of the problems encountered include; the actual data points not being in the first row of the relevant Excel sheet, very wide data (more columns than rows) with each column representing quarters for each year & uneven non-missing data for the countries included. In order to overcome these data deficiencies functions were written in Python and applied to data that were shaped in a similar format. For example a ‘clean\_eurostat\_df’ function was created which allowed for the manipulation of the datasets obtained from Eurostat in a format that is comprehensive in a Python environment for analysis. Steps included; removing the first tens rows, where information such as; time frequency, business trend indicator and unit of measure are included. This information is not relevant for the data points to be analysed.

A frequent issue with the time series data obtained for the purpose of this research paper is the misaligned time frequencies between datasets. The datasets for each country were decided to be in a quarterly format as this allows for monthly information to be easily converted, by taking each third month of the year. This manipulation was performed on multiple datasets (**GIVE EXAMPLES).**

Another frequent issue/decision to be made with the data obtained was to select a relevant starting point on which to begin analysis. The Eurostat tables had uneven starting points for each country’s statistics that were obtained. The decision was made to have the analysis originate from **XXX**, as this is when Ireland’s data entries began at this time point for most datasets included.

Exploratory Data Analysis for Final Datasets (Country Level)

Full final datasets were created for each of the aforementioned countries by joining on each of the relevant statics by the respective starting time points. This section summarises the structure of each final dataset constructed at a country level. Shape, summary statistics, relationships between variables and missing information for each of the country datasets is reported below.

*Ireland*

*United Kingdom*

*Germany*

*France*

*Netherlands*

*Eurozone (27 countries)*

*Euro Area (20 countries)*

**Statistical Insights**

**Machine Learning Methods**

The primary purpose of this section is to train supervised models to accurately predict different targets which comprise the datasets. Machine learning is a type of artificial intelligence whereby an algorithm or method extracts patterns from data (**Thoughtful**). Two machine learning methods are explored. One of the explored methods is a supervised learning machine learning algorithm utilised for the accurate prediction of the production levels in the construction sector in Ireland. According to (**AUTHOR NAME OF PRACTICAL ML)**, supervised learning is a category of algorithms classified under the supervised learning category focus on establishing a relationship between the input and output attributes, and use this relationship speculatively to generate an output for new input data points. The method used to build a predictive model to forecast future productivity levels in the construction sector is the Gradient Tree Boosting or Gradient Boosted Decision Trees (GBDT) method.

Supervised Learning Method

*Gradient Boosting*

Gradient boosting is described by Friedman, 1999 as the construction of additive regression models by sequentially fitting a sample parametrised function (base learner) to current ‘pseudo’-residuals by least-squares at each iteration. The ‘GradientBoostingRegression’ function in used from the ensemble library in scikit-learn. This estimator builds an additive model in a forward stage-wise fashion; it allows for the optimisation of arbitrary differentiable loss functions. In each stage a regression tree is fit on the negative gradient of the given loss function (**SCIKIT-LEARN).**