**MSc in Data Analytics – “Comparative Analysis of Construction Sectors in Ireland, Portugal and UK: A Data Mining Approach.”**

Omprakash Vootla

7 Parklands Row, Parklands

Saggart, Dublin 24 Ireland.

Email: [sbs23020@student.cct.ie](mailto:sbs23020@student.cct.ie)

Student Number: Sbs23020

***Abstract***

*Objective of this Data Mining project is to compare the Construction sector in Ireland and other countries. This study conducted on multiple datasets based on the topics, first Ireland dataset contains the information of average earnings per week in construction industry of various employment types, Second Ireland dataset contains the information of incidents in construction industry.*

*Third dataset contains the information of Number of employees per quarter in Portugal construction industry.*

*Fourth dataset contains the information of UK incidents in construction industry.*

*Approach is to apply EDA (Exploratory Data Analysis), data processing, data preparation and data visualisation to understand the earnings, employment, and incidents trends in construction industry over the time.*

*In EDA we renamed columns, removed unwanted columns, and fixed missing values, we explored Supervised and Unsupervised Machine learning Models.*

***Introduction***

Ireland is facing a housing crisis at this moment of time, there has been lot of construction going on all over the country to meet the housing requirement. This study is focused on finding some insights like employees Avg earnings per quarter over the period, type of employment compared to Portugal type of employment per quarter over the period. Also find some sentiment in Incidents in construction industry.

***Background***

Datasets are derived from Data.Gov.ie, CSO in the Construction domain. The study covers four components of data analytics: Data Preparation, Statistic Analysis, Machine learning algorithms and Data Visualization techniques.

Our aim is to analyse data using EDA, Data cleaning, Data preparation, and apply Machine learning models to predict target variable.

Ireland Average Earnings per Week Datasets has the labelled data, with one time series columns which is year and quarter. Regression models will be best fit for labelled data. We explore regression models like Linear Regression, Decision Tree, and Random Forest and Classification models like Logistic Regression, Decision Tree, Random Forest Classifier and Support vector machine models to compare the results and find best model.

To apply Machine learning models, we split the data set as train and test sets with 70%-30%. We evaluate the results by various methods like, MSE, R-Square, F1-Score, Accuracy and Confusion Matrix etc.

# **Scope and Methodology:**

For this study we are going to implement CRISP-DM Methodology. CRISP-DM is well-established and widely used for data mining and Machine learning projects. This framework consists of six phases.

Business Understanding, Data Understanding, Data Preparation, Modelling, Evaluation, Deployment.

Diagram

Description automatically generated

## 1.1 Business Understanding

Our first dataset is about Average Earnings of various employees in construction industry for a period of years 1969-2008. this dataset derived from publicly available on Central Statistics Office (CSO), Ireland.

Second dataset is about the Portugal Quarterly Employment in construction industry by gender. This dataset is derived from International Labour Organisation <https://ilostat.ilo.org/data/>

Third and fourth datasets are Ireland and UK workplace incidents in construction industry derived from <https://www.hse.gov.uk/statistics/fatals.htm> and <https://www.hsa.ie/> . these two data sets were used in sentiment analysis.

## 1.2 Data Understanding and Data Preparation

Jupyter Notebook: “1. Ireland Avg Earnings per Week - Data Cleaning, Data Preparation and Loading to MySQL”.

As first step we imported Pandas, NumPy, Seaborn, Matplotlib etc libraries.

We read “Earnings and Hours Worked in the Building and Construction Industry.csv” file as Pandas data frame.

This data frame has 160 rows, and 10 columns. With column name like “‘Unnamed: 0', 'All manual employee grades', 'Skilled operatives', 'Apprentices', 'Unskilled and semi-skilled - adults', 'Unskilled and semi-skilled - juveniles', 'Clerical employees - female', 'Clerical employees - male', 'Foremen and supervisors', 'All employees (clerical and operatives)'”.

Graphical user interface

Description automatically generated  
This data frame shows each variable data type as below, we have two float types and rest of them are object types.

A picture containing graphical user interface

Description automatically generated

Next step we renamed columns as python friendly. For example, renamed 'Unnamed: 0': 'Year\_Quarter' etc. than split the “Year\_Quarter” in to two new columns “Year” and “Quarter”. And dropped the original “Year\_Quarter”.

Found 39 missing values in few columns, which are less than 20% the total data, so we imputed them with function “ffill” and “backffill”.

Next step replace “,” “” , and converting all variable data type as integers.

We export cleaned data frame to MySQL data base table “avg\_earnings\_week”. First, we convert all rows into a list of tuples. And save the list as .txt file.

Next step, using pandas MySQL connector we connect to the database. And created a database and table with columns as exactly as in our data frame. Now log in to MySQL and connect to data base table and using insert statement. We load entire .txt file into the table. Now we can use this database table to user for our EDA and ML modelling.

We have implemented same steps to prepare the Portugal Quarterly Number of Employees in Construction Industry, Ireland, and UK workplace Incidents in Construction Industry datasets to use in Machine Learning Modelling and Sentiment Analysis.

## 1.3 EDA and Feature Selection

Jupyter Notebook: 2. EDA - Ireland-Avg Earnings per Week in Construction Industry.

We Connect to MySQL Database “ireland\_avg\_earnings\_week” and query for selecting the table where the cleaned data is stored from Jupyter Notebook 1.

Scatter plot is a graphical representation of data points. Plot indicates that there is spike in the average earning of “Manual\_employee\_grades” and “Clerical\_and\_Operatives” in years 1980-1985. Fallowed by “Semi skilled adults” and “Skilled operatives”. Over all high earnings for “Manual\_employee\_grades” and “Clerical\_and\_Operatives” for entire time periods.

Chart, histogram

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Graphical user interface, application, Word

Description automatically generated

Below Pair plot showing the linear relationship in most of the variables, except few in nonlinear relation between manual employ grades, skilled operatives, apprentices and clerical employee female, clerical employee male and foreman and supervisors.

A picture containing text

Description automatically generated

### 1.3.1 Features selection Using Pearson Correlation

In this step we explore and identify important features using Pearson correlation method. We take “Manual\_employee\_grades” as target variable and other features as independent variables. We split the dataset as train and test sets using “sklearn.model\_selection” “train\_test\_split” method. With test size 0.30. We apply correlation to train set and create heat map.

Graphical user interface, chart

Description automatically generated with medium confidence

This method identified 5 highly correlated features. “'Apprentices', 'Clerical\_and\_Operatives',

'Clerical\_employees\_male', 'Foremen\_and\_supervisors', 'Semi\_skilled\_adults'”. after removing them, two features are identified as important features from this method. “'Skilled\_operatives', 'Clerical\_employees\_female'”.

### 1.3.2 Features selection using Lasso regression method.

We implement lasso regression to identify important features, we import lasso from “sklearn.linear\_model” and fit the model with alpha 0.020. this model identified 7 important features “'Skilled\_operatives', 'Apprentices', 'Semi\_skilled\_adults', 'Clerical\_employees\_female', 'Clerical\_employees\_male', 'Foremen\_and\_supervisors', 'Clerical\_and\_Operatives'”.

### 1.3.3 Outliers detection using IQR Method.

We identified outliers IQR method and visualise them as box plot as below.

Chart, box and whisker chart

Description automatically generated

There are few outliers in Manual employee grades, Skilled Operatives, Apprentices, Semi-skilled adults and Clerical and Operatives. No outliers found others. We can understand as construction industry is more adults and male workers domain. Some time they need to work overtime, so they get paid overtime payment as double. We found them 14 data point overall, after removing those outliers. Store the dataset with new name “df1\_no\_outlier”.