**Introduction**

Understanding the different elements that affect teachers' and educators' compensation in these nations is necessary for a sophisticated examination of salaries in the education sectors of Ireland, Sweden, and Canada. The degree of education, years of experience, kind of institution (public or private), as well as the economic climate and government policies of the nation, all have an impact on the remuneration in the education sector.

In Ireland, a teacher's salary is impacted by their education and years of experience. The beginning pay for primary teachers is often less than that of secondary teachers. Teachers' salary in Sweden is based on their credentials, the degree of education they teach, and the number of years of experience. The educational system in Canada is diverse, and the wages of teachers can differ greatly between provinces and territories. Teachers in Canada typically earn competitive salaries, and their income is influenced by things including experience, education, and local cost of living.

The real worth of salaries may fluctuate due to variations in living expenditures, so it is important to take these aspects into consideration when comparing incomes in these nations. Additionally, it's crucial to consider non-financial aspects like perks, work-life balance, and job satisfaction as these have an impact on how appealing the teaching profession is globally.

The contribution and the purpose of this project is

1. To Analyze the Salary comparison for education sector in Ireland Sweden and Canada
2. Utilize python to create dashboard and further statistics towards understanding the class correlations.

**Data Preparation & Preprocessing**

The data has been retrieved from various country specific datasets and Irelands jobs and compensation value data available.

A table with numbers and text

Description automatically generated

For a study to be reliable and useful, data preparation is essential. To ensure data quality and dependability, it entails cleaning, converting, and organizing data to remove errors and inconsistencies. This process ultimately results in more reliable insights and well-informed decision-making. For EDA, missing value imputation is done for prepare the raw data.

Because real-world datasets frequently contain incomplete information because of numerous factors such data collecting failures or participant non-response, missing value imputation is essential. Missing values can cause biased analyses and incorrect findings if they are not considered. Imputation methods assist close these gaps, maintaining the accuracy of the data and boosting statistical power. A more thorough depiction of the underlying patterns is ensured by properly imputed data, enabling more robust modelling, lowering uncertainty, and delivering accurate results, ultimately assisting in improved decision-making and more insightful data analysis. For this project, the missing numerical values have been imputed by median and categorical values by mode as shown below.

For the data below, the missing value imputation has been performed.

A number and a symbol

Description automatically generated with medium confidence

A white background with black text

Description automatically generated

In addition to that, since the data is time-series in nature, the library of darts has also been installed. The open-source Python module Darts is used for time series forecasting and analysis. Researchers, data scientists, and analysts can undertake a variety of time series-related jobs more easily thanks to the vast range of functions and tools it offers to work with time series data. The following image showcases the visualization of the data.

A graph showing the growth of the stock market

Description automatically generated

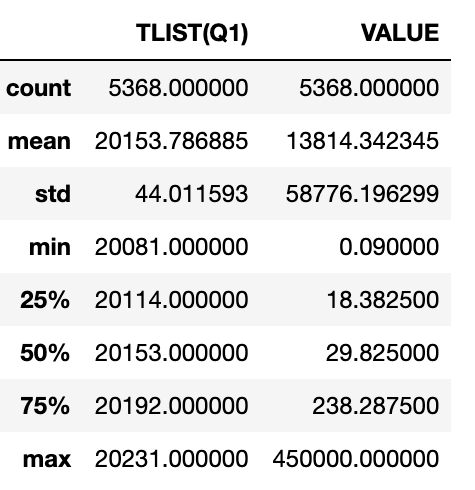
Here, you can see that Education salaries have stagnated since the last recession. We can also see that data is time-series in nature.

Along with this, mix\_max scalar and standard scalar has also been used. Data normalization requires the use of the Min-Max scaler, which reduces characteristics to a particular range (often [0, 1]). It ensures that no variable dominates others due to differences in scale and maintains the relative relationships between data points. For machine learning algorithms sensitive to feature magnitudes, this preprocessing step is essential because it promotes fair comparisons, improved model convergence, and enhanced model performance. StandardScaler is a data preparation method that transforms features by dividing by the standard deviation and subtracting the mean. By standardizing data to have a zero mean and unit variance, features can be compared on an equal footing. By ensuring that features with higher variances do not dominate the model, normalization speeds up convergence and improves algorithm performance. For machine learning models like SVMs, Adaboost, KNNs, and neural networks where feature scaling greatly affects performance, StandardScaler is especially helpful.

**Statistics For Data Analytics**

* 1. Data

The data has been described which showcases the values of the salaries in quarter 1.



* 1. Statistical Analysis

The library of Scipy.stats has been used. The Scipy library's Scipy.stats module is essential for scientific computing and data processing. To do hypothesis testing, probability calculations, and parameter estimation, it provides a thorough collection of statistical functions and distributions. This library offers a solid foundation for carrying out challenging statistical analyses, allowing users to confidently validate hypotheses, get deeper insights from their data, and make decisions based on facts. It is a vital tool in the field of statistical analysis and study.

* + 1. T-Test

The t-test is essential for comparing salaries in the education sector across Ireland, Sweden, and Canada. By conducting t-tests between each pair of countries, researchers can determine if significant salary differences exist and which country's education sector stands out in terms of remuneration. This statistical significance assessment is crucial before applying more complex algorithms like linear regression, decision trees, or AdaBoost.

The t-test helps validate assumptions and provides a solid foundation for subsequent analyses. Once significant differences are confirmed, more sophisticated techniques like linear regression can be used to explore the relationship between education level and salary, accounting for country-specific effects. Decision trees and AdaBoost can then delve into complex interactions between multiple factors influencing salaries in each country. Overall, the t-test serves as a vital first step, ensuring reliable and meaningful comparisons of salary data among the three countries' education sectors.

The hypothesis is the following,

Null Hypothesis (H0): The means of two independent samples are equal.

Alternative Hypothesis (Ha): The means of two independent samples are not equal.

A number with numbers on it

Description automatically generated with medium confidence

Here, we can see that since t-test is greater than p-value, we reject the null-hypothesis hence showcasing that the samples are independent from one another.

* + 1. Chi-Square Test

When comparing categorical variables, such as wage ranges or employment positions, in the education sector across Ireland, Sweden, and Canada, the chi-square test is crucial. Researchers can ascertain whether there are significant relationships between pay scales and nations by using the chi-square test, which also reveals whether the distribution of salaries changes considerably between the three regions.

With the help of this statistical analysis, patterns and differences in salary distributions can be found, emphasising any areas where there may be regional differences. When working with non-numerical data, it complements other statistical methods like t-tests or regression analysis. The chi-square test offers a more thorough knowledge of the elements that contribute to wage discrepancies by revealing insights into the salary landscape across the education sectors in each nation.

The hypothesis for chi-square test is

Null Hypothesis (H0): The null hypothesis states that there is no significant association between the two categorical variables.

Alternative Hypothesis (Ha): The alternative hypothesis states that there is a significant association between the two categorical variables.

A black text on a white background

Description automatically generated

Here we can see that since test statistic is greater than p-value, we reject the null hypothesis showcasing that there is significant association between two categorical variables.

* + 1. ANNOVA

When working with several groups or more than two categories, ANOVA (Analysis of Variance) is crucial for comparing earnings in the education sector across Ireland, Sweden, and Canada. ANOVA enables researchers to compare salaries across all three countries at once, unlike the t-test, which only analyses two groups at a time.

Researchers can find out if there are statistically significant differences in average incomes across the three regions by using ANOVA. This test determines whether country's education sector outperforms others in terms of pay and whether variations are more likely the result of chance or actual inequities. When analysing pay differences, ANOVA offers a strong statistical framework that takes into account both within-group and between-group variability. Policymakers and stakeholders can resolve pay disparities and promote a more equal education sector by taking data-driven actions after understanding these variations. The addition of ANOVA to other statistical techniques improves the thoroughness and precision of analyses of wage comparisons, which eventually results in more sensible and equitable salary policies and practises.

The hypothesis is the following:

Null Hypothesis (H0): The means of all groups are equal.

Alternative Hypothesis (Ha): At least one group mean is significantly different from the others.

A number with numbers on it

Description automatically generated

Since F-statistic is greater than p-value, we reject the null hypothesis and that there is at least one group that is significantly different than the others.

* + 1. Pair-Wise Turkey Test

A useful statistical method for comparing salaries in the education sector across several nations, including Ireland, Sweden, and Canada, is the pairwise Tukey test. The Tukey test can be used to identify particular pairwise differences after an ANOVA has been used to determine whether there are significant earnings differences among these nations.

The Tukey test identifies whether nations have significantly different wages by comparing all combinations of national means. Understanding the distinctive qualities and differences in these countries' education sector profits is vital. Researchers can identify the range in which significant differences exist because the test gives a critical value based on the selected significance level. With the use of this data, stakeholders and policymakers can create targeted plans to increase pay in the education sector across these nations and solve any potential wage gaps.

A white paper with black text

Description automatically generated

* + 1. Normality Test

If the data is normally distributed, it allows researchers to use parametric statistical techniques, such as t-tests or ANOVA, with greater confidence. On the other hand, if the data deviates significantly from normality, non-parametric methods may be more appropriate. By conducting a normality test, researchers can ensure the validity of their statistical analyses and make informed decisions on which methods to employ for salary comparison. This ensures that any findings or conclusions drawn from the data are based on solid statistical grounds, enhancing the reliability and credibility of the study.

Researchers can more confidently utilise parametric statistical methods like t-tests or ANOVA if the data is regularly distributed. On the other hand, non-parametric approaches might be more appropriate if the data considerably deviates from normality.

Researchers can validate the accuracy of their statistical analyses and decide which techniques to use for wage comparison by doing a normalcy test. This increases the study's credibility and reliability by ensuring that any findings or conclusions taken from the data are supported by reliable statistical evidence.

**Machine Learning Tools**

* 1. Machine Learning Algorithms

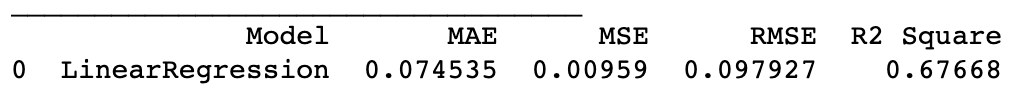
The project uses Linear Regression, Decision Tree and Adaboost.

1.1.1 Linear Regression

When applying linear regression for salary comparison in the education sector across Ireland, Sweden, and Canada in a time series context, the analysis becomes more intricate. Time series linear regression aims to model salary variations over time within each country separately or jointly.

Researchers can use time as an independent variable to capture the temporal patterns in salary changes. This allows for the identification of trends and seasonality, indicating if salaries in the education sector are increasing or decreasing over time and whether there are recurring patterns in salary fluctuations.

Time series linear regression provides insights into the long-term salary trends and how they differ among the three countries. By understanding these dynamics, policymakers can devise strategies to address salary fluctuations over time, considering country-specific factors and ensuring a more stable and equitable education sector. However, it's essential to account for potential autocorrelation and stationarity issues in the data for reliable and accurate time series regression results.

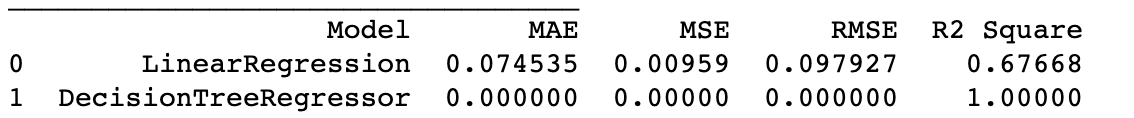


In this project, LR is used to analyses the models performance in forecasting education salaries across countries.

1.1.2 Decision Tree

The methodology becomes useful for capturing complicated temporal interactions when decision trees are applied for salary comparison in the education sector across Ireland, Sweden, and Canada in a time series environment. Decision trees are capable of managing non-linear relationships and automatically spotting linkages between shifts in time and pay.

Decision trees divide the data into subsets based on time and other pertinent criteria, such as country-specific variables, by creating a tree-like structure. This makes it possible to identify specific salary patterns and trends across time, giving us a thorough grasp of how salaries change in the education sector of each nation. The identification of time-dependent elements that affect compensation discrepancies is made easier in time series using decision tree analysis. In order to promote a more equitable and sustainable education sector in each nation, this information aids policymakers and educators in developing targeted initiatives to overcome variations in compensation over time. Deeper and more complicated decision trees, however, may provide interpretation issues that call for thorough examination and consideration of the findings.



In this project, DT is used to analyses the models performance in forecasting education salaries across countries. It is seen that DT performs better.

1.1.3 Adaboost

The methodology becomes effective for improving predicted accuracy over time when used for salary comparison in the education industry across Ireland, Sweden, and Canada in a time series context. AdaBoost is an ensemble learning technique that combines a number of weak learners (usually decision trees) to produce a reliable and precise model.

AdaBoost's ability to capture complicated temporal correlations in a time series context makes it possible to identify key characteristics and the effects they have on changes in income over time. AdaBoost concentrates on difficult time points that might have a large impact on wage fluctuations by iteratively increasing the weights of misclassified cases.

Researchers can learn more about how time interacts with other nation-specific elements to affect salaries in the education sector by using AdaBoost. AdaBoost's improved predictive capability makes it possible to forecast salaries more accurately, which is essential for each nation's resource allocation and policy planning. To avoid overfitting and guarantee reliable model performance, as with any ensemble technique, rigorous parameter tweaking and validation are required.

A number on a white background

Description automatically generated

In this project, Adaboost is used to analyses the model’s performance in forecasting education salaries across countries. It is still that DT still performs better.

1.2 Training and Running the Algorithm

The ML models are trained on GPU tX2 model and running the models in in distributed space with 14 threads and 4 iterative points.

* 1. Hyperparamter Tuning

Finding the ideal set of hyperparameters for a machine learning model is referred to as hyperparameter tweaking or hyperparameter optimisation. Hyperparameters are settings that must be made before the training process starts but are not learned during training.

By choosing hyperparameter values that produce the best results on a validation dataset, hyperparameter tweaking aims to increase the model's performance. Grid search, random search, Bayesian optimisation, and genetic algorithms are frequently used in this procedure to efficiently explore various hyperparameter setups.

In this project, Gridsearch CV has been used, where Cross-validation is used to discover the optimal model performance using the Grid Search CV hyperparameter tuning technique, which systematically searches through a predefined grid of hyperparameter combinations.

A close-up of a number

Description automatically generated

* 1. Crossvalidation

A resampling method called cross-validation is used to assess how well a machine learning model is working. The dataset is divided into various subsets (folds), some of which are used to train the model and others of which are used to test it. To produce a more solid and trustworthy estimate of the model's generalisation ability, this process is performed numerous times. The model's performance is then averaged across all iterations. Cross-validation aids in evaluating a model's performance, guards against overfitting, and helps choose the optimal model for the available data.

In this project, ‘sklearn’ library has been used, to transform data, preprocess and achieve ML results.

A computer screen shot of a number

Description automatically generated

**Visualisation and Dashboard**

The following are the dashboards using three algorithms

A graph showing a line graph

Description automatically generated with medium confidenceA graph showing a line graph

Description automatically generated with medium confidence

**Programming**

* 1. Programming – The paper uses python, numpy, tensorflow and keras moduls as the pretext of the analysis.
  2. Data Structure – The paper uses data frame and pandas list objects to process and transform the data.
  3. Documentation – General documentation is followed
  4. Testing & Optimisation – The dataset is split across 20% set for testing and validation

**Conclusions**

It is seen that for the purpose of forecasting and analysing salaries in the education sector using time-series data, Decision Tree performs better with an impressive R^2 of 1. The following the results shown.

A screenshot of a calculator

Description automatically generated