Prediction of Unemployment Rate using Machine Learning

Balasubramaniym Muthuramalingam(L00157060)
Letterkenny Institute of Technology.

Abstract

Over the past decade, Machine Learning (ML) has been widely used in different domains especially in Economy and Finance, for example, analyzing public data to determine unemployment rate, loan underwriting and GDP calculation etc... Prediction of unemployment rate for incoming years differentiated by gender. Now, ML plays a vital role in providing an accurate result in order to understand the unemployment rate. Using various Regression techniques and Classification techniques a model is built that can provide promising results for the country or public organization to evaluate the economy and financial state of the country. The following algorithms are applied; Linear Regression, Decision Trees, Polynomial Regression and KNN Regression. Analyzing the dataset and pre-processing the data to apply different models and comparing the value with other models to show the difference between each other. The prediction accuracy is determined by finding the R square value for each model and the value should be lesser than 1 but greater than 0, i.e., the value is considered to be fit if it is closer 1, if not, the applied model is not compatible with the dataset. The behavior of each model is recorded and compared for further study.

Introduction

2.1 Unemployment

Unemployment means joblessness, it prevails when people are without a job and actively seeking any sort of employment. It is mathematically calculated as a percentage by dividing the number of unemployed individuals by the number of all the individuals currently employed in the workforce. When unemployment rates are high in number and steady, there are negative impacts on the long-run economic growth.

Unemployment rate can determine the financial crisis involved in a country, constant increase in unemployment can affect the wealth overall nation and creates a negative impact on world trade. Based on historical data of unemployment, Machine Learning (ML) will help to understand the problem and provide a better accurate model and breakdown the complex implications involved within datasets. Every technique involved in this model helps us to identify missing pieces of information through constant iteration thus improving the level of accuracy.

2.2 Machine Learning

Machine learning (ML) was developed in 1959 by Arthur Samuel and it is a subset of Artificial Intelligence. ML can be described as a computer trying to learn from experience. Experience is accumulated by users feeding the data to the machine with predefined algorithms specifically designed to understand certain problems and providing a result relevant to the problem but the accuracy of the

result won't be correct or adequate thus by constant testing and tweaking of mathematical models it will provide improved accuracy. Hence through experience the machine tries to adapt to the problem based on the input provided. The learning process is also called training data. ML is widely used in many applications because there is no need to add or modify code explicitly every time for different sets of inputs. ML is under constant development and it is yet to reach every domain and applications. It still shows potential areas to be explored and widening its scope of usage. Applying ML in finance and economy sectors provided positive feedback over the years and hence creating more opportunities. Providing accurate rates of unemployment based on different sexes using machine learning has been widely implemented in various countries and organizations due to its fast computational method and accuracy. Even though ML does constitute a negative impact on certain scenarios like insufficient data (suitable data) leading to inaccurate results. Hence the objective is to analyze and evaluate the techniques of machine learning involved in calculating rate. It also seeks to explore every aspect of the topic to show there is potential growth in increase or decrease in rate using machine learning.

Literature Review

A paper published by Christos Katris on Prediction of Unemployment Rates using Time stamps and Machine Learning from Athens University of Business and Economics, showed the prediction of rate of unemployment in several countries. Model used in exploring and analyzing the data is FARIMA because of long memory that exists in a time series and has been successfully applied in predicting unemployment. The paper explained how to overcome the issue faced because of heteroskedasticity, further exploration is done by using FARIMA with GARCH errors to achieve more accurate results. Three machine learning techniques have been applied to forecast unemployment rates, fully connected feed forward neural networks, support vector regression and multivariate adaptive regression splines.

Another paper reviewed by Tapio Pahikalla on Prediction of Unemployment using Machine Learning on Registry Data, the overview of the paper is to predict the labour market state of a person based on machine learning trained with a large administrative unemployment registry. Every individual is specified as Markov chains with person specific transition rates. Three aspects are considered while predicting the rate, escaping unemployment, becoming unemployed and being unemployed for a long period. Final objective is to show significant differences that can be learned by utilizing labour market histories.

Datasets

Datasets are contents of information stored in a tabular form. It is represented in .csv format. In this model one dataset is used and split into two categories, one is used for training the model and other is used to test the model. The training dataset is used to check if it is compatible with the respective algorithms, Decision tree and Linear Regression, Polynomial Regression and Naïve Bayes. All the four techniques are compared with each other to list down the differences and similarities among them.

Important links:

https://data.gov.ie/dataset/qlf02-ilo-participation-and-unemployment-rates-by-quarter-sex-and-statistic

Training Dataset

In this dataset, the model is trained. The raw file is uploaded and analysed to understand the contents, total size of the data is 95 entries and further categorized into rows and columns (91 rows and 4 columns). There are 7 classes of different data types. 3 Float 64 types and 1 Object types.

	Year	Both sexes	Male	Female
0	1998Q1	59.5	72.7	46.8
1	1998Q2	59.6	72.9	46.8
2	1998Q3	61.3	74.6	48.3
3	1998Q4	59.5	72.8	46.6
4	1999Q1	60.4	73.2	48.0

```
Year          object
Both sexes    float64
Male          float64
Female        float64
dtype: object
```

The above snippet is a sample dataframe from “train” dataset which is recorded from the year 1998-2020.

The following attributes in the dataset are

Year : Represent year with quarterly

Both Sexes : Rate of unemployment for both the gender

Male : Rate of unemployment for male category

Female : Rate of unemployment for female category

```
(91, 4)
```

Snippet of total number of elements in the dataset categorized by rows and columns.

Methodology

5.1 Overview

Machine learning (ML) is a mathematical model that tries to understand the algorithm, so when machine learning is applied to predict the rate of unemployment, it is important to determine the type of model needed to be used. Evaluate the dataset by using the Classification model. Under this model we shall use compatible techniques like Decision tree, Linear Regression, KNN Regression and Polynomial Regression. These techniques on implementing will help to evaluate the data and provide accurate results. Building a ML model has several steps involved and after the construction, improving the model also involves multiple steps. Once the model is built, datasets are fed into the machine to evaluate them. Datasets are sets of information with relevant details that are suitable to the model while evaluating. They are often shown in tabular format and the extension of the file is .csv. Once they are fed, it is then analyzed and then evaluated by the techniques implemented (training the model). The

model may take some time to evaluate the data to provide a result. Each phase of testing helps the machine to understand the objective and thus improve its accuracy.

Data Collection ⇔ Data Modelling⇔ Deployment

Collection of data (datasets) as your input and modelling of data which comprises several steps like problem definition, data, evaluation and experimentation (testing of model). Finally deploying the model after constant testing. Once the framework is established, we need to find out the correct model, in this topic as mentioned above Classification model is the suitable approach to solve the problem.

Problem definition->Data->Evaluation->Modelling->Experimentation

5.2 Pre-Processing Data

This process is an iterative process and it is carried out until the machine understands the objective and provides an accurate result. Problem definition or type of problem, it is a Classification problem or Classification model. Type of data involved is Structured data and evaluation is the precision of the result (how accurate the result should be). Modelling has three different parts, Choosing and training a model, Tuning the model and Model comparison. All these steps are crucial during the experimental phase as these parts tweak the techniques implemented to provide positive feedback while evaluating.

5.3 Outlier Analysis

Values that exceeds the defined range are dropped along with the missing values. This phase is called the missing value analysis, where the dataset is observed to find entries that does not enclose within the defined range. All the classes are analysed to check for NaN values and out of range values, the missing value percentage will increase if all the out of range values are defined as NaN which will lower the standard of dataset so to maintain the percentage of the missing values, out of range values are marked as outliers and removed using is.null() method.

```
Both sexes    0
Male          0
Female        0
Years         0
Quaterly      0
dtype: int64
```

5.4 Data Manipulation

The dataset attribute “Year” has two different values, year and quarterly, so the column is split into two different columns “Years” and “Quarterly”, then the datatype of both columns is converted to int64.

```
Both sexes    float64
Male          float64
Female        float64
Years           int64
Quaterly        int64
dtype: object
```

	Both sexes	Male	Female	Years	Quaterly
0	59.5	72.7	46.8	1998	1
1	59.6	72.9	46.8	1998	2
2	61.3	74.6	48.3	1998	3
3	59.5	72.8	46.6	1998	4
4	60.4	73.2	48.0	1999	1

As shown, the old column “Year” has been dropped.

5.5 Data Splitting

The “train” dataset is split into x_train y_train and x_test and y_test. 80% of the data is for training and the remaining 20% is for testing.

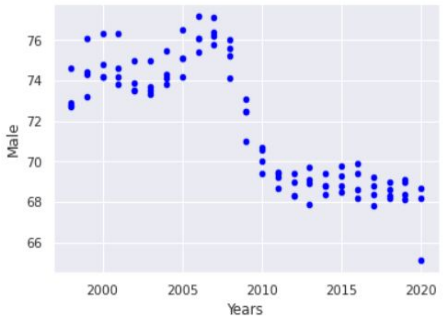
```
# Splitting the dataset and applying the model
x = train['Years']
y= train['Both sexes']
x_train, x_test, y_train, y_test = train_test_split(x,y,test_size=0.2)
```

```
x_train = x_train.values.reshape(-1, 1)
x_test = x_test.values.reshape(-1, 1)
```

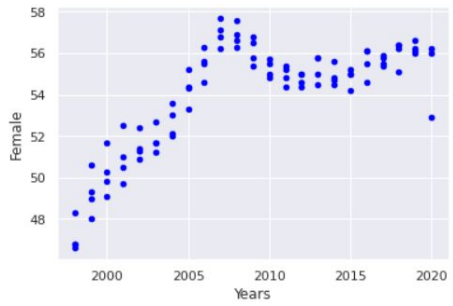
5.6 Data Visualization

The data is represented in graph format, this is used for analytical and comparing purposes. Firstly, the relationship between rate of employment for a specific gender is associated with year.

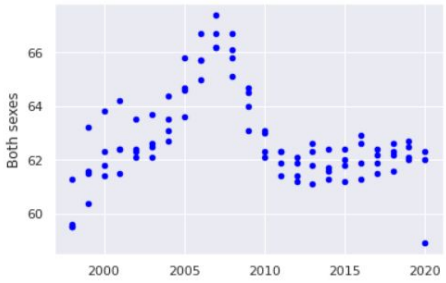
Taking X as year and Y as Rate of unemployment for male and plotting a scatter type graph



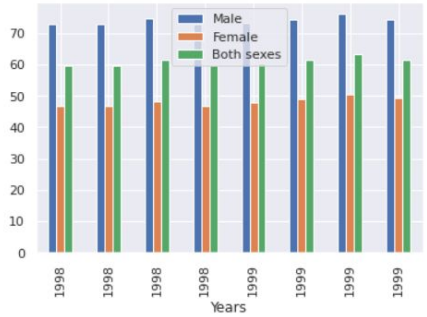
Taking X as year and Y as Rate of unemployment for Female and plotting a scatter type graph



Taking X as year and Y as Rate of unemployment for Both Sexes and plotting a scatter type graph



Showing a representation of all the gender grouped together compared with respect to the year 1998 to 1999.



Algorithm Used

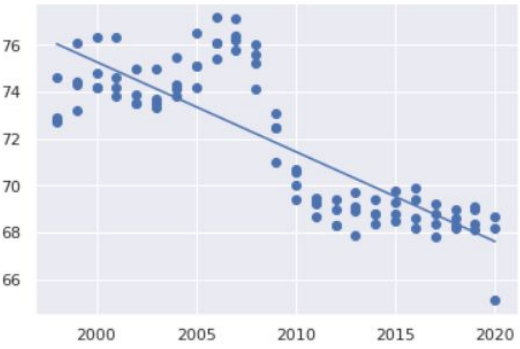
6.1 Linear Regression

Linear regression is a statistical model. It analyzes the linear relationship between two variables with a given set of dependent variables and independent variables. Linear relationship between variables is defined as the value of one or more independent variables will increase or decrease, the value of dependent variable will also increase or decrease accordingly.

It can also be represented in mathematical terms : Y = mX+ b.

The linear relationship between two variables can be positive or negative.

Here, two attributes are assigned to X and Y coordinates, X as Years and Y as Male. A function is defined where the slope and intercept for the model is calculated. A graph is plotted for the following



R value is the coefficient value which lies between 0 and 1 regardless of its sign.

```
-0.8174840624518861
```

Since the value is closer to 1[negative] it is not a perfect fit but linear regression can be applied between two variables for prediction of values.

```
male_prediction = calc(2020)
print(male_prediction)
```

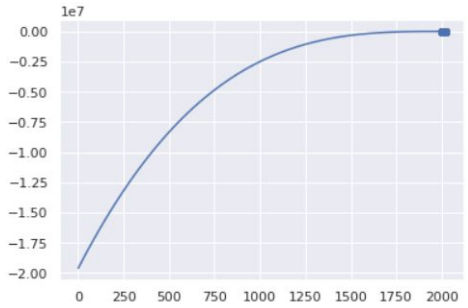
```
67.61558073654385
```

Predicting value for the male gender in the year 2020 and printing it.

6.2 Polynomial Regression

It is another subset of linear regression in which the relationship between two variables one is independent variable x and another one is dependent variable y is computed as an Nth degree polynomial. Polynomial regression attaches a nonlinear relationship between the two different variables, value of x and the corresponding conditional mean of y.

In polynomial regression assigning Female attributes for Y and Years for X, creating a graphical plot for both the coordinates using polynomial regression method. Then calculating the R value for the polynomial model to check if the model is best fit or worst fit for the dataset, the value resides between 0 and 1 , the closer to 1 is better the accuracy of the model



R value

```
0.8565594226301781
```

Prediction of Female unemployment rate for the year 2021

```
female_prediction = poly_model(2021)
print(female_prediction)
```

```
56.05677451938391
```

6.3 Decision Tree

Decision tree (DT) is a predictive approach and it is constructed through an algorithmic approach that can split the data into two parts. Each part holds a condition and for each condition it can result either true or false. DT falls under supervised learning and it is quite efficient when applied in evaluating credit score. The decision tree has a parent node and each node has their own branches or child nodes. Each node can result in two branch nodes, and the result can be either true or false for the problem.

After splitting the data, create a variable where it can store the decision tree regressor with a fit method that captures both x_train and y_train as it's argument.

```
# Prediction on train data

prediction_train_dt = deci_tree_model.predict(x_train)

# Prediction on test data

prediction_test_dt = deci_tree_model.predict(x_test)
```

The dataset is compatible with the applied model, using the entries R squared values were determined for both test and training dataset. The resulted value for both the datasets were below 1 for random forest and decision tree. Mean squared error and Mean absolute error are calculated to evaluate the performance on the dataset before cleaned and processed.

Mean Squared Error MSE, Mean Absolute Error MAE

R square for both test and train dataset. R squared value is calculated after finding the mean squared error and mean absolute error for both the datasets.

```
Mean Squared Error for Train data = 1.0966825856605276
Mean Absolute Error for Train data = 0.7759749455337687
Mean Squared Error for Test data = 1.8700031214614719
Mean Absolute Error for Test data = 1.0503405572755464
```

R value for the model using training dataset[Both sexes]

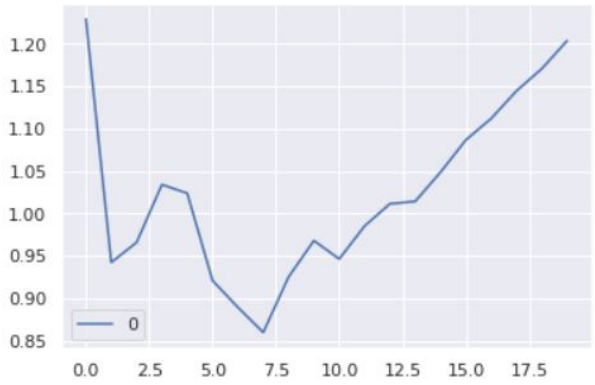
```
0.6120160508350624
```

R value for the model using the test dataset[Both sexes]

```
0.4383300383995419
```

6.4 KNN Regressor

K-nearest neighbors algorithm falls under the type of supervised algorithm which can be used for both classification as well as regression predictive problems. However, it is widely used for classification problems. Determining the K value for the knn regressor



Choosing 7 as the optimal k value because at k= 7, the RMSE is approximately .88, and shoots up on further increasing the k value. It is optimal to choose k=7 as it can provide the best suited result. Now finding the mean square error and R value.

```
The Mean square error is 0.7910633727175099 & the Error is 0.8894174344578084
```

Result and Conclusion

The R squared values for all 4 models were similar to each other which shows the dataset completely fits with the model. Since the size of the data is small and it contains very low attributes it was simpler to predict the unemployment rate for different gender and groups. Finally, tuning the model using a rigid model and using grid search CV for optimization. Before tuning the model using rigid , the parameters have to be defined by the user.

```
model_ridge.get_params
```

```
<bound method BaseEstimator.get_params of Ridge(alpha=1.0, copy_X=True, fit_intercept=True, max_iter=None,
      normalize=False, random_state=None, solver='auto', tol=0.001)>
```

As shown the user can define the parameters by changing the boolean conditions for arguments provided. Using the grid search CV , the R mean squared error is determined.

```
RMSE: 1.8709370492996946
```

The value for RMSE concludes that the model is not fit for the dataset after tuning, the predicted value range will exceed the expectation resulting in inaccurate results.

Reference

1. Sune Karlsson Farrukh Javed June, 2016,Forecasting the Unemployment of Med Counties using Time Series and Neural Network models. Örebro University School of Business.

2. Christos Katris,2020.Prediction of Unemployment Rates with Time Series and Machine Learning Techniques. Athens University of Economics.

3. Tapio Pahikkala and Markus Viljanen,2020, Predicting Unemployment with Machine Learning Based on Registry Data. University of Turku.
4. Tuning Model Link, https://alfurka.github.io/2018-11-18-grid-search/
5. GitHub algorithms, https://github.com/TheAlgorithms/Python/tree/master/machine_learning
6. W3schools link, https://www.w3schools.com/python/python_ml_getting_started.asp
7. Cheng Cheng, Tingting Zheng.2013, March.Data mining for unemployment rate prediction using search engine query data. Zhejiang University of Technology.