CMP310 Fundamentals of Machine Learning Project Report

Nikos Ledakis Engonopoulos

Introduction

The dataset used for this project contained data on the World Population categorised by country with measurements from 1970 up to 2022. These measurements taken in all countries of the world allowed to increasingly rank countries within the dataset based on the amount of population they had. This dataset was firstly chosen due to its ability to allow its user to visualise the population growth within the measured years in a fascinating manner while also for its significance in the planning of humanity's future where using a machine learning model and the above-mentioned dataset, one could accurately predict the population of the world or even specific countries in the following years.

Based on the research that was done for this dataset it has been used in the past by multiple researchers. Either for visualisation processes (Widiatama, 2022) allowing the researcher to demonstrate the vast population of certain countries compared to other countries or even the rest of the world. Or for data analysis purposes calculating a different data point, like the calculation of GDP per country (Shopulatov, 2023). However, no previous research was found on applying a machine learning model to base predictions on the world's population in future years.

Once the above-mentioned research it was determined that, since the dataset has as outputs a multitude of numerical values, the best predictive model to be used would be a regression model. Following that certain tasks will have to be performed on the data to determine the ideal model to predict population growth in future years correctly. For example, determine the correlation between the data in the dataset or determine if any missing values are in the dataset. Once the most fitting model is found it will be trained using a percentage of the dataset and will be tested with the rest to determine its performance. Finally, when the ideal performance is found the model would have the capability to accurately approximate the world population in future years. This process can be repeated to fit more specific scenarios such as the population of continents or even of countries. This project will elaborate further on the above-mentioned process with further detail in the following chapters.

Data Specification

"World Population Dataset" (Banerjee, 2022) is the dataset used for this experiment. It was also chosen because the dataset is created in such a way that can be efficiently manipulated, and the data is expected to update annually thus allowing for the application to be run again in the following years with additional data. The dataset consists of a total of 17 columns (Fig. 1.1). To begin with 8 of those columns are each country's populations from the years 1970 to 2022 storing integer values. While, the rest, consists of either columns tied to the countries identification information, such as country name, capital, continent and three-letter abbreviation of the name, or additional information on the countries namely, total country area, density, growth rate, percentage of the total population and the country's ranking compared to the rest of the world. These columns contain a range of values stored in the form of objects or floats. The first step of Exploratory Data Analysis (EDA) allowed to determine that the dataset did not contain any null or duplicate values. Following that step the correlation between columns had to be visualised, which was achieved through the use of a heatmap (Fig. 1.2). As seen in Figure 1.2 the heatmap demonstrates high positive correlation between the following columns, all the population numbers within the recorded years and the "World Population Percentage" (WPP) column. Continuing with the EDA, pair plots between all the numerical columns against the 2022 population were used and allowed to determine the type of correlation between related columns, while also validating the findings of the heatmap. Figure 3 (Fig. 1.3) is a pair plot between the countries' populations in 2022 against 1970 and demonstrates a positive linear correlation that was consistent throughout all the other plots containing population numbers through the years. Whereas the 2022-WPP plot (Fig. 1.4) demonstrated a one-to-one relationship. This meant that one was derived from the other and thus would have no use in our predictive model. Finally, the rest of the pair plots validated the findings of the heat map showing no or negative correlation with the rest of the columns.

```
Data columns (total 17 columns):
 #   Column                        Non-Null Count  Dtype
---  ------                        --------------  -----
 0   Rank                          234 non-null    int64
 1   CCA3                          234 non-null    object
 2   Country/Territory             234 non-null    object
 3   Capital                       234 non-null    object
 4   Continent                     234 non-null    object
 5   2022 Population               234 non-null    int64
 6   2020 Population               234 non-null    int64
 7   2015 Population               234 non-null    int64
 8   2010 Population               234 non-null    int64
 9   2000 Population               234 non-null    int64
 10  1990 Population               234 non-null    int64
 11  1980 Population               234 non-null    int64
 12  1970 Population               234 non-null    int64
 13  Area (km²)                    234 non-null    int64
 14  Density (per km²)             234 non-null    float64
 15  Growth Rate                   234 non-null    float64
 16  World Population Percentage   234 non-null    float64
dtypes: float64(3), int64(10), object(4)
memory usage: 31.2+ KB
```

Fig. 1.1 Data Description

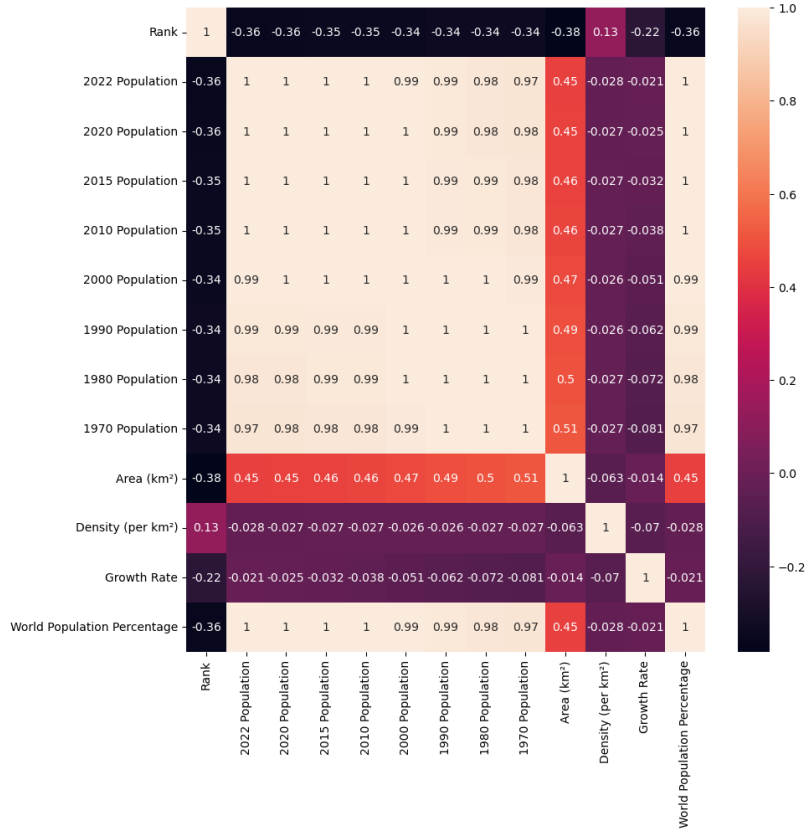

Fig. 1.2 Heatmap

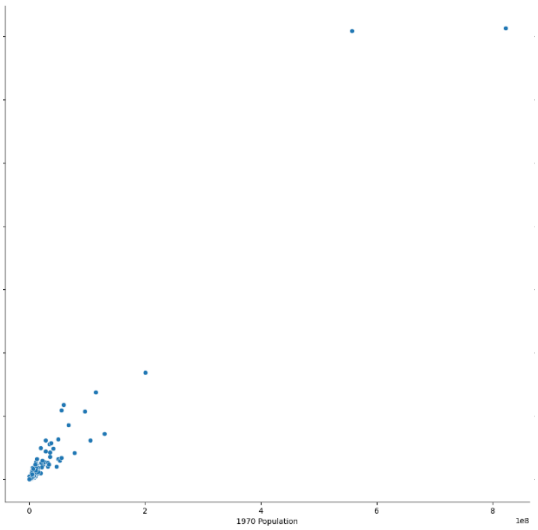

Fig. 1.3 2022-1970 pairplot

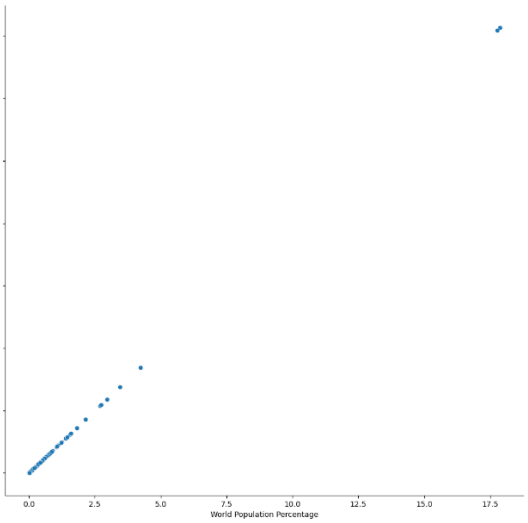

Fig. 1.4 2022-World Popolation Percentage pairplot

Methodology

Once the EDA was completed and the appropriate information to find the ideal model for the dataset (Banerjee, 2022) was found, the most suitable model had to be determined. Based on the goal of this project, creating a model that can predict the future population of the world, the machine learning technique that needed to be used was supervised learning. This decision is supported by the fact that supervised learning, allows for a model with known input and output to determine new ensuing outputs (ElSayed, 2022). Furthermore, it was determined that the technique that needed to be used for the model was of regression method since the data predicted was of numerical value (Beers, 2022). The following step consisted of distinguishing the ideal regression technique for the model. Based on the linear correlation of the data found in the EDA the ideal technique needed to be used was linear regression. This decision is enforced by the usual use of the technique in predicting future values (IBM, 2022). Finally, from the general equation of linear regression ($y=\theta_0+\theta_1x_1+\theta_2x_2+\ldots+\theta_nx_n$) and the nature of the data used it can be determined that simple linear regression is the ideal model needed for predicting future world population. This signifies that the equation requires one feature, in this case, the year, to find the population otherwise known as the prediction variable. Thus, the equation used is adapted to $\mathrm{P}=\theta_0+\theta_1\mathrm{Y}$ where P signifies the population and Y the year.

Before the model can begin the training process the appropriate data needs to be extracted from the dataset. This is done through the renaming of significant columns in the current dataset from the year with the word population next to it to solely the year to allow the data to be easily manipulated. Following that step, a list of strings is created containing the name of the columns that contain the data to be used in the model. Then a new DataFrame is created that is populated with two columns one taken from the list of strings containing each year the data was collected named 'Year'. The other is named 'Total' and contains the world population corresponding to the year, this was achieved by finding all the columns in the original dataset that match strings on the list and retrieving the sum of all values assigned to it and allocating it to the new DataFrame.

The process of building the model began with the allocation of dependent variables from the contents of the column "Population" and independent variables from the ones of column "Year". Following that the "train_test_split" library from sklearn (Scikit-learn.org, 2018) was imported and used to allow for randomly splitting a percentage, that can be alternated, of the data into training and testing. With the above-mentioned library, another library was imported namely "LinearRegression" once again from sklearn (Scikit-learn.org, 2019). This library is used for the training the model and was selected due to the many methods it contains and due to the vast amount of libraries sklearn has and which will allow for more efficient quantification of results in the following steps.

Results

Following the model's creation, the first test case was set to determine the application's general performance. However, before the test case could be run the input for the percentage of the train-test split needed to be determined. Since the usual ideal results for this input is between 70/30 per cent and 80/20 per cent split (Gholamy, Kreinovich and Kosheleva, 2018) the latter was chosen. Once the model was trained using this data split two plots were first used on the training set for a general visualisation of accuracy. The first one was a regression plot using the seaborn library (Waskom, n.d.) (Fig 2.1) and featured the training data fitted with the linear regression fit. The second plot used was a scatter plot from the same library (Waskom, n.d.) with data both predicted from the model and the data given from the dataset (Fig2.2). Both plots demonstrated an upwards line signifying that firstly the regression line fit correctly the data given and the training predictions with the given data had a one-to-one relationship thus showing that the model had made correct predictions. Following that the model is set to predict the test values. A line plot (Waskom, n.d.) (Fig2.3) is used to evaluate the predictions by comparing them to the data given taken from the dataset. Finally, MSE, MAE and R2 are calculated using the metrics library from sklearn (scikit-learn.org, 2019).

Since the first test case produced positive results two more were run with varying the data split from 80/20 for test case 1 to 70/30 for test case 2 and 60/40 for test case 3 for the purpose of determining the ideal split that produced the most accurate result. As seen on (Fig2.4) the ideal model is model 2 with the highest accuracy combined with the smallest mean error. A final test case is run using the best-performing model to determine the future population in the years 2030, 2050 and 2100. The results are compared with numbers predicted by (United Nations, 2022) (Fig2.5) and thus demonstrating the accuracy of the application.

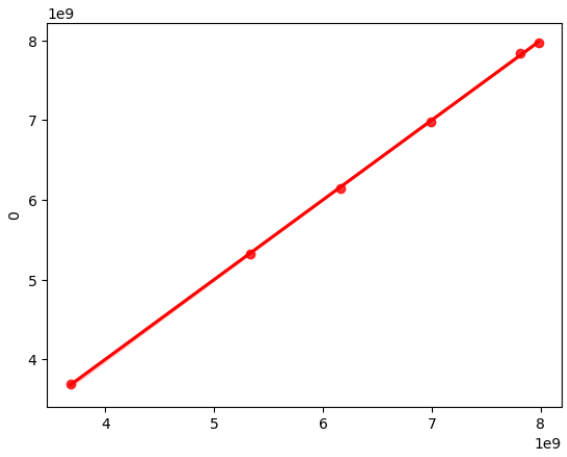

Fig. 2.1 Test case1 Regression plot plot

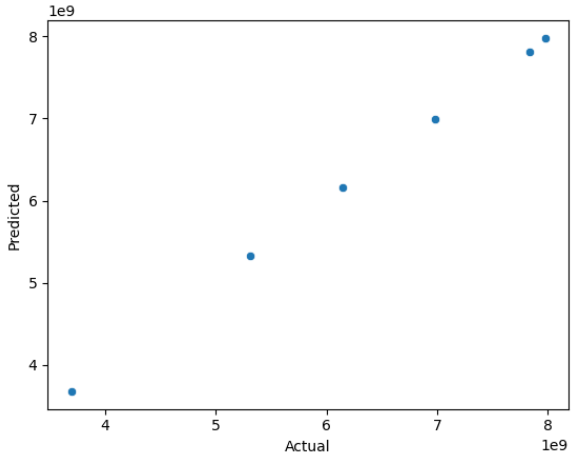

Fig. 2.2 Test case1 Predicted-Actual Training Scatter

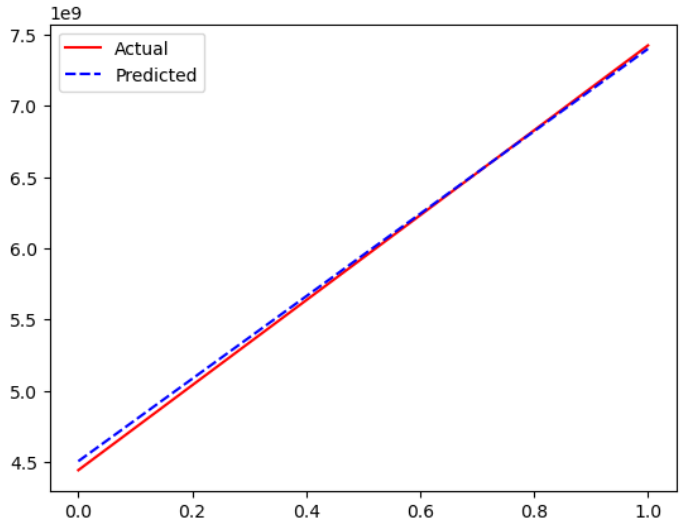

Fig. 2.3 Test case1 Actual-Predicted Lineplot

	Testcase1	Testcase2	Testcase3
MSE	2.248018e+15	1.493328e+15	1.526050e+15
MAE	4.327374e+07	3.642037e+07	3.328926e+07
R^2	9.989891e-01	9.994250e-01	9.995516e-01

Fig. 2.4 Testcase Table

	Actual	Predicted
2030	8.5 billion	8.657390e+09
2050	9.7 billion	1.032947e+10
2100	10.4 billion	1.450968e+10

Fig. 2.5 Actual-Predicted future population (United Nations, 2022)

Discussion and Conclusion

With the best possible model found and applied the application is complete. The process used to find the model started with acquiring as much information on the dataset (Banerjee, 2022) as possible with the purpose of narrowing down the ideal method that the machine learning model would use. Narrowing down the supervised learning technique to specifically a linear regression model. Meanwhile, the information gathered as well, allowed to determine the data in the dataset that were positively correlated and thus would need to be included for training the model. A DataFrame was then created containing solely the data that would be used to train and test the model containing two columns one for the given year while the other containing the total world population. Therefore, ending up with a DataFrame of just the world population through the years 1970 to 2022. The data was assigned to independent and dependent values and split with varying train-test split percentage to determine which percentage would produce the ideal model. Finally, once the ideal split was determined to be 70/30 and the model was run using specific years in the future as new independent variables achieving the project's goal.

Regression as a general method allows to discover patterns within numerical values and suits the nature of the project's objective which could be categorised as discovering the pattern of population growth. Additionally, linear regression is a technique that worked successfully with the data due to the strict linearity that the data follows. Meaning that the population number cannot fluctuate in high volumes between the years that it is measured allowing for a linear function to easily represent the relationship between the two data points. Finally, the disadvantages of linear regression, mainly its sensitivity to outliers and proneness to multicollinearity (Waseem, 2019), don't concern the current dataset used based on the nature of the data provided.

However, there are some limitations that hinder the final accuracy of the model. Mainly the small size of the dataset coupled with the technique's disadvantage of being prone to overfitting. This can lead to results that may not be of the highest accurate precision. The testing of the model could, additionally, benefit from the inclusion of more analytic plots such as a residual plot for the purposes of analysing the model in greater depth.

When evaluating the results through the use of plots, the accuracy of the predicted values closely resembles the set values in the dataset. In Figure 2.4 MSE and MAE appear to be very large numbers potentially indicating an error in the model, However, this phenomenon is common in this scenario because the data itself is of very high numerical value. Finally, from Figure 2.5 where the predicted data is compared to data found by (United Nations, 2022) can be seen that even though the predictions are accurate they do not match as closely the given number as when the year is set far into the future. This can be attributed to the model potentially being overfitted which would be supported by the 99% accuracy the R2 produced.

To conclude the overall purpose of the project, predicting the population based on a given year in the future was achieved and the outputs produced are quite accurate when predicting the population within the few following years up to 2030. However, the accuracy

requires improvement for years further than that. This can be achieved through the use of a dataset containing more frequent population counts and from further in the past.

References

Banerjee, S. (2022). *World Population Dataset*. [online] www.kaggle.com. Available at: https://www.kaggle.com/datasets/iamsouravbanerjee/world-population-dataset.

Beers, B. (2022). *What Regression Measures*. [online] Investopedia. Available at: https://www.investopedia.com/terms/r/regression.asp [Accessed 11 Apr. 2023].

ElSayed, S. (2022). *Lec04a-supervised-learning*. [online] Abertay Learning Space. Available at: https://mylearningspace.abertay.ac.uk/d2l/le/content/27499/viewContent/632973/View [Accessed 11 Apr. 2023].

Gholamy, A., Kreinovich, V. and Kosheleva, O. (2018). Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation. *Departmental Technical Reports (CS)*. [online] Available at: https://scholarworks.utep.edu/cs_techrep/1209/ [Accessed 12 Apr. 2023].

IBM (2022). *About Linear Regression | IBM*. [online] www.ibm.com. Available at: https://www.ibm.com/topics/linear-regression [Accessed 11 Apr. 2023].

Scikit-learn.org. (2018). *Sklearn.model_selection.train_test_split — scikit-learn 0.20.3 documentation*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html [Accessed 11 Apr. 2023].

Scikit-learn.org. (2019). *sklearn.linear_model.LinearRegression — scikit-learn 0.22 documentation*. [online] Available at: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html [Accessed 11 Apr. 2023].

scikit-learn.org. (2019). *3.3. Metrics and scoring: quantifying the quality of predictions — scikit-learn 0.22.1 documentation*. [online] Available at: https://scikit-learn.org/stable/modules/model_evaluation.html [Accessed 12 Apr. 2023].

Shopulatov, A. (2023). *Estimating GDPs: Only Linear Regression*. [online] kaggle.com. Available at: https://www.kaggle.com/code/abrorshopulatov/estimating-gdps-only-linear-regression [Accessed 9 Apr. 2023].

United Nations (2022). *Population*. [online] United Nations. Available at: https://www.un.org/en/global-issues/population Chapter 'The world in 2100'.

Waseem, M. (2019). *Linear Regression for Machine Learning | Intro to ML Algorithms*. [online] Edureka. Available at: https://www.edureka.co/blog/linear-regression-for-machine-learning/ [Accessed 13 Apr. 2023].

Waskom, M. (n.d.). *seaborn.regplot — seaborn 0.11.1 documentation*. [online] seaborn.pydata.org. Available at: https://seaborn.pydata.org/generated/seaborn.regplot.html.

Waskom, M. (n.d.). *seaborn.scatterplot — seaborn 0.11.1 documentation*. [online] seaborn.pydata.org. Available at: https://seaborn.pydata.org/generated/seaborn.scatterplot.html [Accessed 12 Apr. 2023].

Waskom, W. (n.d.). *seaborn.lineplot — seaborn 0.11.2 documentation*. [online] seaborn.pydata.org. Available at: https://seaborn.pydata.org/generated/seaborn.lineplot.html [Accessed 12 Apr. 2023].

Widiatama, F. (2022). *WORLD VISUALIZATION FROM WORLD POPULATION DATA*. [online] kaggle.com. Available at: https://www.kaggle.com/code/fahrilianwidiatama/world-visualization-from-world-population-data [Accessed 9 Apr. 2023].