International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 04 Issue: 05 | May -2020 ISSN: 2582-3930

USING PREDICTIVE MACHINE LEARNING REGRESSION MODEL TO PREDICT THE POPULATION OF INDIA

Author: Harshitha D
Co-Author: DR. Samitha Khaiyum
AUTHOR M.C.A. Post Graduate Students D.S.C.E
CO-AUTHOR M.C.A Associate Professor & HOD D.S.C.E

---***---

Abstract:

A population consists of all individuals of the same species that occupy a specific geographical area at a given time. There are small varieties. A species can have a single population or limit many populations to a separate area.In any country, the government is always concerned about the management of the people of the country and for effective governance should be a plan based on the number of citizens (population) in that country.The government therefore spends a lot on census exercises that count the population of the country. The census exercises and India were marked by following duties, and results were inaccurate and the results obtained. A number of research papers have been published on how to solve this problem. In this study, different types of predictable models have been developed to characterize the population of India using machine learning regression method. The best of the models were selected and used to predict India's population by 2050. The tool used for implementation is the Mat lab modeling toolbox. The research work will serve as a very useful tool for forecasting the population and will help the government in her future plan.

KEYWORDS: Algorithm, Artificial Intelligence, Government, Machine Learning, Population, Prediction.

1. Introduction:

Artificial Intelligence (AI) is one of the main areas of the computer that has been used to solve various kinds of real problems. Its scope is so wide that it now transcends various disciplines. For example, AI has proven to be a useful tool for solving problems in medicine, engineering, science, applied science, agriculture, languages, etc. We can hardly identify a particular field or field where AI is not found useful. One of the most important branches of AI is Machine Learning (ML).John, Brian and Aoife [JBA14] defined ML as an automatic process extracting a pattern of data. Bontempi [Bon13] has defined machine learning as the domain of computer intelligence that cares about how to build a computer program that is automatically enhanced by experience. In order to have a working ML algorithm, there must be enough data where the algorithm gains its experience to solve similar problems in the future. Thus, machine learning is in the field of AI, where the program derives a variety of experiences from historical data and uses that experience to solve future problems. Obviously, the optimal performance ML algorithm depends on the historical data. In our opinion, this problem can be successfully tackled by developing new methods of fertility regulation and implementing voluntary family planning widely and rapidly around the world. One of the traits of a man is that he has a natural desire to ask questions about probable future events. In fact it is a natural human desire to predict future events. After decades of research and development, computer science is now a point where predictable algorithm is becoming more important and indispensable as one of the ML areas. Therefore, many technologies have been put together, including fast computing, cloud computing, speech recognition, and mobile computing to make this possible. Although only a few countries have made concerted efforts in this direction, responsible groups in the social, economic and scientific communities in many countries have become increasingly aware of the problem and the need for intelligent and fair action .A predictive algorithm is a scientific idea to empirically establish a relationship between the historical database, which can then be used to make future decisions while trying to solve some problems in practice. Therefore the research has shifted to the field of prediction or prediction of the probability of future events by the range of available data (big data) available through modern technologies. The predictable model is the product of a predictable algorithm and big data.Big data + predictive algorithm = predicted model. Predicate modeling has been used to solve different problems by different researchers. For example, it has been used to predict weather conditions, the rate of disease spread, birth rate, mortality, percentage of road accidents, etc. and to predict population. Our research work is using machine learning of a predictive regression model to predict the population of India.

2. Problem Statement:

There are many countries in the world, and in order for a country to continue to exist, there must always be a government with exclusive competence to govern its citizens. To manage effectively, one of the most important steps the government has to take is to have a comprehensive idea of the number of people they control, namely that the people of the country must be known to the government. The current population of our country is 102.7 billion. The population of our country is currently growing. This is very dangerous and when our natural resources are depleted. The main reason for the high growth rate is the increasing gap between births and deaths. The growing urban population created many problems for both urban and rural areas. In urban areas this has led to the storage of food, raw materials and a great diversity of raw materials. This has caused pollution and imbalance in the environment. In rural areas, the growing urban population has caused declines in forested areas and Left foot affects soil fertility. Unfortunately, despite the billions of money spent on training, there has always been some problem with obtaining accurate census data or census results. How do we solve or minimize this problem now so that the government can predict the number of its citizens without having to spend millions or billions of money on a census that will ultimately give false results.

3. Literature Review:

John ([Joh17]) presents a research work entitled Models and Reliable Projections. In his research work, the population of Nigeria was projected for the interval from 1991 to 2050 using the experimental growth model (EGM) and logistic growth model (LGM). The two models were combined using mean projection (AP).The result obtained shows that the AP is better than the actual or official projection. The aim of this study was to compare the different models. Machine learning algorithm was not used.

Andre et al [A+17] was able to compare a machine learning algorithm to create a predictive model to detect unrecognized diabetes. They were able to develop a predictable model for detecting undiagnosed diabetes and were able to compare the performance of different machine learning algorithms. The instrument used was an artificial neutral net and logistical suppression. The research developed a machine-learning prediction model for diabetes and not population prediction.

Vilalta et al ([V+02]) developed predictive algorithm for computer system monitoring. A system was then set up to alert the user to a specific error. Designed to alert the user to special errors. The result shows that a predictable algorithm can identify critical events. The system was unable to predict the population in any form. It is clear from the literature that seasonal writers nevertheless operated within the predictive algorithm; not enough has been done in the area of population projections.

Olatayo and Adeboy ([OA13]) predicted the Nigerian population by birth and death. Uses regression analysis. Research Deaths and Births Nigeria has a significant impact on population growth. This is a statistical approach without using the machine learning algorithm.

Folorunso et al ([F + 10]) used the neural network to estimate Nigeria's population. How to practice. It's an AI-based method, but it's very different from the downside. The few who worked in the area did not use a regression model / algorithm to predict machine learning, but rather statistical or other models. For this work, we used a prediction model for learning the regression machine.

4. Methodology:

We developed machine leaning regression model with actual population data from the year 1900 to 2006, to obtain different models and the best of them was selected and extrapolated to predict population growth till 2050. Mat lab Machine Learning toolbox was used for implementation.

Steps to be followed:

1) STEP 1: DEFINING THE PROBLEM

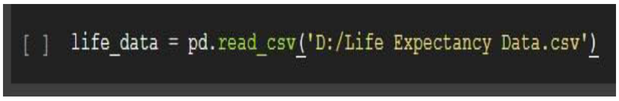
- Life expectancy from birth is a frequently utilized and analyzed component of demographic data for the countries of the world.
- It represents the minimum life span of a child and is an indicator of the overall health of a country.
- Life expectancy rate can fall because of problems like bad habits, war, disease and unhealthy. Improvements in health, country development and welfare increase life expectancy. The higher the life expectancy of people, the better a country is in.

STEP 2: COLLECTION OF DATA

- The data was collected from WHO and United Nations website. https://www.kaggle.com/kumarajarshi/life-expectancy-who

STEP 3: PREPARE THE DATA

- First you must upload the data to the jupyter notebook for analysis.



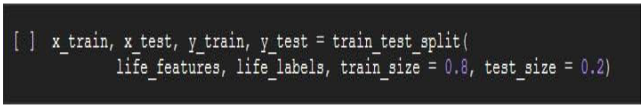
```
[ ] life_data = pd.read_csv('D:/Life Expectancy Data.csv')
```

- Check are there any null rows or columns, if any remove the rows or fill the columns with the mean value of that particular column.

2) STEP 4: SPLITTING THE DATA INTO TRAINING AND TESTING

- Here in this step we split the data into two separate tables training and testing respectively.

- Training table is used for analysis and alteration purpose, whereas testing table is used to test the data in the final step.
- We will also split the table data into 80% and 20%.



```
[ ] x_train, x_test, y_train, y_test = train_test_split(
            life_features, life_labels, train_size = 0.8, test_size = 0.2)
```

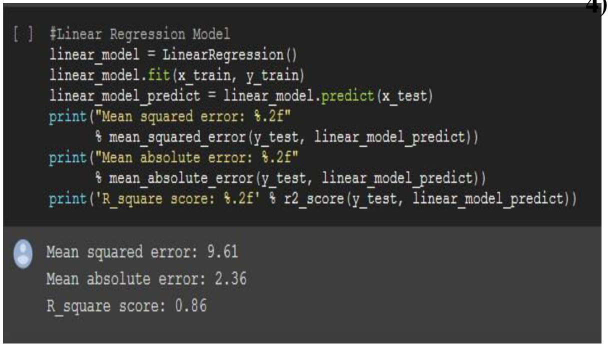
STEP 5: ALGORITHM SELECTION

Here we select the appropriate algorithm/model that is necessary for the analysis purpose, we have selected the following models for processing the dataset.

- Linear Regression
- Linear Regression with polynomial features
- Decision Tree
- Random Forest

STEP 6: TRAINING THE ALGORITHM WITH DATA FOR MACHINE

- Here the data is divided into xtrain,ytrain,xtest and ytest where 80% of data is taken as train data and remaining 20% of data is taken as test data.
- Then if there are any nulls in the rows or column, remove those and clean the data and also if you have any high range values take the mean of those and reduce the range of those values.
- Then the algorithm is trained with 80% cleaned data of xtrain and ytrain.

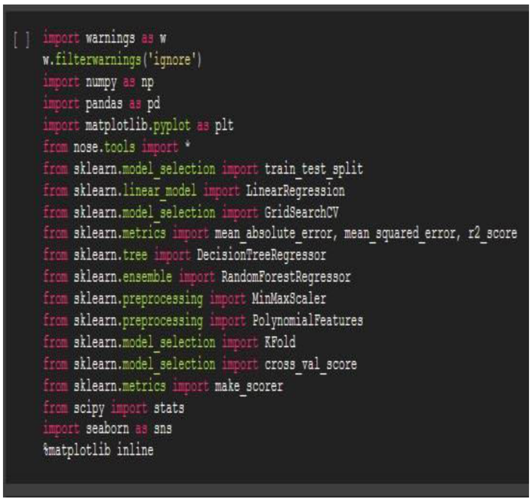


```
[ ] #Linear Regression Model
    linear_model = LinearRegression()
    linear_model.fit(x_train, y_train)
    linear_model_predict = linear_model.predict(x_test)
    print("Mean squared error: %.2f"
          % mean_squared_error(y_test, linear_model_predict))
    print("Mean absolute error: %.2f"
          % mean_absolute_error(y_test, linear_model_predict))
    print('R_square score: %.2f' % r2_score(y_test, linear_model_predict))
```

```
Mean squared error: 9.61
Mean absolute error: 2.36
R_square score: 0.86
```

STEP 7: EVALUATE TEST DATA

- After the algorithm is trained with 80% of data, the algorithm is to be tested with remaining 20% of data.
- Import all the packages and functions that are necessary for the analysis process.



```
[ ] import warnings as w
    w.filterwarnings('ignore')
    import numpy as np
    import pandas as pd
    import matplotlib.pyplot as plt
    from nose.tools import *
    from sklearn.model_selection import train_test_split
    from sklearn.linear_model import LinearRegression
    from sklearn.model_selection import GridSearchCV
    from sklearn.metrics import mean_absolute_error, mean_squared_error, r2_score
    from sklearn.tree import DecisionTreeRegressor
    from sklearn.ensemble import RandomForestRegressor
    from sklearn.preprocessing import MinMaxScaler
    from sklearn.preprocessing import PolynomialFeatures
    from sklearn.model_selection import KFold
    from sklearn.model_selection import cross_val_score
    from sklearn.metrics import make_scorer
    from scipy import stats
    import seaborn as sns
    %matplotlib inline
```

- We can analyses the data quality using scatter plots
- Before that check there are any null values in test data, if present remove all the null values and obtain clean data

3) STEP 8: PARAMETER TUNING

- A tuning is to be done for algorithm in order to control the behavior of the algorithm.
- There are many tuning methods available, here we applied linear regression.
- Linear regression is suitable for life expectancy because its most widely used predictive model.

4) STEP 9: START USING YOUR MODEL

- After the algorithm is trained with test data, it gives the prediction that is it gives the accuracy of our model.
- If the prediction is below 70% then the model is failed, there may be a mistake in choosing the data, cleaning the data or methods etc.
- If the prediction is above 70% then the model is good and ready for usage.

5. Results and Discussions

Figure 1 shows the actual and the predicted population side by side. The model shows that the predicted population for India in 2050.

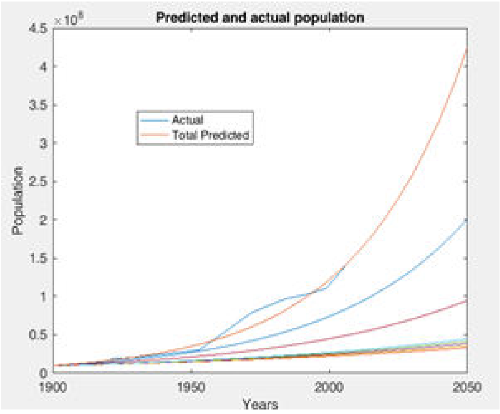


Fig. 1. Predicted and actual population

Figure 2 shows the predicted population by age group. The age distribution from 0 to 15 years is the largest and the age distribution from 105 to 120 years is the smallest.

Figure 3 explain further the comparison of age distribution in percentage annually, from the year 1900 to 2040 with an interval of 20 years. For each year, children under 15 have the largest population.

Figure 4 shows the total prediction of the India's population by taking male attribute data, from the year 1900 to 2040.

Figure 5 shows the average prediction of the India population by comparing the population with the Location attribute.

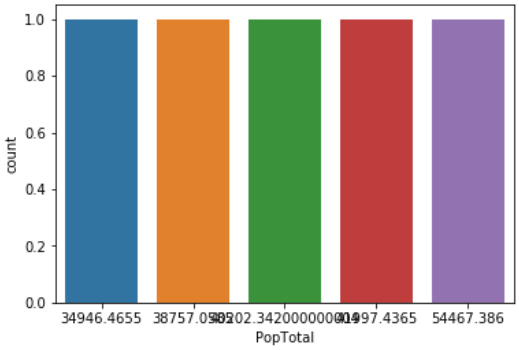


Fig. 2. Predicted population according to age group

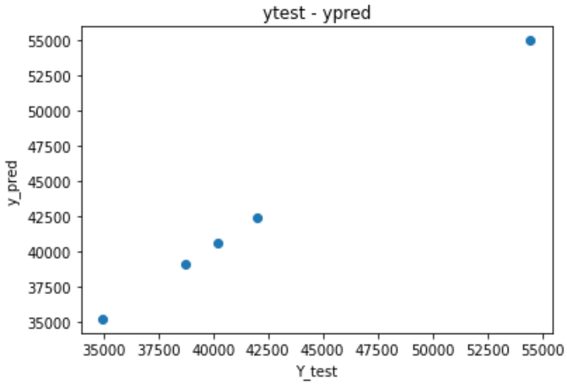


Fig. 3. Age distribution in the year 1900

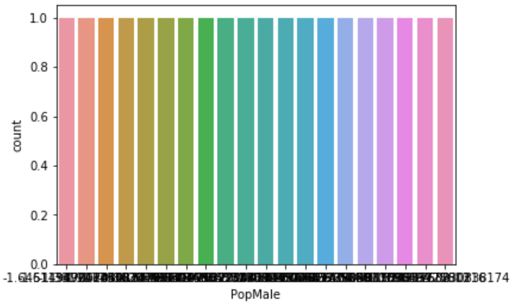


Fig. 4. Predicted population according to Population of Male

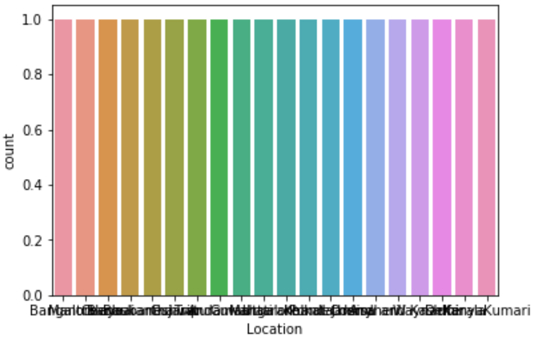


Fig. 5. Predicted population according to Population of Location

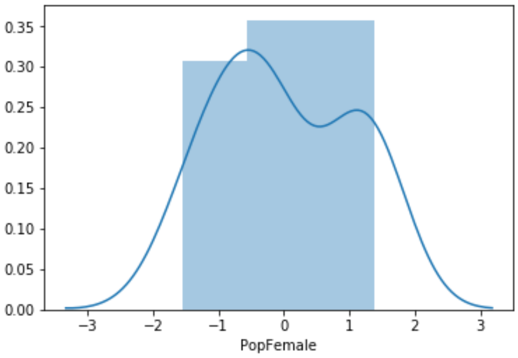


Fig. 6. Predicted population according to Population of Male

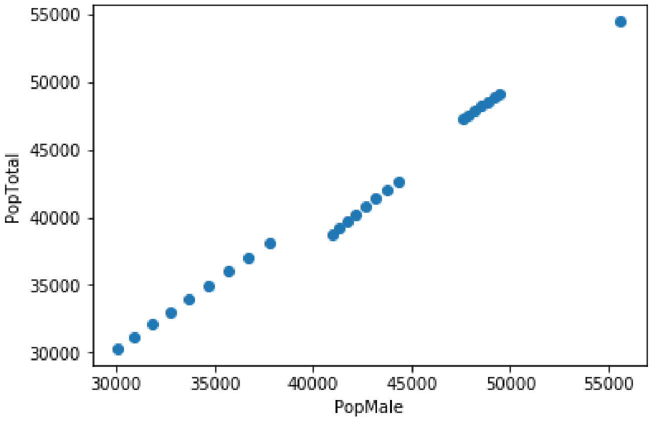


Fig. 7.Average Predicted population according to Population of Male

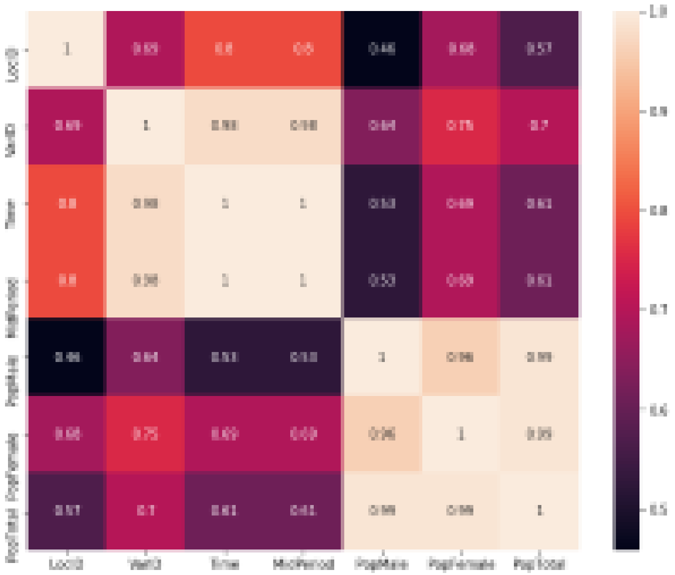


Fig. 8. Axis Subplots for Population Data

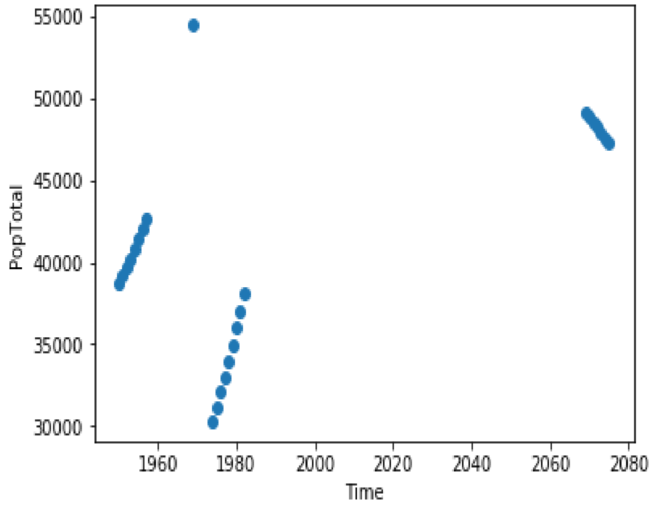


Fig. 9. Predicted population according to Population of Time

Figure 6 shows the total prediction of the India's population by taking female attribute data, from the year 1900 to 2040.

Figure 7 talks about the average male population of India according the database available.

Figure 8 Axis sub plots which explains about all the attributes of the database and it has a color code which represents the density of the population for each attribute.

Figure 9 shows the average prediction of the India population by comparing the population with the Time attribute.

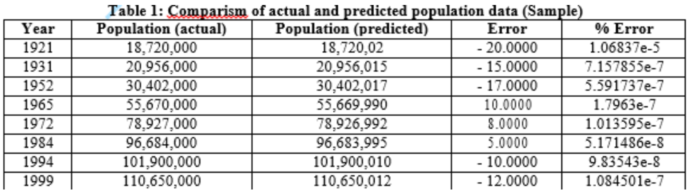


Table 1: Comparism of actual and predicted population data (Sample)

Year	Population (actual)	Population (predicted)	Error	% Error
1921	18,720,000	18,720,02	- 20.0000	1.06837e-5
1931	20,956,000	20,956,015	- 15.0000	7.157855e-7
1952	30,402,000	30,402,017	- 17.0000	5.591737e-7
1965	55,670,000	55,669,990	10.0000	1.7963e-7
1972	78,927,000	78,926,992	8.0000	1.013595e-7
1984	96,684,000	96,683,995	5.0000	5.171486e-8
1994	101,900,000	101,900,010	- 10.0000	9.83543e-8
1999	110,650,000	110,650,012	- 12.0000	1.084501e-7

Table 1: Compares of actual and predicted population data (Sample)

Table 2: Sample Error obtained

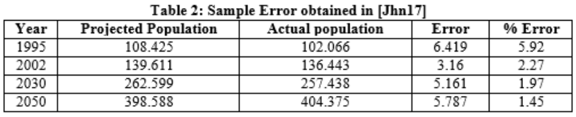


Table 2: Sample Error obtained in [Jhn17]

Year	Projected Population	Actual population	Error	% Error
1995	108.425	102.066	6.419	5.92
2002	139.611	136.443	3.16	2.27
2030	262.599	257.438	5.161	1.97
2050	398.588	404.375	5.787	1.45

6. Conclusion:

In this research work, attempts were made to model the people of India using a regression model of Machine Learning. Various models have been developed and the best that has marked the people of India. The research work will provide the government with very useful information and help prevent billions of money from being spent on censuses that ultimately yield no useful results. Therefore, research needs to make a significant contribution to the country's development. This research will increase the capacity of governments at all levels to provide adequate supplies to a growing population. Therefore, research must make an important contribution to the development of the country. This research will increase the capacity of governments at all levels to provide adequate supplies to the growing population. The population under 15 is high. Therefore there should be acceptable plans to create good schools, water with pipes and well-equipped hospitals. Creating jobs for young school children should be taken seriously. Again, the few who enter the age (over 70 years old) must be well united and, like those who belong to the very old age group (over 100 years old), must have special incentives from the government.

7. References:

[Joh17] N. I. John - Befolkningsspørgsmålet i Nigeria, Asian Research Journal of Mathematics, bind 5: 1 - 10, ISSN: 2456-477X, 2017.

[JBA14] D. K. John, M. N. Brian, D. Aoife - Grundlæggende om maskinlæring til forudsigelig dataanalyse. MIT Press, Cambridge, England, 2014.

[V + 02] R. Vilalta, C. V. Abte, J. L. Hellerstein, S. Ma, S. M. Weiss - Forudsigelig algoritme i computersystemstyring, IBM System Journal, vol. 41, 2002.

[A + 17] R. O. Andre, R. Valter, L. Cirano, I. D. Bruce - Sammenligning af maskinlæringsalgoritmer til at opbygge en forudsigelig model til påvisning af udiagnosticeret diabetes. Sao Paulo Med. Tidsskrift, 135 (3), 234-246, 2017.

[Bon13]G. Bontempi - Machine Study Strategies for Project Time Series, Machine Learning Group, Computer Science Department. Boulevard of Triumph - CP212, 2013.

[OA13] T.O. Olatayo, N.O. Adeboye - Prediction of Population Growth by Birth and Mortality in Nigeria, Mathematical Theory and Modeling, Volume 3, ISSN: 2224-5804 (print).