International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 03 Issue: 08 | August -2019 ISSN: 2590-1892

LIFE EXPECTANCY PREDICTION USING PYTHON FOR DATASCIENCE AND MACHINE LEARNING TOOLS

VINAYAKA RAJU M: A1, CHIRANTANA K: A2, TEJUS GOWDA: A3, PUNITH M.S: A4, RAJESH C: A5, MAHENDRA KUMAR B: CO-AUTHOR

A1 – A5 M.C.A. Post Graduate Students D.S.C.E
CO-AUTHOR M.C.A Assistant Professor D.S.C.E

Abstract - Overlap of lifespans depends on variation in survival across ages and can be high or low independently of high or low life expectancies (*ref[7]). Here we develop formal demographic measures to study the complex relationships between attributes like adult mortality, infants' death, alcohol consumption, GDP and population of a country, various diseases like Hepatitis B, Measles, Polio, Diphtheria, HIV/AIDS etc., which yields accurate life expectancy. Our work involves analyzing an open source data set by inserting data into multiple machine learning models and training the model to predict age of a given country.

1. INTRODUCTION

Life expectancy is one of the most important factors in end-of-life decision making. Good prognostication for example helps to determine the course of treatment and helps to anticipate the procurement of health care services and facilities, or more broadly: facilitates Advance Care Planning. Advance Care Planning improves the quality of the final phase of life by stimulating doctors to explore the preferences for end-of-life care with their patients, and people close to the patients. Physicians, however, tend to overestimate life expectancy, and miss the window of opportunity to initiate Advance Care Planning. This research tests the potential of using machine learning and natural language processing techniques for predicting life expectancy from electronic medical records. (* ref [3])

2. Body of Paper

The project relies on accuracy of data. The project is aimed to develop a machine learning model based on data given by the World Health Organization to determine the life expectancy for different countries in years. The data offers a timeframe from 2000 to 2015. Here we select the appropriate algorithm/model that is necessary for the analysis purpose, we have selected the following models for processing the dataset. The output algorithms have been used to test if they can maintain their accuracy in predicting the life expectancy for data they haven't been trained. Four algorithms have been used: (* ref [1]).

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

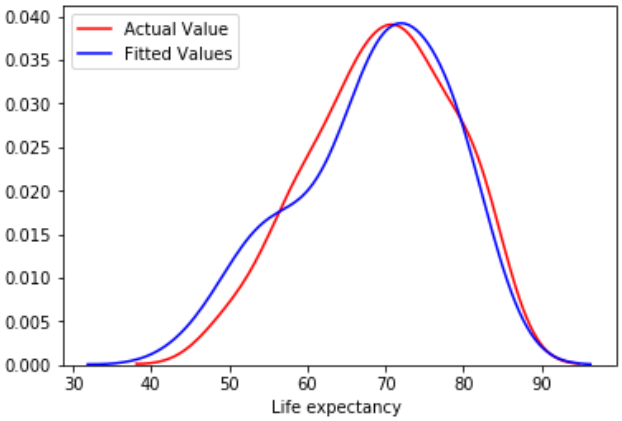
LINEAR REGRESSION

Several algorithms will be tried out. First the classical linear regression. The model is fitted first on the trained data the R square is 0.92 on the training data. Later its R square which is coefficient of accuracy determination is checked on the testing data. The score is 87 % in the iteration of writing. We also calculate the MAE, the modulus between the predicted and the real value at 2.32 and the MSE (the same only put to the power of 2) at 9.8 (* ref [1]).
The goal of regression analysis is to model the expected value of a dependent variable *y* in terms of the value of an independent variable (or vector of independent variables) *x*. In simple linear regression, the model. (* ref [2]).

$$y = b_0 + b_1 x + e$$

The below distplot figure factory displays a combination of statistical representations of numerical data, such as the relationship between actual values and predicted values using Linear Regression

Fig 1 – Output of Linear Regression Model



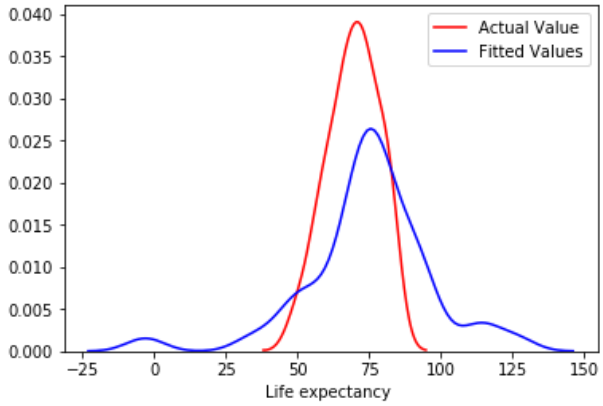
LINEAR REGRESSION WITH POLYNOMIAL FEATURES

The Linear Regression is being tested on the training data with the new Polynomial Features. The Polynomial Features function has been used to get the interactions of the input variables only to the power of 2. (* ref [1])
For infinitesimal changes in *x*, the effect on *y* is given by the total derivative with respect to *x*: The fact that the change in yield depends on *x* is what makes the relationship between *x* and *y* nonlinear even though the model is linear in the parameters to be estimated. (* ref [2])

$$y = b_0 + b_1x + b_2x^2 + b_3x^3 + \ldots + b_nx^n + e$$

All the errors were significantly higher than the previous model, and the R square is in this case negative and this happens only when fits a nonlinear model to the data, as a statistical estimation problem, it is linear, in the sense that the regression function E($y \mid x$) is linear in the unknown parameters that are estimated from the data. This is the worst performing model for now.

Fig 2 – Output of Linear Regression Model with Polynomial Features



DECISION TREE

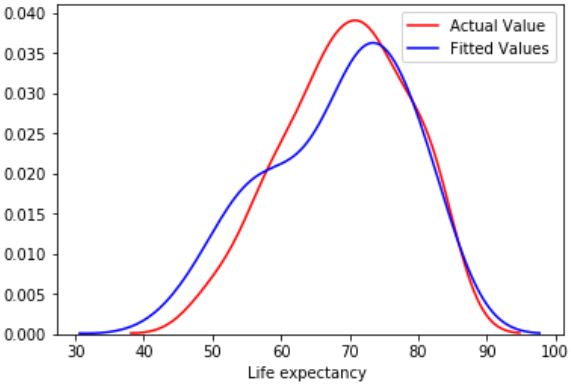
In a standard characterization tree, the thought is to part the dataset dependent on homogeneity of information. In a relapse tree the thought is this: since the objective variable does not have classes, we fit a relapse model to the objective variable utilizing every one of the autonomous factors. At that point for every autonomous variable, the information is part at a few part focuses. At each part point, the "mistake" between the anticipated worth and the genuine qualities is squared to get a "Whole of Squared Errors (SSE)". The split point mistakes over the factors are looked at and the variable/point yielding the least SSE is picked as the root hub/split point. This procedure is recursively proceeded and cross Validation has been performed. The R square on the preparation information is 1 implying that the calculation has taken in the information by heart, with the cross approval the figure decreases to 77% and utilizing the test date we get 80%.After performing lattice search with least examples split in the range somewhere in the range of 2 and 10 we get the best split of 3. The R square on the preparation information is 98%, the calculations has almost inclined the information by hearth. On the test information we get R square of 81 %, the MAE is 2.71 and MSE is 16.59 (* ref [1]).Data comes in records of the form:

$$(x,Y) = (x_1, x2, x3\ldots x4, Y)$$

Decision trees are very helpful visualizing aid for analysing a series of predicted outcomes for a specific model. In that capacity, it is regularly utilized as an enhancement (or even option in contrast to) relapse examination in deciding how a progression of logical factors will affect the reliant variable.

In this specific model, we examine the effect of illustrative factors of age, sex, miles, obligation, and salary

Fig 1 – Output of Decision Model



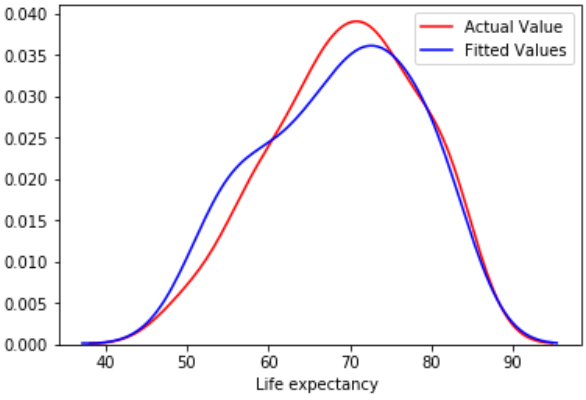
RANDOM FOREST

A Random Forest is an ensemble technique capable of performing both regression and classification tasks with the use of multiple decision trees and a technique called Bootstrap Aggregation, commonly known as bagging (* ref [8]). The basic idea behind this is to combine multiple decision trees in determining the final output rather than relying on individual decision trees (* ref [9]). The algorithm has learned 98% on the training data without cross validation and 88% with, the value is 92 % on the test data.

$$\text{Random Forest Prediction } s = \frac{1}{K}\sum_{K=1}^{K} K^{th} \text{ tree response}$$

The result of each tree relies upon a lot of anticipated qualities picked freely with supplanting and with a similar appropriation for every one of the trees in the model, which is a subset of the indicator estimations of the first informational index. The ideal size of the subset of indicator factors is given by log2 M+1, where M is the quantity of sources of info. Arbitrary Forest strategy characterizes an edge work that estimates the degree to which the normal number of decisions in favour of the right class surpasses the normal decision in favour of some other class present in the needy variable. This measure furnishes us not just with a helpful method for making forecasts, yet additionally with a method for partner a certainty measure with those expectations. For regression problems, *Random Forests* are formed by growing simple trees, each capable of producing a numerical response value (* ref [6])

Fig 1 – Output of Random Forest Model



3. CONCLUSIONS

Life expectancy usually declines with age. According to the our research we got to know that, in 2003 life expectancy at birth for both men and women of all races in the United States was 77.5 years. At age 65, life expectancy was 18.4 years; at age 75, 11.8 years; and at age 85, 6.8 years. These calculations imply a 65-year-old could expect to live until 83, a 75-year-old until age 86, and an 85-year-old until age 91. Generally, in conjunction with declining fertility rates, improvements in life expectancy have resulted in the "greying" of many national populations. (* ref [4])

REFERENCES

1. www.datasciencesociety.net [1]
2. en.wikipedia.org [2]
3. bmcmedinformdecismak.biomedcentral.com [3]
4. dl.taq.ir [4]
5. Rehm J., Mathers C., Popova S., Thavorncharoensap M.,Teerawattananon Y. Global burden of disease and injuryand economic cost attributable to alcohol use and alcohol-use disorders. Lancet 2009; 373; 2223-33
6. documentation.statsoft.com [6]
7. journals.plos.org [7]
8. Zatonski W., Manczuk M., Sulkowska U., HEM ProjectTeam. Closing the Health Gap in European Union. Warsaw:Cancer Epidemiology and Prevention Division, The MariaSklodowska-Curie Memorial Cancer Center and Institute of Oncology; 2008
9. WorldHealthAssembly(WHA).Resolution61.13.Strategiesto Reduce the Harmful Use of Alcohol . Geneva: World Health Organization; 2008.
10. www.thesisscientist.com [10]
11. Warren C. W., Lee J., Lea V., Goding A., O'Hara B. Evolution of the Global Tobacco Surveillance System (GTSS) 1998– 2008. Glob Health Promot 2009; 16: 4–37.