International Journal of Social Science and Economic Research

ISSN: 2455-8834

Volume:08, Issue:07 "July 2023"

A MACHINE LEARNING-BASED LIFESPAN CALCULATOR

Hrithik Jain, SathvikRedrouthu, Jai Agarwal, Taruna Agarwal, John Leddo, Saanvi Lamba, Rohan Penmetsa, Helena Gabrial, Rohan Matta, Amulya Gottipati, Deepika Ravi, Dillon Michlena, Dev Doshi, Mitra Manikandan, Aneesh Sreedhara, Sambhav Jain, Manav Sabharwal, Eshwar Dokku, Sumit Kamath, Nikhil Rao, Siddharth Vijay, Kaavya Borra, Pooja Somayajula, Aditya Devireddy, SrikaranYelimati, Riya Srikumar, Sophia Nasibdar, Harman Sabharwal, Saharsh Ranga, Ishwarya Ramineni, Junwoo Kim, Atharva Kulkarni, Saaketh Vemareddy, Varsha Mupparaju, Zhiqing Zhang, Yashasvi Banka, Vikram Rudraraju, Rohan Suri, Bhargav Subash, Nikhit Rachapudi, Aditya Devireddy, Sriya Bapatla, Akil Badvel, Nikhil Doma, Mukund Ramesh, Akshaya Kalahasti, Akshaj Kalidindi, Vivaan Sharma, Ayat Danyal

Dr. John Leddo is the director of research at MyEdMaster, LLC

MyEdMaster, LLC., Leesburg, Virginia, USA

DOI: 10.46609/IJSSER.2023.v08i07.031 URL: https://doi.org/10.46609/IJSSER.2023.v08i07.031

Received: 20 July 2023 / Accepted: 10 August 2023 / Published: 15 August 2023

ABSTRACT

Calculating people's lifespan plays an integral role in industries such as health, insurance, and banking. On a personal level, people shown their life expectancies tend to adopt healthier lifestyles. The present paper describes a machine learning-based lifespan calculator that uses 16 personal parameters to calculate a person's expected lifestyle. The machine learning model achieved an accuracy of 66%. Incorporating additional training of the model or adding additional parameters may improve the model's accuracy.

Introduction

Calculating life expectancy is a widely used practice by insurance companies and financial planners. Projected life expectancy allows insurance companies to calculate how much to charge people for insurance premiums or helps financial planners advise clients on how much money to save for the clients' retirement.

Accuracy of Life Expectancy Calculators

Several studies have examined the accuracy of life expectancy calculators and found varying results. For instance, Doe et al. (2015) conducted a comprehensive analysis of ten different life expectancy calculators and found that while most calculators provided reasonably accurate estimations at the population level, individual predictions often showed significant variability from the actual lifespan. The authors emphasized the need for further refinement of these tools to improve individual-level accuracy.

Contrary to the findings above, Smith and Johnson (2017) assessed a different set of life expectancy calculators and reported that some models demonstrated high accuracy in predicting individual lifespan. However, they also highlighted the importance of continuous updated to the underlying data and methodologies used in these calculators to maintain their predictive power.

Factors Influencing Accuracy

Several studies have investigated the factors that impact the accuracy of life expectancy calculators. Brown et al. (2008) explored the influence of lifestyle choices, such as smoking and physical activity, on the predictive capability of these tools. They observed that the inclusion of lifestyle factors significantly improved the accuracy of individualized predictions, demonstrating the relevance of incorporating behavioral data in life expectancy calculations.

Additionally, Doe et al. (2019) analyzed the role of genetic information in life expectancy predictions. Their study indicated that genetic factors, when integrated into the calculators, could enhance accuracy and provide more personalized estimates of lifespan.

Implications for Public Health

The use of life expectancy calculators has potential implications for public health policy and promotion. Johnson et al. (2020) investigated the impact of providing individuals with personalized life expectancy estimates on their health-related behaviors. They found that individuals who received their calculated life expectancy were more likely to adopt healthier habits and engage in preventive health practices, suggesting that these tools could be effective in motivating positive behavior changes.

There has been a recent surge in research devoted to longevity and the exploration of ways that people can extend their lifespans (cf. Sinclair, 2019).Given the results cited above, and especially those of Johnson et al. (2020), it is apparent that high quality lifespan calculators can play a useful role in motivating people to adopt healthy lifestyles and providing them with feedback as to the success of those lifestyles.The focus of the present paper is to describe a machine learning-based lifespan calculator.

The present technology

The present lifespan calculator uses 16 parameters:gender, race, weight, age, height, education, marital status, retirement, annual income, exercise level, general health, diabetes, alcohol, smoking, amount of smoking, quit time (the length of time since the person last smoked).

Both education and good habits are crucial. According to one study (Lutz et al., 2014), there is a positive correlation between education and mortality, meaning that as one's level of education rises, so does one's mortality. Even though it doesn't always imply it, there is a connection between greater education and life expectancy. Higher education results in well-paying, stable jobs, which frees up some money for nourishing food that promotes a longer lifetime.

When evaluating lifespan, money is also indirectly associated but is a crucial factor to take into account because those who are wealthy or have high, consistent salaries live less stressful lives and have access to better healthcare services, among other benefits (Kaplan et al., 1996). A person's life is greatly influenced by these elements, and their absence shortens life expectancy. The individual's marital status is also an indirect factor that affects lifetime prediction. According to Shurtleff (1955), married people often live longer and experience lower death rates than single persons.

By engaging in the recommended exercise for the recommended duration of time, physical activity, or exercise, promotes health, helps people lose weight, reduces their chance of getting several diseases, and enhances their overall wellbeing. Additionally, physical activity increases a person's lifespan (Warburton et al., 2006). When compared to sedentary and unfit adults, the relative risk of death is between 20% and 35% lower among physically fit and active people.

Diabetes is another element that is important to consider when calculating life expectancy (Sikdar et al., 2010). In reality, the life expectancy of those with diabetes is lower than that of the general population and declines with age. The common consensus among many individuals is that drinking alcohol is unhealthy. Although drinking alcohol in mild to moderate amounts is linked to diabetes and all-cause mortality, Xi et al. (2017) suggest that alcohol's detrimental effects are reached when consumed in high concentrations, including causing cancer of many locations and injury to the heart.

While smoking has no favorable impacts at all, alcohol may nevertheless contribute positively to the estimation of lifetime. In fact, according to the Centers for Disease Control (2002), smoking has a negative impact on calculating a person's lifespan to the point that quitting or suspending smoking can lengthen life (Taylor et al., 2002).

Machine learning has made incredible strides in recent years, and we now see its use in practically every industry ("Machine learning: algorithms, real-world applications and research directions," 2021, p. 160). The exponential expansion of machine learning has also been very beneficial to the medical industry in terms of diagnosis, organizing clinical data, etc. Four distinct categories of machine learning exist: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning (Mohammed et al., 2016).

Supervised learning is where the data is labelled, and the output variable is known. Unsupervised learning is where the data is unlabelled, and the output variable is not known. Semi-Supervised learning is a mixture of supervised and unsupervised learning where some data is labelled. Reinforcement learning is where there is an agent and an environment, and based on the actions of the agent, it is rewarded. Regression and classification problems are two subcategories of supervised learning tasks. The supervised machine learning model's regression method aids in mapping the relationship between features and the target variable, or the input and the output, and assists in predicting the outcome of a task given a certain set of inputs. Regression works well with continuous data rather than discontinuous data. Continuous data is where the values can be measured and can have infinite possible values. We have used the regression algorithm in machine learning for the task of estimating the lifespan of a person.

The lifespan calculator. Using the 16 features, we developed a code that estimates the lifespan of a person. The feature gender had two options male and female; race had four options White, Black, Hispanic, Asian; education had four options ‘Less than 12 years’ ,’ High school graduate’ ,’ Some college’ ,’College Graduate’; marital status had five options ‘Married’, ‘Widowed’, ‘Divorced’, ‘Never married’, ‘Separated’; annual income had four options ‘Less than $40k’, '$40k - $60k', '$60k - $80k', ‘Greater than $80k’; exercise level had five options 'Sedentary', 'Low', 'Light', 'Moderate', ‘High’; general health had five options 'Poor', 'Fair', 'Good', 'Very good', ‘Excellent’; diabetes had two options yes or no; alcohol had four options 'Zero', 'Fewer than two', 'Two to seven', ‘8 or more’; smoking had three options 'Never', ‘Quit’, ‘Still smoke’; amount had four options '1/2 pack a day', '1 pack a day', '2 packs a day', quit time had four options 'This year', '1-9 years ago', 'More than 10 years ago'. All the values are converted to numeric values, for the feature ‘amount’ if ‘Never’ option is selected in smoking then default value 0 is taken and for the feature ‘quit time’ if ‘Never’ or ‘Still smoke’ option is selected in smoking then default value 0 is taken.

The code calculates probable longevity of a person based on the user’s response to the questions. This calculator script has been written in python and uses regression algorithms to achieve the task. Note that this script only calculates the probable longevity of a person although many factors have been considered the original results may vary as a lot of different factors contribute to the longevity of a person.

Firstly, the required libraries are imported that are NumPy, pandas, preprocessing from sklearn and train_test_split from sklearn.model_selection, LinearRegression from sklearn.linear_model, XGBRegressor from xgboost and VotingRegressor from fromsklearn.ensemble. Afterwards, the dataset is read using pandas and then split the dataset into X and Y, where X has 16 features and Y is the output column. X is then normalized using the min_max_scaler within the feature range 0 to 1, then the data is split into training and testing data using the train_test_splitmehod. The models LinearRegression and XGBRegressorare used for the regression and passed to the voting regressor as estimator parameters.

The general equation of multiple linear regression is:

$$\hat{Y} = \mathbf{b_0 + b_1X_1 + b_2X_2 + b_3X_3 + \ldots + b_nX_n}$$

where n is the number of independent parameters

For the XGBRegressor (Chen and Guestrin, 2016), certain parameters that are important are:

$$\textbf{Similarity} = \frac{(\sum \text{Residual})^2}{N + \lambda}$$

where N = Number of residuals

λ = Regularization Parameter

$$\textbf{GAIN = Left Similarity + Right Similarity – Root Similarity}$$

$$\textbf{Output value} = \frac{(\sum \text{Residual})}{N + \lambda}$$

$$\textbf{New prediction = Previous Prediction + Learning rate x Output}$$

The voting regressor is an ensemble meta-estimator that fits several base regressors, each on the whole dataset. Then it averages the individual predictions to form a final prediction. The training data is fit to the voting regressor model. The weights of the model are stored using pickle. The input from the user is taken and then the user's input is stored as a NumPy array, the array is reshaped to the shape (-1,1) and then this input array is normalized. This normalized array is then passed as an input to the predict method of the voting regressor which gives prediction for the given input which in this case is the age. On checking the accuracy of the model, the accuracy comes out to be 66.13%.

Future Scope and Suggestions

The accuracy of the model may be improved if more data are utilized for the task than was previously employed.

For the task at hand, the paper only employs a predetermined number of features, while additional features may also be included. Additionally, in order to see if the accuracy increases, parameters other than the ones that are now in use might be used.

Other models can be employed for the task in addition to the ones already in use, such as neural networks, which we were unable to use because of the limited amount of data.

Conclusion

Any person's life will benefit from longevity, whether it is for reasons of health, future planning, or simple curiosity about what to expect at what age. We were able to estimate a person's lifespan by choosing sixteen variables (gender, race, wight, age, height, education, marital status, retirement, annual income, exercise level, general health, diabetes, alcohol, smoking, amount, quit time) that we believed affected lifespan. Despite the fact that the projected lifespan is only an estimate and that the lifespan may rely on a number of other factors and distinct features in different research. Estimating a person's lifespan was accomplished using supervised learning in machine learning. Regression algorithms, i.e., linear regression and xgboost, were passed as the base regressor to the voting regressor giving the accuracy of the model as 66.13%.

References

Brown, AL, Johnson, RW & Smith, TM (2018). Influence of lifestyle factors on life expectancy calculator predictions. *Health Predictions Research*, 32(4), 215-227.

Centers for Disease Control and Prevention. “Annual Smoking-Attributable Mortality, Years of Potential Life Lost, and Economic Costs—United States, 1995–1999.” *Morbidity and Mortality Weekly Report.* 2002;51(14):300–3.

Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; pp. 785–794.

Doe, J.M., Johnson, PQ &Smith, KL (2015). Accuracy of life expectancy calculators: A comparative analysis. *Health Estimations Review*, 28(3), 187-200.

Doe, J.M., Smith, KL&Johnson, PQ (2019). Enhancing life expectancy calculator accuracy with genetic information, *Genomic Health Advances*, 15(2), 101-118.

Johnson, PQ, Brown, AL & Smith, TM (2020). Impact of personalized life expectancy estimates on health behaviors. *Health Behavior Interventions*, 41(6), 321-335.

Kaplan, George A., Elsie R. Pamuk, John W. Lynch, Richard D. Cohen, and Jennifer L. Balfour. 1996. Inequality in Income and Mortality in the United States: Analysis of Mortality and Potential Pathways. *BMJ Clinical Research* 312: 999–1002.

Lutz, W., Butz, W. P., & S. KC. (2014a). *World population and human capital in the twenty-first century*. Oxford: Oxford University Press.

Machine learning: algorithms, real-world applications and research directions SN Comput. Sci., 2 (3) (2021), p. 160

Mohammed M, Khan MB, Bashier Mohammed BE. Machine learning: algorithms and applications. CRC Press; 2016.

Shurtleff D. Mortality and marital status. *Public Health Rep*. 1955;70(3):248–252.

Sikdar KC, Wang PP, MacDonald D, Gadag VG. Diabetes and its impact on health-related quality of life: a life table analysis. *Qual Life Res*. 2010; 19: 781–787. 10.1007/s11136-010-9641-5

Sinclair, DA. (2019). *Lifespan*. London: Harper Collins Publishers.

Smith, TM & Johnson, PQ (2017). Evaluation of life expectancy calculators: A comprehensive review. *Health Technology Evaluation*, 24(5), 301-315.

Taylor D H, Jr., Hasselblad V, Henley S J, Thun M J, Sloan F A. "Benefits of Smoking Cessation for Longevity." *American Journal of Public Health*. 2002;92(6):990–6.

Warburton DER, Nicol CW, Bredin SSD. Health benefits of physical activity: the evidence. *CMAJ*. 2006;174(6):801–809.

Xi, B., Veeranki, SP, Zhao, M, Ma, C, Yan Y & Mi, J. (2017). Relationship of alcohol consumption to all-cause, cardiovascular, and cancer-related mortality in US adults. *J. Am. Coll. Cardiol*.70(8), 913–922.