Machine Learning: Reasons Behind Average Lifespan

CS4991 Capstone Report, 2022

Rishal Jagtap
Computer Science
The University of Virginia
School of Engineering and Applied Science
Charlottesville, Virginia USA
rj4rc@virginia.edu

Abstract

The quality of life across Virginia is largely volatile due to its unique property of having both rural and urban cities, attributes that can increase the gap in average lifespan for its residents. Machine learning models and tools were used to better understand the strengths and limitations of these cities along with common characteristics that can be used to relate them to one another. This process requires the use of the Virginia Life Expectancy and Health Opportunity Index database which holds key information regarding Education, Health Access Opportunity and their corresponding Life Expectancy values. T-SNE (t-distributed Stochastic Neighbor Embedding) is then used to model the higher dimensional data on a two dimensional graph that is more easily interpretable. From the data, major attributes that lead to a longer lifespan were wealth, education, and accessibility to needed health care, mostly associated with urban cities. Further research would stretch the range of cities to cover all of the United States and gather more information to create a dedicated plan of action, tailor-made for each city, rather than implementing a blanket solution.

1. Introduction

Machine learning algorithms are one of the newest and most powerful methods of analyzing big data in today's world, where most datasets are very dense. During the COVID-19 pandemic, two of the biggest concerns for the general public were the impact of the virus on the average citizen, and the unexpected side effects of being exposed to it.

However, COVID-19 alone was not the cause of death. In addition to the sickness, COVID-19 would also lower the immune system and cause people to succumb to other diseases. This problem prompted the main research question of this technical report: Predicting the general life expectancy in Virginia with regard to its diverse geography. According to the CDC, Virginia has the 18th highest life expectancy rate at roughly 77.6 years as of 2020. Additionally, COVID-19 ranked third, right below cancer at second and heart disease at first place as the leading cause of death for Virginians.

From these initial values, we sought to determine the root causes of lower life expectancy in Virginia. Virginia is unique in that it sits at the cusp between urban and rural lifestyles. Although 88% of its geography is considered rural, largely in the middle and southern section, Virginia's highest density cities are located in its northern section, which is completely urbanized. This makes for a more complex dataset, as the state can effectively compare the rural and urban lifestyle.

2. Background

We propose the implementation of multiple algorithms as a way to determine possible causes for lower life expectancy in Virginia. Our hypothesis was that rural areas of Virginia were more likely to have a lower average life expectancy due to the lower access to health care opportunities. Additionally, economic factors play a major role in this hypothesis, as urban cities offer more job opportunities and have a richer education system. In contrast, the cause of concern for urban cities was pollution rates and spread of diseases, which could potentially influence lung-related cancer and other respiratory issues leading to death.

3. Related Works

The most relevant sources of information for this report were the Centers for Disease Control and Prevention (2022) graphical charts which highlighted trends in life expectancy for male and female Virginians. The graphs highlight data captured in 2020, when the effects of the pandemic were at their peak. Additionally, the CDC's National Center for Health Statistics provided information regarding common causes of death in Virginia. This heavily influenced which features in the dataset would be relevant to the research question. Features such as education, annual income, and population density seemed to provide a clearer image for life expectancy.

Smith (2022), with the Virginia Department of Health, expands on the rural culture of Virginia's mid and southern sections. This document was key for understanding why clusters in the dataset appeared most closely to the urban areas of Virginia.

4. Process Design

Our team split the project into two major phases of machine learning models, both of which provide their own distinct graphs so that attributes and variables were testable as by themselves and in tandem. This method was used so that legislative members can have a whole picture on the variables as there may be cases in which a solution designed to tackle the most dangerous factor may not effectively handle the lesser factors that build into a greater threat.

4.1 K-Means Clustering

The first of the phases utilizes the K-means clustering model, which attempts to cluster data based on life expectancy against the variables provided. For the model, we tried to identify three different clusters with respect to life expectancy. This model is useful because it will take into account all 23 variables associated with each point, then connect each to the closest centroid. The objective is to have the data split into three distinct sections on the graph.

The next step would be to find the commonality within the data points of each cluster. This means that data points around a centroid should have similar features that resulted in their being close to one another. An example would be that for a single cluster, every point may have a high average yearly income that resulted in higher life expectancy.

Last, as part of the K-means clustering model, if the clusters are calculated with respect to the longitude and latitude of the State (in this case Virginia), it is possible to visually examine the clusters with respect to communities throughout Virginia.

4.2 PCA and T-SNE Modeling

The second phase of the project uses both Principal Component Analysis (PCA) and T-distributed Stochastic Neighbor Embedding (T-SNE) as a means of determining how many factors are actually relevant to ensuring a higher life expectancy for communities. Objectively, PCA is used to remove the

weaker correlations in the data set. If for example, a 95% accuracy is required, then PCA will take the higher dimensional data of 23 variables, and systematically remove them one-by-one while trying to maintain the overall correlation of the data set. This could result in 23 dimensions being reduced to ten where the model determined that the lowest 13 correlations did not have a meaningful effect on the outcome.

Likewise, T-SNE also attempts to remove variables, but does so by removing them based on their effects on local rather than global clusters. Like the K-means algorithm, this model checks the distance of points within a cluster, but then removes variables and rechecks the distance to find drastic differences.

5. Results

This project designed a tool to help members of the Virginia Legislatures to pinpoint major causes of lower life expectancy across all of Virginia. Our aim was to enable legislators in creating tailor-made solutions rather than having to use blanket policies might not meet the needs of struggling communities.

We devised the various studies and associated tools to be easy to understand, with graphs as visual aids. While the models may show correlations within the data, the tools will not act as a standalone aid that can be used to create solutions. It will be up to the legislators to use the information appropriately. For example, the model may find correlations in lower life expectancy with regards to rural communities, but will not provide a solution to such a problem.

The Initial K-means data determined that education and community environment ratings were both integral for higher life expectancies. As positive factors, increases in both education and community environmental ratings were present as life expectancy increased, indicating correlation and potential causation among the population of Virginia. Along with the Initial K-means data indicating that education and community environment ratings were positive factors, the Second K-means data, which included the latitudinal and longitudinal data for Virginia also highlighted other significant factors that affect life expectancies. Based on the inclusion of positioning, suburban areas appeared to have the greatest life expectancy among the various socio-economic stratas present across the state. By comparison, both rural and urban locations appeared to have lower life expectancies compared to suburban areas.

Geographic clusters produced the highest life expectancies when they scored highly on walkability and low on material deprivation. Suburbs are supposed to be in the middle round in terms of economic and environmental purposes which explains their higher-than-average life expectancy.

Last, the T-SNE and PCA indicated that geographic features are strong enough to affect the other features used for the analysis. Accordingly, life expectancy increased as the algorithms transitioned from the left portion of Virginia (rurally dominated) into the upper right section (urban dominated).

6. Conclusion

From the analysis performed on the dataset, the key takeaway from the project was more aligned with reaching a greater understanding of how to utilize the data, rather than finding solutions suitable for urban and rural locations. By trying different methods on Virginia, our objective was to gain precise insight into the complex relationships between and among the factors that can affect life expectancy. Being able to use the PCA and T-SNE algorithms helped group together

the most important features of the dataset while also removing the least impactful causes.

By projecting these methods on other states, it would be possible to pin-point the threats that can bring down their respective life expectancy rates. Local government officials may find this tool extremely beneficial in creating more educated strategies for their communities.

7. Future Work

This project did allow us to gain some key insight on the types of issues most closely related to life expectancy. However, it does not provide any direct solutions to the problems at hand. Future work should include comparing and contrasting tried methods in local and inter-state level programs. This additional section would allow lawmakers the ability to more easily consume the big-data on their municipalities.

Furthermore, the project would look to spread its reach outside of Machine Learning alone and help devise actual strategies for tackling low life expectancy. Methods such as surveying citizens and performing outreach to groups that are more impacted by causes of lower life expectancy would be a long-term goal.

References

Arias, E. Centers For Disease Control Prevention. 2022. Retrieved December, 2022 from https://www.cdc.gov/nchs/data/nvsr/nvsr71/nvsr71-02.pdf

Centers for Disease Control Prevention. 2022. Retrieved December, 2022 from https://www.cdc.gov/nchs/pressroom/sosmap/life_expectancy/life_expectancy.htm

Smith, B. 2022. Virginia Department of Health. Retrieved December 2022 from https://www.vdh.virginia.gov/content/uploads/sites/76/2022/01/Virginia-Rural-Health-Plan_2-Defining-Rurality.pdf