Predicting Virginian Health and Lifespan

Byron Xu
Department of Computer Science
University of Virginia
Charlottesville, Virginia
bx7ugx

Khoi D. Pham
Department of Computer Science
University of Virginia
Charlottesville, Virginia
kdp3gh

I. Abstract

For this project, we are working with two datasets, one from the Virginia Open Data Portal and one from the Centers for Disease Control and Prevention (CDC). Our goal is to figure out the factors involved in the lifespan and overall health of the Virginia population. We will be looking at the strongest correlating features with HOI, Health Opportunity Index, and life expectancy. With this we can pinpoint if HOI and Life Expectancy are both being affected by similar features. We hope to predict things such as lifespan and HOI dependency on location, seeing variables are most important. With this information, Virginia will hopefully be able to better allocate resources towards the areas that would most benefit the overall health of Virginians. In further research, we may be able to correlate these features with other variables like socio-economic status and demographic information in order to find whether there are any hot-spots of lower HOI.

A. *Motivation*

Health and safety should always be the number one priority. We believe that good health is a human right. Everyone should be able to live comfortably and healthily. That is why we decided to go with a topic pertaining to health. Since Virginia only has a limited amount of resources, we want to figure out the the most effective areas of interest that Virginia should prioritize in order to improve the well-being of the average resident.

II. Data Collection

The primary dataset (https://data.virginia.gov/Family-Health/Health-Opportunity-Index/6q6u-dcz7) is taken from the Virginia Open Data Portal and measures HOI. The data contains 1875 data entries, with no null entries. Out of the 20 features, one is a categorical feature.

As the Virginian government has already created their own HOI predictor, we would like to expand on their work by utilizing the official CDC Virginian life expectancy data (https://www.cdc.gov/nchs/data-visualization/life-expectancy/index.html). Since both datasets include the feature of census tract, it is possible to ultimately compare the features involved in the HOI to the average life expectancy of each census tract. In the CDC dataset, there are 1907 data entries that include both the census tract and life expectancy.

III. Hypothesis

We believe that out of the 11 features, Economic Opportunity Profile will have the greatest influence on Lifespan and overall health.

IV. Related Work

Almost every other state has a quantitative health value, whether it be an index or a ranking. For instance, California has the Healthy Places Index https://www.healthyplacesindex.org/ to quantify the quality of health in regions in the state. The US census also has a couple of health related datasets, https://www.census.gov/topics/health.html. Researchers are also trying to find the main factors that affect the lifespan of Americans, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4112220/. All of these datasets try to quantify individual factors like income inequality and pollution, factors that we will use in our dataset to predict the lifespan and overall health of Virginians.

V. Method

Our task of linear regression falls under a regression task. To evaluate the ability of our machine learning algorithm, we used Sklearn and other Python frameworks to check for any null values and categorical variables. We also did some preliminary visualization of the data. Then, we split the dataset into testing, training, and validation sections.

Additionally, since the census tracts for the two datasets were written in different formats, one in an 11-digit GEOID number and one as a 4-digit, 2-decimal census tract number, we had to have one identical feature to combine the two data frames together on. In order to do so, we integrated a third data-set $/(https : //unicede.air - worldwide.com/unicede/unicede_virginia_fips_3.html/)$, which allowed us to map each county into its 3-digit identifier. This had to be done because census tracts could be the same number but located in a different county.

We used pandas to drop data entries that were faulty, such as having a county that was actually a city. Then, we added a new feature, which was a 9-digit GEOID that involves the 3-digit county identifier and the 6-digit census tract number. Thus, we created the GEOID feature in both datasets, dropped the redundant Census Tract, and merged the two together to create one singular DataFrame to apply our techniques on.

This led to our total dataset having 1117 datapoints. From there we split the data. 893 for training, 112 for validation, and 112 for testing.

Example machine learning techniques that we used in our experiment included Grid Search in order to find the best hyperparameters, Stochastic Gradient Descent as the model to predict the life expectancy, and RMSE to see how much our error is. The hyper parameters being tuned are the alpha value, loss equation, penalty, and learning rate schedule. The model and grid search is provided by Sklearn.

VI. Experiments

After figuring out how to combine the two datasets using the census tracts, the lifespan dataset and the Virginia Department of Health HOI dataset, we utilized a Stochastic Gradient Descent Regressor to predict the average lifespan of a census region using factors such as food availability, job opportunities, and other census data. The best hyper-parameters of the SGD regressor are as follows:

{'alpha': 0.01, 'learning_rate': 'constant', 'loss': 'epsilon_insensitive', 'penalty': 'l1'}

Then we ran the model on the test set and got a root mean square error of 3.84 years, which isn't too bad considering average lifespan of Virginians range from as low as 56.9 years and as high as 91.5 years. Though there is room for improvement.

In addition to using the test set, we ran 5-fold cross-validation and got the following scores:

[4.15985837 3.62128078 4.53033727 5.17782948 3.50801712]

We then tried again with a Support Vector Regression model. Again, we used grid search on the regressor, tuning hyperparameters such as the kernel, C value, degree, and gamma. The best hyper-parameters of the Support Vector Regressor are as follows:

{'C': 1, 'degree': 3, 'gamma': 'auto', 'kernel': 'rbf'}

The root mean square error was 3.76 years, slightly better than the 3.84 years of the linear regression model. Running 5-fold cross-validation, we got the following scores:

[3.81018597 4.10425674 4.14651759 5.05995893 3.83497548]

VII. Results

Of the 20 features, the most impactful aspects of a census tract on lifespan were education, food accessibility, and community environment profile. These were found using the coefficients of the SGDRegressor. These attributes make sense as wealthier communities tend to have higher lifespans and these features are usually tied with those communities. Surprisingly, access to care and walkability had a negative impact on the lifespan of a census tract. These values can be found in the Jupyter Notebook file. Income Inequality and Wellness Disparity Profile also had slightly positive coefficients with lifespan.

This could be explained by the model focusing on areas (Northern Virginia and Richmond) where there are high lifespans (due to the wealthy communities in those areas), but also have inequality problems, making the model correlate more inequality with increased lifespan.

Rural and urban communities had around the same lifespans, which was also surprising.

VIII. Conclusions

The model needs a lot of work, though this might not be because of any specific machine learning algorithms or any aspects of the tuning process, but rather the data. While a root mean square error of 3.6 years isn't bad, it is hardly accurate. The features that affected lifespan the most also did not make intuitive sense. There are a couple reasons.

Boil down something as complex as lifespan to 20 features is not enough. It doesn't consider features such as diet, exercise, wealth, crime, and other aspects of life. The lifespan of someone correlates heavily with the region they live in. Counties such as Fairfax and Alexandria (Northern Virginia communities) have a significantly higher lifespan than the rest of Virginia. The model fixates on these regions and tunes its coefficients accordingly. Though features such as income inequality and community environment profile are heavily correlated with increased lifespan, it does not mean those features cause a higher lifespan.

What was learned from this experiment was that you can't rely on a small set of features to increase lifespan. Judging from the lifespan dataset and visualization from the CDC, lifespan is mostly location dependent. At a glance, locations with higher lifespan are generally in areas with a lot of wealth. If the governor of Virginia really wants to increase lifespan, they should send general aid to poorer areas with lower lifespan, instead of looking at our model and making the entirety of Virginia have features similar to these wealthier areas.

IX. Contributions

One of our original group members dropped the course, leaving us with two members left in our team. Both members of the project contributed equally to the project, looking over the entire project and its subsections.

Byron wrote the majority of the "Method", "Abstract", and "Data Collection" section and formatted the document, while Khoi write the majority of the "Experiments" and "Results" section. Both contributed to the "conclusion" section. Byron did most of the data visualization, documentation, data-cleaning, and finding the data-sets. Khoi implemented most of the machine learning model, such as the pipeline, Grid Search, Stochastic Gradient Descent, and hyperparameter tuning.

X. Sources

Centers for Disease Control and Prevention. (2020, March 9). Life expectancy data vizualization. Centers for Disease Control and Prevention. Retrieved December 8, 2022, from https://www.cdc.gov/nchs/data-visualization/life-expectancy/index.html

Virginia Department of Health Office of Health Equity. (2020, August 31). Health opportunity index: Virginia Open Data Portal. Health Opportunity Index — Virginia Open Data Portal. Retrieved December 8, 2022, from https://data.virginia.gov/Family-Health/Health-Opportunity-Index/6q6u-dcz7