Life Expectancy

Amanda Fu-Kalilikane

Background

Over the years, due to improved management of sicknesses and better access to food and healthcare, life expectancy has improved greatly. But there is always room for improvement, what kind of factors would make the largest differences on life expectancy with the United States.

Other projects have compared life expectancy around the world, seeing how life expectancies differ with various socioeconomic status' of various countries. In this project, the focus will be just on the United States to see if we can pinpoint specific factors or even states that could use improvement.

Client

Many organizations may find this kind of information interesting. Many industries use things like life expectancy as a marketing angle, how to prolong life and how to add years to your life. Healthcare organizations or insurance companies may use this information to their benefit in determining places that require more improvement or policy changes.

Data Wrangling

Source

The data used for this project was accessed on Kaggle, and generated in 2021.

"2021 County Health Rankings National Data"

County Health Rankings | Kaggle

This database originates from University of Wisconsin, Population Health Institute, School of Medicine and Public Health at the following web address:

Rankings Data & Documentation | County Health Rankings & Roadmaps

There were additional features on the database that were not used in this project but could potentially be used for further evaluation and to increase accuracy of the model.

Summary

This database consists of several sheets that contain the data as well as feature descriptions with further information about where the data came from or how it was derived. There was also a sheet that contained the county rankings throughout the country.

The main sheet of interest, containing all of the features and data except for the life expectancy column which was on a different sheet. So the two sheets were merged to ensure that life expectancy was associated with the correct state and county.

Description

After joining but before cleaning the data, the dataset had 3193 rows and 523 rows.

After cleaning, the data would be down to 2908 rows and 20 columns to use for modeling.

Data Cleaning

Duplicated Values

There were no duplicate values found, so no adjustments had to be made here.

Missing Values

There were 14 features with missing values ranging from 1 to 287 missing values. Out of the 14 features, 11 features had under 5% missing values and had those rows dropped.

There were 51 missing values within the “County” column. As there are 50 states and the District of Columbia in the United States, the missing value could be explained if this row were to hold the max, min or averages for the particular state or district. The empty county in Alabama had the value 3187 for PCP Number, and a value of 5310 for MHP Number. This was compared with the sum calculated for the total PCP number in Alabama at 3187 and the total MHP Number at 5307.
Given these values, it was assumed that the null county line for each state is holding the average or mean value for that particular feature for each state. So the empty County rows were dropped.

Only the features 'Premature Deaths', 'MHP Rate', and 'MHP Number' were above 5% and these missing values were filled with the means.

Biased Features

Some features from the database were not included such as "Poor or fair health" or "Poor physical health days" were based on self-reporting by people and not well defined. The excel sheet description simply stated “Percentage of adults that report fair or poor health” or “Average number of reported physically unhealthy days per month” so it seemed that this feature would be too subjective and biased to include it for this project. As one person who thinks that they have good health could be eating and doing the same thing as someone who thinks they only have fair health.

Unreliable

There were several columns labeled 'Unreliable', which when referred back to the original data was found that unreliable meant that there was a value reported but since it was twenty counts or less, was considered unreliable. Further inspection of the original data showed that the column unreliable is a subgroup of certain features, such as unreliable for the premature death counts. This was left alone as the main features of interest would be isolated from this data anyways.

Columns

For most of the features, the feature would have the main feature information, but then is also broken into subgroups. The breakdown included 95% CI - low, 95% CI - High, quartile, and some would include rate. The columns were named with the feature and then the subgroup afterwards (i.e. premature_deathDeaths is the feature Premature Deaths, subgroup Deaths). Many columns were dropped due to this as it was only necessary to see the main feature information and not the breakdowns. The remaining columns with the main feature would also be renamed for clarity.

Table 1: Features with original name, after renaming and feature description

Feature Original Name	Feature Renamed	Feature Description
'Life expectancy'	'Life Expectancy'	Average number of years a person can expect to live
'population_Population'	'Population'	Number of people living in a specified area
'%_rural_% Rural'	'% Rural'	Percentage of population living in a rural area
'premature_deathDeaths'	'Premature Deaths'	Number of premature deaths (a death is considered premature if the individual is younger than 75)
'adult_smoking_% Smokers'	'% Smoking'	Percentage of adults that report currently smoking (age-adjusted)
'adult_obesity_% Adults with Obesity'	'% Obesity'	Percentage of adults (age 20 and older) that report BMI >=30
'physical_inactivity_% Physically Inactive'	'% Physical Inactivity'	Percentage of adults (age 20 and older) that report no leisure-time physical activity
'excessive_drinking_% Excessive Drinking'	'% Excessive Drinking'	Percentage of adults that report excessive drinking (age adjusted)
'uninsured_% Uninsured'	'% Uninsured'	Number of people under age 65 without insurance
'primary_care_physicians_# Primary Care Physicians'	'PCP Number'	Number of Primary Care Physicians in patient care
'primary_care_physicians_Primary Care Physicians Rate'	'PCP Rate'	Primary Care Physician per 100,000 population

'mental_health_providers_# Mental Health Providers'	'MHP Number'	Number of Mental Health Providers
'mental_health_providers_Mental Health Provider Rate'	'MHP Rate'	Mental Health Provider per 100,000 population
'preventable_hospital_stays_Preventable Hospitalization Rate'	'Preventable Hospital Rate'	Discharges for Ambulatory Care Sensitive Conditions per 100,000 Medicare Enrollees
'mammography_screening_% With Annual Mammogram'	'% Mammogram'	Percentage of female Medicare enrollees having an annual mammogram (age 65-74)
'flu_vaccinations_% Vaccinated'	'% Flu Vaccine'	Percentage of Medicare enrollees having an annual flu vaccination
'unemployed_% Unemployed'	'% Unemployed'	Number of people ages 16+ unemployed and looking for work
'Median household income'	'Median Household Income'	The income where half of households in a county earn more and half earn less

Exploratory Data Analysis

Potential Outliers

With the Life Expectancy column sorted, there seemed to be some outliers or incorrectly labeled data points. There are several points that are above 90, with two points going as high as 102 and 101, with several of the high life expectancy data points attributed to counties in Colorado. These particular points were looked into, and another source was found stating that those were the correct life expectancies of those particular counties.

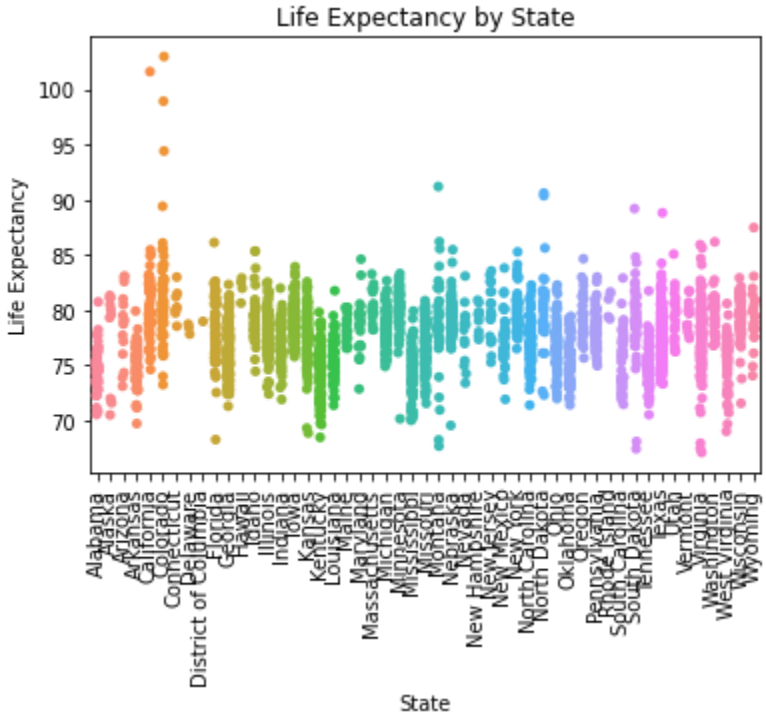

Figure 1 : Life Expectancy of each county organized by State.

Life Expectancy

Overall, taking the average life expectancy by state, the lowest mean was Mississippi with a life expectancy at 74.17, the highest with Hawaii at 82.006 and the second highest Colorado at 81.308.

Feature Correlation

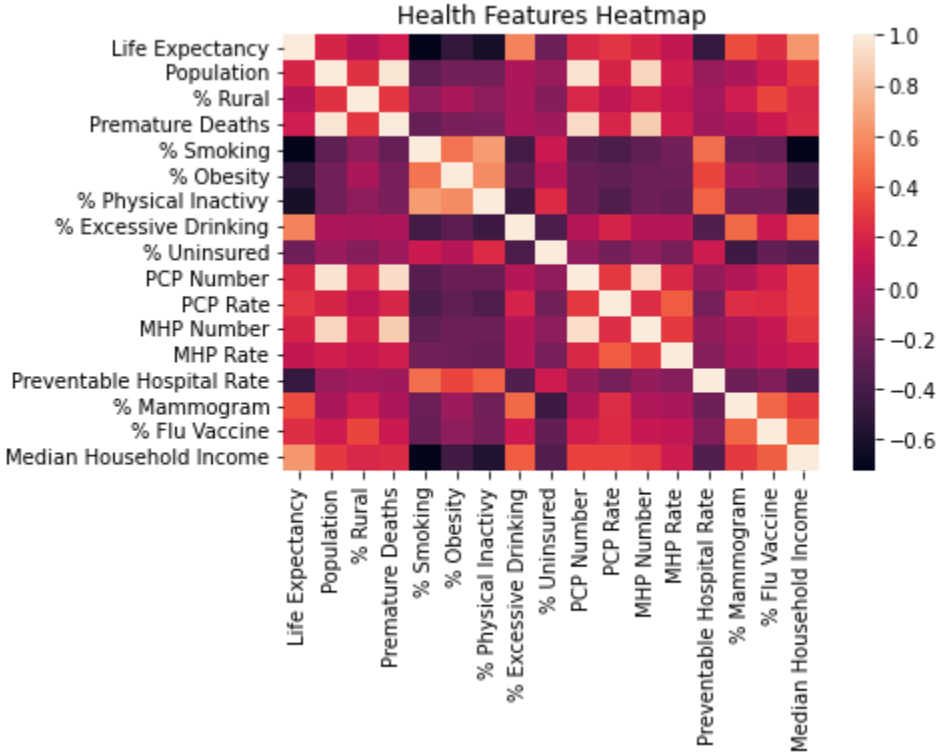

Figure 2: Health Features Correlation Heatmap where lighter color is more positively correlated and darker is more negatively correlated

When considering the other features, many of the correlations were expected, such as obesity percentage, smoking percentage and physical inactivity percentage having a negative correlation with life expectancy, as these are not considered healthy lifestyle choices.
There was not as large of a correlation between mammograms and flu vaccine with life expectancy as expected as these aspects are things that would be considered positive influences on life expectancy as they would help prevent illness.
Another correlation that was noted, the % smoking, % obesity and the % excessive drinking were all negatively correlated to the median household income as well.

Features that were found to be highly positively correlated with one another included the population, % rural, MHP number and the PCP number. This would be expected because of course the number of providers would increase with the population.

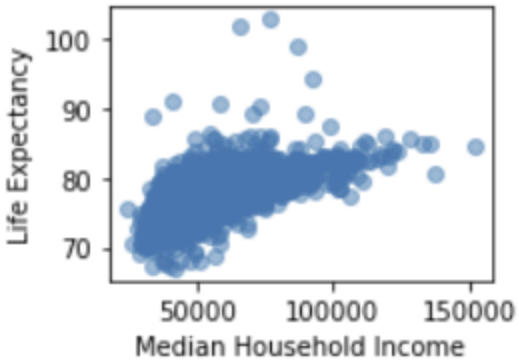

Figure 3: Life Expectancy vs Median Household Income

In comparing the separate features to life expectancy, it seems that while households that have a lower median income are correlated to a lower life expectancy, the life expectancy evens out as income increases. This can provide an assumption that the middle class has moderate to good access to healthcare, but that the lower class may require better access.

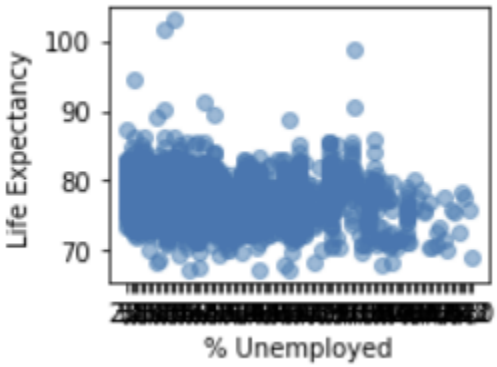

Figure 4: Life Expectancy vs % Unemployed

In comparison, the % Unemployed to Life expectancy shows that the populations that have a higher percentage of unemployed people have a small range that tends in the slightly lower range on the life expectancy scale. But overall, the range looks very flat.

Model Selection

Feature Selection

To improve accuracy, feature selection was used with Random Forest Regressor as the estimator to determine the best features and improve model performance. Features were reduced from 17 to 10. The following features were used for all models.

- Life Expectancy
- Population
- % Rural
- Premature Deaths
- % Smoking
- % Obesity
- % Physical Inactivity
- MHP Rate
- Preventable Hospital Rate
- % Unemployed

Data Scaler

For both modeling the Linear Regression and K-Nearest Neighbors, a Standard Scaler was used to transform the data. No scaling was necessary for Random Forest Regressor

Random Forest Regressor

With the default parameters, the random forest regressor performed as follows:

- Train Score: 0.946199
- Test Score: 0.59203
- MSE: 3.98182
- RMSE: 0.68108

The model seems to perform well in the lower ranges but has points that fall further away from actual value as predicted life expectancy increases. The train score was the highest of all the models with the test score much lower, suggesting that this model overfit.

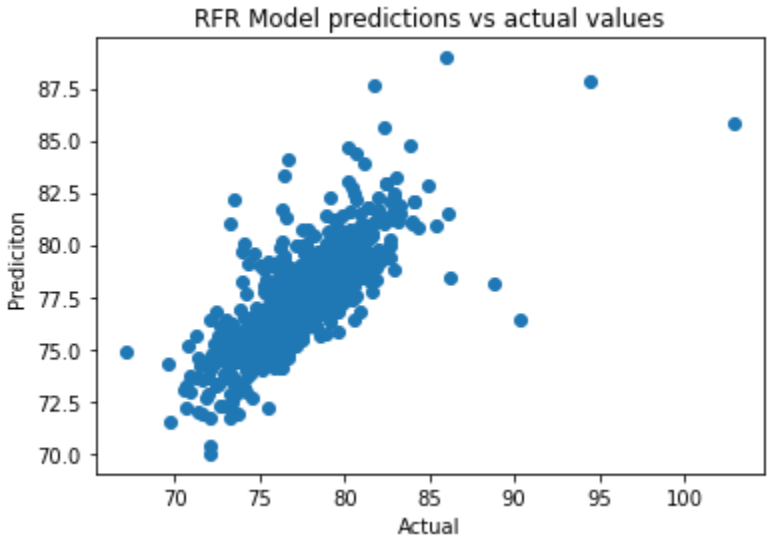

Figure 5: Comparison of actual Life Expectancy Values to predicted values for Random Forest Regression using default parameters and feature selection

Linear Regression

With the default parameters, the Linear Regression model performed as follows:

- Train Score: 0.58808
- Test Score: 0.56262
- MSE: 4.26891
- RMSE: 2.0661

This model was much improved in that the training and test score were much closer together. The train and test score are much closer together showing a much more accurate model.

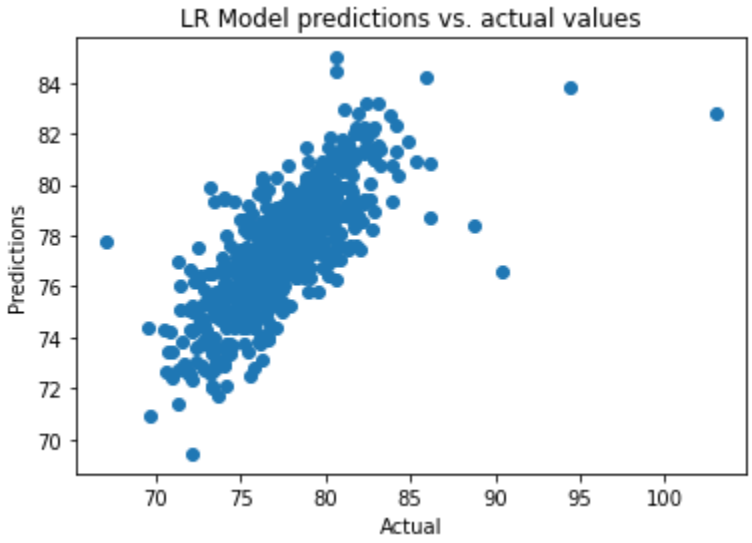

Figure 6: Comparison of actual Life Expectancy Values to predicted values for Linear Regression using default parameters and feature selection

K-Nearest Neighbors

With the default parameters, the Linear Regression model performed as follows:

- Train Score: 0.63405
- Test Score: 0.56545
- MSE: 4.26891
- RMSE: 2.0661

While this model was better in the overfitting than Random Forest Regression, it still overfit more than the Linear Regression model.

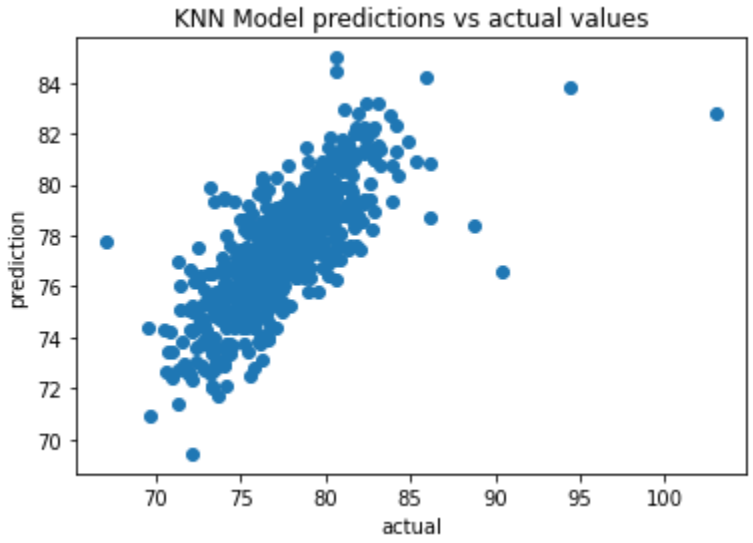

Figure 7: Comparison of actual Life Expectancy Values to predicted values for K-Nearest Neighbors using default parameters and feature selection

Table 2: Comparing Model scores for train, test (R^2), MSE, RMSE

Model	Train	Test (R^2)	MAE	MSE	RMSE
RFR	**0.94619**	**0.59203**	**1.36307**	**3.98182**	**0.68108**
LR	**0.58088**	**0.56262**	**1.39997**	**4.26891**	**2.0661**
KNN	**0.63405**	**0.56545**	**1.39997**	**4.26891**	**2.0661**

Future Considerations

One thing to consider was that this data was derived from 2021, in the midst of COVID pandemic. Whether or not that had an effect on these life expectancy numbers would be very interesting to look into. As people were quarantined and COVID spread was plateauing, perhaps life expectancy improved during this year. On the other hand, there is the potential that because COVID claimed the lives of so many, that this particular year, the life expectancy could be lower than others. It would definitely be interesting to see how this compares with previous years and future years.

Conclusions

Although the Random Forest Regression Model performed best in testing scores, the Linear Regression model was most accurate for predicting Life Expectancy with the least amount of overfitting. In the future, the model could still be improved by considering other features as this project did remove many columns in the initial stages.