Life Expectancy

Amanda Fu-Kalilikane

Problem

What factors affect the average life expectancy in the United States the most, and how can they be improved for the upcoming year?

Solution

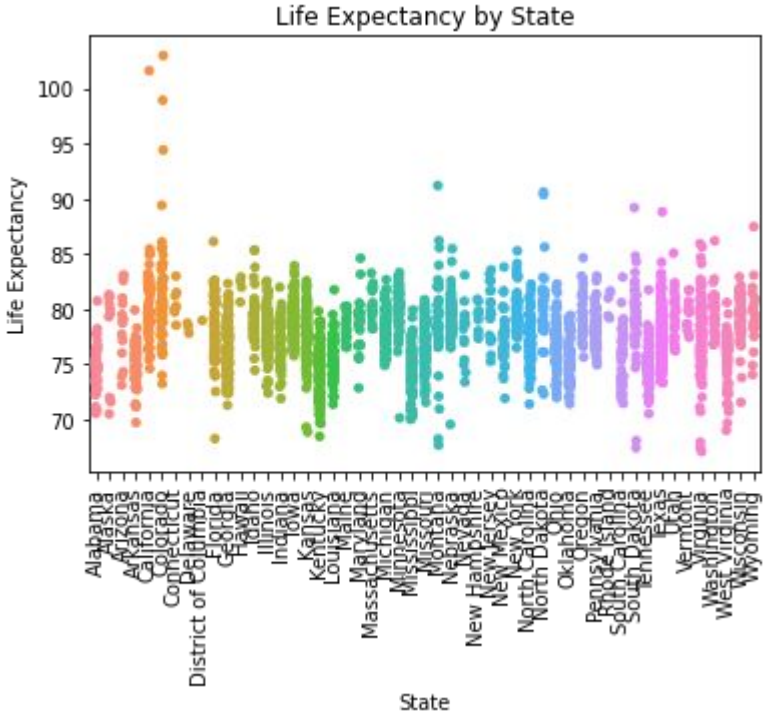
After data cleaning and exploration, feature selection was completed to find the features that would affect the model process the most and produce the most accurate result.

By reducing the counties numbers on %smoking, %obesity and preventable hospital rate could improve life expectancy as these were features shown to be negatively correlated and were chosen by feature selection to be features that are most relative to the dataset.

Outliers

There are several points that are above 90, with two points going as high as 102 and 101

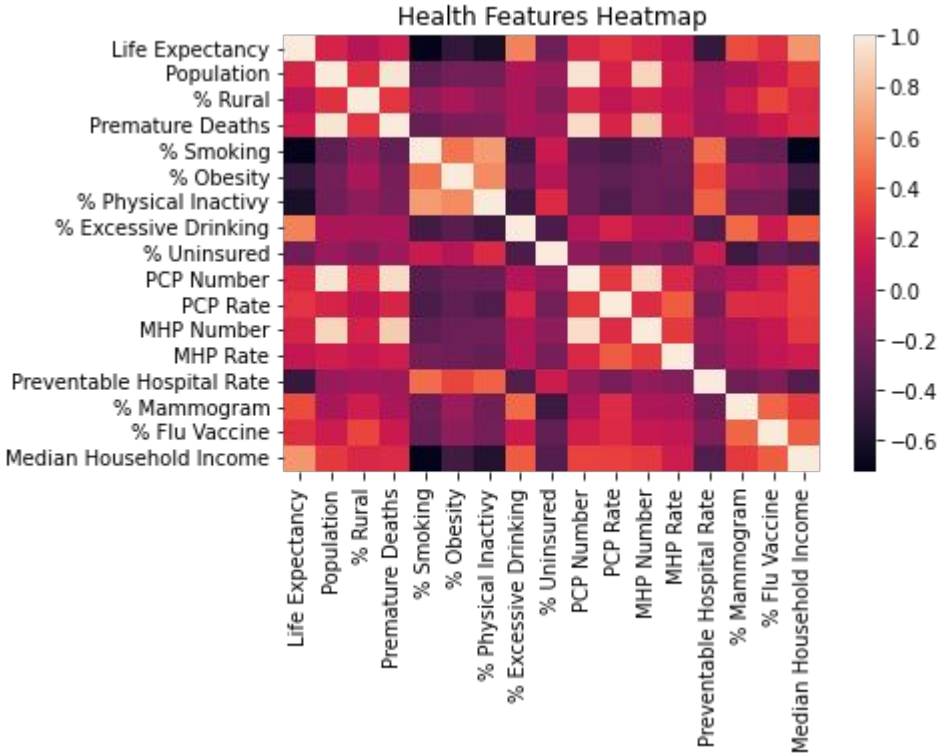
Further research showed that several of these points were in Colorado and there are other sources that stated the same life expectancy for these counties.



Features Heatmap

Expected

- Features considered to not be healthy lifestyle choices would have a negative correlation with life expectancy



Feature Selection

Feature Selection for Accuracy

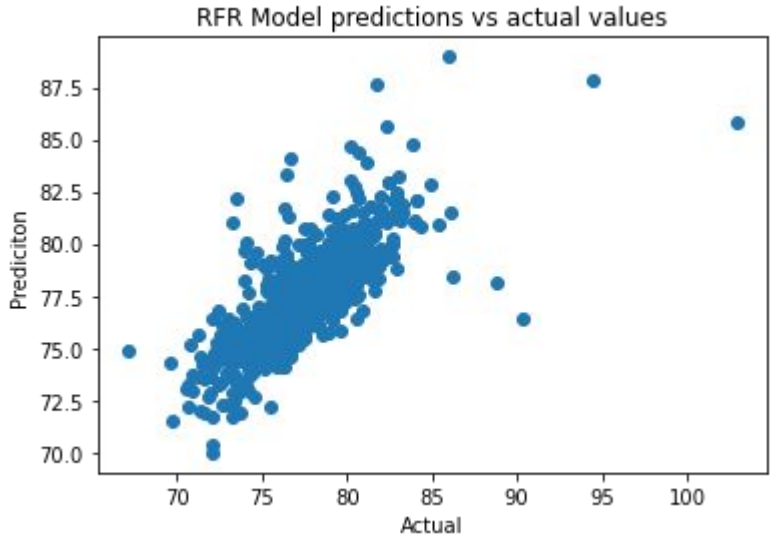
Feature selection was used with Random Forest Regressor as the estimator to determine the best features and improve model performance. Features were reduced from 17 to 10. The following features were used for all models.

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

Model - Random Forest Regressor

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

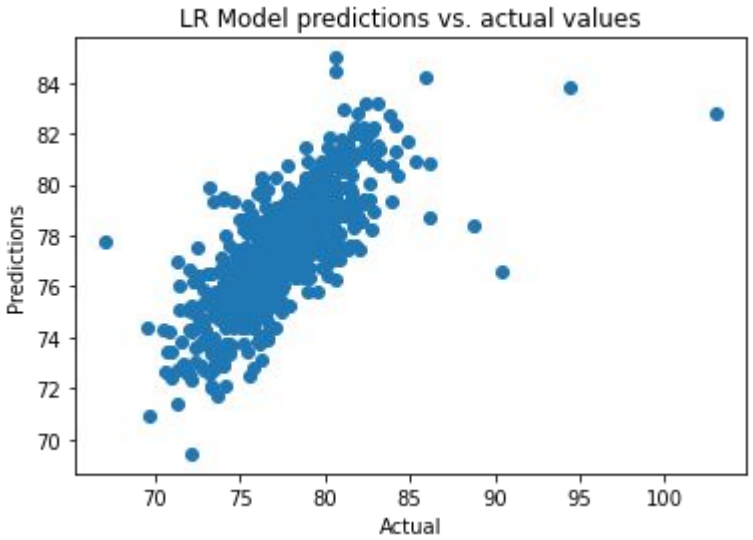
The model seems to perform well in the lower ranges but has points that fall further away from actual value as predicted life expectancy increases. The train score was the highest of all the models with the test score much lower, suggesting that this model overfit.



Model - Linear Regression

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

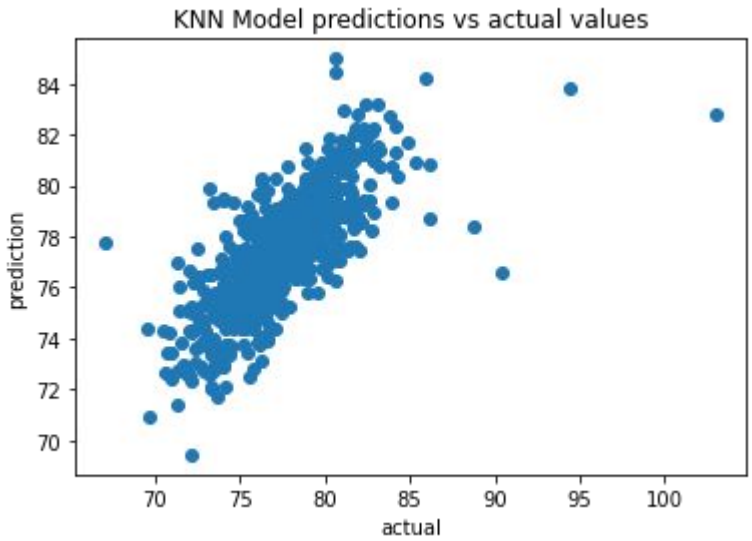
This model was much improved in that the training and test score were much closer together. The train and test score are much closer together showing a much more accurate model.



Model - K-Nearest Neighbor

- 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.



Comparing Model Scores

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

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.