Machine Learning Applications for Life Expectancy Prediction

ECON590 Research Paper

Ray Carpenter, Jason Jiminez, Nick Kirkman

UNC Chapel Hill 17
November 2022

Contents

1 Introduction

One of the most important metrics for evaluating the overall health of a population is life expectancy. In the pre-modern, poor-world life expectancy was around 30 years old. Since then, life expectancy has increased rapidly and now sits at around 73 years globally. This paper seeks to examine the factors that contribute to an individual's life span. Using data from the Health and Retirement Study (HRS) conducted by the Institute for Social Research at the University of Michigan, this paper utilizes various machine learning techniques to identify the main contributors to lifespan.

Not only is it important to know what contributes to a long life, identifying key factors for life expectancy has a number of policy implications as well. Certain U.S. state policies are specifically crafted to improve life expectancy, such as a tax on tobacco or environmental policies. The same concept applies in the opposite direction, where legislation may have negative impacts on life expectation unintentionally. Identifying the factors that contribute to determining life expectancy may help policy makers make better choices for their constituents. Additionally, information on life expectancy is invaluable to insurance companies. Life expectancy is the most important metric insurance companies use in determining life insurance premiums. Therefore, life expectancy calculation is a critical practice in actuarial science. With the multitude of machine learning tactics available and their proven predictive power, many of these insurance companies still rely on traditional methods of underwriting. While relatively accurate, this paper seeks to expand upon these predictive methods for life expectancy while also identifying some of the important factors in determining a person’s age at their time of death.

This paper uses a total of 9 different machine learning techniques to predict an individual’s lifespan. More details on each of these models will be shown in the methodology and results section. As expected, certain models performed better than others. The best tactic, support vector machines, was to predict an individual’s lifespan within 2.24 years on average on out-of-sample data.

2 Literature Review

Given the importance of the topic, there is a robust amount of literature examining life expectancy estimation using a variety of different methods. Shin, Lee, and Kim (2020) used principal component analysis to extract comorbidity patterns, utilizing data from a 10-year follow up analysis of the Korean Longitudinal Study of Aging. They identified four main principal components: principal component 1 (Disease of the circulatory system: diabetes, heart disease, and hypertension), principal component 2 (Disease of visual and musculoskeletal system: difficulty in daily activities due to visual impairment, arthritis and rheumatism, and fall during the last 2 years), principal component 3 (Disease of mental disorder: psychiatric and cerebrovascular diseases), and principal component 4 (Disease of the respiratory and digestive system: liver disease except fatty liver, diagnosis of malignant tumor, and chronic lung disease). They find that principal component 4 is associated with an increased mortality rate in the population aged between 45 and 64 years.

Li et. al (2018) used data from the Nurses' Health Study and the Health Professionals Follow-up Study to find that individuals who adopted a collection of five "low-risk factors" defined as never smoking, body mass index (BMI) 18.5–24.9 kg/m2, 30+ minutes/day moderate to vigorous physical activity, moderate alcohol intake, and a high diet quality score had a projected life expectancy of 43.1 years at age 50 for females and 37.6 years for males at age 50. Comparatively, the life expectancy at age 50 of individuals who adopted none of the low-risk lifestyle factors was 29 years for females 25.5 years for males.

Beeksma et. al (2019) used clinical data and a supervised machine learning task, mainly involving long short-term memory recurrent neural networks on the medical records of deceased patients, to predict life expectancy. They found that while doctors overestimated life expectancy in 63% of the incorrect life expectancy prognosis, the model overestimated life expectancy in only 31% of the incorrect prognoses. They suggest that machine learning and natural language processing techniques offer a feasible and promising approach to predicting life expectancy

A large portion of the research in mortality revolves around a specific cohort of individuals, such as those diagnosed with certain illnesses. This paper seeks to take a more general approach, looking at both lifestyle and socioeconomic factors along with medical factors to predict an individual's lifespan. Beeksma et. al (2019) specifically relies on clinical data that lacks information on a person's lifestyle habits and only contains medical records. We believe the predictive power of supervised machine learning tactics can be enhanced with the inclusion of this data.

3 Data

3.1 Data Source

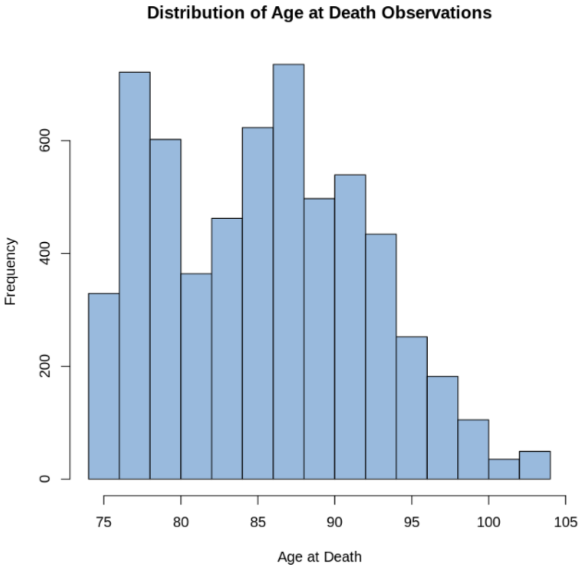
We used the data presented by the Health and Retirement Study (HRS), which is a longitudinal panel study that surveys a representative sample of approximately 20,000 people in America, supported by the National Institute on Aging and the Social Security Administration. The study, published by the University of Michigan, has followed respondent in-depth interviews for 30 years. HRS began in 1992 when the very first participants agreed to share their stories. It is one of the leading sources for information on the health and economic circumstances of adults over age 50 in the United States. The HRS collects information on physical and mental health outcomes, socioeconomic information such as an individual's income, and demographics. Research can use the dataset to address important questions about the challenges and opportunities of aging.

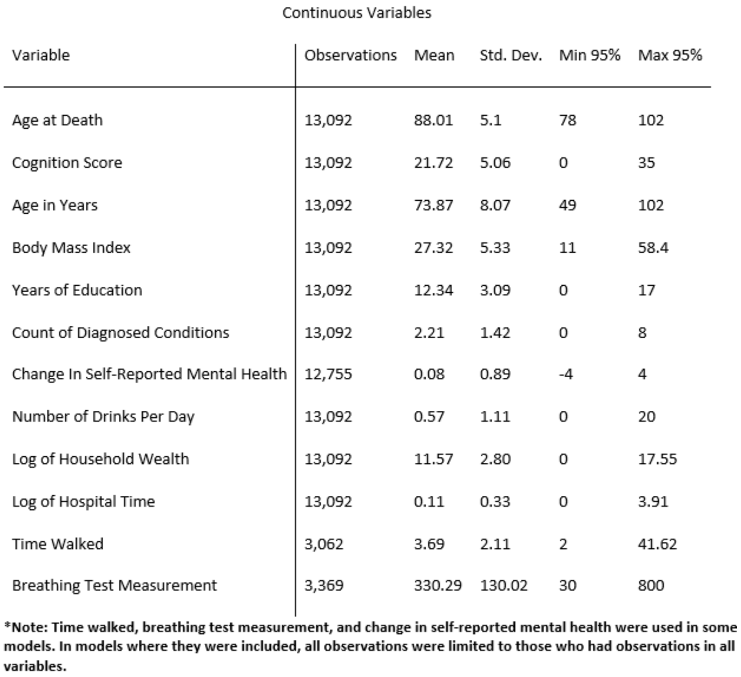
The original dataset came with 150,000 observations and thousands of variables. It was also in a wide panel format. We converted it to a long format using Stata. Due to it being panel data, we dropped individuals who had more or less than 7 waves. This made it so we didn't have to add weights to the each observation. However, in future studies we would use weights to use a much larger portion of the data. We then switched over to R to finish the cleaning. Next, we read through all 1200 unique variables and picked 150 variables that we thought may have an impact on life expectancy. We picked variables that were used in other similar studies and added other variables we thought might explain the outcome. From here we cleaned the data. We removed categorical variables that made up less than one percent of the observations and removed observations with missing values (addressed below). This step saw our observations drop from about 150,000 to around 75,000. The panel nature of this dataset adds a correlation between observations. However, here we do not account for this which is a limitation of our study. We treated each observation as separate from another.

Missing data was prominent within the survey. A drawback of the HRS is that since it was recorded periodically over the course of decades, not all the questions were answered by respondents due to a multitude of reasons. For example, some respondents had passed away during part of the study, or some did not know the answer to characteristic questions, others simply chose not to respond to the survey and send in their results. Subsequently, variables that contained a large portion of missing observations were omitted in our model as well. This still left a robust number of observations for analysis. Although this may leave out relevant information, it greatly simplified the process for our analysis. Variables with a large number of misplaced values do not represent the dataset well overall and are omitted.

3.2 Summary Statistics

Summary statistics of some key continuous variables are displayed below in Figure 1 along with a histogram of our outcome variable. Summary statistics for categorical variables can be found in the appendix. The variable age at death had 13,092 observations. The mean was 88.01. In other words, the average life expectancy of our data was 88.01. The youngest observation at death was 78 and the oldest was 102. We used all observations for most of our models, which was 13,092. Time walked and the breathing test scores were left out of most models. As they limited our observation. They are included in the summary statistics table and have different observation counts for this reason. We used economic, mental health, and physical health variables to explain the outcome.





Continuous Variables

Variable	Observations	Mean	Std. Dev.	Min 95%	Max 95%
Age at Death	13,092	88.01	5.1	78	102
Cognition Score	13,092	21.72	5.06	0	35
Age in Years	13,092	73.87	8.07	49	102
Body Mass Index	13,092	27.32	5.33	11	58.4
Years of Education	13,092	12.34	3.09	0	17
Count of Diagnosed Conditions	13,092	2.21	1.42	0	8
Change In Self-Reported Mental Health	12,755	0.08	0.89	-4	4
Number of Drinks Per Day	13,092	0.57	1.11	0	20
Log of Household Wealth	13,092	11.57	2.80	0	17.55
Log of Hospital Time	13,092	0.11	0.33	0	3.91
Time Walked	3,062	3.69	2.11	2	41.62
Breathing Test Measurement	3,369	330.29	130.02	30	800

***Note: Time walked, breathing test measurement, and change in self-reported mental health were used in some models. In models where they were included, all observations were limited to those who had observations in all variables.**

Figure 1: Summary Statistics

4 Methodology

The following methods were used for prediction (with limited inference as well): simple linear regression, lasso, ridge, elastic net, regression trees and random forests, boosting, SVM, KNN, and Partial Least Squares (PLS). To select optimal parameters and to evaluate the ability of each model, cross-validation was used to test the models on out-of-sample data. In certain cases, a 50/50 training and testing split was used and in others K-fold cross-validation was implemented using 10 folds. The key metric this paper uses for comparison among models is the root mean squared error, or RMSE. Further information on each model will be given under the specific results.

5 Results

5.1 Simple OLS

The first model we ran was a simple Ordinary Least Squares model. This was just the regression of age at death on all of the linear predictors. There were no interactions, higher order terms, or any other non-linear terms. This was to set a bar for our root mean squared error to be compared against. The Root mean squared error was 11.86. This was using cross-validation on a validation set. Without doing anything to our model, we were almost 12 years off on predicting life expectancy. This is worse than guessing the mean life expectancy for the sample. Which yielded an RMSE of 9.37.

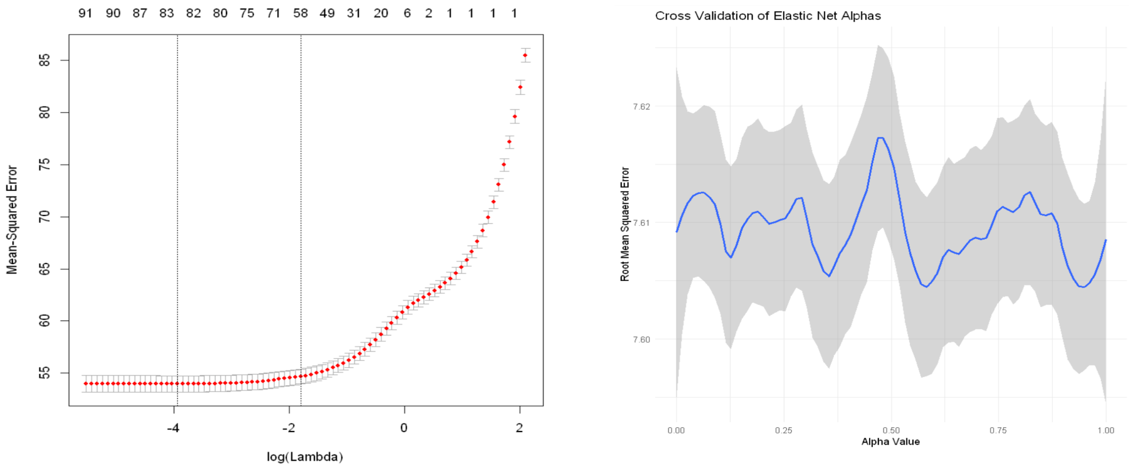
Even though this is a model studying prediction, this model can give us coefficients that may be interesting for inference. Age had a positive effect on life expectancy. Age will be squared in future models to see how it changes the prediction as one ages. Being a veteran had a positive marginal effect of 0.43 years. Being Hispanic had a negative marginal effect of -1.73 years. Being Female was not statistically significant. All waves were statistically significant. When measured against New England, the Mid-Atlantic region had a positive marginal effect of 1.23. Having Arthritis had a negative marginal effect of -0.67. The log of income from social security, hospital time, and household wealth were all statistically significant. With marginal effects of 0.03, 0.74, and .12 respectively. Being Jewish was statistically significant with a negative marginal effect of -1.59 when measured against the base group of being Protestant.

None of the health statuses were statistically significant. These included depression, sleep being restless, body mass index, change in mental health, and reporting of psychological problems. Drinks per day was also not statistically significant. All other variables had a statistical significance below a 99 percent confidence interval.. There were a lot of insignificant variables that most likely introduced too much variation. This is why we needed to cut down on the number of predictors. That is why we used Lasso, Elastic-Net, and Ridge regressions to do so next.

5.2 Elastic-Net/Ridge/Lasso

A function that used elastic-net with cross-validated test error, cross-validated alpha parameters, and cross-validated lambdas was used. This prevented us from having to choose between lasso, ridge, and every alpha parameter in between 0 and 1 in the elastic-net. All three of these models add a parameter that penalizes the number of predictors. They shrink the non-significant variables toward 0. In the case of lasso, it may remove these coefficients. Ridge regression brings the coefficients close to 0 but it will not entirely remove them. The elastic-net model we used cross-validated the lambdas each time it was run. The lambda cross-validation plot on the chosen alpha can be seen below in the graph to the left in Figure 2.

Figure 2: Lambda and Alpha Selection



Above to the right in Figure 2 is the plot of recording the RMSE as the alpha value was increased from 0 to 1 in increments of 0.01. It shows that there was not much difference between any of the alpha values. This also means there is not much difference between the performance of lasso and ridge regression in this case. As they take on an alpha value of 1 and 0. However, the model that produced the lowest RMSE had an alpha of 0.60. This is what was used to produce the graphs here and the coefficients.

The best elastic-net model had an RMSE value of 7.61. This is much better than the OLS model found before. It is also about two years more accurate than just guessing the average age at death in this sample. The lambda value that was one SE from the minimum MSE was 0.11. This is statistically the same as the lowest point, but we chose the more simple model in a prediction scenario.

We also used the reduced feature space generated from the elastic-net model and used subset selection to test every single possible interaction of predictors. This is known as best subset selection or the brute force approach. It tests every possible interaction in the feature space. We stopped the program after 1.2 million models because it most likely converged to the best model. We tried this multiple times and it produced the same results. It suggested that there are statistically significant interactions between education and change in self-reported mental health, education and BMI, having high blood-pressure and cognition score, having high blood-pressure and BMI, having high blood-pressure and smoking now, having high blood-pressure and number of drinks per day, having high blood-pressure and past memory, having high blood-pressure and time walked, and number of conditions and time walked. If we were to move forward with building a regression model to predict life expectancy we would use these interactions.

5.3 Support Vector Machine: Radial Kernel

Next, a support vector machine (SVM) using a radial kernel was created. SVM divides the feature space using a hyperplane. An SVM model with a radial kernel applies a non-linear transformation to the data points before applying that hyperplane. This kernel can account for infinite dimension in the predictor set which helps to control the variance.. This is impossible to actually perform computationally so a parameter is added to the model to control this. The program we used finds this parameter for us. The SVM model produced had the lowest error so far.

Figure 3: Distribution of
Observed Age at Death and Predicted Life Expectancy

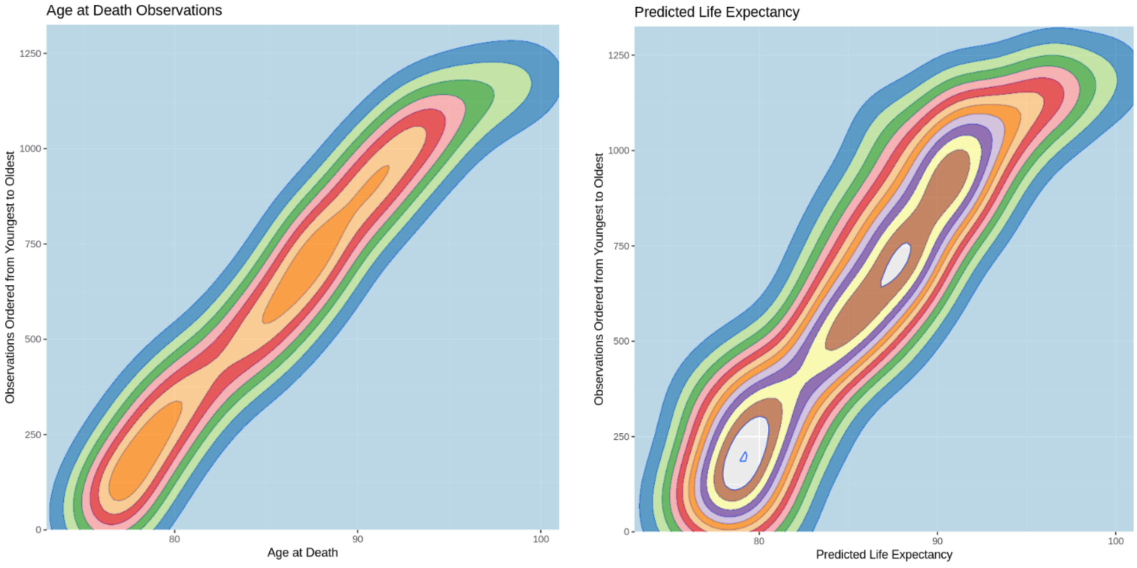


Figure 3 can be used to see how well we predicted life expectancy on our test set. The left graph shows age at death by observations ordered from youngest to oldest. The graph on the right shows the predicted life expectancy for those same individuals. One can see that our model predicted the two separate groupings and retained the overall shape well. These images may also suggest overfitting, as ours has higher peaks in the graph and is not as evenly distributed.

The model produced here had a test RMSE of 2.24. This is the lowest RMSE produced so far. This means on average we were roughly 2.24 years off when predicting life-expectancy for our test set. This model can be useful in real-world applications.

5.4 Decision Tree

Next, we move into the world of decision trees. Tree-based methods stratify the feature space into different regions used for predictions. The predictor space is split along internal nodes, which are the branches of the tree. We first fit a simple decision tree with age at death as the response variable, as seen in Figure 4. The predictors used in the tree are *agey_m* (individual's age in the middle of the interview period) and *conde* (sum of indicators for whether a doctor has ever told the Respondent that s/he has ever had a particular disease)[1]. The tree has a total of 6 terminal nodes.

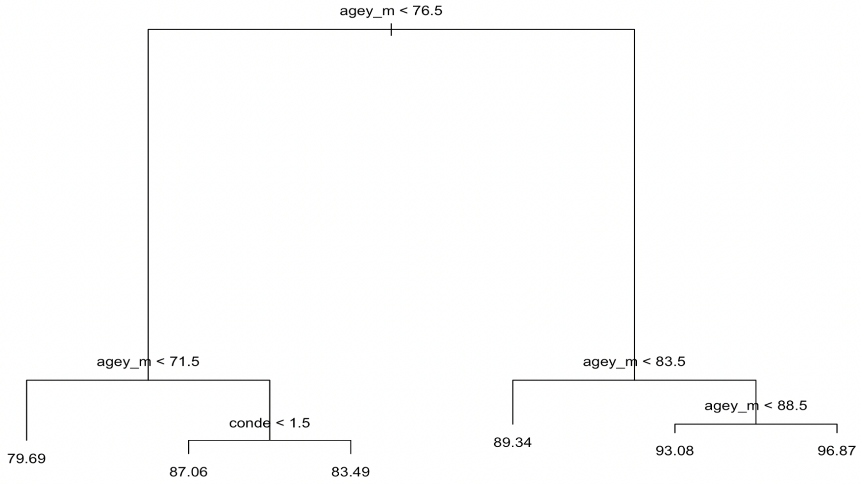


Figure 4: Initial Decision Tree

[1] high blood pressure, diabetes, cancer, lung disease, heart disease, stroke, psychiatric problems, and arthritis

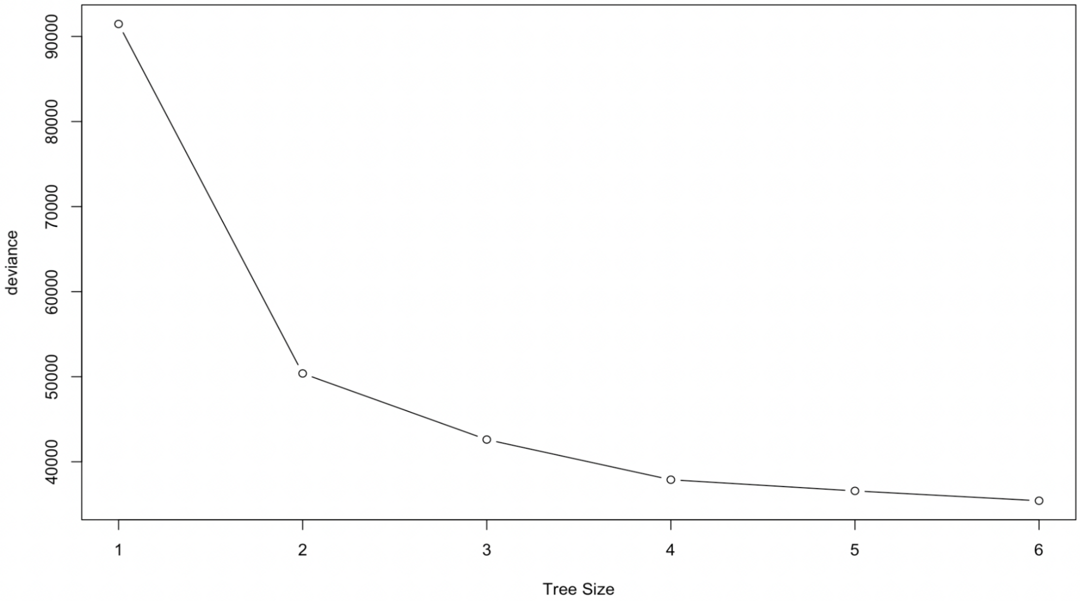
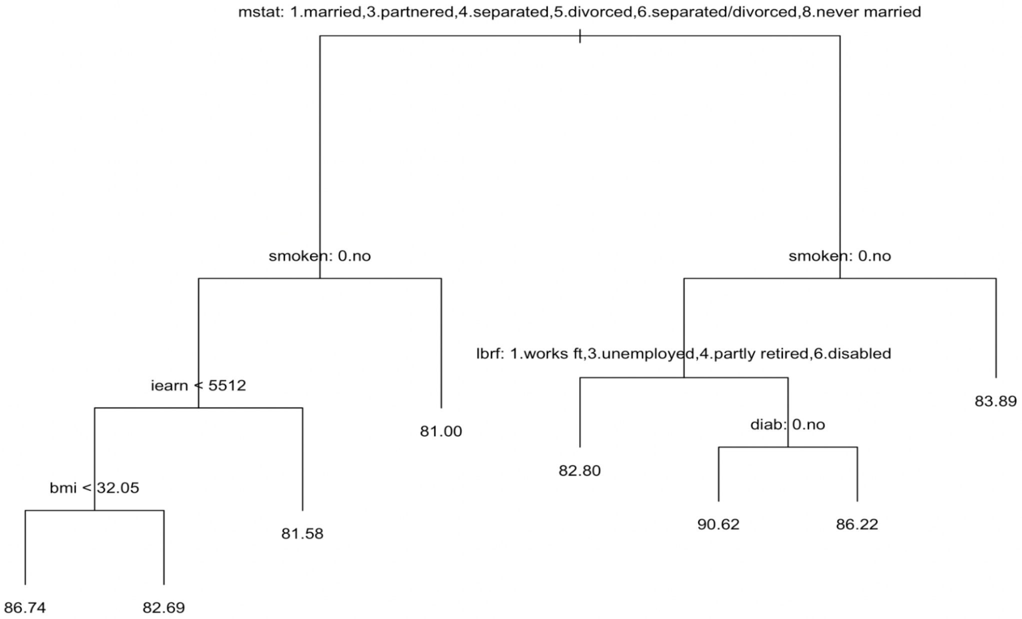


Figure 5: Tree Size and Deviance

Figure 5 highlights that this tree size produces the lowest residual deviance. This decision tree was then pruned to avoid overfitting the data, but this resulted in the same decision tree. The decision tree displayed in Figure 5 produced a cross-validated test RMSE of 4.1 years. When removing *agey_m* from the predictor space (to see how well we can predict life expectancy *without* knowing an individual's current age), the decision tree in Figure 6 is created.

Figure 6: Decision Tree with Age at Time of Interview Removed


mstat: 1.married,3.partnered,4.separated,5.divorced,6.separated/divorced,8.never married
smoken: 0.no
smoken: 0.no
iearn < 5512
lbrf: 1.works ft,3.unemployed,4.partly retired,6.disabled
83.89
81.00
diab: 0.no
82.80
bmi < 32.05
90.62
86.22
81.58
86.74
82.69

This decision tree was then pruned using cost complexity pruning to produce the decision tree in Figure 6. This tree has 5 terminal nodes and utilizes the variables *mstat* (marital status)*, smoken* (currently smoking indicator)*,* and *iearn* (income) and generates a test RMSE of 5.92.

5.5 Bagging

To extend upon the concept of decision trees, we then utilize bagging to build trees. Bagging is mainly used to reduce the variance of the decision tree method. The bagging method bootstraps the data generating a large number of available training sets to get many different decision trees, which are then averaged to obtain predictions. This comes from the idea that aggregating many decision trees may result in better predictive performance. In this case, we utilize 500 trees. It's important to note that bagging uses all predictors in the feature space. The predictions produced from the bagging method are plotted against the validation set lifespans in Figure 7. This bagging method produces a cross-validated test error of 3.41 years, an improvement from the simple decision tree. However, since this method aggregates many decision trees, we cannot directly see which variables are the most important terminal nodes.

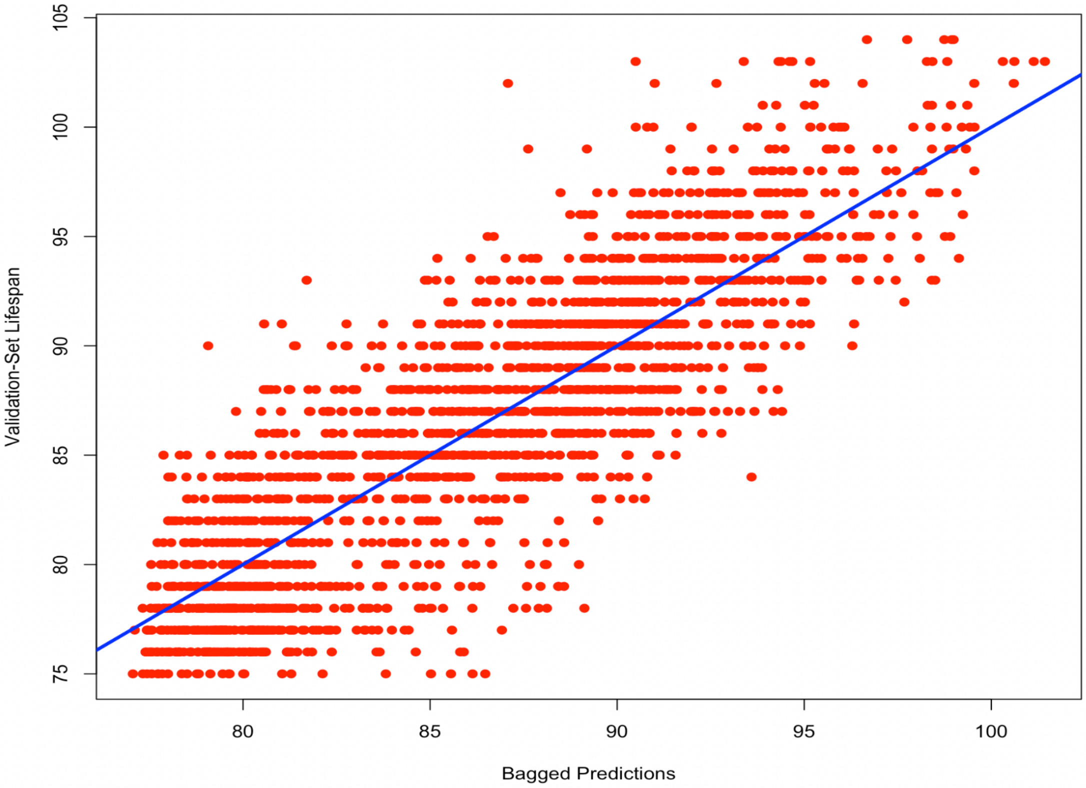


Figure 7: Bagged Predictions vs. Validation-Set Lifespan

5.6 Random Forest

An extension of the bagging procedure is the random forest method. Random forests improve upon the bagged trees by decorrelating the aggregated trees which helps to reduce the variance when the trees are averaged. The procedure follows that of bagging but when building each decision tree, a random selection of *m* predictors is chosen as split candidates from the full feature space. Typically, m is set to be $\frac{p}{3}$ in fitting regression trees, which is what is used here. The random forest procedure produces a cross-validated test RMSE of 3.35 years, a slight improvement upon bagging. Despite the lack of interpretability, we can see how the test error would increase when removing variables from the predictor space. This is displayed in Figure 8. From this we can immediately see the importance of *agey_m*, which is intuitive as a person's current age will obviously be an important predictor for their lifespan. Beyond that, we see the next 3 most important variables are *conde*, *logisret* (log of social security income*,* and *cendiv* (Census Division)[2].

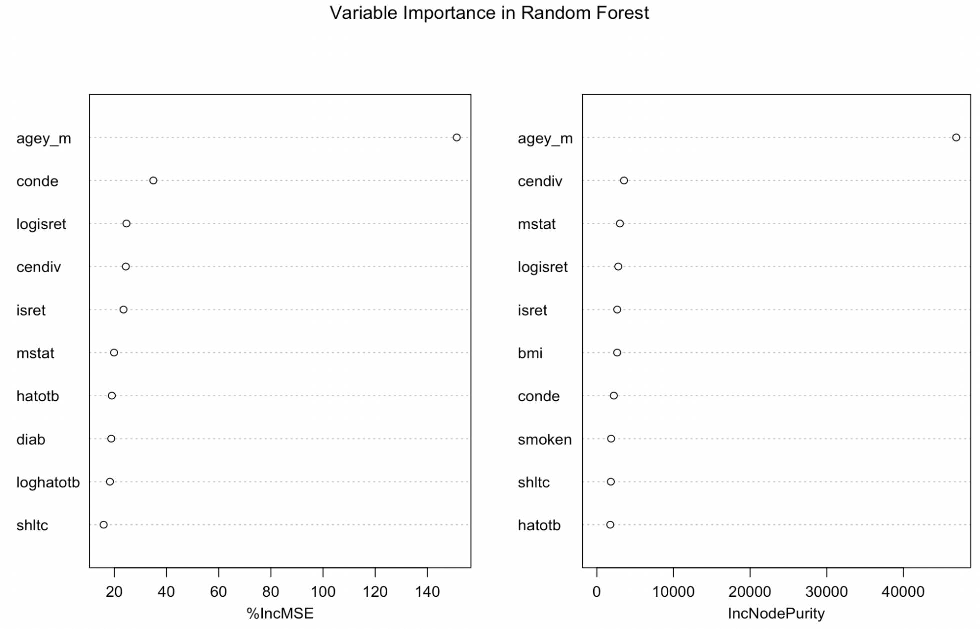
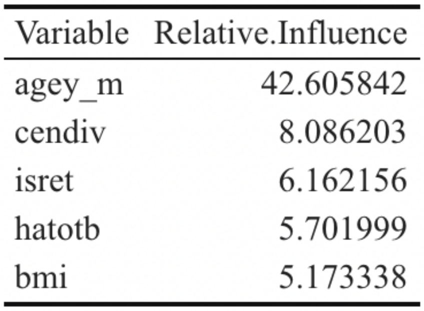


Figure 8: Variable Importance in Random Forest

5.7 Boosting

The last method we use for generating decision trees is boosting. Boosting is an extension of the bagging method. In this case, however, each tree is grown sequentially using information from the previously grown trees. This boosting approach learns slowly which avoids fitting the data hard. After running the boosting model with 5000 trees and a shrinkage parameter of 0.001 (parameter for how quickly the tree learns), the relative influence of the 5 most important variables are displayed in Figure 9.

[2] Place at time of death: New England, Mid Atlantic, EN Central, WN Central, S Atlantic, ES Central, WS Central, Mountain, Pacific, Not US/inc US territory



Variable	Relative.Influence
agey_m	42.605842
cendiv	8.086203
isret	6.162156
hatotb	5.701999
bmi	5.173338

Figure 9: Variable Importance in Boosting

Again, it is no surprise to *agey_m* as the most important predictor. We also see that a person's body mass index (BMI) is relatively important in predicting life expectancy. The marginal effect of *agey_m* and *cendiv* on the response after integrating out other variables are displayed in Figures 9 and 10, respectively.

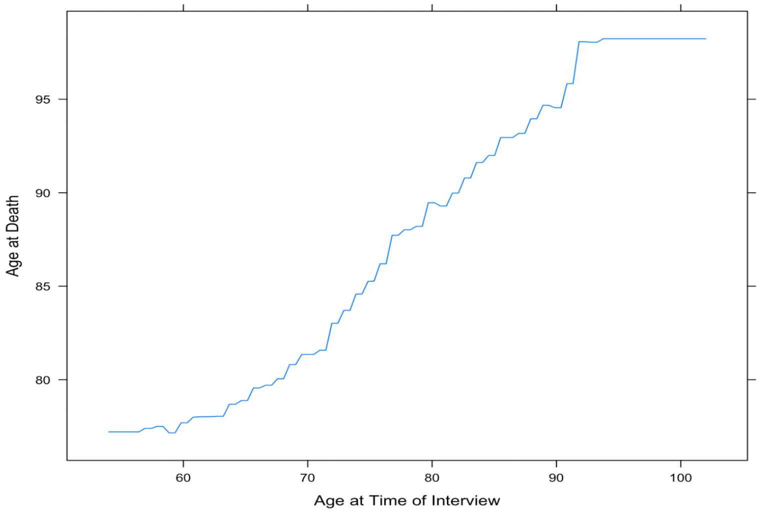


Figure 9: Marginal Effect of Current Age

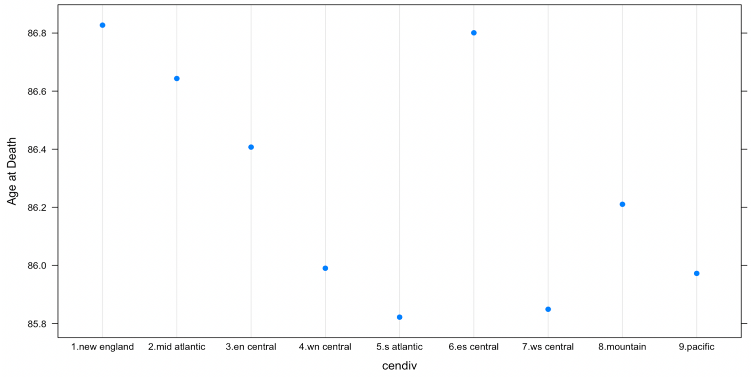


Figure 10: Marginal Effect of Census Division

The cross-validated test RMSE for this boosting model is 3.31 years. When using a different shrinkage parameter of 0.2, the test RMSE increased to 3.42.

5.8 K-Nearest Neighbors

K-Nearest Neighbors is a non-parametric method that approximates the association between independent variables and the continuous outcome by averaging the observations in the same neighborhood that are similar. The K-NN algorithm can be used for classification and regression. We used regression in the case of studying our continuous outcome variable age as a function of life factors. We chose the number of neighbors (K) as 5, 10, and 25 iterations in order to showcase the optimal number of neighbors to reach the smallest root mean squared error. The lowest RMSE of 5.45 results from using 10 nearest neighbors.. From 5 to 10 neighbors, the RMSE decreases, and then from 10 to 25 neighbors, we see the RMSE increase again. Figure 10 illustrates that the lowest RMSE comes when using 10 neighbors.

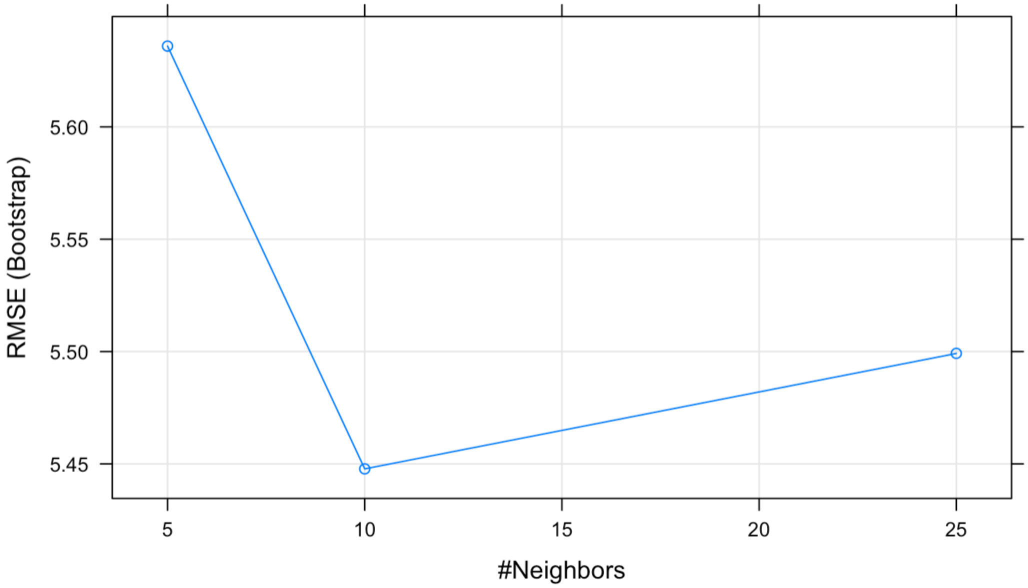
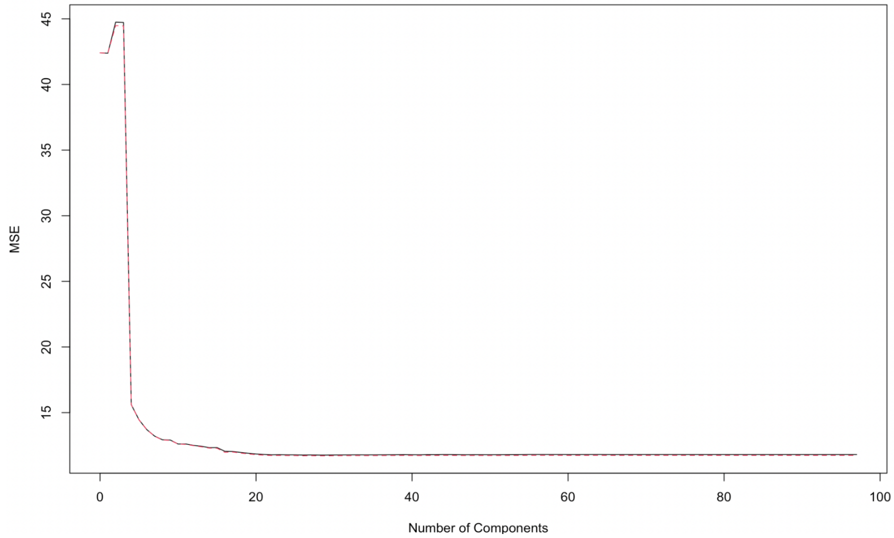


Figure 11: Optimal Number of Neighbors

5.9 Partial Least Squares

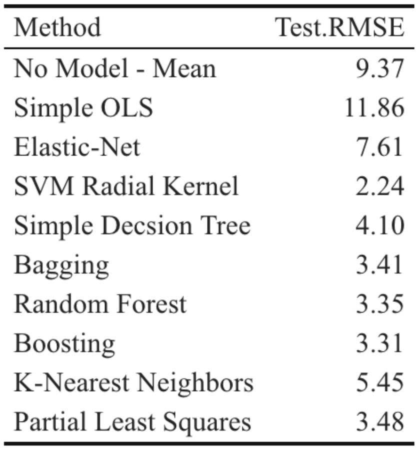
Partial least squares, or PLS, is a dimension reduction method which creates a new set of features Z which are linear combinations of the original predictors, and then fits a linear model using OLS from this new feature space. The features are identified in a supervised fashion to ensure that they are related to the response. After considering 97 different components, Figure 12 displays that the lowest cross-validating error occurs when the number of components is approximately 20. Using more components than that does not seem to reduce the test error by a significant amount. Using 20 components, the cross-validated test error from the PLS model is 3.48 years.

Figure 12: Number of Components and MSE



6 Conclusion

To summarize our findings, the test RMSE for all machine learning techniques used in this paper are displayed below in Figure 13. We find that the machine learning tactic that was most accurate in predicting lifespan is the support vector machines with a radial kernel, producing a highly accurate test RMSE of 2.24 years. Other methods that performed well were various forms of the decision tree along and PLS, all producing RMSEs under 3.5 years.



Method	Test.RMSE
No Model - Mean	9.37
Simple OLS	11.86
Elastic-Net	7.61
SVM Radial Kernel	2.24
Simple Decsion Tree	4.10
Bagging	3.41
Random Forest	3.35
Boosting	3.31
K-Nearest Neighbors	5.45
Partial Least Squares	3.48

As we gain predictive power with the complexity of the machine learning tactic, we tradeoff with interpretability. This paper attempted to include both models that can be used for their predictive prowess like support vector machines along with those that are more suited for inference, such as the simple OLS model. For example, from an inference perspective, the simple decision tree showed that marital status was a key predictor in determining an individual's lifespan, along with their smoking habits. Given the panel nature of the data, we limited individuals who had exactly 7 observations. In future studies, we would create a weighting variable to add those who had more or less observations. Another limitation of our study would be the time correlation between certain observations. We would use fixed effects or other methods beyond the scope of this course to deal with this issue.

Given the accuracy of the support vector machine model, we believe it can be used by insurance companies to help calculate lifespans (and therefore life insurance premiums) in the future. Certain models can also be used for inference. To determine what factors lead to one's death and how we may address them. Perhaps even more advanced machine learning techniques involving neural networks and deep learning that were not examined in this paper can provide even more predictive power in the future. As these long-term studies on individuals like the HRS continue in the future, more data will be collected that will only improve the predictive accuracy of these models.

7 References

1. Beeksma, M., Verberne, S., van den Bosch, A. *et al.* Predicting life expectancy with a long short-term memory recurrent neural network using electronic medical records. *BMC Med Inform Decis Mak* 19, 36 (2019). https://doi.org/10.1186/s12911-019-0775-2

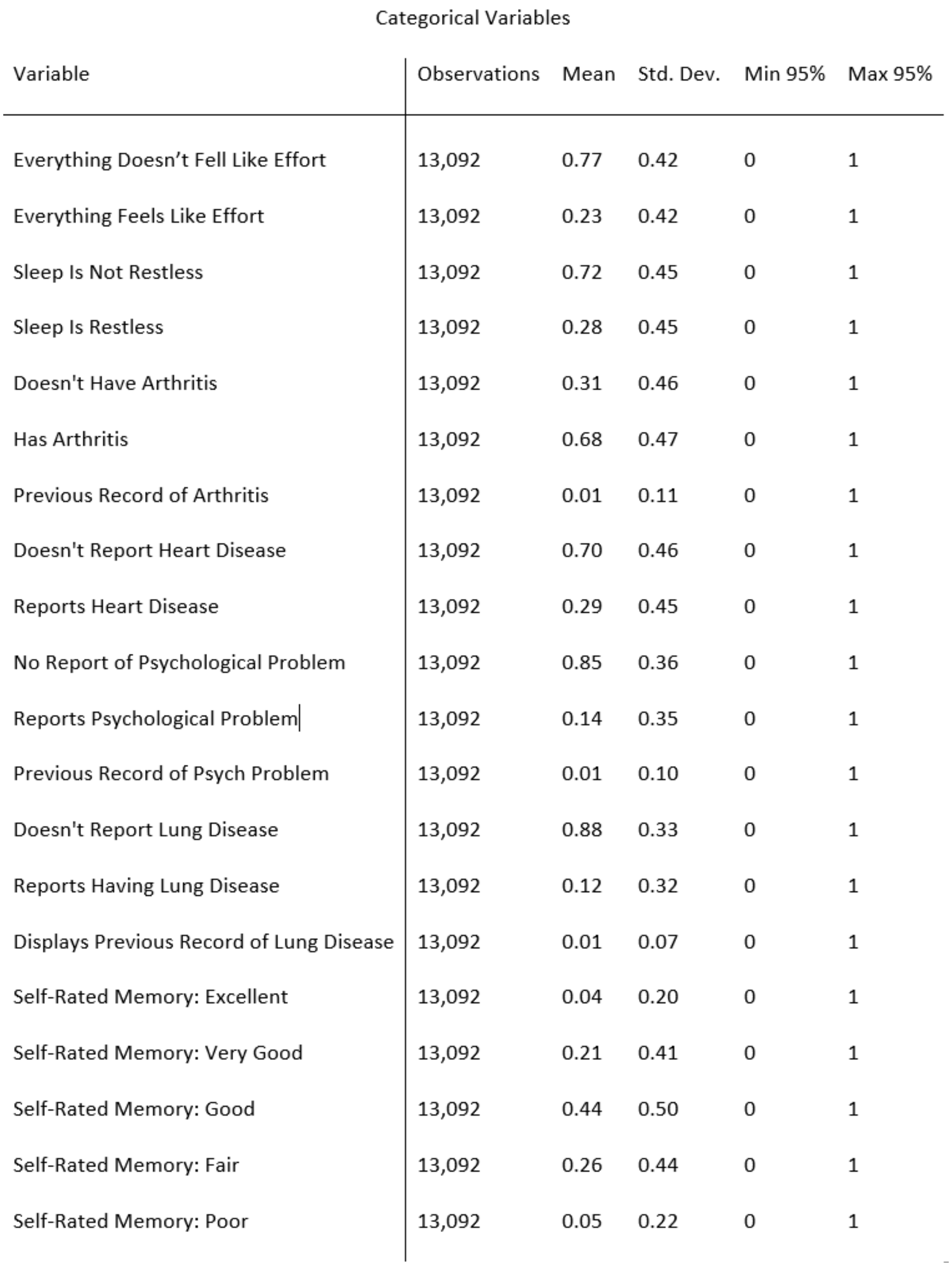
2. Li, Yanping et al. "Healthy lifestyle and life expectancy free of cancer, cardiovascular disease, and type 2 diabetes: prospective cohort study." *BMJ (Clinical research ed.)* vol. 368 l6669. 8 Jan. 2020, doi:10.1136/bmj.l6669

3. Montez, Jennifer Karas et al. "US State Policies, Politics, and Life Expectancy." *The Milbank quarterly* vol. 98,3 (2020): 668-699. doi:10.1111/1468-0009.12469

4. Shin, Jaeyong et al. "Predicting Old-age Mortality Using Principal Component Analysis: Results from a National Panel Survey in Korea." *Medicina (Kaunas, Lithuania)* vol. 56,7 360. 18 Jul. 2020, doi:10.3390/medicina56070360

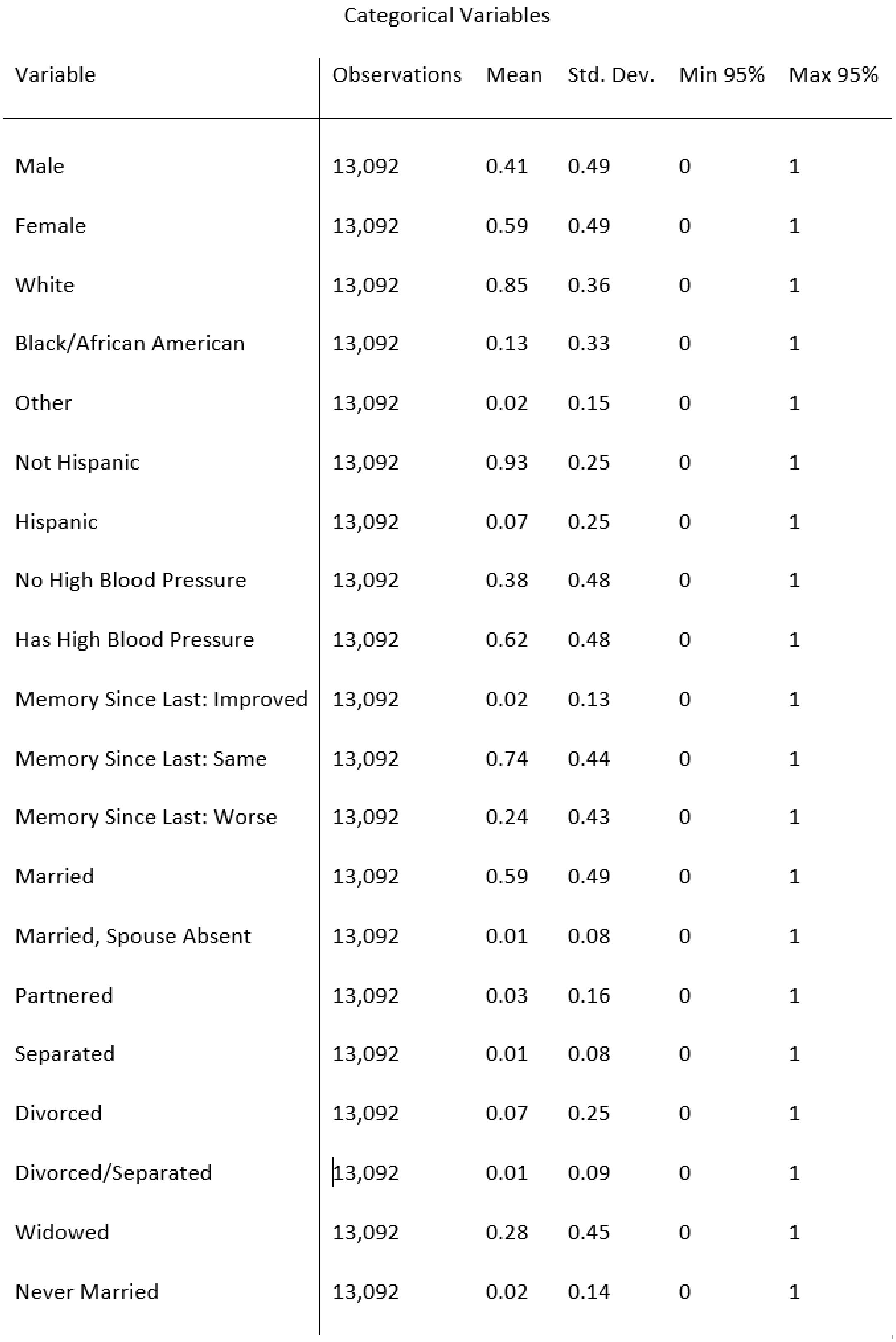
Appendix

Additional Summary Statistics



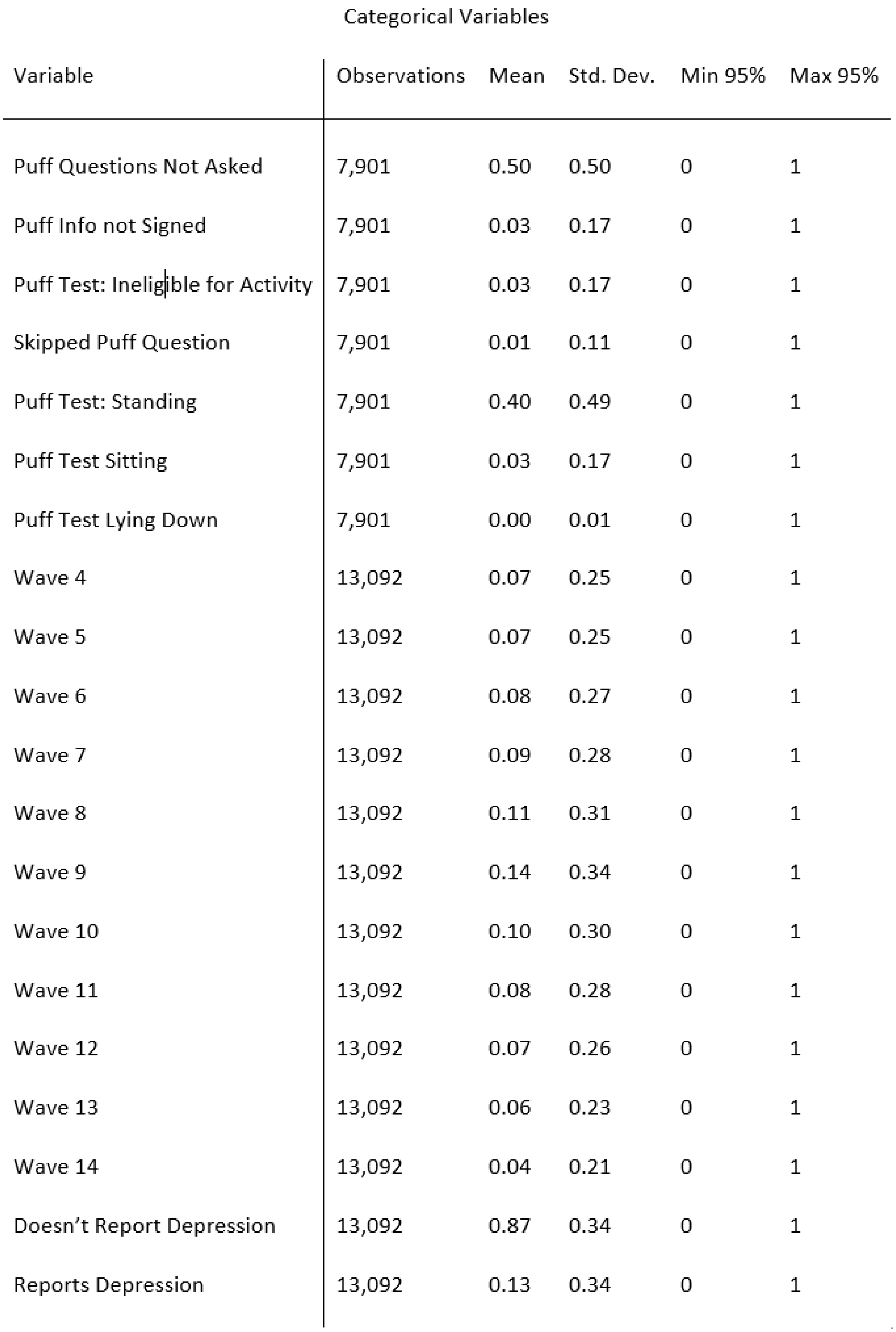
Categorical Variables

Variable	Observations	Mean	Std. Dev.	Min 95%	Max 95%
Everything Doesn’t Fell Like Effort	13,092	0.77	0.42	0	1
Everything Feels Like Effort	13,092	0.23	0.42	0	1
Sleep Is Not Restless	13,092	0.72	0.45	0	1
Sleep Is Restless	13,092	0.28	0.45	0	1
Doesn't Have Arthritis	13,092	0.31	0.46	0	1
Has Arthritis	13,092	0.68	0.47	0	1
Previous Record of Arthritis	13,092	0.01	0.11	0	1
Doesn't Report Heart Disease	13,092	0.70	0.46	0	1
Reports Heart Disease	13,092	0.29	0.45	0	1
No Report of Psychological Problem	13,092	0.85	0.36	0	1
Reports Psychological Problem	13,092	0.14	0.35	0	1
Previous Record of Psych Problem	13,092	0.01	0.10	0	1
Doesn't Report Lung Disease	13,092	0.88	0.33	0	1
Reports Having Lung Disease	13,092	0.12	0.32	0	1
Displays Previous Record of Lung Disease	13,092	0.01	0.07	0	1
Self-Rated Memory: Excellent	13,092	0.04	0.20	0	1
Self-Rated Memory: Very Good	13,092	0.21	0.41	0	1
Self-Rated Memory: Good	13,092	0.44	0.50	0	1
Self-Rated Memory: Fair	13,092	0.26	0.44	0	1
Self-Rated Memory: Poor	13,092	0.05	0.22	0	1



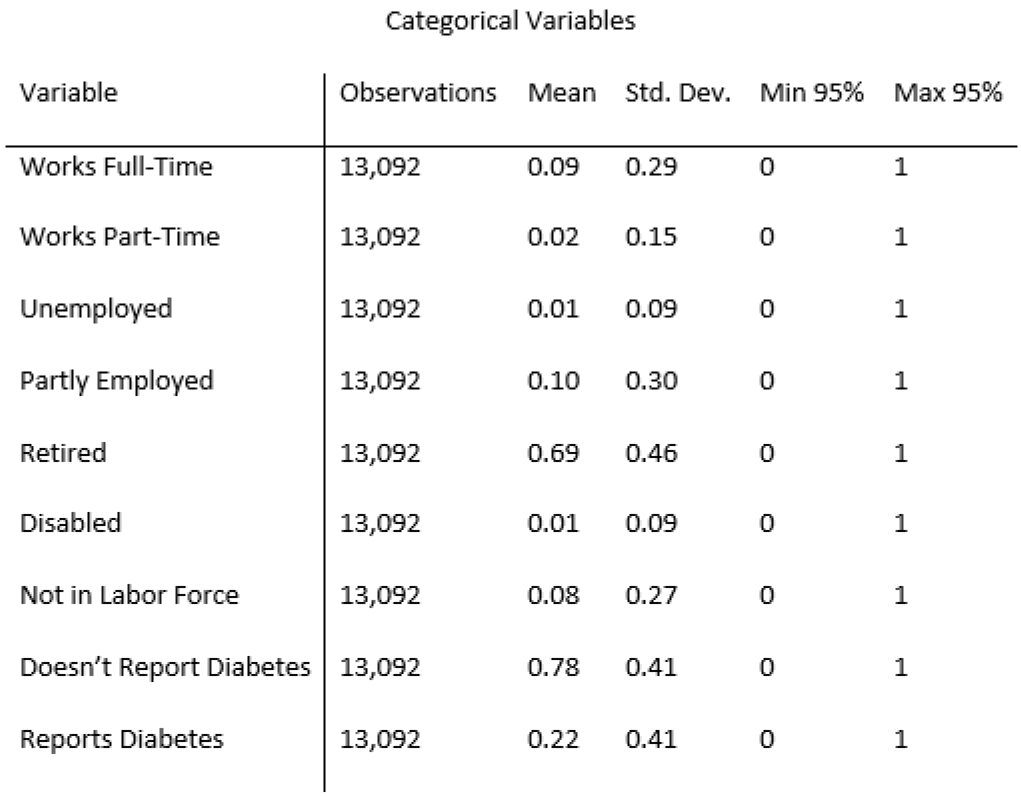
Categorical Variables

Variable	Observations	Mean	Std. Dev.	Min 95%	Max 95%
Male	13,092	0.41	0.49	0	1
Female	13,092	0.59	0.49	0	1
White	13,092	0.85	0.36	0	1
Black/African American	13,092	0.13	0.33	0	1
Other	13,092	0.02	0.15	0	1
Not Hispanic	13,092	0.93	0.25	0	1
Hispanic	13,092	0.07	0.25	0	1
No High Blood Pressure	13,092	0.38	0.48	0	1
Has High Blood Pressure	13,092	0.62	0.48	0	1
Memory Since Last: Improved	13,092	0.02	0.13	0	1
Memory Since Last: Same	13,092	0.74	0.44	0	1
Memory Since Last: Worse	13,092	0.24	0.43	0	1
Married	13,092	0.59	0.49	0	1
Married, Spouse Absent	13,092	0.01	0.08	0	1
Partnered	13,092	0.03	0.16	0	1
Separated	13,092	0.01	0.08	0	1
Divorced	13,092	0.07	0.25	0	1
Divorced/Separated	13,092	0.01	0.09	0	1
Widowed	13,092	0.28	0.45	0	1
Never Married	13,092	0.02	0.14	0	1



Categorical Variables

Variable	Observations	Mean	Std. Dev.	Min 95%	Max 95%
Puff Questions Not Asked	7,901	0.50	0.50	0	1
Puff Info not Signed	7,901	0.03	0.17	0	1
Puff Test: Ineligible for Activity	7,901	0.03	0.17	0	1
Skipped Puff Question	7,901	0.01	0.11	0	1
Puff Test: Standing	7,901	0.40	0.49	0	1
Puff Test Sitting	7,901	0.03	0.17	0	1
Puff Test Lying Down	7,901	0.00	0.01	0	1
Wave 4	13,092	0.07	0.25	0	1
Wave 5	13,092	0.07	0.25	0	1
Wave 6	13,092	0.08	0.27	0	1
Wave 7	13,092	0.09	0.28	0	1
Wave 8	13,092	0.11	0.31	0	1
Wave 9	13,092	0.14	0.34	0	1
Wave 10	13,092	0.10	0.30	0	1
Wave 11	13,092	0.08	0.28	0	1
Wave 12	13,092	0.07	0.26	0	1
Wave 13	13,092	0.06	0.23	0	1
Wave 14	13,092	0.04	0.21	0	1
Doesn’t Report Depression	13,092	0.87	0.34	0	1
Reports Depression	13,092	0.13	0.34	0	1



Categorical Variables

Variable	Observations	Mean	Std. Dev.	Min 95%	Max 95%
Works Full-Time	13,092	0.09	0.29	0	1
Works Part-Time	13,092	0.02	0.15	0	1
Unemployed	13,092	0.01	0.09	0	1
Partly Employed	13,092	0.10	0.30	0	1
Retired	13,092	0.69	0.46	0	1
Disabled	13,092	0.01	0.09	0	1
Not in Labor Force	13,092	0.08	0.27	0	1
Doesn’t Report Diabetes	13,092	0.78	0.41	0	1
Reports Diabetes	13,092	0.22	0.41	0	1