Mental Health at All Ages:
Identifying Risk Factors for Anxiety in Older Adults

Abstract
Anxiety disorders in the elderly population often go undiagnosed, even though they affect 10-20% of older adults. Our analysis investigates factors associated with risk of anxiety in older adults. We analyzed a dataset containing medical information about adults aged 70-85, with metrics including measures of physical fitness, history of medical incidents, social interaction, and cognitive ability. We first performed penalized lasso regression on the data to select the most promising predictors for anxiety. Then, our subsequent analysis fits multiple models on these predictors to explore the significant factors and predictive performance of each model. We use methods including stepwise regression, a regression tree, and tree-based ensemble methods. Based on 5-fold cross validation, the stepwise regression models yielded the best predictions for anxiety scores. Additionally, we found that younger age, more pre-existing comorbidities, higher depression score, and more time spent on phones were all associated with higher anxiety scores.

1. Background & Significance

Anxiety disorders affect 10-20% of older adults. However, anxiety in older adults often goes undiagnosed since patients often don't recognize symptoms or are simultaneously suffering from other medical conditions (American Association for Geriatric Psychiatry 2022). If left untreated, anxiety can lead to cognitive impairment, poor physical health, and poor quality of life. Thus, our analysis aims to understand the factors that are associated with risk of anxiety in older adults. Our research question asked, "Which medical and lifestyle factors are most important for predicting anxiety among older adults?" Anxiety was measured through a visual analogue scale in which participants rate how anxious they felt over the past 24 hours. A 0 indicates not at all anxious, and 10 indicates extremely anxious (Williams et al 2010).

2. Methods

2a. About the dataset

Our analysis uses a dataset available on Zenodo that contains medical information about older adults aged 70-85 (Ellul et al 2019). The dataset includes medical data for 30 patients collected across four checkups, with one check-up approximately every 6 months. The medical data includes 56 metrics, including measures of physical fitness, physical activity, history of medical incidents, social interaction, and cognitive ability.

2b. Data cleaning

Our dataset includes time series data collected across four medical visits, but longitudinal analysis methods are outside the scope of this project, so our analysis does not consider temporal relationships. Therefore, to avoid violating the assumption of independence between observations, we only used data from the patients' third medical visits. We made this decision because the third visit contained the least missingness out of all the other visits. Additionally, the distribution of anxiety scores in the third visit has a similar shape to the distribution of anxiety scores overall (see Appendix A1). Next, we handled missing data. One patient had values that were missing in the third visit but present in the fourth visit; for this patient, we imputed data from the fourth visit. Another patient had two missing values across all visits; for this patient, we performed regression imputation (see Appendix A2). Finally, two continuous numeric variables with one missing value each were resolved using mean imputation.

2c. Variable selection via penalized lasso regression

After cleaning our variables, we aimed to reduce our total number of predictors (58) to a lower dimension feature space. Variable selection is necessary to avoid severe multicollinearity problems and to ensure that there are more observations than predictors in our analysis, since we have a small sample size of n=30. Since penalized lasso regression can estimate coefficients to be zero, we used it to perform both variable selection and parameter estimation at once. We accomplished this by first choosing the optimal lambda (tuning parameter) value for lasso regression through cross-validation using the cv.glmnet function in R. The resulting lasso model (with optimal lambda = 0.012) had non-zero coefficients for the following predictors: age, comorbidities_most_important, screening_score, depression_total_score, social_visits, social_phone, health_rate, comorbidities_count, and medication_count (see Appendix A3). We also ran VIF screening on these predictors to confirm the absence of a severe multicollinearity problem. We thus reduced the number of predictors from 58 to 9 significant predictors, which we used for further analysis.

2d. Checking assumptions

Before analysis, three crucial assumptions were checked: linearity, constant variance, and normality. The linearity and mean 0 assumption held, but the normality and constant variance assumption did not for residuals (see row 1 of Appendix A4). As well, anxiety scores were

shown to be skewed right (see row 2 of Appendix A4). Thus, to improve normality and the constant variance assumption, we performed a box-cox power transformation (with optimal lambda = 0.5; see Appendix A5). After, the transformed anxiety scores were much more normally distributed (see Appendix A6).

2e. Modeling: Stepwise Regression, Regression Tree, and Ensemble Methods
Our analysis considers three types of modeling methods: 1) stepwise linear regression, 2) regression tree, and 3) tree-based ensemble methods. We fit each of these models on our cleaned dataset, using the nine significant predictors identified by lasso regression in part 2c.

Model 1: Stepwise Regression
Our first model utilized stepwise regression using both AIC and BIC criterion to examine the statistical significance of the predictors chosen via lasso regression and create models that penalized for the number of predictors. The stepwise AIC and BIC models were identical and included three predictors: age, depression_total_score, and medication_count (see Appendix A7 for summary output).

Model 2: Regression Tree
Our second model included a full and pruned regression tree. To find the optimal size for the pruned tree, a simulation was performed 1000 times to determine the most common size with the lowest deviance based on the cross validation score, which was five (see Appendix A8).

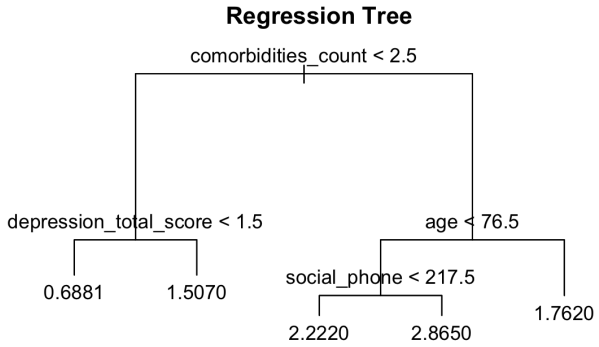


Fig 1: Regression Tree. The full and pruned tree both have 5 terminal nodes. Higher anxiety scores are related to more comorbidities, more time spent on phones, higher depression scores, and younger age.

Model 3: Tree-Based Ensemble Methods
Next, we used tree-based ensemble methods to further explore important predictors and determine whether ensemble methods could outperform a simple regression tree. We implemented two different ensemble methods: bagging and random forest. Both of these methods aim to reduce variance in the model by aggregating the results from many different trees that are built using different training samples. In order to determine the ideal number of aggregated trees to use for each ensemble method, we plotted the error in each ensemble method using up to 500 trees. We found that the error score in both the bagging and random forest methods leveled off at around 100 trees, so we used this number of trees for our analysis (see Appendix A9). For both of the ensemble methods, we computed 5-fold cross validation scores to compare predictive performance. We also identified the most important predictors in the bagging and random forest models by ranking each variable's score on the Gini impurity index, which is based on the algorithm used to calculate the splits in trees.

3. Results
The modeling methods described in the previous section offered results for each model's predictive performance, with cross validation scores and significant variables in each model. These results can offer insights into the most important considerations for predicting anxiety scores in our sample of older adults.

3a. Model predictive performance: comparison using 5-fold cross validation
A 5-fold cross-validation score was calculated for each of the five models: stepwise AIC-best regression, stepwise BIC-best regression, regression tree, bagging ensemble method, and

random forest ensemble method. Each of the cross validation scores were calculated using a random number generator with the same seed, to ensure that the random folds were consistent across CV scores. Both AIC and BIC stepwise regression had a CV score of 0.029. The regression tree had a CV score of 1.06. For the ensemble methods, the bagging model had a CV score of 0.64, while random forest had a CV score of 0.61. Note that all of these CV scores are on the transformed scale, according to the Box-Cox transformation in section 2d. AIC and BIC stepwise regression yielded the lowest CV score, suggesting that the model created by stepwise regression created the most accurate predictions for anxiety scores.

3b. Comparison of significant predictors

Model	Significant predictors
Stepwise regression (AIC and BIC best)	age, depression_total_score, medication_count
Regression tree	age, depression_total_score, comorbidities_count, social_phone
Bagging and random forest (importance rankings: see Appendix A10 and A11)	age, depression_score, medication_count, comorbidities_count, social_phone

Our models all yielded similar significant factors in predicting anxiety — all of the predictors were found in at least two models. All three of our models included depression and age as significant predictors, suggesting that there is a relationship between depression and self perceived anxiety as well as age and anxiety.

4. Discussion & Limitations

Based on 5-fold cross validation, the stepwise regression models yielded the best predictions for anxiety scores. This model indicates that a younger age, higher levels of depression, and more medications are associated with higher anxiety scores. Since the AIC and BIC best stepwise regression models had the best predictive performance, we can infer that these are some of the most significant risk factors for anxiety in older adults. Although the other models did not perform as well on cross validation, they showed similar findings that younger age and higher levels of depression are associated with anxiety. The regression tree and ensemble methods also indicate that more pre-existing comorbidities and more time spent on phones are associated with higher anxiety scores. These factors should also be considered as possible indicators that could alert medical professionals to the risk of high anxiety scores in older adults.

One of the limitations of this project was our treatment of the data as static data, when in fact the dataset as a whole was time-series data, with check up information being taken from the same patients over a two year period of time. Thus, we were unable to capture any temporal relationships within the data. Future work could be done to apply time-series methods to this data, including longitudinal analysis (such as multivariate ANOVA or change score analysis) and other time series methods, to get a fuller picture of the relationship between time, anxiety, and the other predictors. Furthermore, our dataset was very small (n = 30), so our results may not be very generalizable and may instead be overfit to this small population of older patients. It would be valuable to work with a larger dataset with more observations so we could create even more robust models that are generalizable to the larger elderly population.

References

American Association for Geriatric Psychiatry (2022). “Anxiety and older adults: Overcoming worry and fear.” Retrieved May 6, 2023, from https://www.aagponline.org/patient-article/anxiety-and-older-adults-overcoming-worry-and-fear/

Ellul, J., Polycarpou, M., Kotsani, M., & Zacharaki, E. I. (2019). “Aggregated Virtual Patient Model Dataset.” Zenodo. https://zenodo.org/record/2670048#.ZE7B7-zMIbm

Kalogiannis, S., Deltouzos, K., Zacharaki, E. I., Vasilakis, A., Moustakas, K., Ellul, J., & Megalooikonomou, V. (2019). “Integrating an openEHR-based personalized virtual model for the ageing population within HBase.” *BMC medical informatics and decision making*, 19(1), 1-15. https://doi.org/10.1186/s12911-019-0745-8

Williams, V. S. L., Morlock, R. J., & Feltner, D. (2010). “Psychometric Evaluation of a visual analog scale for the assessment of anxiety.” *Health and Quality of Life Outcomes*, *8*(1), 57. https://doi.org/10.1186/1477-7525-8-57

Appendix

Figure A1: Distribution of anxiety scores across visits

Our analysis focuses on the third visit, which had a similar distribution of anxiety scores as the dataset overall.

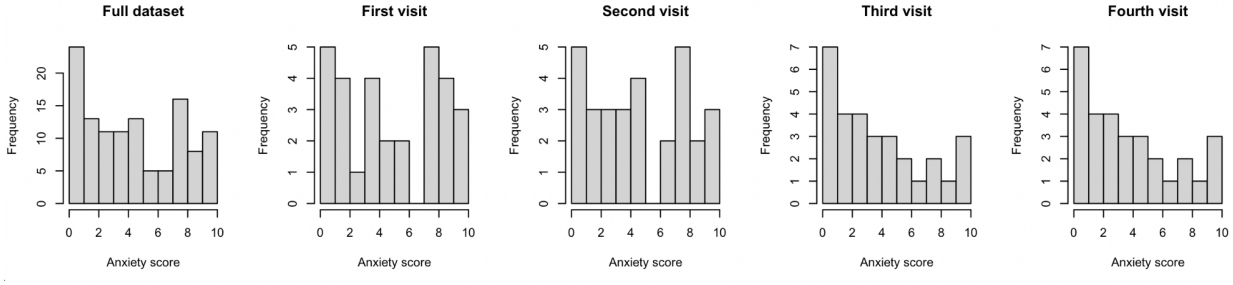


Figure A2: Regression imputation

Single linear regression was used to impute missing values for body fat percentage (left) and lean body mass percentage (right), based on each variable's correlation with BMI score.

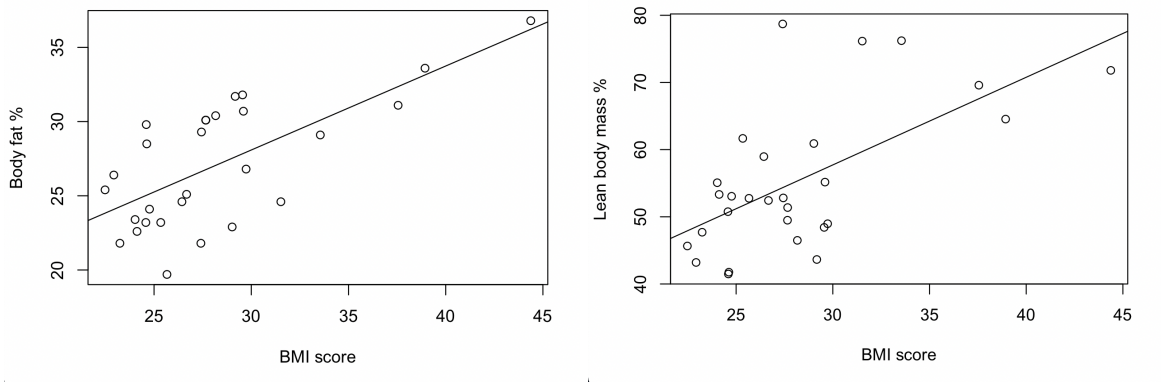


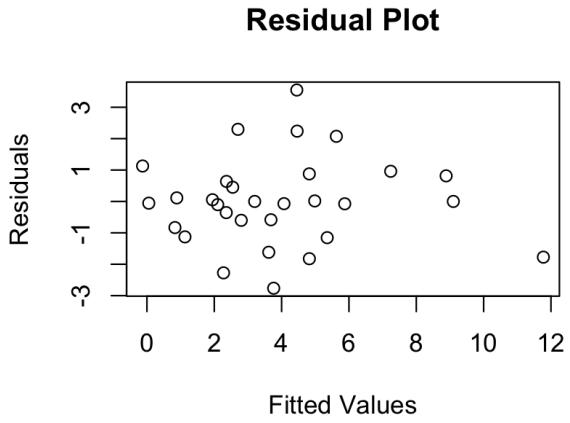
Figure A3: Lasso regression model coefficients

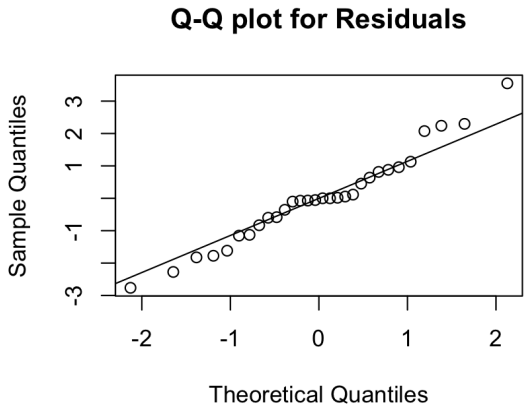
Non-zero coefficients from the penalized lasso regression model output are shown below. The coefficients for all other predictors were estimated to be zero, and predictors with a zero coefficient were eliminated from further analysis.

	Coef.
(Intercept)	39.1029876651
age	-0.1679481971
comorbidities_most_important	0.4323228150
screening_score	-1.6906640541
depression_total_score	0.2332328684
social_visits	-0.0299951576
social_phone	0.0004674094
health_rate	-0.0646271050
comorbidities_count	0.0144137518
medication_count	0.0680667500

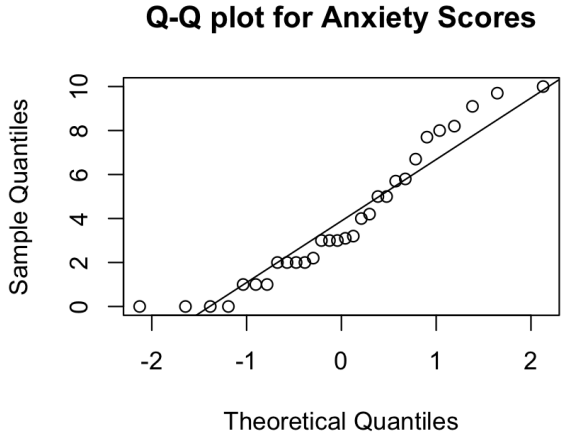
Figure A4: Plots to check normality and constant variance assumptions

Row 1: The following two plots show the residual plot and the Q-Q plot for the residual values. In the residual plot, the points seem to be randomly scattered enough around the 0 value to be linear and hold the mean 0 assumption. The spread of the points do not seem to be roughly equal (left). From the Q-Q plot, the points seem to roughly follow the 45 degree line, but are not quite normal (right).





Row 2: The following two plots show the Q-Q plot and histogram for anxiety scores, the response variable. In the Q-Q plot, the points do not follow the 45 degree line well (left) and the scores seem to be rightly skewed in the histogram (right).



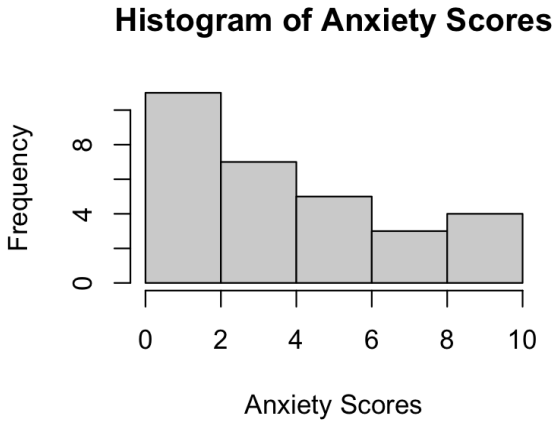


Figure A5: Box-cox plot for optimal lambda selection

The following plot indicates the 95% confidence interval for the optimal lambda λ which is defined as the lambda value with the highest likelihood, as measured by the y axis. Since the confidence interval did not include 1, it was appropriate to perform the transformation. For easier interpretation, the lambda value of 0.5 was chosen within the interval.

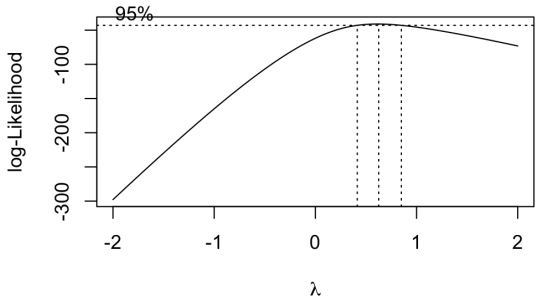


Figure A6: Anxiety score distribution after power transformation

Before power transformation, the anxiety scores were skewed right (left). After power transformation, the values followed a more normal distribution (right).

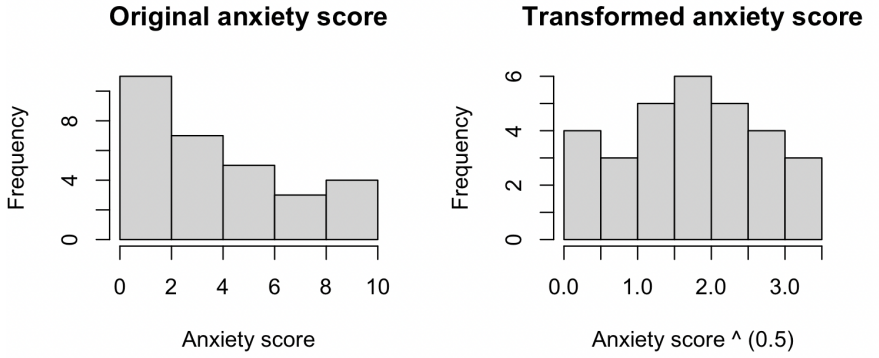


Figure A7: Stepwise Regression

The AIC-best and BIC-best stepwise models were identical. They included three predictors: age, depression_total_score, and medication_count. The output, with coefficients and significance levels, is shown below.

```
Call:
lm(formula = anxiety_perception ~ age + depression_total_score +
    medication_count, data = new)

Residuals:
     Min       1Q   Median       3Q      Max
-0.37552 -0.07291  0.02574  0.13117  0.24855

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)             3.17711    0.72530   4.380 0.000172 ***
age                    -0.02955    0.00934  -3.163 0.003944 **
depression_total_score  0.03576    0.01820   1.965 0.060202 .
medication_count        0.02489    0.01070   2.327 0.028026 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.171 on 26 degrees of freedom
Multiple R-squared:  0.5207,    Adjusted R-squared:  0.4654
F-statistic: 9.416 on 3 and 26 DF,  p-value: 0.0002194
```

Figure A8: Regression Tree

The first plot shows that size 5 was the most common optimal size when pruning the full regression tree based on cross-validation (left). The second plot shows the regression tree constructed where the full and the pruned tree is the same (right).

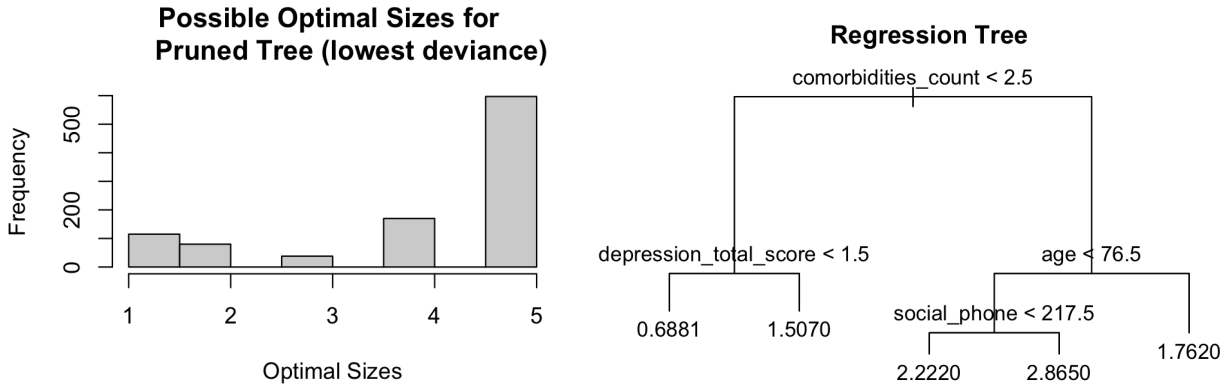


Figure A9: Plots for selecting number of trees

The following plots demonstrate the ideal number of aggregated trees to use for bagging (left) and random forest (right) ensemble methods. The error score in both plots levels off at around 100 trees.

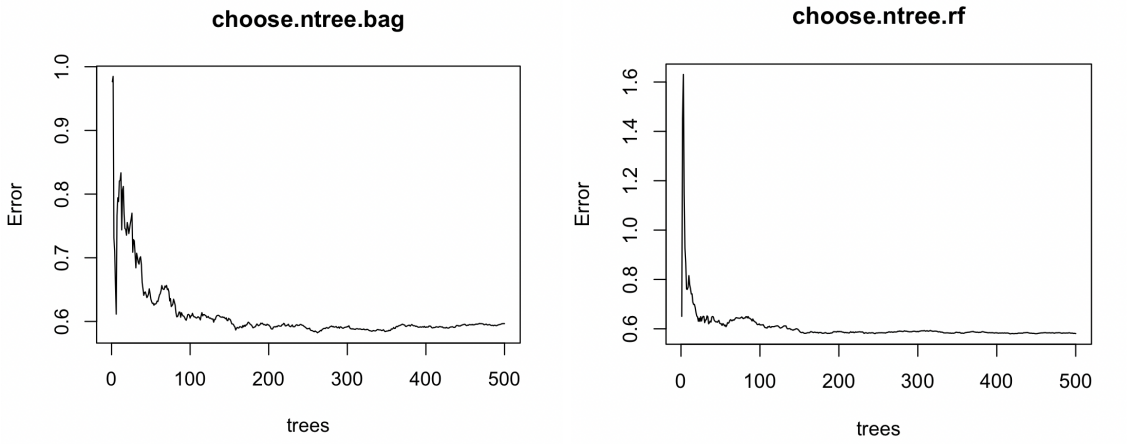


Figure A10: Bagging importance ranking

Predictor importance ranking is measured by IncNodePurity, which quantifies the average decrease in node heterogeneity from splitting on the variable. For regression, it is measured by the residual sum of squares. A higher IncNodePurity corresponds to more importance for predicting the outcome variable, anxiety score.

	IncNodePurity
comorbidities_count	4.3641761
age	4.3269527
social_phone	3.7822939
medication_count	3.2056726
depression_total_score	2.6371255
social_visits	1.8398925
screening_score	1.0650967
health_rate	0.5980447
comorbidities_most_important	0.3058075

Figure A11: Random forest importance ranking

Predictor importance ranking is measured by IncNodePurity, which quantifies the average decrease in node heterogeneity from splitting on the variable. For regression, it is measured by the residual sum of squares. A higher IncNodePurity corresponds to more importance for predicting the outcome variable, anxiety score.

	IncNodePurity
age	4.1912497
comorbidities_count	3.6541413
social_phone	3.5149757
medication_count	2.7297331
depression_total_score	2.3166355
social_visits	2.0922932
screening_score	0.9899869
health_rate	0.9429367
comorbidities_most_important	0.4755632