Early social and structural determinants of adolescent wellbeing

Hexiang Liu Yuxuan Zhang Stacey Bevan Yaoyu Wang

05/01, 2022

Contents

Executive Summary

Background

In late 2021 the US Surgeon General's Advisory released a public health statement about the alarming assaults on youth wellbeing. Even before the COVID pandemic, 20% of US youth reported a mental health disorder, and only half received appropriate treatment. In the last two years, emergency room visits for psychiatric crises have doubled and youth have reduced access to friends, teachers and healthcare providers that can recognize early signs of mental health challenges. This report called for coordinated action for community support of adolescents, including increased attention in schools and healthcare systems. There is a need to identify mechanisms that underpin mental health symptoms. This is especially true for adolescents who disproportionally experience social conditions that predispose them to poor outcomes, like poverty, food insecurity and local environmental factors.(1) While child mental health infrastructure has long been under resourced, there is a renewed urgency to understand how to prevent poor outcomes and promote resilience among at risk youth.

The global burden of childhood mental health symptoms is significant, non-communicable diseases are now the leading cause of disability.(2) This longstanding public health problem has developed a large body of research linking early childhood exposures to longitudinal health outcomes. Specifically, adverse childhood experiences were identified in a landmark study as associated with almost every adult disease. These include neighborhood violence, food insecurity, interpersonal trauma, and other assaults to development.(3) This study, and research that followed, had significant impact for healthcare systems that had not traditionally considered social structures as exposures for health outcomes. This work also unveiled the devastating cumulative impact of experiences that render children vulnerable to poor outcomes later in life.(4) Of particular interest is the prenatal environment, which is a sensitive period for neurodevelopment that impacts both child and maternal health.(5) These exposures are more prevalent, pervasive, and compounded by systemic racism for Black and Hispanic children, which can lead to longstanding sequela for these populations.(6) Investments in child wellbeing are a health equity issue.

Gaps and Study Aims

Where there is a robust body of work about early predictors of adolescent outcomes, there are several gaps in this understanding. Firstly, although early child exposures are well described, less is known about the impact of social conditions prenatally or at time of birth. Secondly, many samples inadvertently overrepresent families with majority identities. This preferentially develops an evidence base that is not sensitive to needs of families who are disadvantaged across stratifications of race and class, among others. Third, most prior research relies on the report of one individual to measure social constructs at the family level. Especially as it concerns child development, mothers are most likely to provide this information and the paternal role is not as well described. Lastly, adolescent wellbeing is often assessed by constructs that are proximate to the individual, like peer relationships and health behaviors. Models that include early social development are not frequently combined with structural determinants that may act as unmeasured confounders.

This study will fill these gaps using a unique application of a longitudinal dataset through the following aims: 1) develop a model of adolescent well-being from early social and structural health determinants and 2) determine which informant report is most predictive in this model.*

Data

To answer these questions, we use data from the Fragile Family and Child Wellbeing Study, which is a collaboration between Princeton and Columbia Universities. This study uses a stratified, multistage design of nearly 5,000 children born in US cities between 1998 and 2000. Sixteen cities were randomly selected of urban areas with populations over 200,000 and weighted to create a nationally representative panel. The study oversampled single parent households, racial-ethnic minority families and low-income caregivers at time

of birth and is well positioned to investigate the mechanisms that contribute to disparities in child outcomes. Follow-up interviews were conducted at ages 1,3,5,9 and 15 and data collection is ongoing. Although there is some attrition, over 70% of families completed surveys at 15 years of age. Included in this panel was self-report from adolescents in addition to primary caregiver participation across many constructs, including health and wellbeing.

The original study sought to understand the nature of unmarried parents, longitudinal outcomes of children and how policies and environmental conditions impact young families. Each wave collected data on attitudes and expectations, childcare, behavioral development, demographics, education, employment, family ties, finances, health, housing, criminal legal systems, parenting, and romantic relationships. For this study, we use the unweighted sample for ease of analysis with the unrestricted data. Adolescent wellbeing is the self-reported outcome variable, constructed through 12 questions and adapted to a 100-point scale. Predictor variables were extracted from those related to social determinants of health and multiformat report of child social skills at birth and preschool age. We preferentially chose variables that are sensitive to evidence based social or health interventions, like neighborhood safety and food insecurity, and questions answered by two caregivers.

Methods

We tested a variety of models to understand the relationship between early life factors and adolescent wellbeing. Model 1 includes linear regression, linear regression with LASSO regularization, relaxed LASSO in linear regression. We then used a 0-1 classification in Model 2 to test logistic regression with LASSO and relaxed LASSO in logistic regression. Model 3 tests a random forest model with the addition of PCA transformed data. Model 4 tests a LASSO logistic model with PCA scores. Finally, we used bagging to build an ensemble model and chose Model 3 with a representative tree with 5 nodes for interpretation.

Findings

In the last model, the first node asks if the child co-sleeps with their parents at age 5. While this is perhaps not the most intuitive result, co-sleeping is common in many cultures although a less accepted practice in the US. In this context, co-sleeping may be associated with anxiety if the child is not able to separate from their parents at bedtime. Notably, there is only one prosocial behavior at age 5 in the tree: peer engagement as reported by caregivers. This is surprising given the literature on early child development. The only other variable related to behavior is having a consistent bedtime routine, which is highly related to co-sleeping. Most of the remainder of the variables are related to social conditions, and therefore preventable with appropriate interventions.

There are two environmental exposures negativly predictive of adolesent wellbeing- household smoking and prenatal alcohol consumption. Community saftey, cohesion (willingness of neighbors to help each other) and avaliability of social services like food stamps are also important contributors of wellbeing. Finally, father's self report of life satisfaction and investment in their child's life are important predictive factors. There are a couple of factors that have a neuanced interpretation. For example, not having a computer in the home is associated with wellbeing. We may expect that family access to technology would facilitate accessing resources that promote healthy development. That said, excessive exposure to screens is well understood to stiffen brain maturation in early life.

Implications

Age zero to five is an important period in child development, interventions administered in this timeframe are likely to have enduring benefits. Of the public health programs that have focused on supporting young families with social vulnerabilities, many pay dividends on initial investments.(7) Insight into early life predictors of adolescent well-being can advise policy aimed at reducing disease burden longitudinally. There is a

particular need to study the implementation of this evidence base in healthcare, educational and community-based settings. Preventative measures are a priority, however understanding longitudinal outcomes may also be helpful for point-in-time interventions. For example, understanding early risk factors could lead to increased allocation of resources to support resilience in adolescents. This is critical during the present youth mental health crisis and in efforts to prevent long term effects on adolescent wellbeing during the pandemic. In short, understanding longitudinal predictors of adolescent wellbeing has direct implications for reducing onset of mental health problems and screening to identify risk factors for early intervention across community settings.

References

1. Roos LL, Wall-Wieler E, Lee JB. Poverty and Early Childhood Outcomes. Pediatrics. 2019;143(6):e20183426.
2. Polanczyk GV. The burden of childhood mental disorders. European Child & Adolescent Psychiatry. 2013;22(3):135-7.
3. Benarous X, Raffin M, Bodeau N, Dhossche D, Cohen D, Consoli A. Adverse childhood experiences among inpatient youths with severe and early-onset psychiatric disorders: Prevalence and clinical correlates. Child Psychiatry and Human Development. 2017;48(2):248-59.
4. Sarkar T, Patro N, Patro IK. Cumulative multiple early life hits- a potent threat leading to neurological disorders. Brain Res Bull. 2019;147:58-68.
5. Dean RS, Davis AS. Relative risk of perinatal complications in common childhood disorders. School Psychology Quarterly. 2007;22(1):13-25.
6. Bernard DL, Smith Q, Lanier P. Racial discrimination and other adverse childhood experiences as risk factors for internalizing mental health concerns among Black youth. Journal of traumatic stress. 2021.
7. Dawson G, Ashman SB, Carver LJ. The role of early experience in shaping behavioral and brain development and its implications for social policy. Dev Psychopathol. 2000;12(4):695-712.

Detailed Analysis

1 Data Preparation and Exploratory Data Analysis

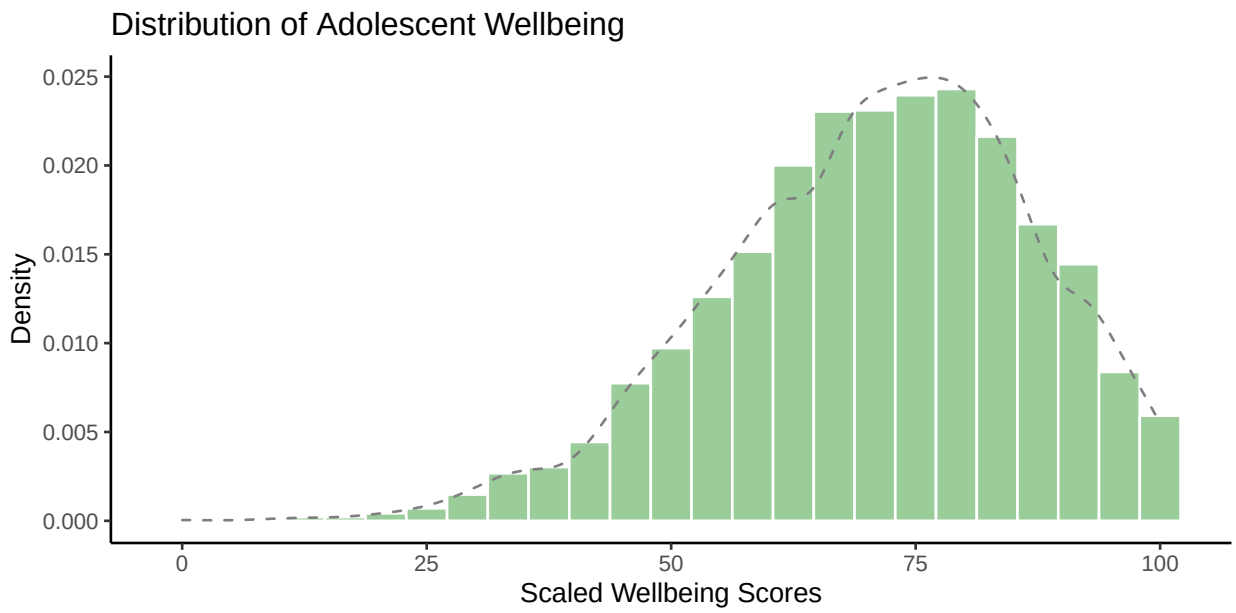
Data Prep

We start our analysis from the cleaned data (see Appendix 1 for data preparation from original data). We merge the features and labels and remove person id to get the data frame.

There are 3,407 records with 53 features and 1 label (wellbeing). Each column stands for one question in the survey and the answers are integers representing their answers. The description for each question is in the appendix.

Distribution of Wellbeing

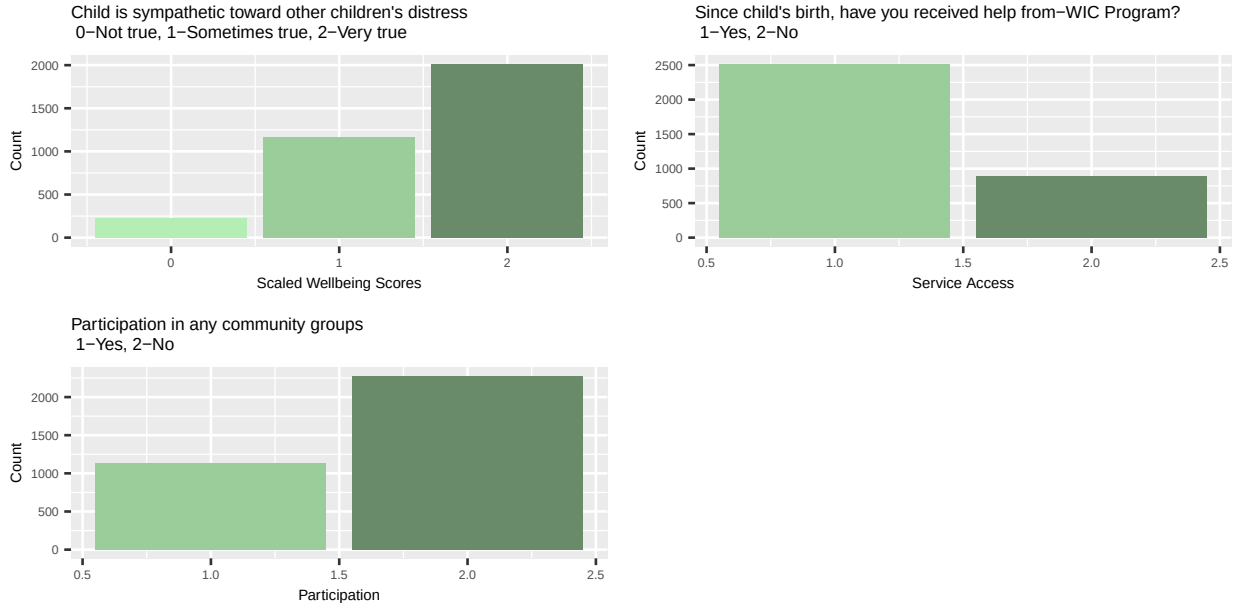
In our study, we want to predict children's wellbeing, so we first take a look at the distribution of their self-reported wellbeing scores



The histogram is little skewed to the left. Most people have scores above 50 and the mean wellbeing is around 70, indicating that in general adolescents in the sample rate their wellness highly.

Distribution of some features

Let's also pick a few features and check their distribution:



As is shown in the above plots, most children are sympathetic toward others' distress. More than half have received help from WIC, which is predictive of access to other government sponsored resources. More than half of the families participated in community social environments. These results confirm that the majority of our participants have good childhood wellbeing. However, there are also a considerable part of people who report negative scores on early life conditions associated with child development. This distribution provides variability in key predictors in the following models.

2 Model Fitting

Next, we fit the data to several models. Our aims are to predict childhood wellbeing based on the answers of the above mentioned questions and assess multiple informant report of similar constructs.

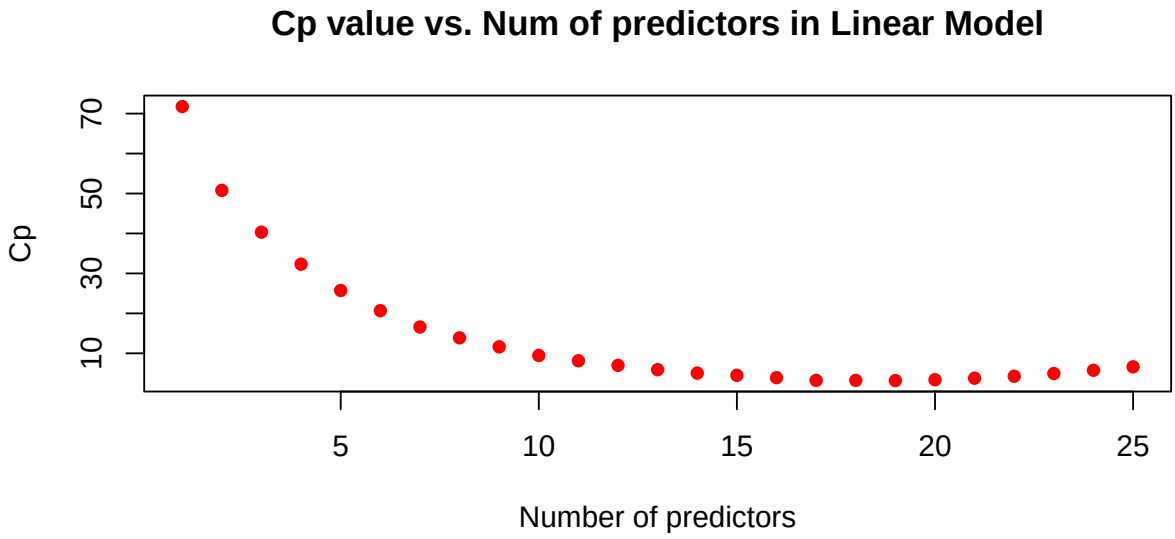
Train/Test/Validation split

We first split the data frame for training, testing and validation of the models. The testing error is used to measure the performance of each model. The model with the least testing error will be further tested on the validation data.

Model 1: Linear Regressions

We start our exploration from fitting all the variables to a linear regression model.

Apparently the linear model with all variables doesn't work well. The R-square is only 0.037, and most features are insignificant in 0.1 level.Non-linear relationship between those questions and wellbeing can be the main cause. We still conducted a backward model selection with Cp to see whether there are some interesting significant variables.



Here, 17 variables give the lowest Cp.

```
##  [1] "(Intercept)" "k5e2a"       "f1e3"        "m1a13"       "m1f5"
##  [6] "m1b18"       "m2d3a"       "m2h8e"       "m2h19b"      "p4b15"
## [11] "p4l11"       "p4l19"       "p4l43"       "m4i0g"       "m4h4"
## [16] "m4i0p"       "f4i8a1"      "f4a2"
```

The final linear model will use these 17 variables

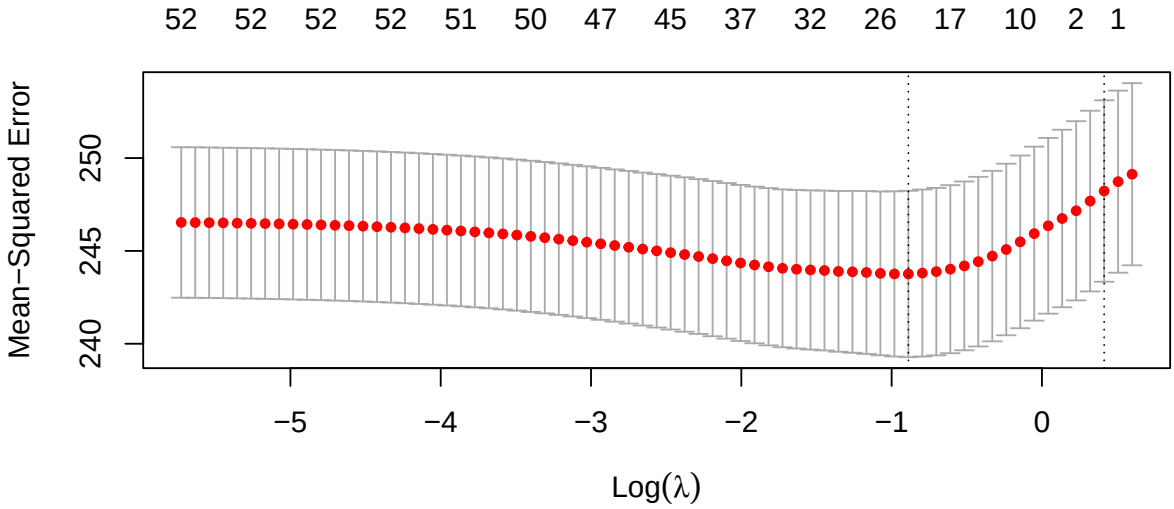
```
##
## Call:
## lm(formula = wellbeing ~ k5e2a + f1e3 + m1a13 + m1f5 + m1b18 +
##     m2d3a + m2h8e + m2h19b + p4b15 + p4l11 + p4l19 + p4l43 +
```

```
##     m4i0g + m4h4 + m4i0p + f4i8a1 + f4a2, data = data.train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.656 -10.399   1.231  10.879  38.948
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  57.3443     8.1123   7.069 1.97e-12 ***
## k5e2a        -0.4319     0.2074  -2.082  0.03744 *
## f1e3         -0.8215     0.3072  -2.674  0.00753 **
## m1a13         4.2726     2.2420   1.906  0.05679 .
## m1f5         -0.7434     0.4322  -1.720  0.08554 .
## m1b18        -2.2972     1.3998  -1.641  0.10090
## m2d3a        -1.8190     0.8206  -2.217  0.02672 *
## m2h8e         1.6965     0.7112   2.385  0.01714 *
## m2h19b        5.7857     3.0751   1.881  0.06001 .
## p4b15         1.4023     0.7112   1.972  0.04872 *
## p4l11         2.3553     0.4801   4.906 9.84e-07 ***
## p4l19        -2.0838     0.6803  -3.063  0.00221 **
## p4l43         0.8009     0.4293   1.865  0.06222 .
## m4i0g        -1.0454     0.6040  -1.731  0.08360 .
## m4h4         -1.7290     0.8603  -2.010  0.04455 *
## m4i0p        -1.2611     0.6431  -1.961  0.04998 *
## f4i8a1        2.6527     1.6074   1.650  0.09900 .
## f4a2         -0.4650     0.2353  -1.976  0.04824 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.45 on 2782 degrees of freedom
## Multiple R-squared:  0.04792,    Adjusted R-squared:  0.0421
## F-statistic: 8.236 on 17 and 2782 DF,  p-value: < 2.2e-16

## [1] 291.6771
```

There's little improvement. With 17 selected variables based on Cp value, the resulting linear model have a R-square of only 0.04 and the MSE on testing dataset is 291.6771.

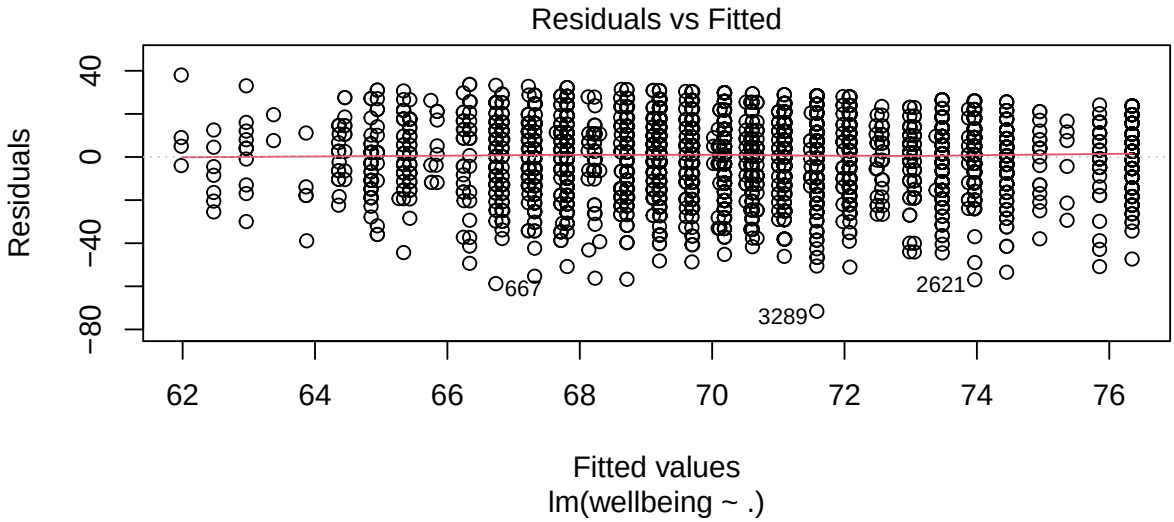
We then tried LASSO regularization on linear regression. The result has a MSE is 299.2521 and is not a good fit.

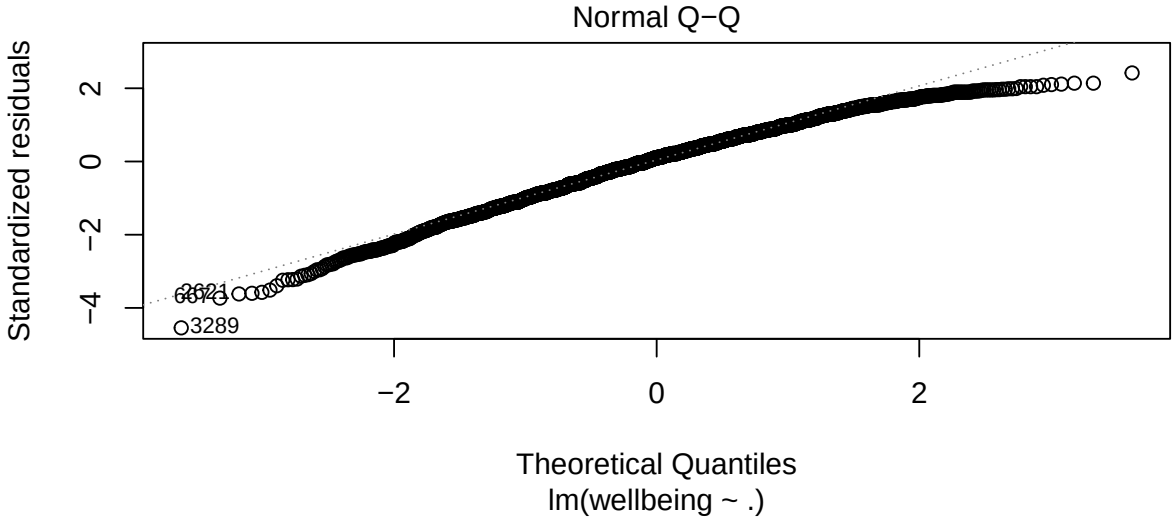



```
## [1] 299.2521
```

A relaxed LASSO with variables from the prior model has a slightly better testing error (287.5523) but R-square is still low (0.03038)

```
##
## Call:
## lm(formula = wellbeing ~ ., data = data.fl.sub)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -71.581 -10.082   1.524  11.418  38.021
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  68.8265     1.8424  37.358  < 2e-16 ***
## m2h8e         2.8685     0.6385   4.493 7.26e-06 ***
## p4l11         2.3768     0.4427   5.369 8.45e-08 ***
## p4l32        -0.4912     0.4578  -1.073  0.28338
## m4h4         -2.4716     0.7579  -3.261  0.00112 **
## m4i0p        -1.8953     0.5851  -3.239  0.00121 **
## p4b9          1.3948     0.5814   2.399  0.01650 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.76 on 3400 degrees of freedom
## Multiple R-squared:  0.03209,    Adjusted R-squared:  0.03038
## F-statistic: 18.79 on 6 and 3400 DF,  p-value: < 2.2e-16
```



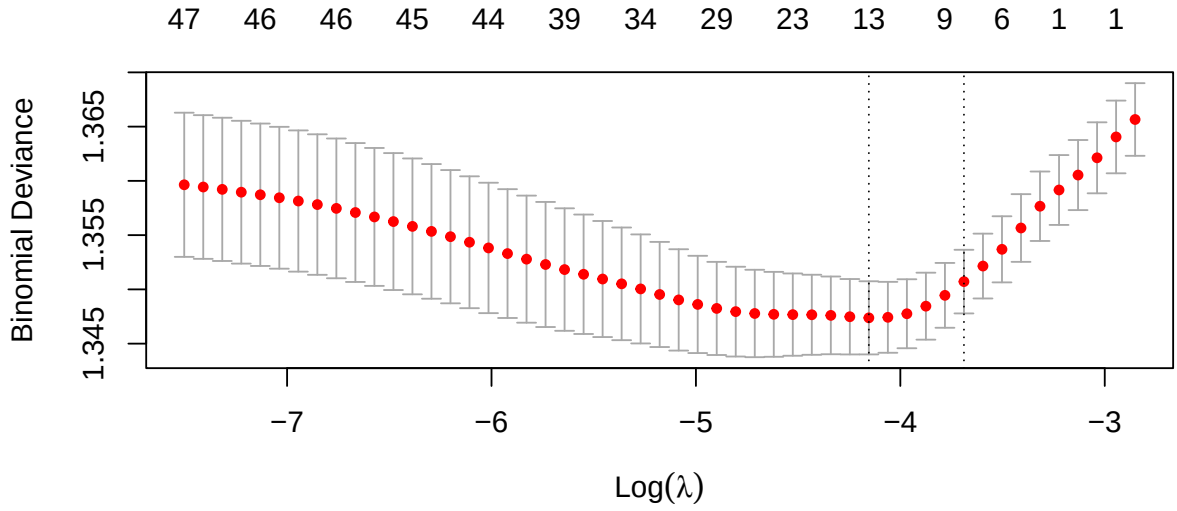



```
## [1] 287.5523
```

Model 2: Logistic regressions

Based on the results so far, we concluded that linear models are not suitable for our analysis. We decided to change the wellness varialbe to a 0-1 classification problem by defining >70+ as good wellbeing (1) and use this to fit a logistic regression model.

Again, most of the features are not significant. Let's try logistic regression with LASSO.




```
## [1] 0.3973941
```

This model gives a testing error of 0.3973941, which is not bad.

Next we fit a relaxed LASSO for logistic regression model.

```
##
## Call:
## glm(formula = wellbeing ~ ., family = binomial, data = data.fl.sub1)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6347  -1.2403   0.8629   1.0490   1.7021
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  0.44043    0.27369   1.609  0.10756
## f1e3        -0.09871    0.04049  -2.438  0.01477 *
## m2h8e        0.22599    0.09312   2.427  0.01523 *
## p4l11        0.32484    0.06293   5.162 2.45e-07 ***
## p4l19       -0.25090    0.08832  -2.841  0.00450 **
## m4h4        -0.33907    0.10955  -3.095  0.00197 **
## m4i0p       -0.23491    0.08492  -2.766  0.00567 **
## p4b9         0.16176    0.08260   1.958  0.05019 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3823.7  on 2799  degrees of freedom
## Residual deviance: 3730.4  on 2792  degrees of freedom
## AIC: 3746.4
##
## Number of Fisher Scoring iterations: 4
```

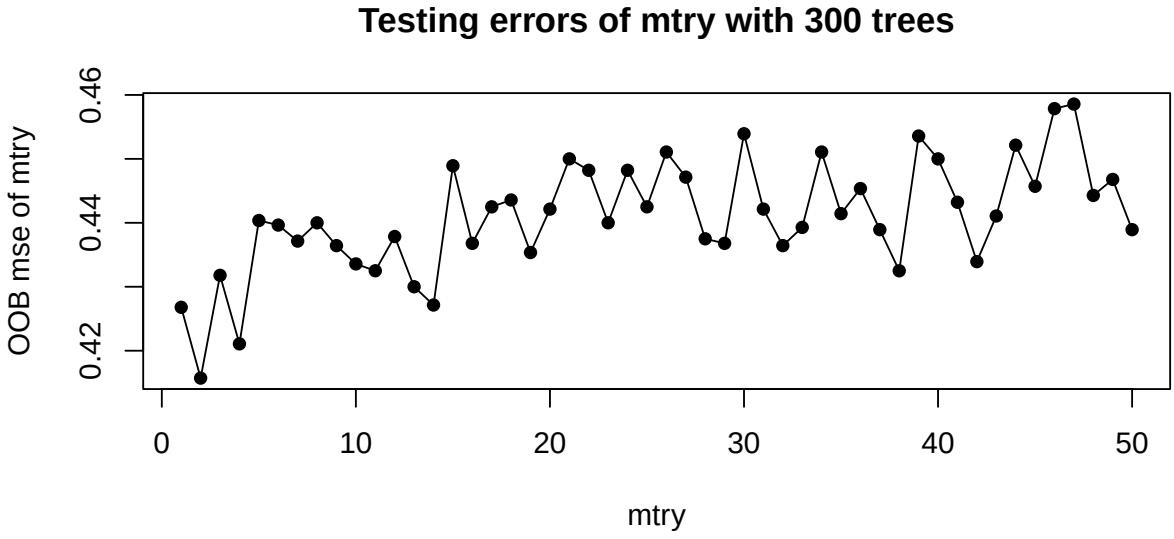
```
##
## fit.1se.glm.pred    0    1
##                0  341  268
##                1  858 1333

## [1] 0.4104235
```

The relaxed LASSO model has a higher testing error (0.4104235).

Model 3: Random forest

An ntree > 100 can mitigate the OOB testing errors. We decide to use ntree = 300.

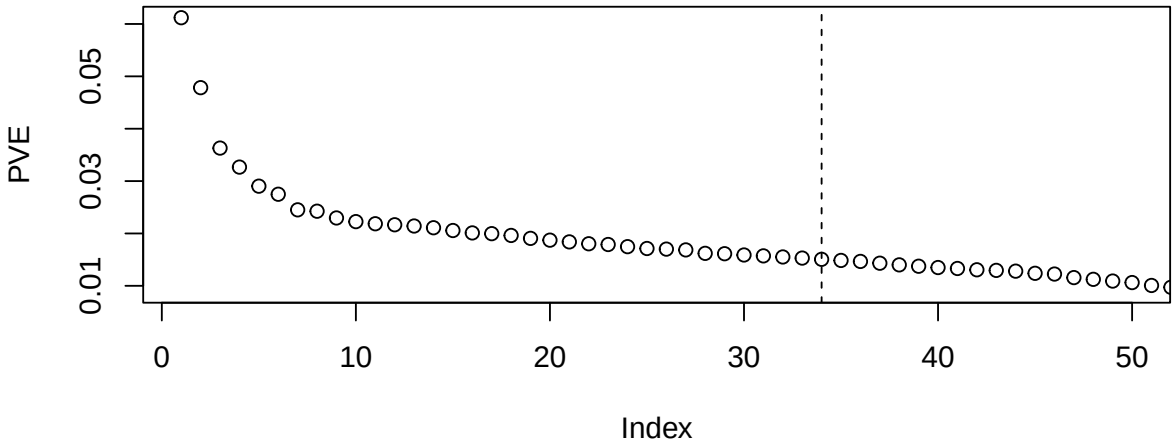


We decide to use mtry=14, so our final model will use ntree=300 and mtry=14.

```
## [1] 0.3843648
```

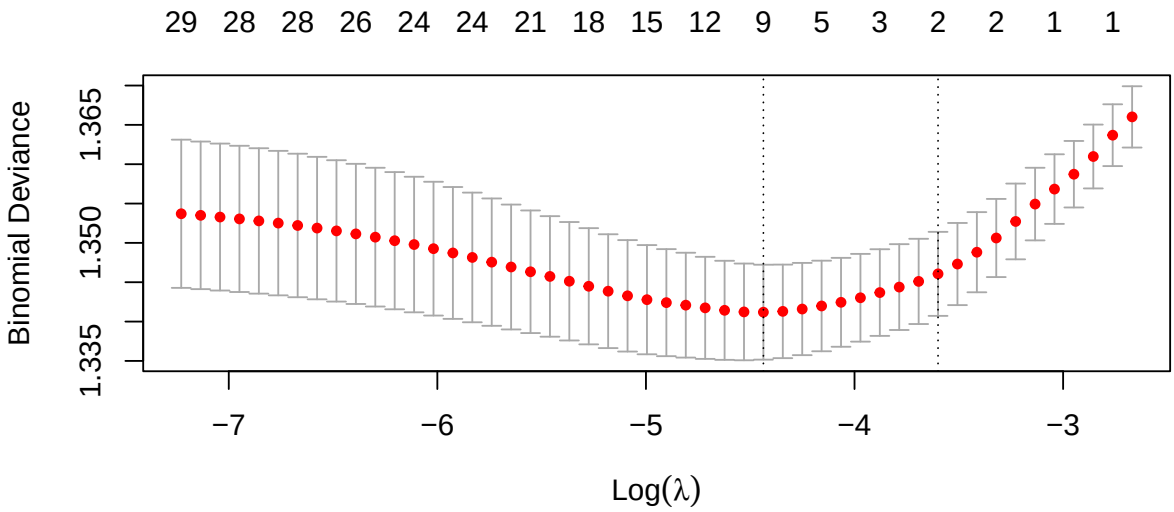
The testing error for random forest is 0.38, which is lower than previous logistic models.

We move to creating PC values with the training data.

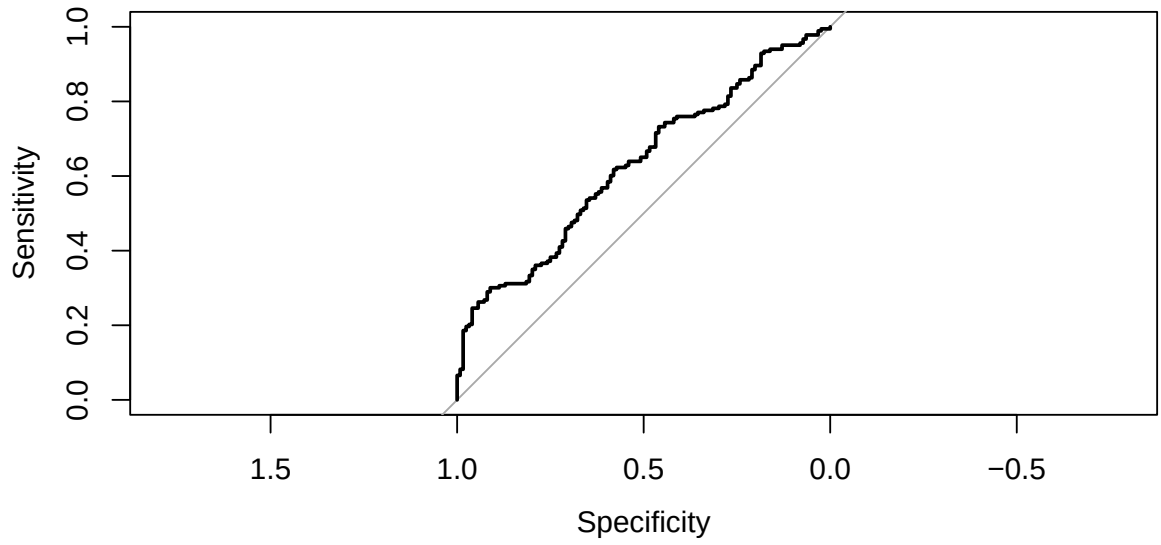


Based on the plots, we decide to use 34 PC scores to represent the original data set, since 34 PCs would capture about 76.5% of the variance. Then, we extract PC scores from these 34 PCs and predict PC scores for testing data.

Model 4: LASSO logistic model with PCA scores.



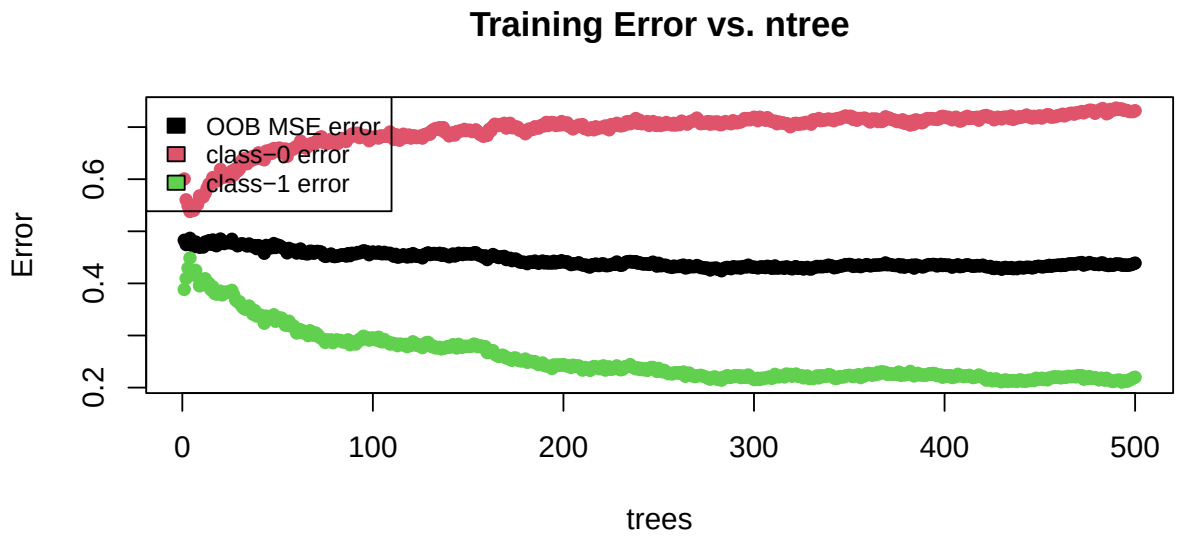

```
## [1] 0.3941368
```



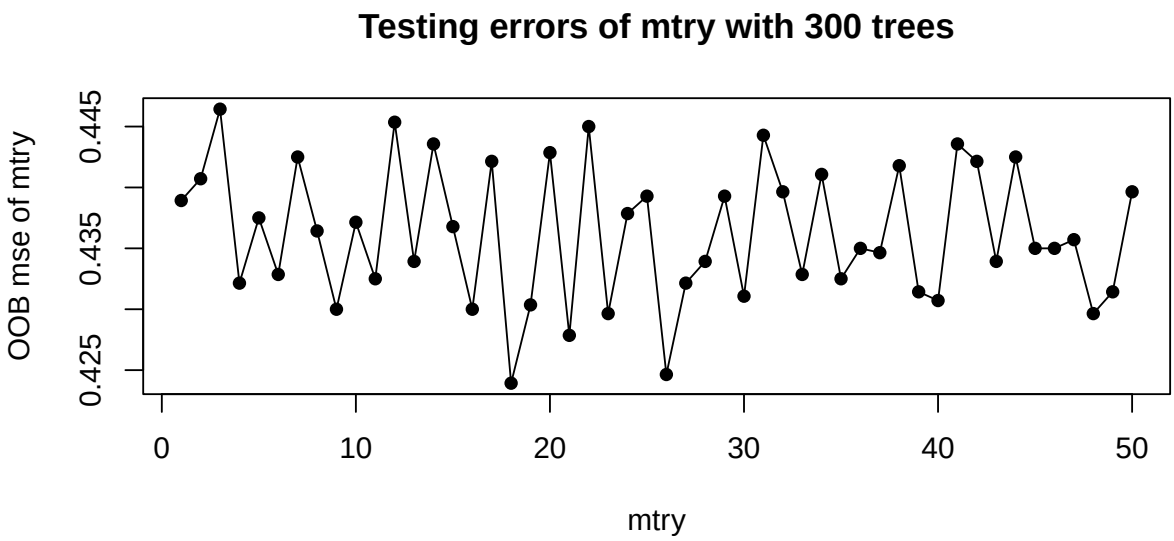

```
##
## Call:
## roc.default(response = data.test1$wellbeing, predictor = predict.glm.pca,     plot = T)
##
## Data: predict.glm.pca in 124 controls (data.test1$wellbeing 0) < 183 cases (data.test1$wellbeing 1).
## Area under the curve: 0.6296
```

LASSO with PCA scores gives a testing error of 0.3941368.

Model 5: Random Forest with PCA scores



An ntree > 200 can settle the OOB testing errors. We will go with ntree = 300.



We will use mtry=18 and our final model will use ntree=300 and mtry=18. (mtry=p/3)

```
## [1] 0.4364821
```

The testing error for randomForest model with PCA data is 0.44, which is higher than the LASSO logistic PCA model.

Model 6: Baggings

So far, we have 6 models: 1- Linear model (not considered); two variations of model 2: LASSO Logistic regression (testing error=0.3973941) and Relaxed LASSO Logistic regression (testing error=0.4104235); 3- Random forest(testing error=0.3843648); 4- LASSO Logistic regression with PCA (testing error=0.3941368); 5- Random forest with PCA (testing error=0.4364821). We then build an ensemble model by taking the average of predicted probabilities from model 2, 3 and 4, and make prediction based on the averaged probability.

```
## [1] 0.3941368
```

Final model and validation

Based on the testing error, our final model will be Model 3: Random forest.

```
## [1] 0.07666667
```

The final validation error is 0.076, which is acceptable.

```
## [1] "Constructing distance matrix..."
## [1] "Finding representative trees..."
```

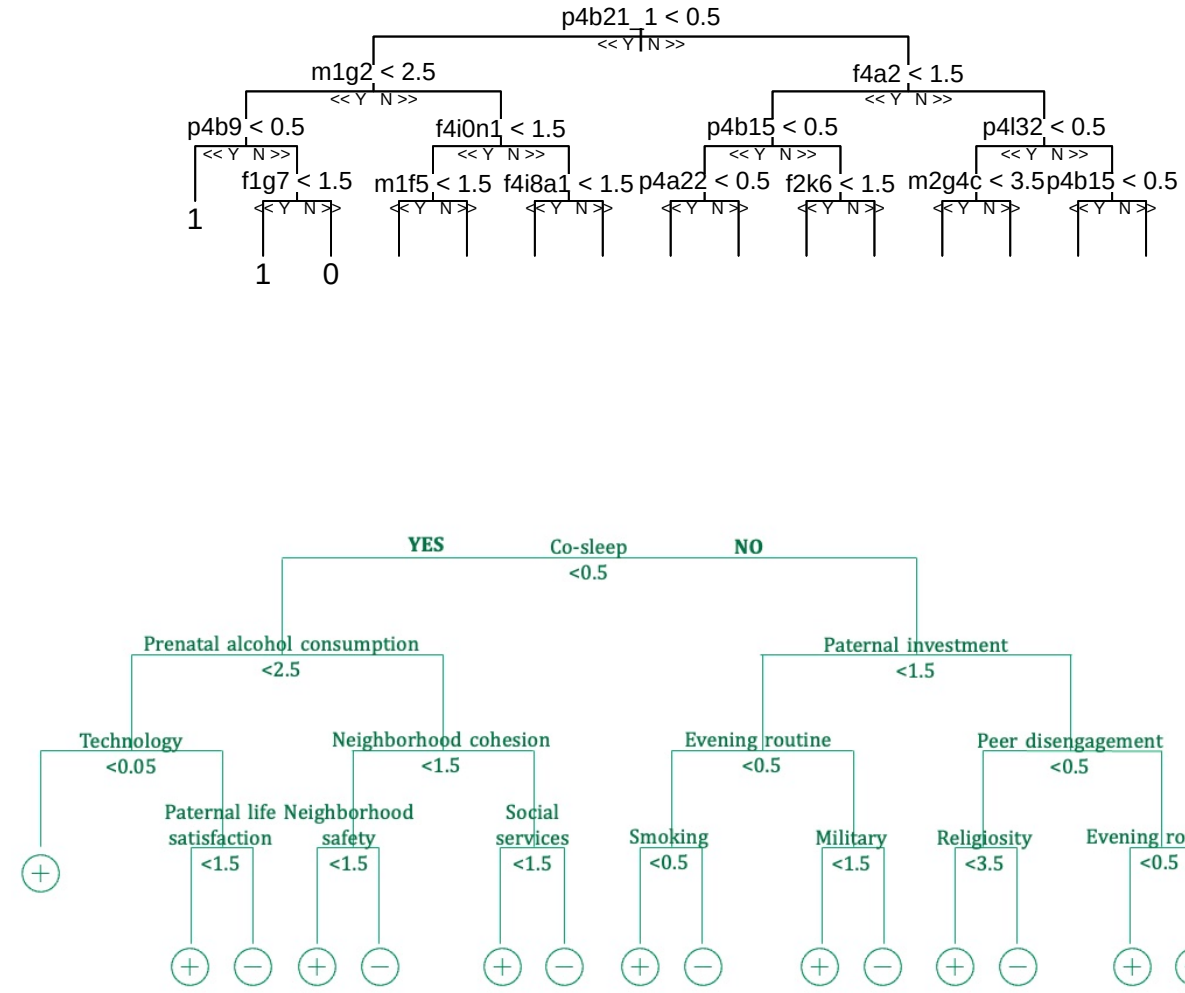


Figure 1: Tree of early predictors of Adolescent wellbeing

This plot shows the best tree using d2 metric with the depth of 5. While it is not one of the trees in the randomForest, it is easier to look at. Because the random forest produced a tree with over 50 features, we elected to limit the nodes to better visualize relationships. This tree provides an interpretable relationship between early predictors of child wellbeing.

Appendix

Appendix 1: Data Preparation

We first select 12 columns based on domain knowledge that are related to children's wellbeing. These items were created by study researchers from the Adaptive Social Behaviora Inventory and the Social Skills Rating System. Both are tools validated to assess adolescent social and psychological wellbeing. These scales ask questions like: I am open and direct about what I want, I make friends easily, I am self-confidenc tin social

situations. For each column, a larger value stands for better self report of wellness. We sum them to create a total score for wellbeing. Finally, we multiplied this score by 4.16 to create a variable scaled to 100 for easier understanding.

We then selected more than 60 columns that are associated in the literature with positive child development at birth and age 5. The total number of children in survey is 4800, so we dropped all columns with >2000 NAs. This is a reasonable decision because the survey has skip patters and not all families were asked these questions. At year 15, the study had an attrition rate of less than 30% and questions were administered by in-person research staff. We are confident the data is not biased based on missingness and therefore did not conduct sensitivity analyses.

For the remainder of the variables, we used a package called "mice" to impute the values. We used method='polr' for ordinal categorical columns and method='pmm' for continuous column. We checked the percentage for each level, which was similar before and after imputing. This indicates that the imputations did not change the distribution for each level, which is a good indicator for the following analyses.

We elected not to list all study questions used as variables in this project, however key constructs include parental life satisfaction, presense of prenatal care, neighborhood saftey, and father's investment in child rearing at the time of birth. At age 5 we included variables associated with prosocial behavior (child gets along with peers, child is nervous, child emotion regulation), participation in school events, recieving social services like supplemental income, exposure to interpersonal violence and family sleep hygiene.

At this stage, we have two choices, leave the categorical as categorical or change it to numeric. We tried both methods and found numeric is a better choice for modeling. If needed, we can change it to numeric directly because the data frame has ordered categorical columns. This cleaned data was used in the study to model adolescent wellbeing from early life predictors.