Project 2

Assessing Risk Factors of Bone Fractures Within the First Year of Treatment for Women with Osteoporosis

Presented by:
Derek Rogers
Garrett Drake
Joshua Hudson

Overview

- Introduction
- Objective Summary
- Data Description
- Exploratory Data Analysis (EDA)
- Objective 1
- Objective 2
- Conclusion
- Appendix

Objective Summary

- Objective 1
  - Modeling approach
  - Feature Selection Summary
  - Final Model
  - Model Interpretation
- Objective 2
  - Summary of approach
  - Prediction metrics and complexity discussion
  - Model comparisons
  - ROC curves
  - Insights

Data Description

- Data taken from the glow_bonemed dataset in the aplore3 R package.
  - This data set looks to assess risk factors and predict if a woman with osteoporosis will have a bone fracture within the first year of joining the study.
    - 500 subjects
    - 18 variables (Next Slide)
    - No missing values

Data Description (Cont.)

- Variable Details
  - **Response:**
    - **Fracture** - Any fracture in first year
  - **Explanatory Variables:**
    - **bonemed** - Bone medications at enrollment
    - **bonemed_fu** - Bone medications at follow-up
    - **bonetreat** - Bone medications both at enrollment and follow-up
    - **priorfrac** - If the patient previously had a fracture
    - **Age** (at enrollment)
    - **weight** (in kilos)
    - **height** (in CM)
    - **BMI** (Kg/m^2)
    - **Smoke** – Subject is a smoker
    - **Premeno** - Menopause before age 45
    - **Momfrac** - Mother had hip fracture
    - **Armassist** - Arms are needed to stand from a chair
    - **Raterisk** - Self-reported risk of fracture
    - **Fracscore** - Fracture Risk Score (Composite Risk Score)

Continuous Numerical

Nominal Yes/No

Ordinal 1: Less than others of the same age
2: Same as others of the same age
3: Greater than others of the

Summary Stats

Continuous

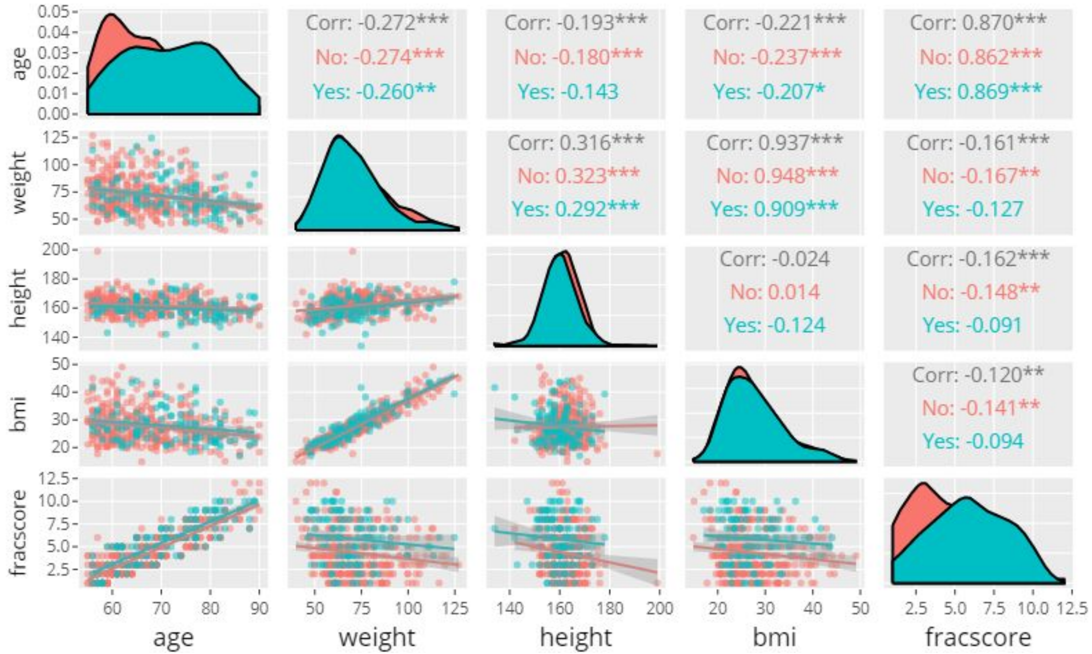
var <chr>	min <dbl>	max <dbl>	mean <dbl>	sd <dbl>	variance <dbl>
age	55.00000	90.00000	68.56200	8.989537	80.811780
bmi	14.87637	49.08241	27.55303	5.973958	35.688178
fracscore	1.00000	12.00000	4.69800	2.495446	6.227251
height	134.00000	199.00000	161.36400	6.355493	40.392289
weight	39.90000	127.00000	71.82320	16.435992	270.141825

Nominal & Ordinal

priorfrac	premeno	momfrac	armassist	smoke	raterisk	fracture	bonemed	bonemed_fu	bonetreat
No :374	No :403	No :435	No :312	No :465	Less :167	No :375	No :371	No :361	No :382
Yes:126	Yes: 97	Yes: 65	Yes:188	Yes: 35	Same :186	Yes:125	Yes:129	Yes:139	Yes:118
					Greater:147				

EDA

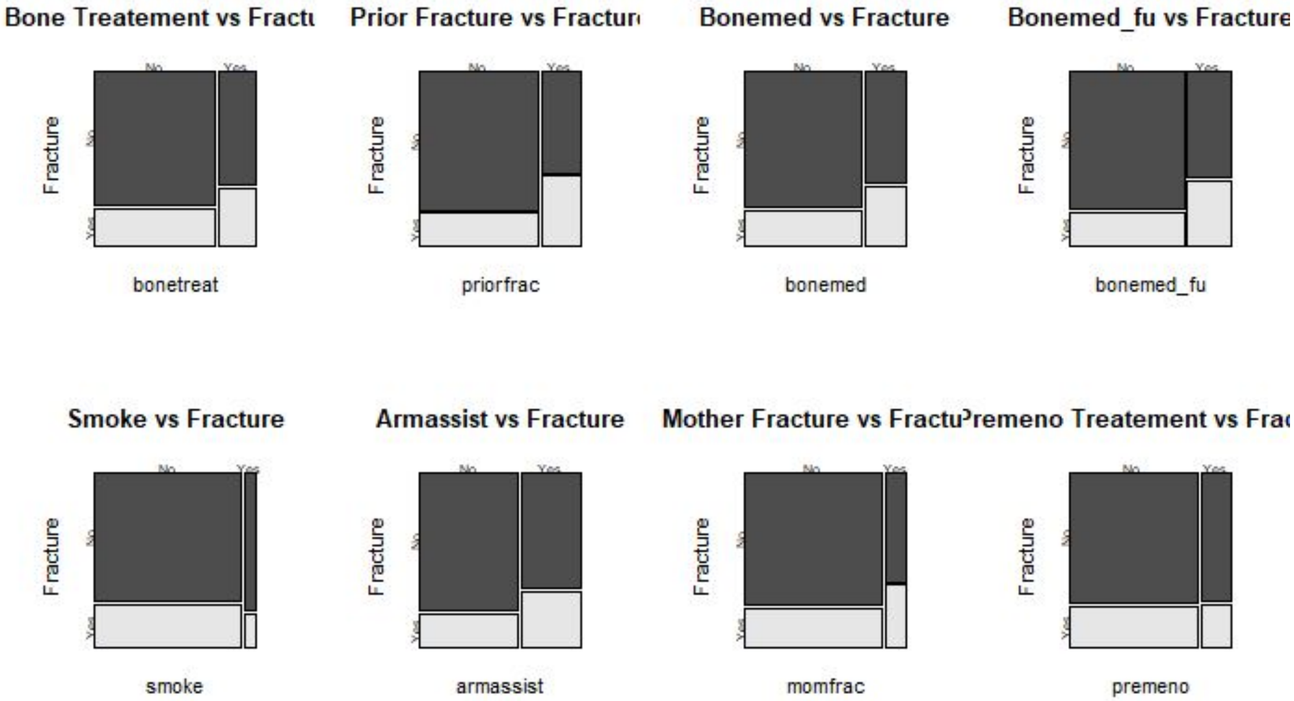
- Relationships – **Continuous** Variables



Collinearities:

- Weight and BMI
- Weight and Height
- Age and Fracscore

Yes/No Categorical Variable Mosaic Plots

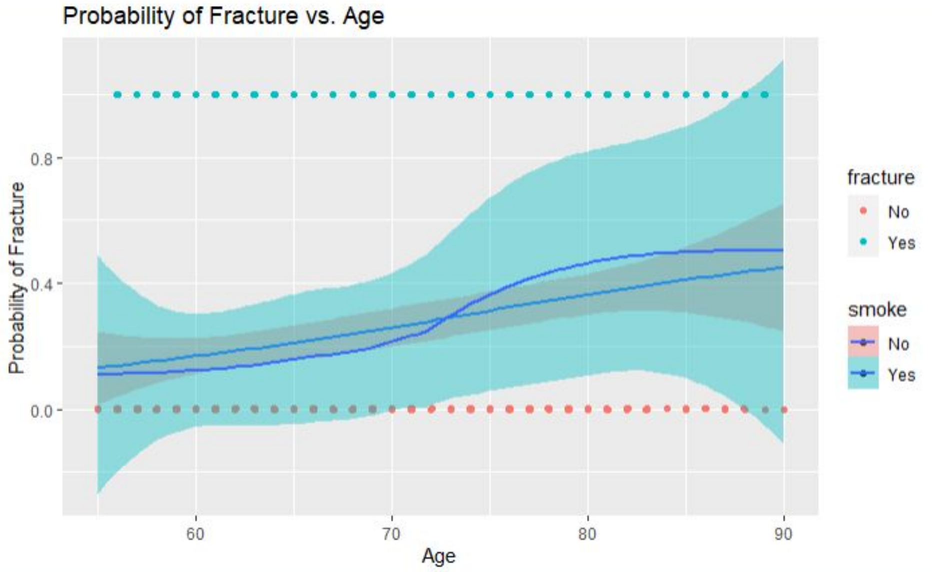


Multiple Correspondence Analysis

```
Link between the variable and the categorical variable (1-way anova)
=============================================
                    R2       p.value
bonetreat  0.87245886 7.754159e-225
bonemed    0.84726742 2.443029e-205
bonemed_fu 0.83803940 5.422994e-199
raterisk   0.19571783  3.118876e-24
fracscore  0.19131230  2.223696e-17
priorfrac  0.10838142  4.201009e-14
fracture   0.04922945  5.400633e-07
armassist  0.02280481  7.047091e-04
premeno    0.01718799  3.315107e-03
smoke      0.01441500  7.194873e-03
```

EDA

- Relationships – Loess Plots Variables



Objective 1

- Modeling approach – Multiple Logistic Regression
  - Wide open – All variables model:
    - fracture = priorfrac + age + weight + height + premeno + momfrac + armassist + smoke + raterisk + bonemed + bonemed_fu + bonetreat
    - Complexity and assumption issues – Variable down selection needed
      - EDA
      - Intuition
      - VIF

Objective 1 (Cont.)

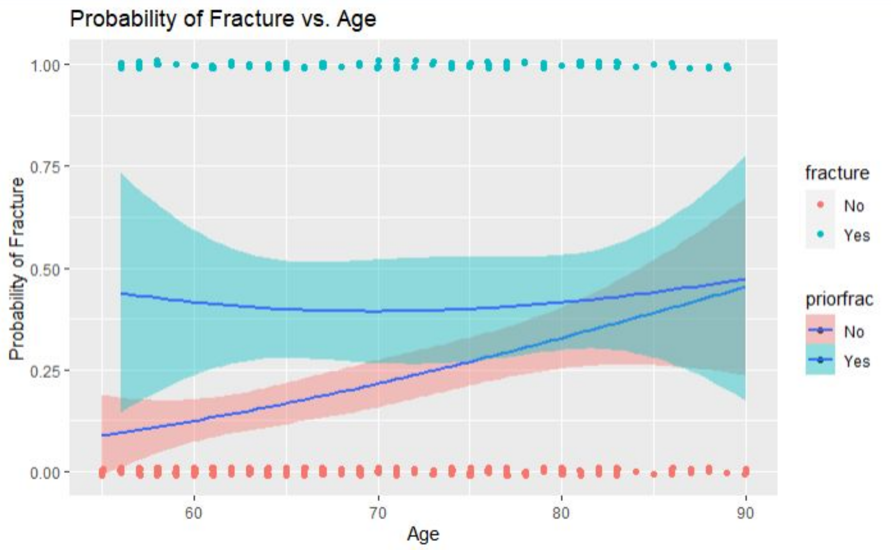
- Reduced model – All variables model:
  - fracture = age + priorfrac + bonemed

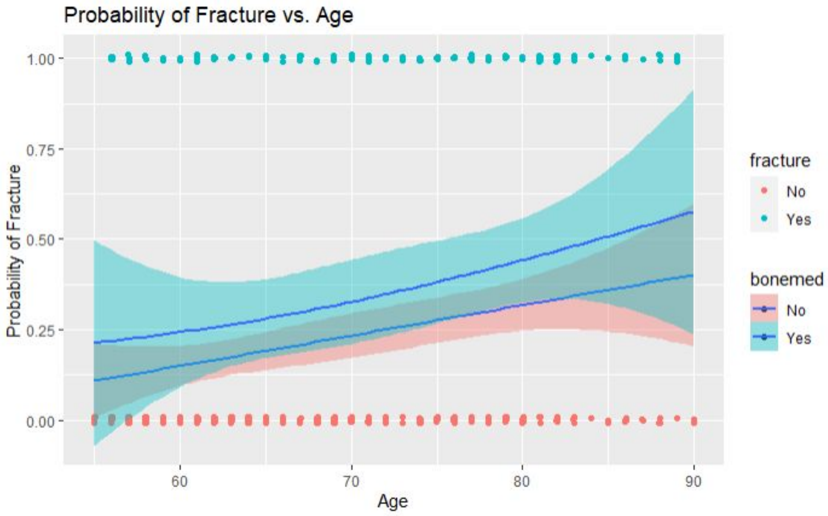
$$P(Fracture) = \frac{e^{-4.089+0.038\boldsymbol{Age}+0.793\boldsymbol{PriorFracture}+0.474\boldsymbol{BoneMedication}}}{1+e^{-4.089+0.038\boldsymbol{Age}+0.793\boldsymbol{PriorFracture}+0.474\boldsymbol{BoneMedication}}}$$

$$Odds(Fracture) = e^{-4.089+0.038\boldsymbol{Age}+0.793\boldsymbol{PriorFracture}+0.474\boldsymbol{BoneMedication}}$$

Objective 1

- Relationships – Loess Plots for Model Variables





Objective 1 (Cont.)

- Model Interpretation:
  - Holding prior fracture history and bone medication status at the time of enrollment fixed, For any 1-year increase in age, the odds a fracture occurring will increases by a factor of 1.04.
  - 95% CI: (1.01, 1.06)

  - Holding age and bone medication status at the time of enrollment fixed, the odds a fracture occurring will increases by a factor of 2.2 for those who have had history of fractures as compared to those who have no history with fractures.
  - 95% CI: (1.4, 3.5)

  - Holding age and prior history of fractures fixed, the odds a fracture occurring will increases by a factor of 1.6 for those who have had bone medication prescribed at the time of enrollment as compared to those who have no history with fractures.
  - 95% CI: (1.0, 2.5)

Objective 2

- More complex models where prediction is more important than interpretation
- Multiple models for this were built
  - Complex and LDA models using interactions
  - Random Forest Model
  - KNN model on the continuous variables

Complex model using interactions

Began by continuing from EDA looking at possible interactions and important variables.

Relevant EDAs and effects plots to follow with effects plots for this model for some insight

Model is:

fracture = age+ bonetreat + fracscore + priorfrac + bonemed + bonemed_fu + priorfrac:fracscore + age:fracscore + fracscore:bonetreat

Complex model coefficients and Hosmer and Lemeshow GOF test

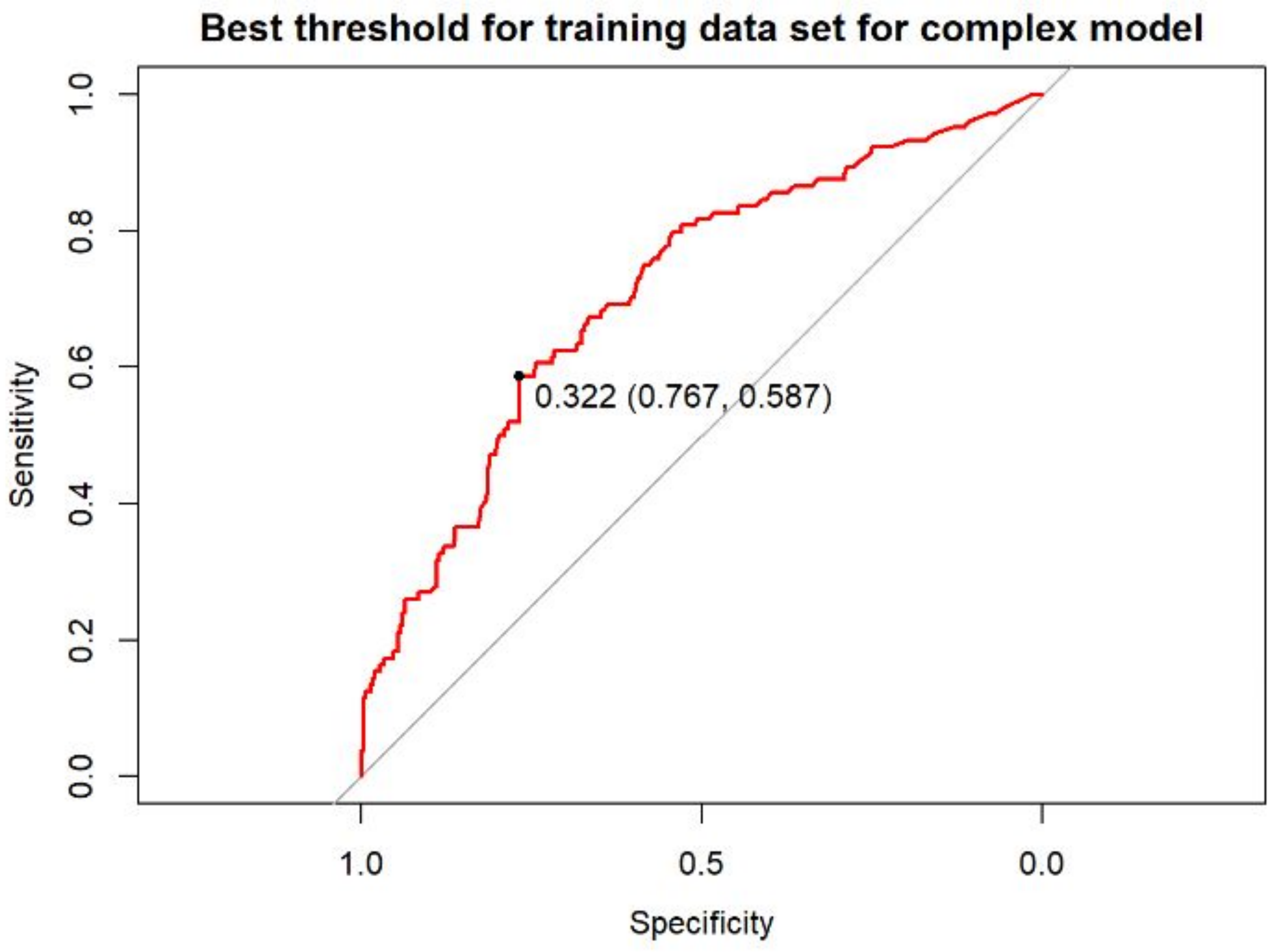
```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  complex1$y, fitted(complex1)
## X-squared = 3.0878, df = 8, p-value = 0.9287
```

```
## glm(formula = fracture ~ age + bonetreat + fracscore + priorfrac +
##     bonemed + bonemed_fu + priorfrac:fracscore + age:fracscore +
##     fracscore:bonetreat, family = "binomial", data = trainingDataframe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6317  -0.7893  -0.5366   0.7834   2.2496
##
## Coefficients:
##                         Estimate Std. Error z value Pr(>|z|)
## (Intercept)             0.417594   2.417307   0.173  0.86285
## age                    -0.049396   0.039618  -1.247  0.21247
## bonetreatYes           -1.631263   1.054259  -1.547  0.12179
## fracscore               0.144325   0.433342   0.333  0.73910
## priorfracYes            1.510602   0.777307   1.943  0.05197 .
## bonemedYes              1.284100   0.722148   1.778  0.07538 .
## bonemed_fuYes           1.525616   0.537536   2.838  0.00454 **
## fracscore:priorfracYes -0.220090   0.136224  -1.616  0.10617
## age:fracscore           0.003440   0.006112   0.563  0.57360
## bonetreatYes:fracscore -0.126240   0.106897  -1.181  0.23762
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 458.45  on 399  degrees of freedom
## Residual deviance: 409.12  on 390  degrees of freedom
## AIC: 429.12
```

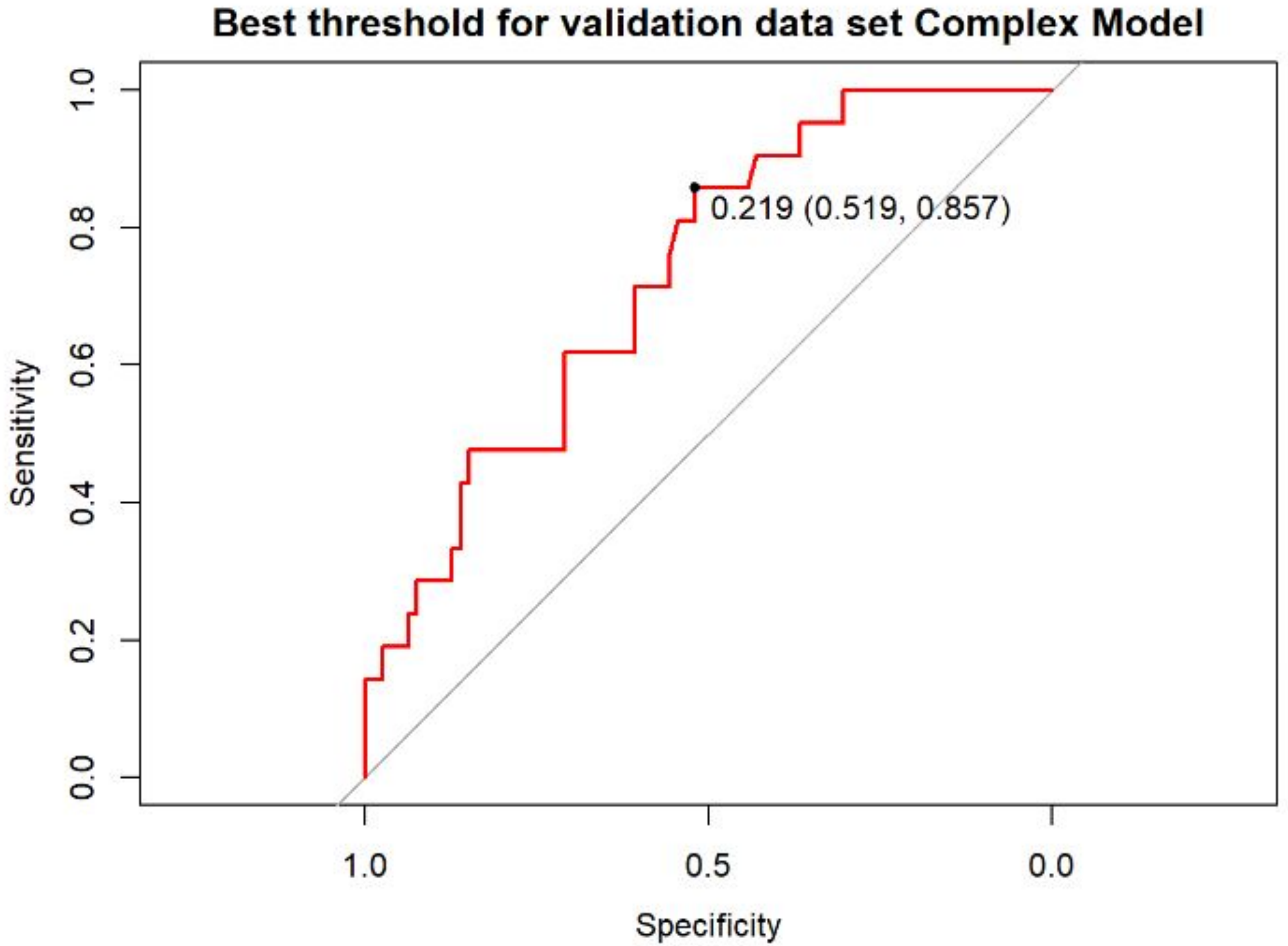
Confidence intervals for complex 1 model

```
##                                 2.5 %      97.5 %
## (Intercept)                0.01318641 177.332120
## age                        0.87981580   1.028116
## bonetreatYes               0.02389475   1.541288
## fracscore                  0.49663165   2.735423
## priorfracYes               0.96742351  20.864617
## bonemedYes                 0.86882032  15.975764
## bonemed_fuYes              1.62347863  13.742479
## fracscore:priorfracYes     0.61370375   1.050575
## age:fracscore              0.99138414   1.015554
## bonetreatYes:fracscore     0.71556364   1.090464
```

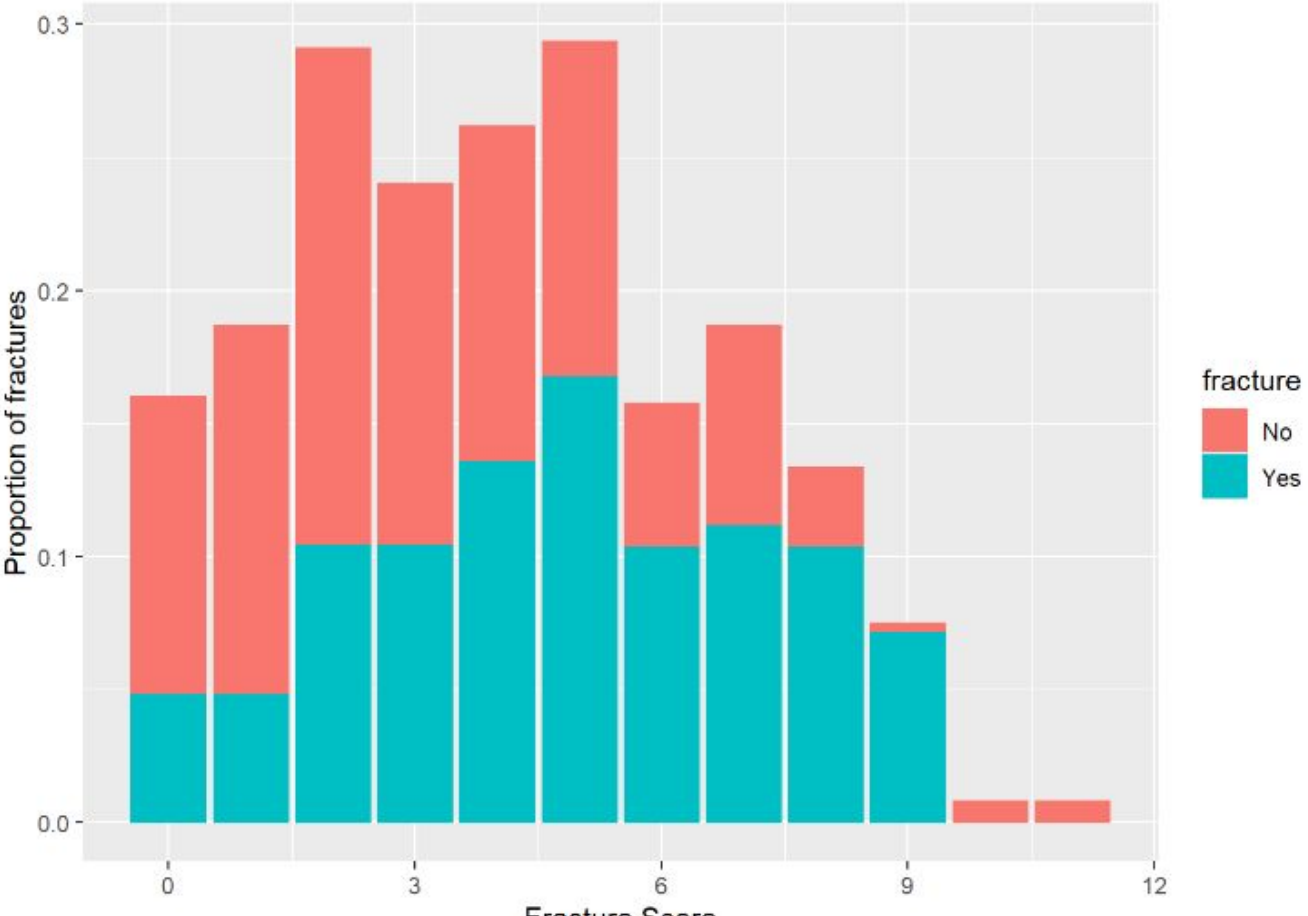
Performance for training set



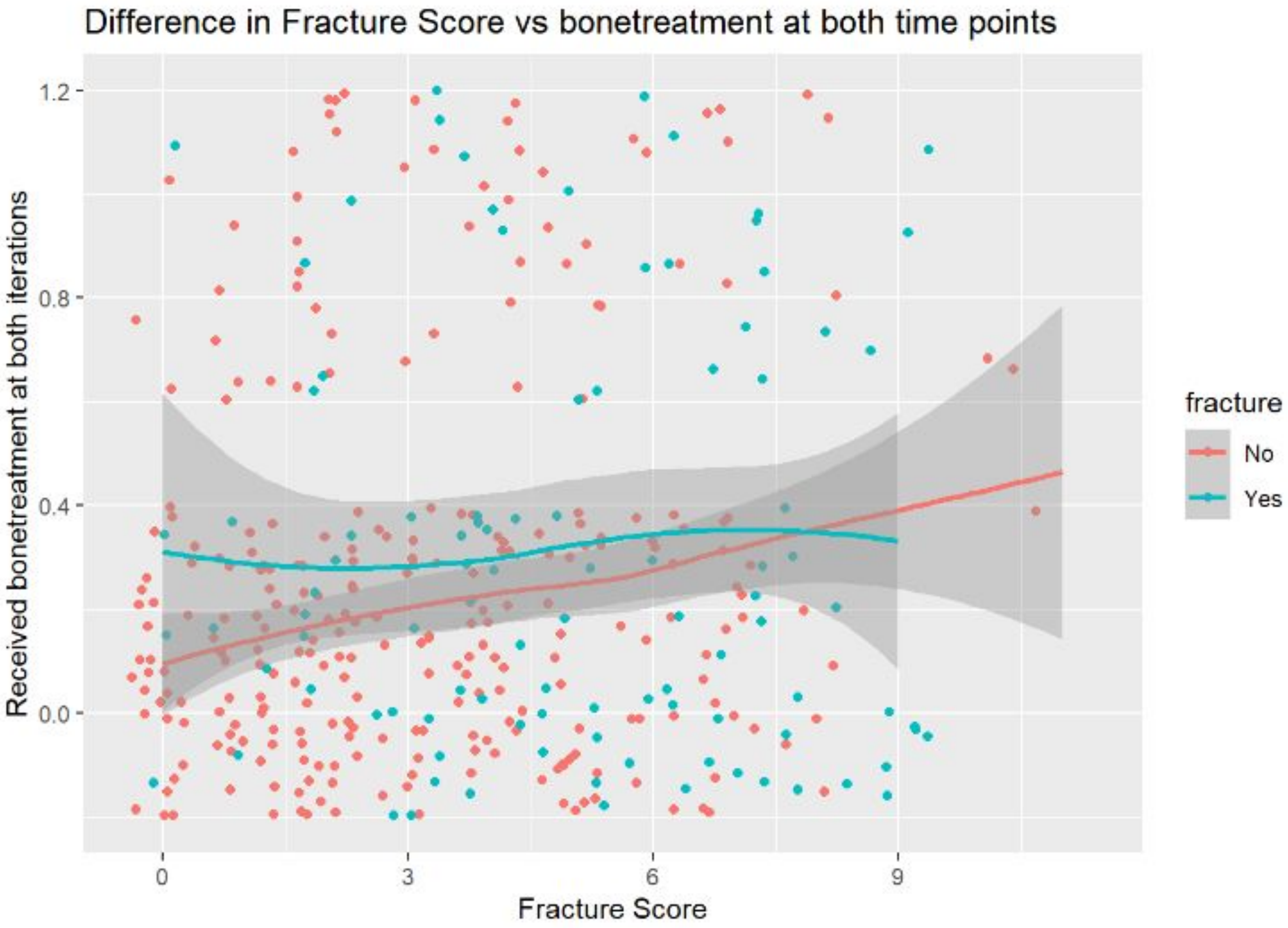
Performance for validation set



As fracture score increase so does the proportion of getting a fracture

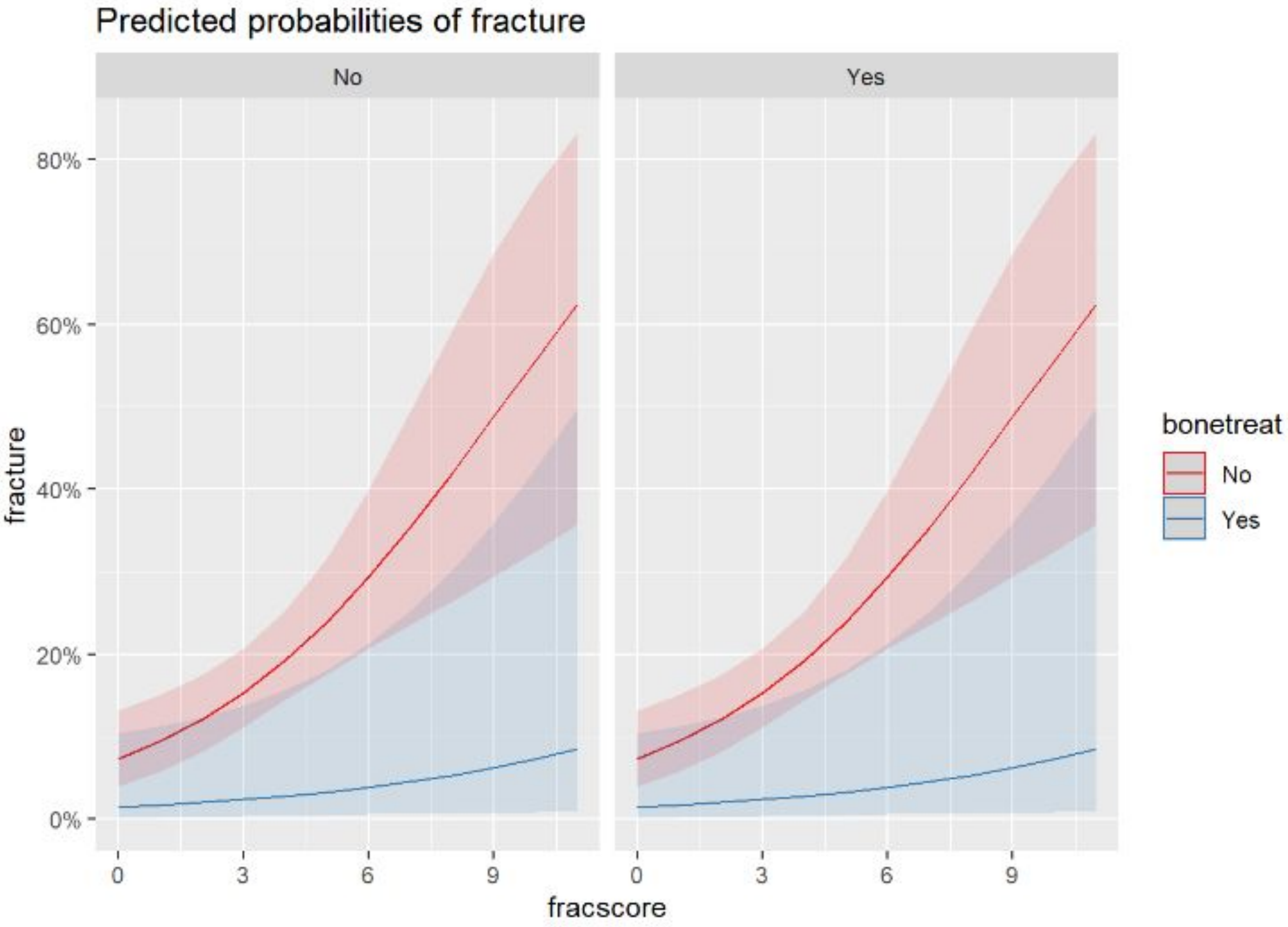


EDA

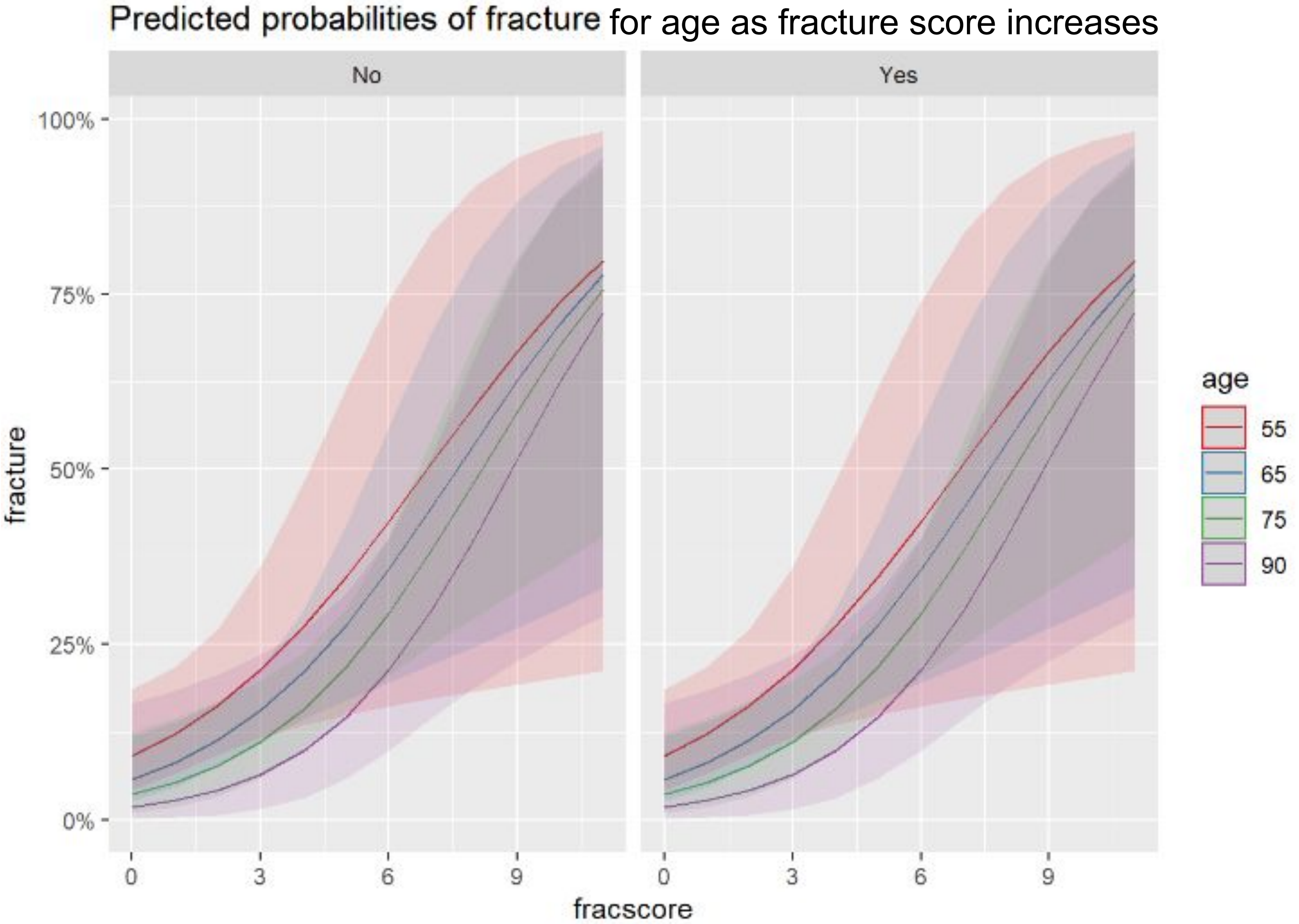


Difference in Fracture Score vs bonetreatment at both time points
Received bonetreatment at both iterations
1.2
0.8
0.4
0.0
0
3
6
9
Fracture Score
fracture
No
Yes

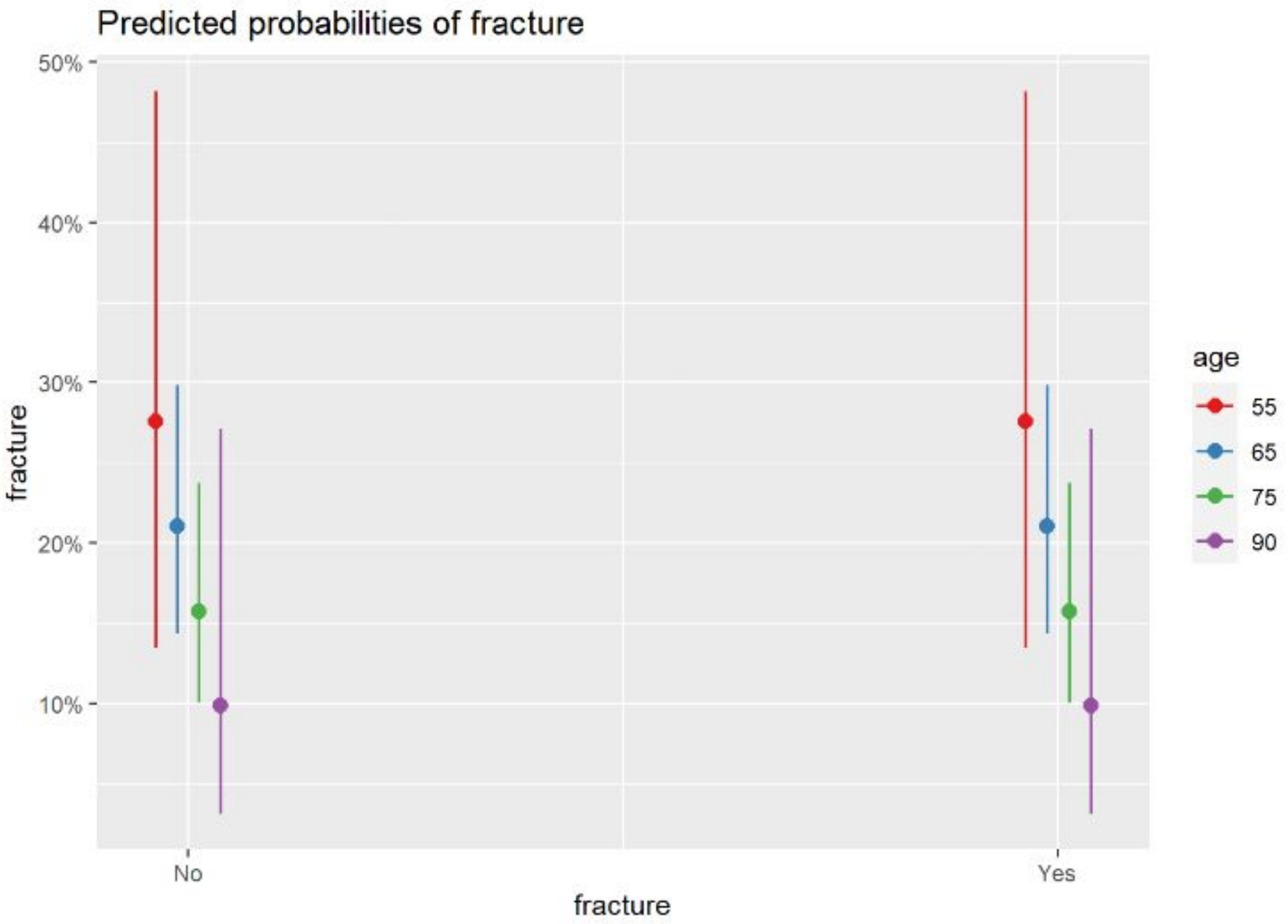
Effects plot of complex model



Effects plot of complex model



Effects plot of complex model



LDA Model

Began by continuing from EDA looking at possible interactions and important variables, unfortunately there did not appear to be very good separation between any variables, likely due to the low prevalence rate of getting a fracture, measuring the wrong data, or not doing the study long enough.

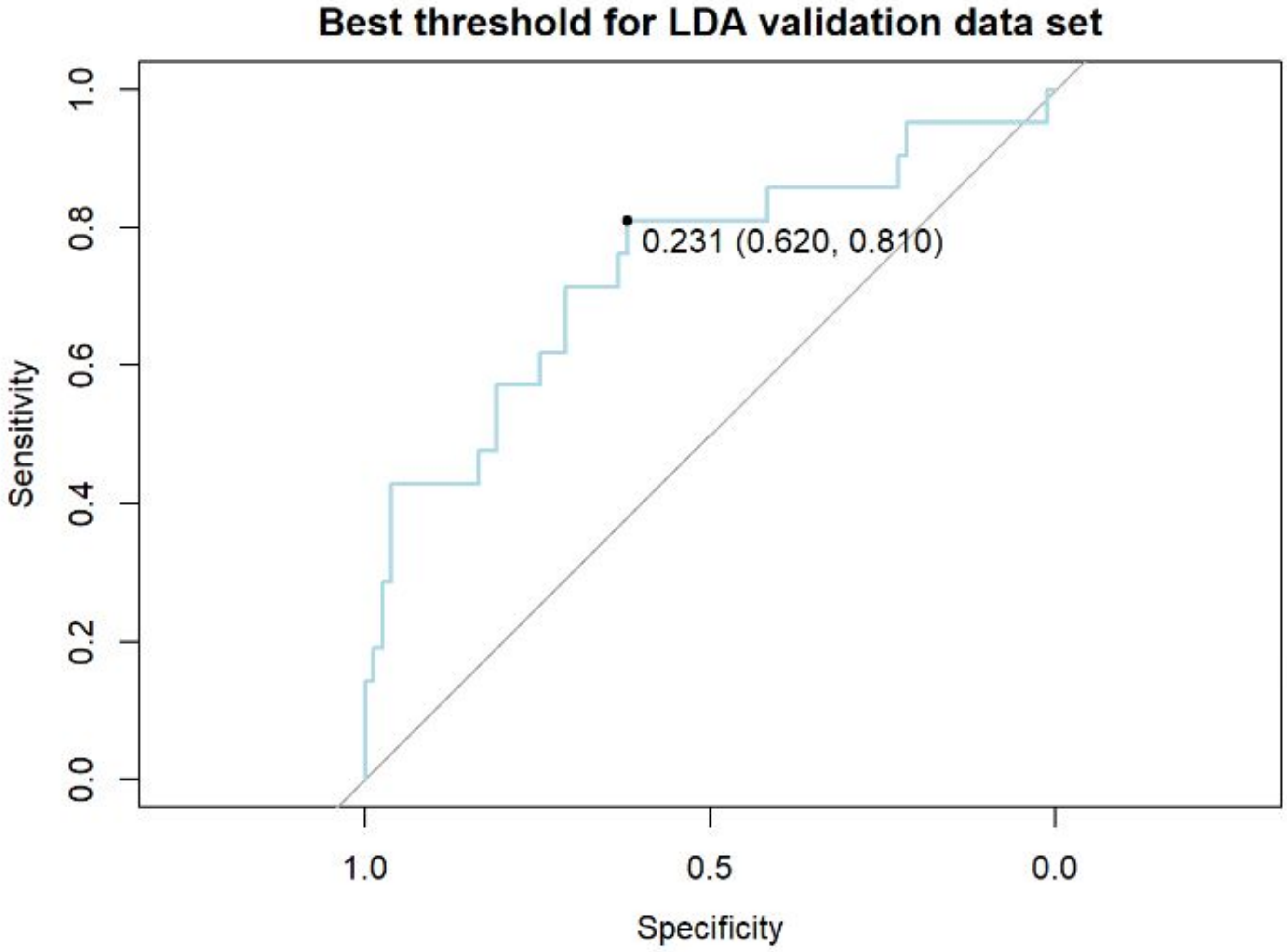
Relevant EDAs and effects plots to follow with effects plots for this model for some insight

Model is

fracture = (all variables and) - sub_id - phy_id - site_id + priorfrac:fracscore + age:fracscore + fracscore:bonetreat

(phy_id) - physician ID was turned into factor for this model

ROC for LDA model



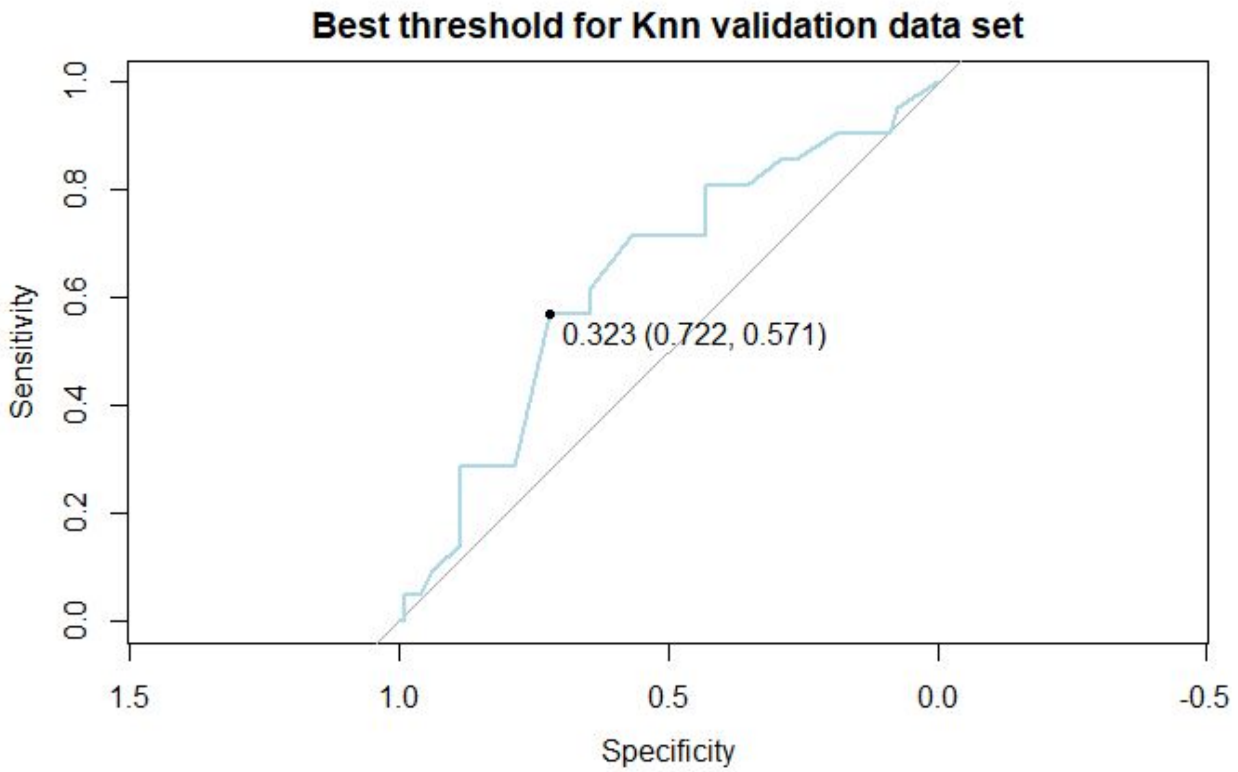
Knn Model

All continuous variables used:

- age
- weight
- height
- bmi
- Fracscore (0-12 categorical)

All Knn models never performs better than always guessing no fracture (75% accuracy) with the best model only using age.

```
k-Nearest Neighbors

400 samples
  5 predictor
  2 classes: 'No', 'Yes'

No pre-processing
Resampling: Cross-Validated (10 fold, repeated 1 times)
Summary of sample sizes: 359, 361, 360, 360, 359, 361, ...
Resampling results across tuning parameters:

  k   Accuracy   Kappa
   1  0.6793168   0.15146923
   2  0.6643074   0.09772090
   3  0.6869418   0.07266771
   4  0.6624969   0.01579128
   5  0.6802533  -0.01485883
   6  0.6999484   0.02039872
   7  0.7047624  -0.01348597
   8  0.7000094  -0.02251761
   9  0.7149515  -0.01036712
  10  0.7298358   0.07106741
  20  0.7350829   0.02719129
  30  0.7400891   0.00000000

Accuracy was used to select the optimal model using the largest value.
The final value used for the model was k = 30.
```



Best threshold for Knn validation data set
Sensitivity
Specificity
0.323 (0.722, 0.571)
0.0
0.2
0.4
0.6
0.8
1.0
1.5
1.0
0.5
0.0
-0.5

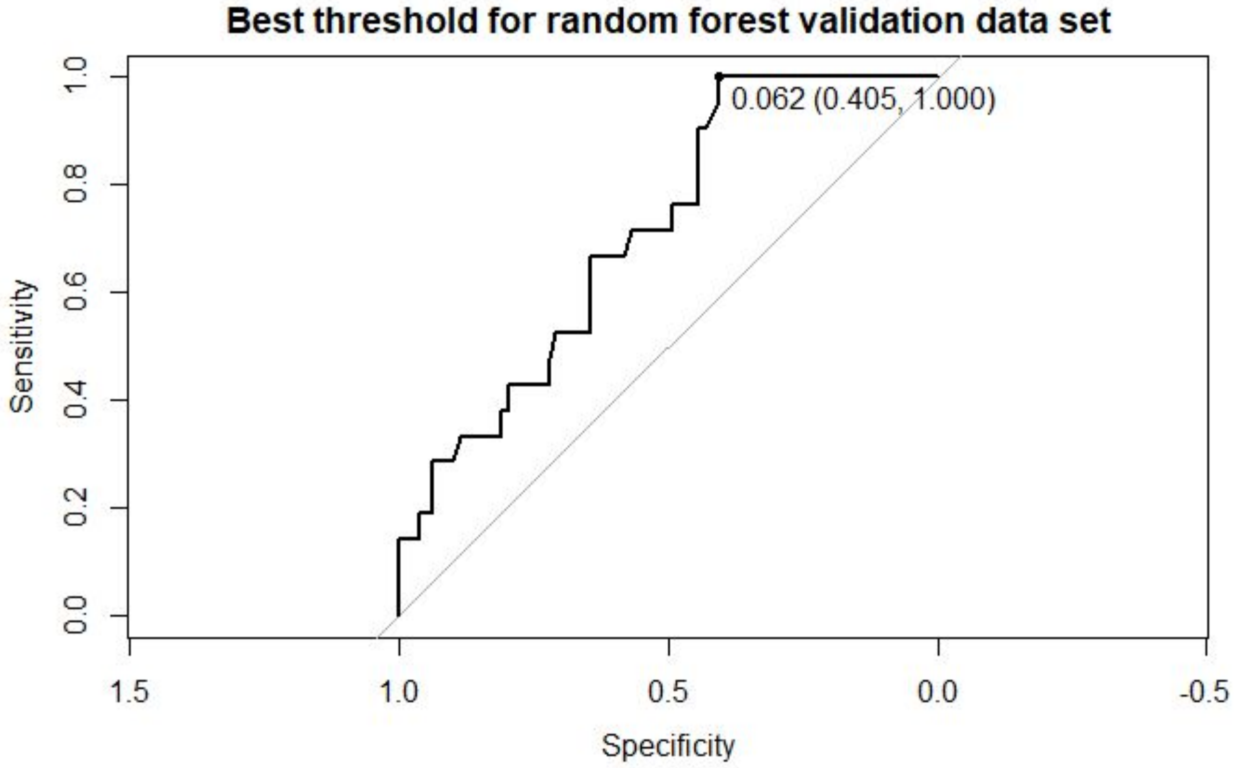
Random Forest Model

Variables used:

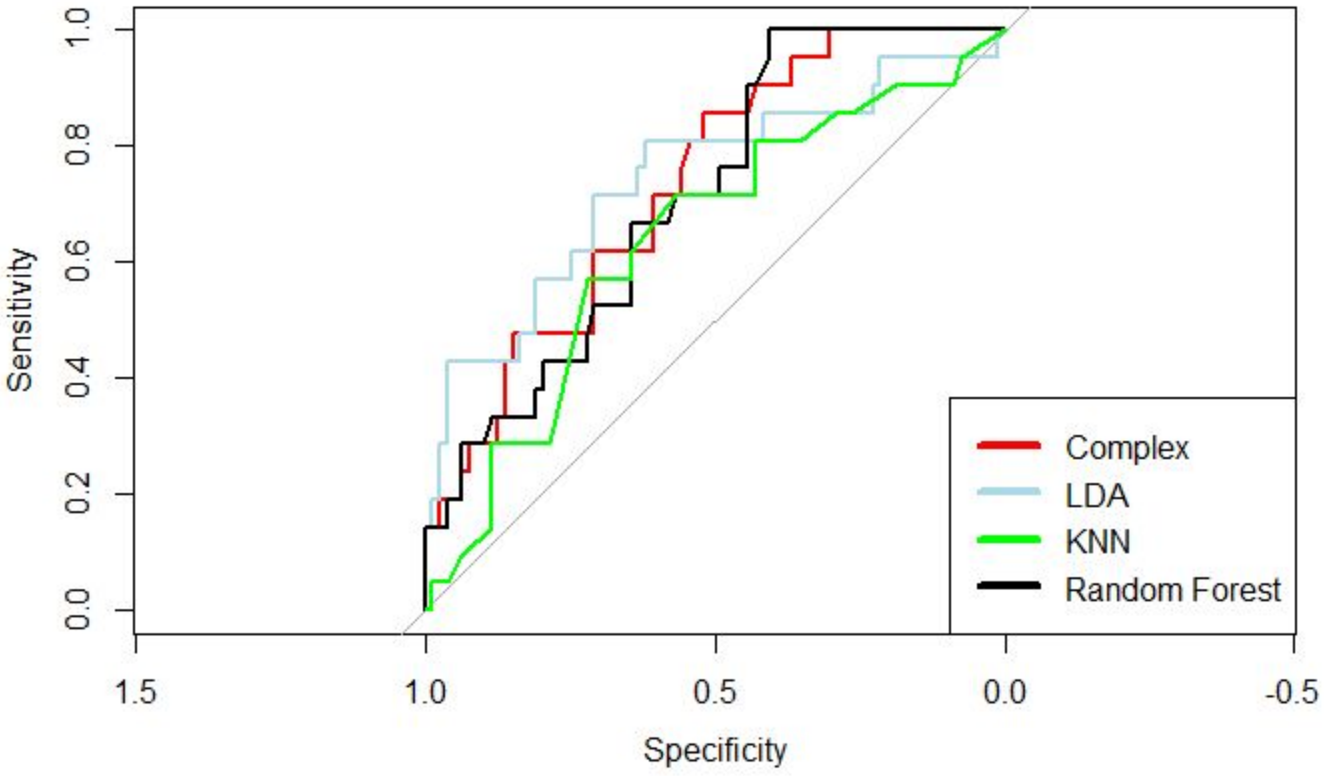
- priorfrac
- age
- premeno
- momfrac
- armassist
- smoke
- fracscore
- bonemed
- bonemed_fu

```
Confusion Matrix and Statistics

          Reference
Prediction No Yes
       No  75  16
       Yes  4   5

               Accuracy : 0.8
                 95% CI : (0.7082, 0.8733)
    No Information Rate : 0.79
    P-Value [Acc > NIR] : 0.46055

                  Kappa : 0.2372

 Mcnemar's Test P-Value : 0.01391

            Sensitivity : 0.9494
            Specificity : 0.2381
         Pos Pred Value : 0.8242
         Neg Pred Value : 0.5556
             Prevalence : 0.7900
         Detection Rate : 0.7500
   Detection Prevalence : 0.9100
      Balanced Accuracy : 0.5937

       'Positive' Class : No
```

Random Forest Model



ROC comparing all objective 2 models



Objective 2 model’s performances

- Complex model had and AUC-ROC of 0.7312
- LDA model had and AUC-ROC of AUC = 0.7414
- Random Forest model had and AUC-ROC of AUC = 0.711
- KNN model had and AUC-ROC of 0.6389
- Best model performance based off of AUC-ROC was the LDA model

Conclusion

- The probability of a fracture for a woman with osteoporosis within the first year of joining the study can be modeled using multiple logistic regression and can be sufficiently explained using factors such as age, history of fractures, and bone medication status at the time of enrollment.

- LDA model outperformed Random Forest, KNN, and complex models
    - Performance of models likely suffered due to imbalance in data set (almost 75% had no fracture during the study)

Scope of inference

Attempts were made to find if the study was observational or experimental in nature, because this would affect the scope of inference. Assumptions will be made that this was an observational study, since there is no mention of random sampling or random assignment and there are the almost the same amount of people at each treatment iteration, i.e. initial, follow-up, and taking treatment at both times; therefore it is not safe to generalize to another population beyond this study, nor is it appropriate to make causality claims. If this is a random sample with random assignment it is appropriate to make generalizable claims to similar groups and make causal claims.

Future Recommendations

Given more time, possibly consider further exploration into EDA to see if there are other interactions or polynomials terms that were missed.

Unfortunately there did not appear to be very good separation between any variables, likely due to the low prevalence rate of getting a fracture, measuring the wrong data, or not doing the study long enough.

Research effects plots when using caret package

Measure different variables

Utilize a PCA model for possibly better predictions - see appendix

References

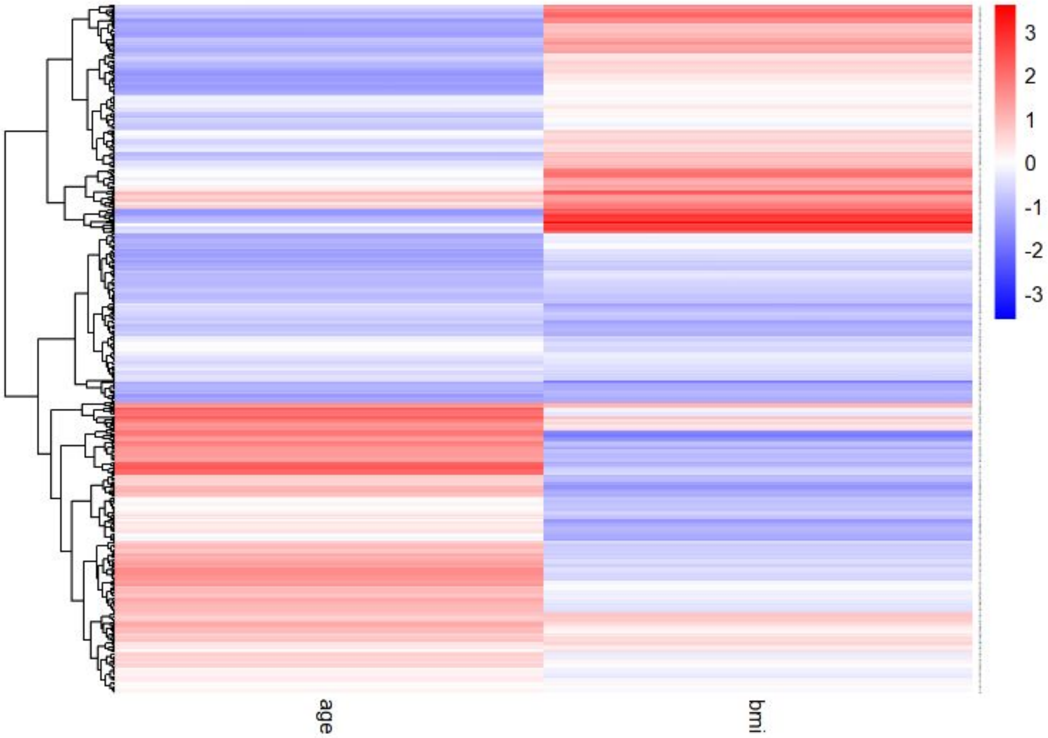
Hosmer, D.W., Lemeshow, S. and Sturdivant, R.X. (2013) Applied Logistic Regression, 3rd ed., New York: Wiley

https://cran.r-project.org/web/packages/aplore3/aplore3.pdf#page=11&zoom=100,132,90

Appendix

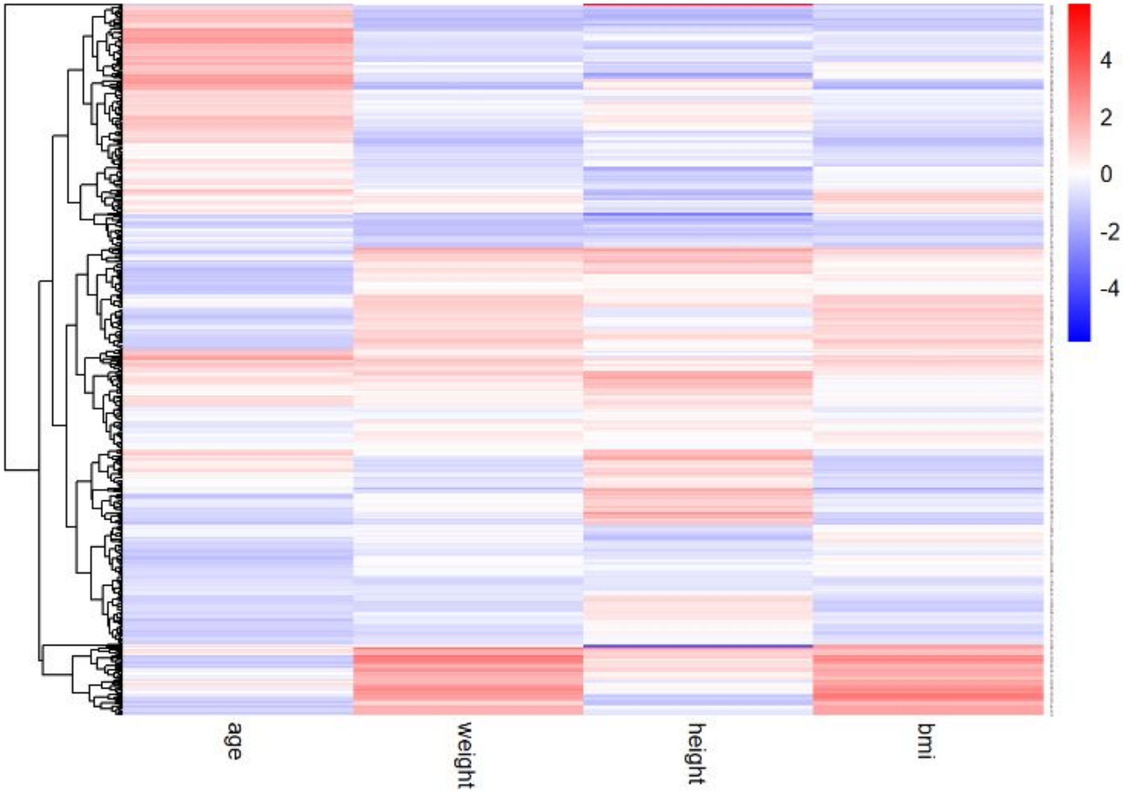
Clustering EDA

```
pheatmap(glow_bonemed[, c(5,8)], scale = "column", fontsize_row = 0.1, cluster_cols = F, legend = T, color = colo
rRampPalette(c("blue", "white", "red"), space = "rgb")(100))
```

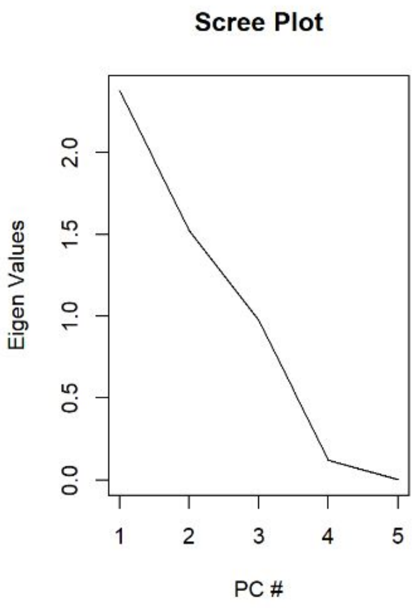


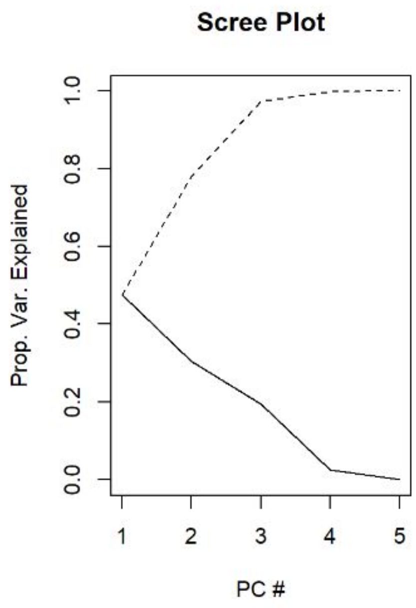
Appendix

```
pheatmap(glow_bonemed[, 5:8], scale = "column", fontsize_row = 0.1, cluster_cols = F, legend = T, color = colorRa
mpPalette(c("blue", "white", "red"), space = "rgb")(100))
```



Appendix



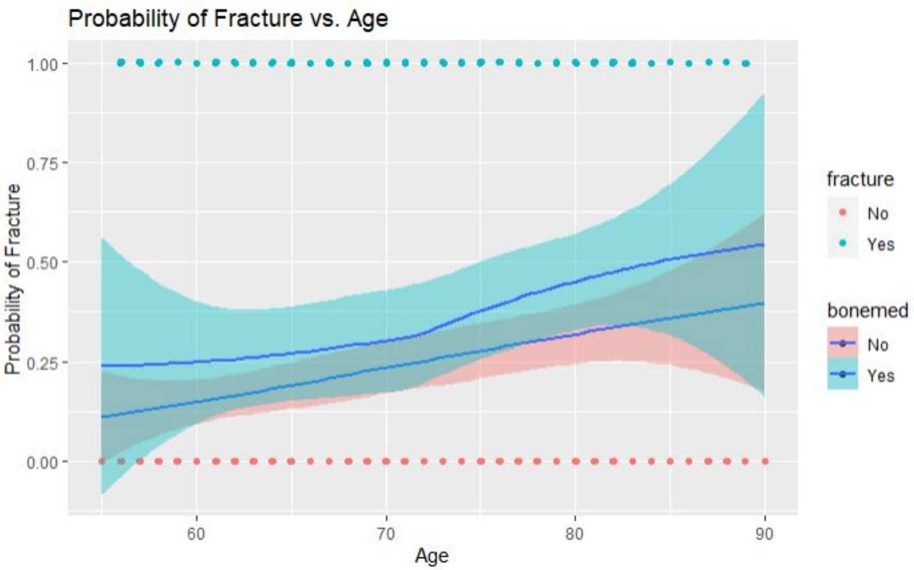


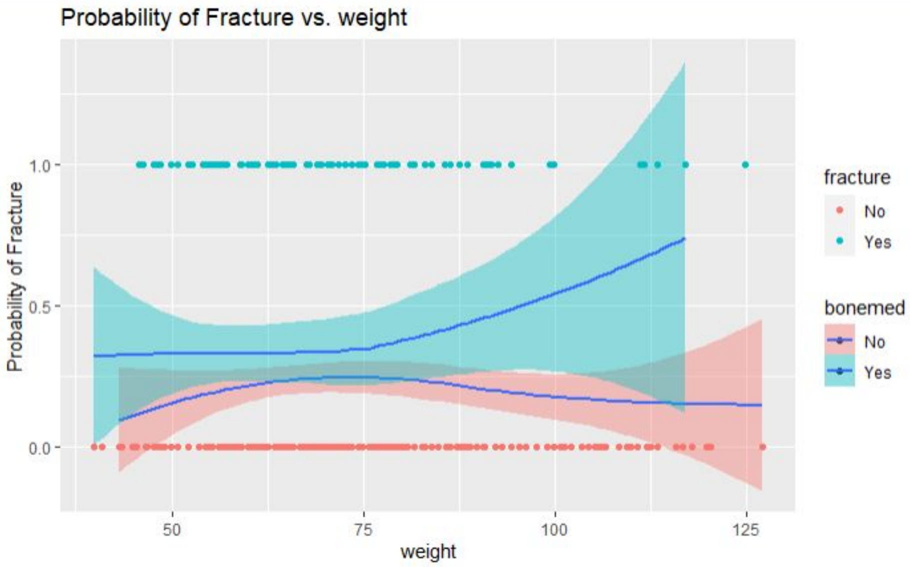

```r
corr_vars <- c("age", "weight", "height", "bmi", "fracscore")
pc.result<-prcomp(glow_bonemed[, corr_vars],scale.=TRUE)
#Eigen Vectors
pc.result$rotation
```

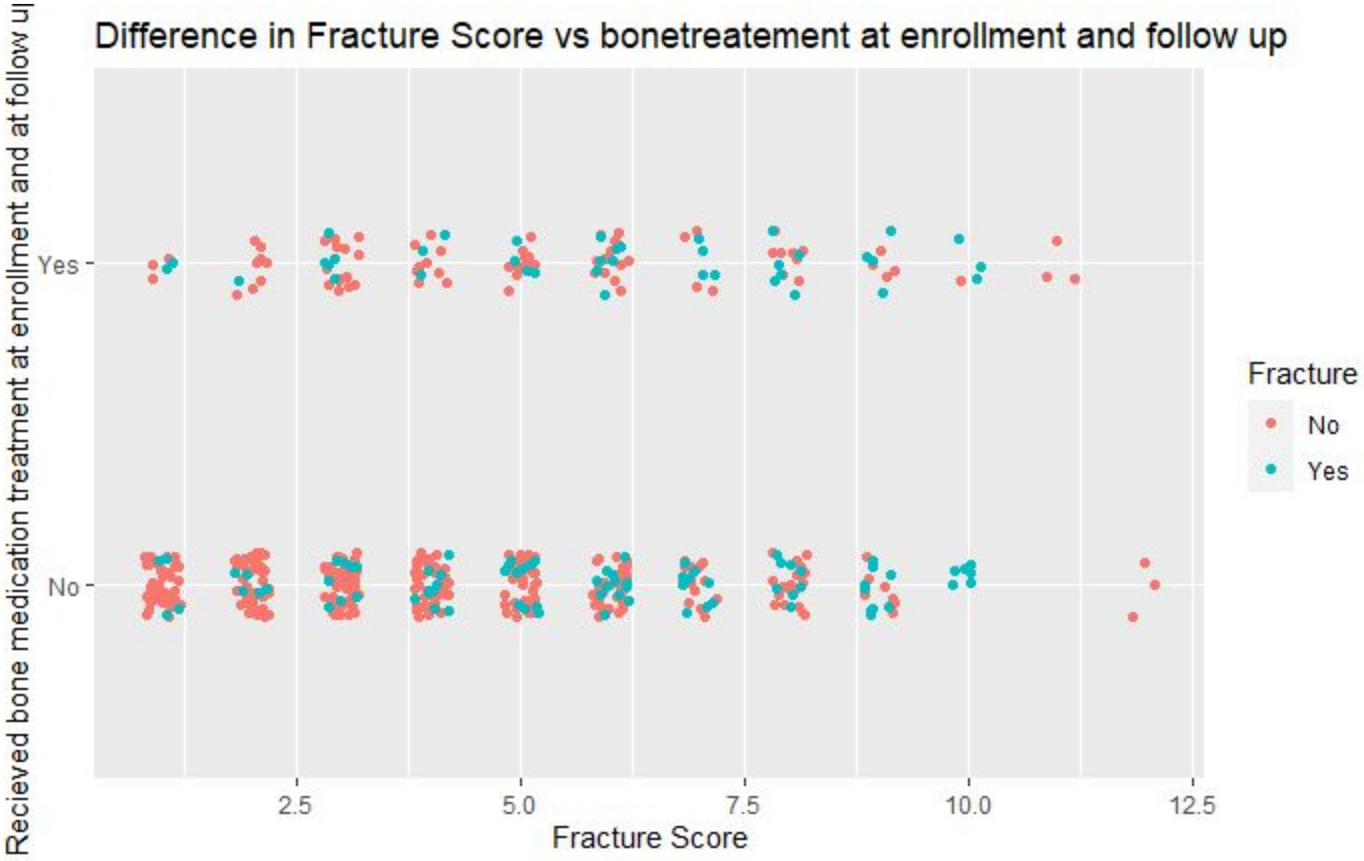
```
##                 PC1         PC2         PC3         PC4          PC5
## age       0.4947219  0.46742140 -0.15246583  0.71654567 -0.009160237
## weight   -0.5273035  0.46578775 -0.08840991  0.03240244 -0.704362523
## height   -0.2345770 -0.08196149 -0.93823245  0.01885633  0.240042129
## bmi      -0.4741030  0.51615173  0.24533677  0.05137380  0.667820563
## fracscore 0.4442985  0.53984137 -0.16872342 -0.69463484  0.013601399
```

EDA

- Relationships – Loess Plots Variables

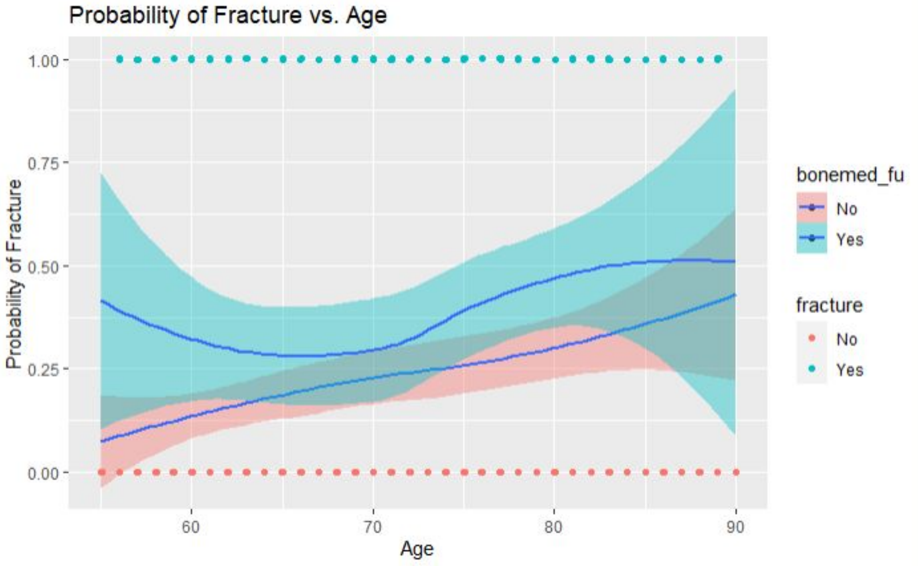


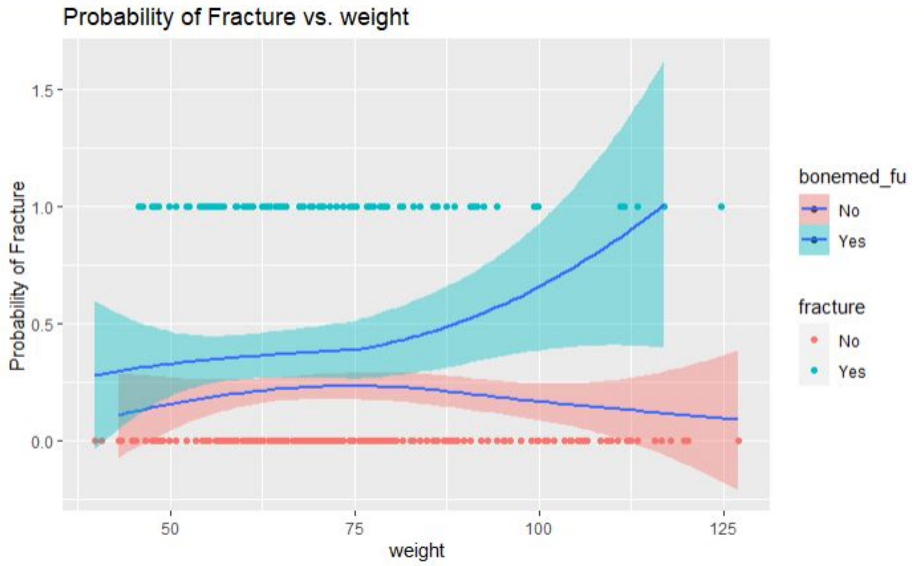



Difference in Fracture Score vs bonetreatement at enrollment and follow up
Recieved bone medication treatment at enrollment and at follow u
Yes
No
2.5
5.0
7.5
10.0
12.5
Fracture Score
Fracture
No
Yes

EDA

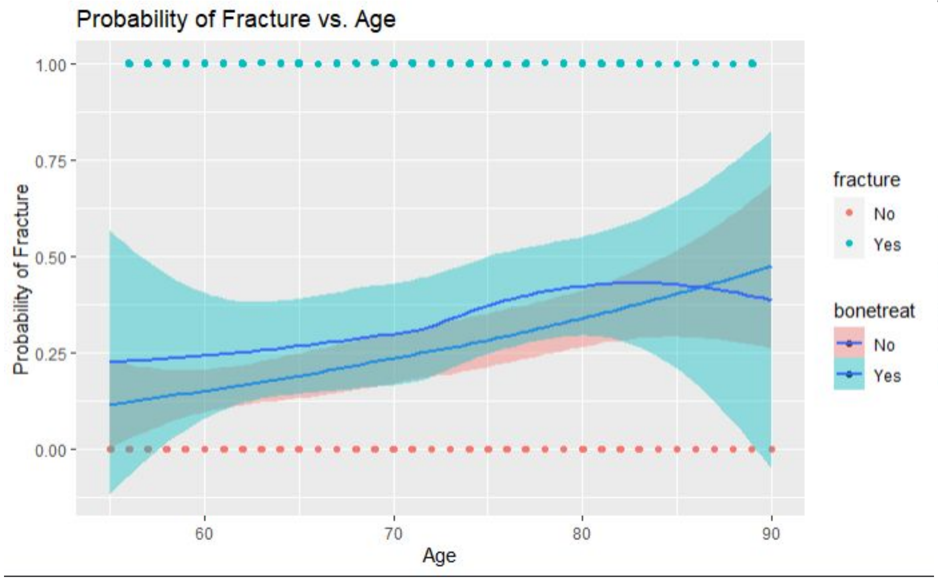
- Relationships – Loess Plots Variables

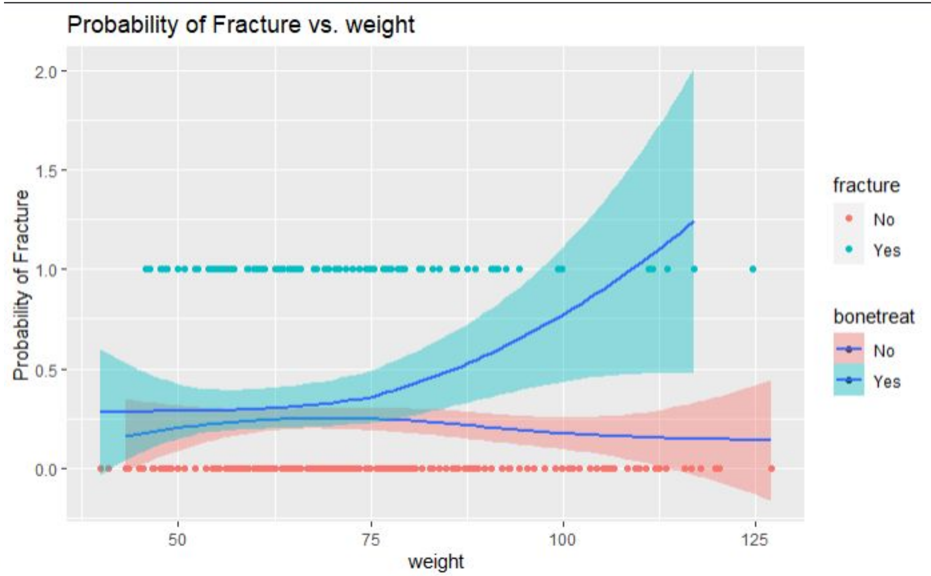




EDA

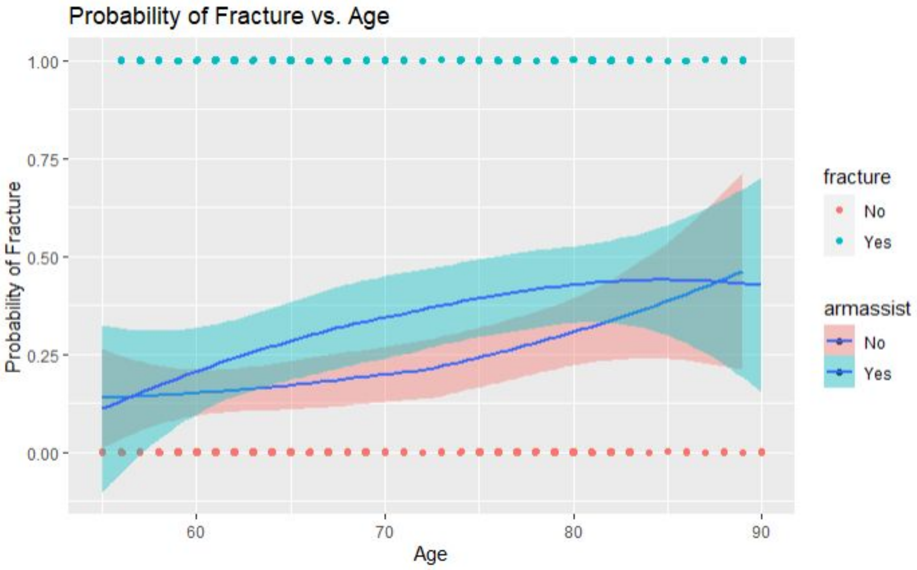
- Relationships – Loess Plots Variables

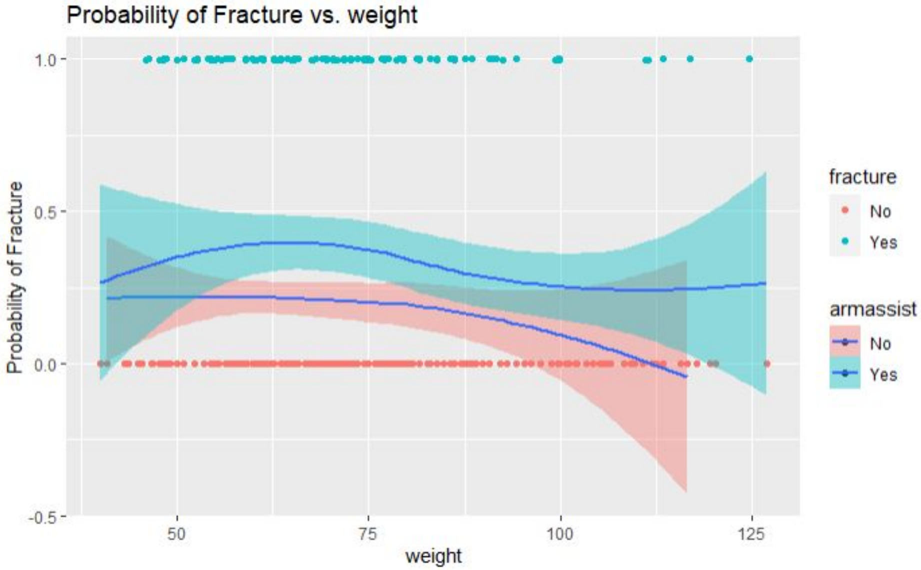




EDA

- Relationships – Loess Plots Variables





DataScience@SMU