Analysis of Osteoperosis and Bone Fractures

MSDS 6372: Anish Bhandari, Rob Lane, & Alex Thibeaux

2023-08-13

# Introduction

We’ve picked the “glow\_bonemed” dataset from the R package “aplore3,” which contains information on 500 cases in 18 columns. Among these, 3 columns help identify cases, 14 offer explanatory information, and 1 indicates whether a fracture occurred. Our main goal is to predict whether women with osteoporosis will have any bone fractures in the first year after joining the study.

We’ll begin by conducting Exploratory Data Analysis (EDA) and constructing a logistic regression model primarily for interpretation.Next, we’ll create 3 advanced models to predict fractures, and finally, we’ll compare these models and provide a brief summary of our findings.

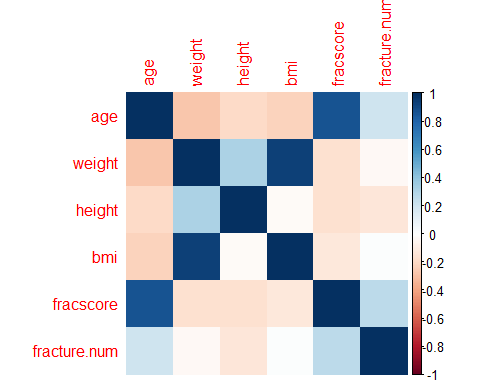
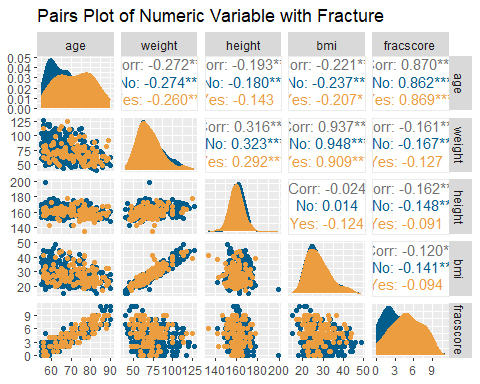
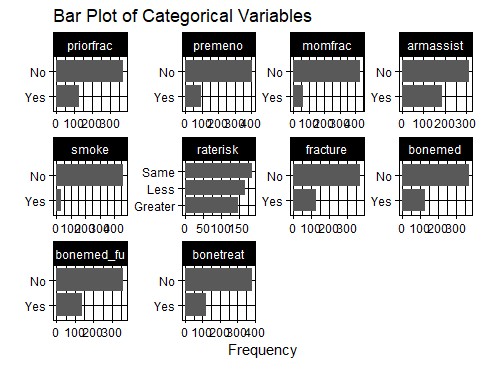
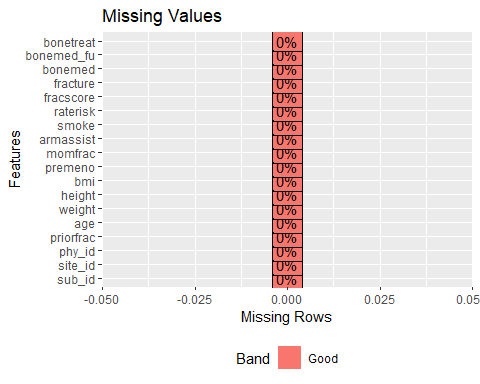
### glow\_bonemed variables:

| **Variable** | **Type** | **Description** |
| --- | --- | --- |
| age | int | Age at Enrollment (Years) |
| armassist | 2-factor | Arms are needed to stand from a chair (1: No, 2: Yes) |
| bmi | num | Body Mass Index () |
| bonemed | 2-factor | Bone medications at enrollment (1: No, 2: Yes) |
| bonemed\_fu | 2-factor | Bone medications at follow-up (1: No, 2: Yes) |
| bonetreat | 2-factor | Bone medications both at enrollment and follow-up (1: No, 2: Yes) |
| fracscore | int | Fracture Risk Score (Composite Risk Score) |
| **fracture** | **2-factor** | **Any fracture in first year (1: No, 2: Yes)** |
| height | int | Height at enrollment (Centimeters) |
| momfrac | 2-factor | Mother had hip fracture (1: No, 2: Yes) |
| phy\_id | int | Physician ID code (128 unique codes) |
| premeno | 2-factor | Menopause before age 45 (1: No, 2: Yes) |
| priorfrac | 2-factor | History of Prior Fracture (1: No, 2: Yes) |
| raterisk | 3-factor | Self-reported risk of fracture (1: Less than others of the same age, 2: Same as others of the same age, 3: Greater than others of the same age) |
| site\_id | int | Study Site (1 - 6) |
| smoke | 2-factor | Former or current smoker (1: No, 2: Yes) |
| sub\_id | int | Identification Code (1 - n) |
| weight | num | Weight at enrollment (Kilograms) |

# Ojective 1

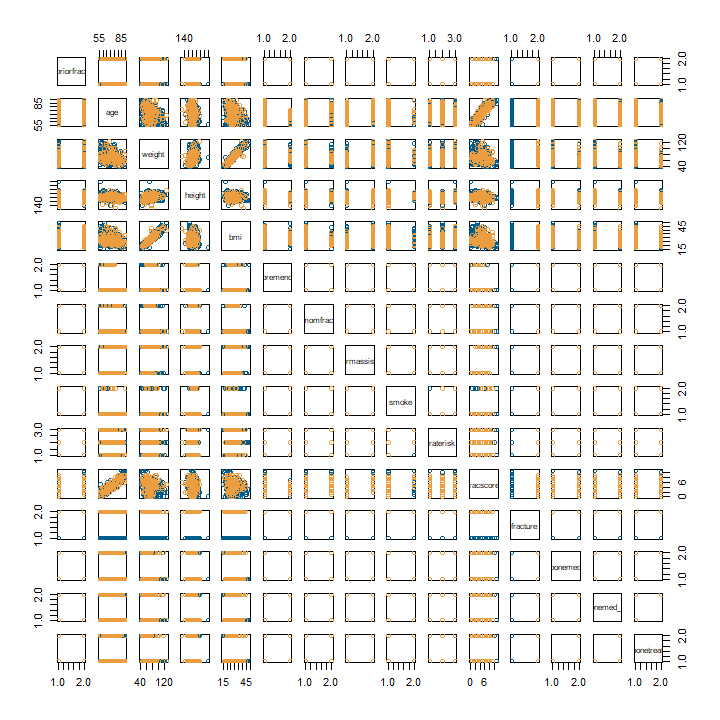
## 1A: EDA

## sub\_id site\_id phy\_id priorfrac age   
## Min. : 1.0 Min. :1.000 Min. : 1.00 No :374 Min. :55.00   
## 1st Qu.:125.8 1st Qu.:2.000 1st Qu.: 57.75 Yes:126 1st Qu.:61.00   
## Median :250.5 Median :3.000 Median :182.50 Median :67.00   
## Mean :250.5 Mean :3.436 Mean :178.55 Mean :68.56   
## 3rd Qu.:375.2 3rd Qu.:5.000 3rd Qu.:298.00 3rd Qu.:76.00   
## Max. :500.0 Max. :6.000 Max. :325.00 Max. :90.00   
   
## weight height bmi premeno momfrac armassist  
## Min. : 39.90 Min. :134.0 Min. :14.88 No :403 No :435 No :312   
## 1st Qu.: 59.90 1st Qu.:157.0 1st Qu.:23.27 Yes: 97 Yes: 65 Yes:188   
## Median : 68.00 Median :161.5 Median :26.42   
## Mean : 71.82 Mean :161.4 Mean :27.55   
## 3rd Qu.: 81.30 3rd Qu.:165.0 3rd Qu.:30.79   
## Max. :127.00 Max. :199.0 Max. :49.08   
   
## smoke raterisk fracscore fracture bonemed bonemed\_fu  
## No :465 Less :167 Min. : 0.000 No :375 No :371 No :361   
## Yes: 35 Same :186 1st Qu.: 2.000 Yes:125 Yes:129 Yes:139   
## Greater:147 Median : 3.000   
## Mean : 3.698   
## 3rd Qu.: 5.000   
## Max. :11.000   
## bonetreat  
## No :382   
## Yes:118   
##   
##   
##   
##



## age weight height bmi fracscore  
## age 1.0000000 -0.27159637 -0.19264861 -0.22125651 0.8699165  
## weight -0.2715964 1.00000000 0.31596915 0.93733603 -0.1613754  
## height -0.1926486 0.31596915 1.00000000 -0.02437689 -0.1619952  
## bmi -0.2212565 0.93733603 -0.02437689 1.00000000 -0.1203472  
## fracscore 0.8699165 -0.16137536 -0.16199525 -0.12034723 1.0000000  
## fracture.num 0.2076535 -0.03625944 -0.13640055 0.01498506 0.2644795  
  
## fracture.num  
## age 0.20765352  
## weight -0.03625944  
## height -0.13640055  
## bmi 0.01498506  
## fracscore 0.26447951  
## fracture.num 1.00000000

We observe mild correlations between height and weight, fracture and age, and fracture and fracscore. Additionally, we identify a high correlation between weight and BMI, as well as a very high correlation between fracscore and age.

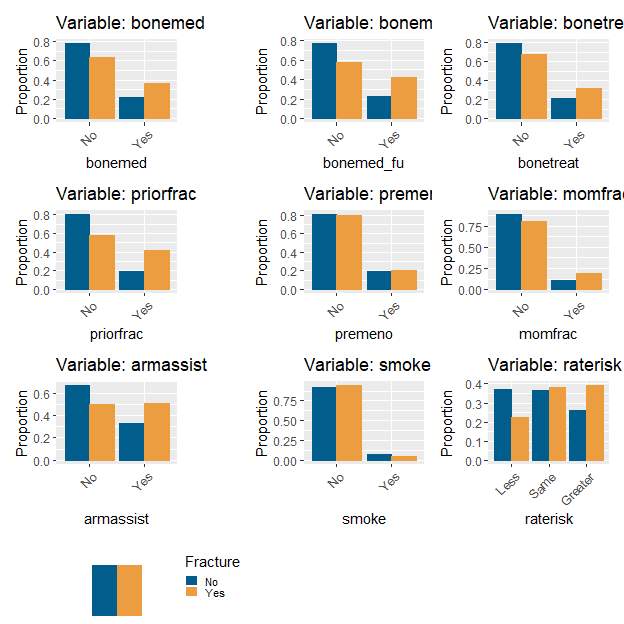


There are a few cases of multicollinearity between:  
*fracscore : age  
bmi : weight  
site\_id : phy\_id*

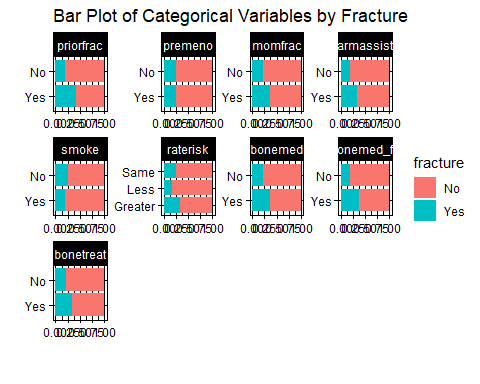
One thing of note with the above scatterplot is that there seems to be a low level of separation on the output variable (fracture).

## p-values for selected variables (t-test):  
## site\_id : 0.1152253  
## phy\_id : 0.1249636  
## age : 6.844017e-06  
## weight : 0.4033704  
## height : 0.004122593  
## bmi : 0.7348124  
## fracscore: 8.213732e-09

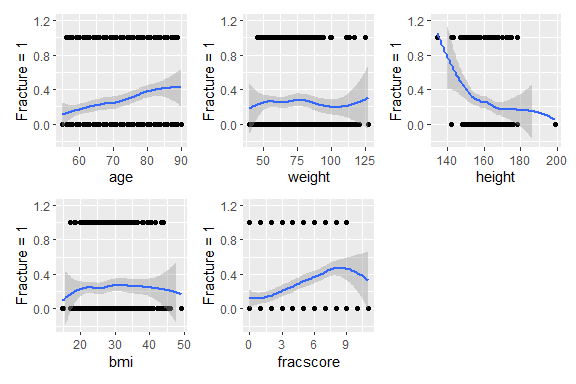
We believe that the ID variables don’t have a meaningful affect on the output variable. Of the above tested variables, only age, height, and fracscore seem to have an impact on the output variable (without checking for the effect of other variables simultaneously).



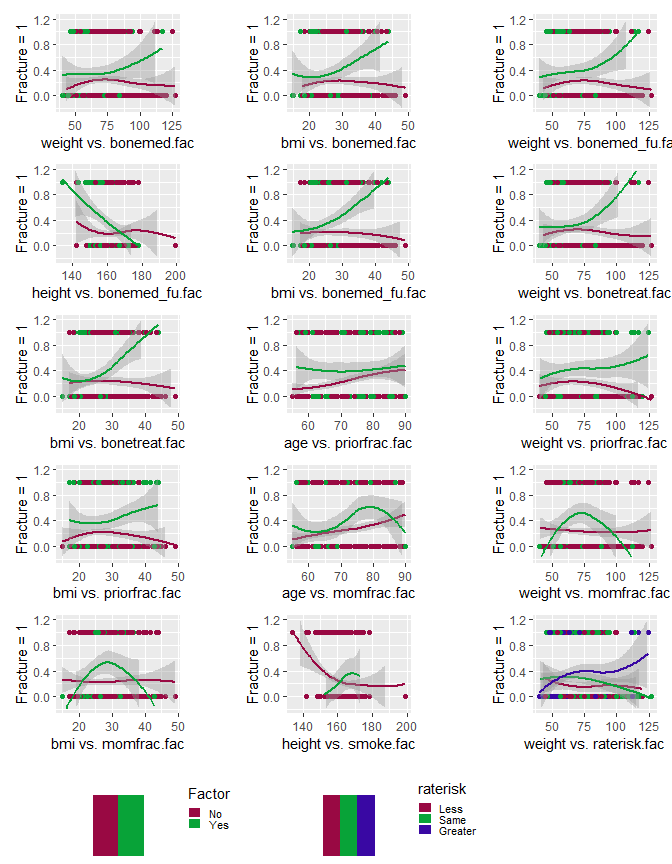
Above are all of our selected categorical variables. One thing we can see is that there are relatively few examples in our sample who smoke, are pre-menopause, or have a mother who had a fracture. As shown elsewhere, most of our variables have rather weak affects alone on the outcome of having a fracture.



One thing we can see in the above chart is that the proportional differences of fracture status are not very different for some of the categorical variables. The largest differences we can see here are found within *priorfrac*, *bonemed\_fu*, *raterisk*, and *bonemed*.



There seems to be a slight positive correlation between age and likelihood of fracture, and between fracscore and likelihood of fracture. Both of these follow reasonable expectations when considering expected trends in osteoperosis cases. The strongest trend shown here, though, is a negative trend with height. All else equal, it would be reasonable to expect more breakages in taller patients because they can fall a slightly larger distance.



This a group of loess curve plots which are set up to explore interactions between selected pairs of factor variables (noted on the right of each x-axis with a suffix of “.fac”) and a numeric variable (noted on the left of each x-axis). The reason these were selected is because these appear to show a difference in the presented trend (frequency of fractures) for the numeric variable when they are separated by a factor.

For the top row, middle curve, when the factor (bonemed) is yes (green), there is a strong positive response as the numeric variable (bmi) increases. There does not seem to be a strong response when bonemed is no (mauve). For this reason, we believe that this variable exhibits enough evidence of interaction to explore its effect in a complex model. Several of the interactions with momfrac seem to show a further posibility of a polynomial interaction. These will be explored later in this paper.

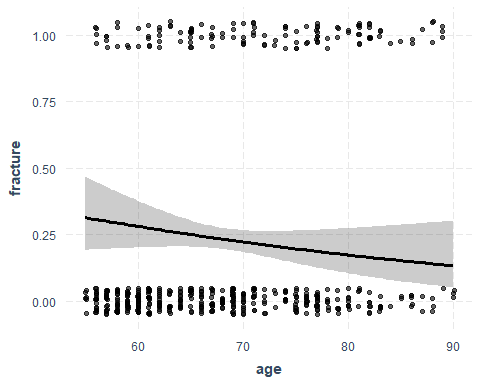
## Multiple Logistic Regression

##   
## Call:  
## glm(formula = fracture ~ age + height + fracscore, family = "binomial",   
## data = bone)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.52891 -0.76170 -0.58449 0.04907 2.17175   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 6.27383 3.30233 1.900 0.05746 .   
## age -0.03108 0.02356 -1.319 0.18715   
## height -0.04073 0.01762 -2.312 0.02079 \*   
## fracscore 0.33097 0.08651 3.826 0.00013 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 562.34 on 499 degrees of freedom  
## Residual deviance: 521.06 on 496 degrees of freedom  
## AIC: 529.06  
##   
## Number of Fisher Scoring iterations: 4

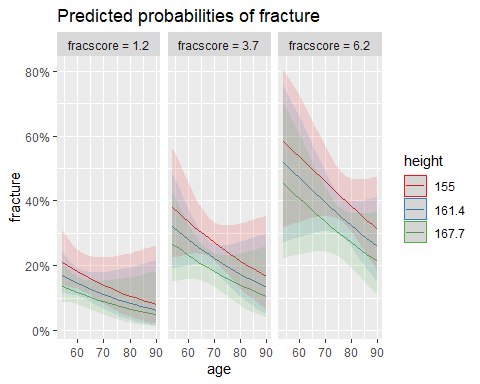
## (Intercept) age height fracscore   
## 530.5043049 0.9693982 0.9600884 1.3923147

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 0.8628055 3.728759e+05  
## age 0.9254176 1.015182e+00  
## height 0.9269801 9.934206e-01  
## fracscore 1.1767093 1.652935e+00



## Data were 'prettified'. Consider using `terms="age [all]"` to get smooth  
## plots.

 Feature Selection

##   
## Step: AIC=511.78  
## .outcome ~ weight + bmi + priorfracYes + momfracYes + fracscore +   
## bonemedYes + bonemed\_fuYes + bonetreatYes  
##   
## Df Deviance AIC  
## <none> 493.78 511.78  
## - priorfracYes 1 497.19 513.19  
## - momfracYes 1 497.62 513.62  
## - bonemedYes 1 498.69 514.69  
## - weight 1 499.07 515.07  
## - bmi 1 501.37 517.37  
## - fracscore 1 503.30 519.30  
## - bonetreatYes 1 503.55 519.55  
## - bonemed\_fuYes 1 506.73 522.73

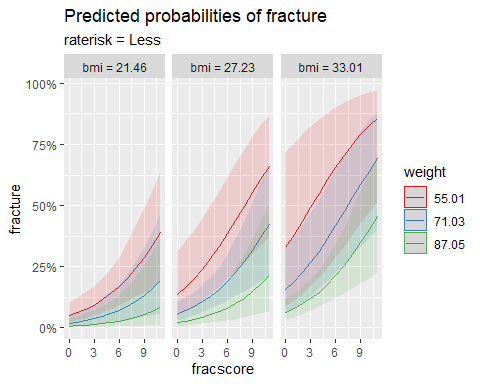
## (Intercept) weight bmi priorfracYes momfracYes   
## -3.10007420 -0.04506251 0.14600942 0.49789690 0.62696661   
## fracscore bonemedYes bonemed\_fuYes bonetreatYes   
## 0.16261249 1.48233049 1.73619701 -2.62466026

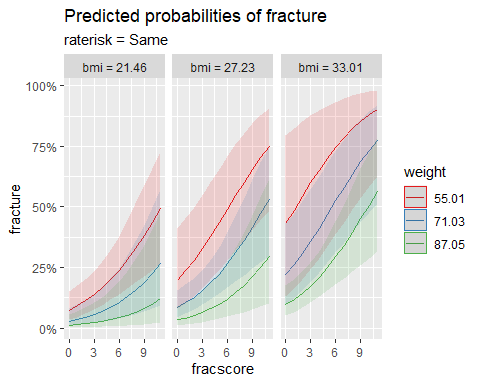
Best AIC Model:

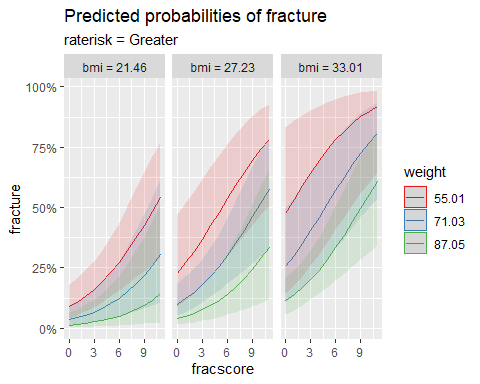
##   
## Call:  
## glm(formula = fracture ~ weight + bmi + momfrac + raterisk +   
## fracscore + bonemed + bonemed\_fu + bonetreat, family = "binomial",   
## data = dftrain)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.6495 -0.6861 -0.4959 -0.2625 2.3464   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.75181 0.83746 -4.480 7.46e-06 \*\*\*  
## weight -0.06121 0.02473 -2.475 0.01332 \*   
## bmi 0.19418 0.06712 2.893 0.00381 \*\*   
## momfracYes 0.56694 0.38057 1.490 0.13630   
## rateriskSame 0.42984 0.35547 1.209 0.22658   
## rateriskGreater 0.61693 0.37131 1.661 0.09661 .   
## fracscore 0.22705 0.05870 3.868 0.00011 \*\*\*  
## bonemedYes 1.36679 1.07140 1.276 0.20206   
## bonemed\_fuYes 1.78882 0.57051 3.135 0.00172 \*\*   
## bonetreatYes -2.69014 1.22394 -2.198 0.02795 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 388.05 on 349 degrees of freedom  
## Residual deviance: 324.73 on 340 degrees of freedom  
## AIC: 344.73  
##   
## Number of Fisher Scoring iterations: 4

## `terms` must have not more than four values. Using first four values  
## now.

## [[1]]







## (Intercept) weight bmi momfracYes rateriskSame   
## 0.02347529 0.94062296 1.21431459 1.76286102 1.53701426   
## rateriskGreater fracscore bonemedYes bonemed\_fuYes bonetreatYes   
## 1.85322317 1.25489859 3.92272583 5.98241457 0.06787168

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 0.004346640 0.1172649  
## weight 0.894692673 0.9861078  
## bmi 1.068195107 1.3909238  
## momfracYes 0.823595867 3.6859789  
## rateriskSame 0.769289890 3.1196027  
## rateriskGreater 0.897984220 3.8748340  
## fracscore 1.120438417 1.4113969  
## bonemedYes 0.414387448 36.3320263  
## bonemed\_fuYes 2.006596995 19.3896584  
## bonetreatYes 0.005580113 0.8208017

##   
## Call:  
## glm(formula = fracture ~ weight + bmi + fracscore + bonemed +   
## bonemed\_fu + bonetreat, family = "binomial", data = bone)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -1.55392 -0.73201 -0.55907 -0.01121 2.25762   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) -3.11590 0.62440 -4.990 6.03e-07 \*\*\*  
## weight -0.04111 0.01974 -2.082 0.037332 \*   
## bmi 0.13632 0.05321 2.562 0.010407 \*   
## fracscore 0.22061 0.04591 4.805 1.55e-06 \*\*\*  
## bonemedYes 1.39321 0.65377 2.131 0.033085 \*   
## bonemed\_fuYes 1.76752 0.48460 3.647 0.000265 \*\*\*  
## bonetreatYes -2.54735 0.82843 -3.075 0.002106 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 562.34 on 499 degrees of freedom  
## Residual deviance: 500.54 on 493 degrees of freedom  
## AIC: 514.54  
##   
## Number of Fisher Scoring iterations: 4

## (Intercept) weight bmi fracscore bonemedYes   
## 0.04433865 0.95972347 1.14604816 1.24683747 4.02775735   
## bonemed\_fuYes bonetreatYes   
## 5.85632214 0.07828915

## Waiting for profiling to be done...

## 2.5 % 97.5 %  
## (Intercept) 0.01277694 0.1485898  
## weight 0.92255625 0.9969995  
## bmi 1.03400547 1.2746025  
## fracscore 1.14078075 1.3662815  
## bonemedYes 1.10933300 15.2564381  
## bonemed\_fuYes 2.30010612 15.7522508  
## bonetreatYes 0.01481554 0.3944644

Probabilities and Predictions

## Setting direction: controls < cases

## Age, Height & Frac Score AUROC

## Area under the curve: 0.6887

## Setting direction: controls < cases

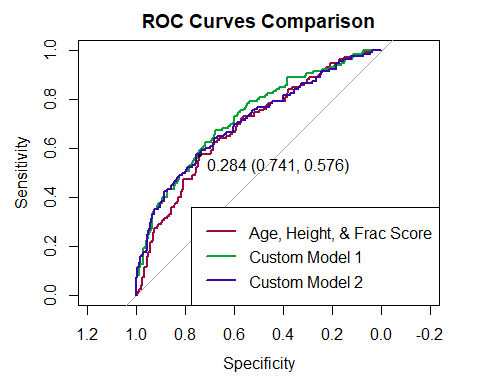
## Custom Model 2 AUROC

## Area under the curve: 0.7271

## Setting direction: controls < cases

## Custom Model 3 AUROC

## Area under the curve: 0.7087

 Confusion Matrices for Simple Logistic Regression

## Confusion Matrices for Simple Logistic Regression Models

## ### Age, Height & Frac Score ### ##### Custom Model 1 ###### ##### Custom Model 2 ######

## Confusion Matrix and Statistics  
##   
##   
## No Yes  
## No 278 53  
## Yes 97 72  
##   
## Accuracy : 0.7   
## 95% CI : (0.6577, 0.7399)  
## No Information Rate : 0.75   
## P-Value [Acc > NIR] : 0.9951701   
##   
## Kappa : 0.284   
##   
## Mcnemar's Test P-Value : 0.0004465   
##   
## Sensitivity : 0.5760   
## Specificity : 0.7413   
## Pos Pred Value : 0.4260   
## Neg Pred Value : 0.8399   
## Prevalence : 0.2500   
## Detection Rate : 0.1440   
## Detection Prevalence : 0.3380   
## Balanced Accuracy : 0.6587   
##   
## 'Positive' Class : Yes   
##

## Confusion Matrix and Statistics  
##   
##   
## No Yes  
## No 213 35  
## Yes 162 90  
##   
## Accuracy : 0.606   
## 95% CI : (0.5616, 0.6491)  
## No Information Rate : 0.75   
## P-Value [Acc > NIR] : 1   
##   
## Kappa : 0.2151   
##   
## Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.7200   
## Specificity : 0.5680   
## Pos Pred Value : 0.3571   
## Neg Pred Value : 0.8589   
## Prevalence : 0.2500   
## Detection Rate : 0.1800   
## Detection Prevalence : 0.5040   
## Balanced Accuracy : 0.6440   
##   
## 'Positive' Class : Yes   
##

## Confusion Matrix and Statistics  
##   
##   
## No Yes  
## No 284 53  
## Yes 91 72  
##   
## Accuracy : 0.712   
## 95% CI : (0.6701, 0.7513)  
## No Information Rate : 0.75   
## P-Value [Acc > NIR] : 0.976697   
##   
## Kappa : 0.3027   
##   
## Mcnemar's Test P-Value : 0.002047   
##   
## Sensitivity : 0.5760   
## Specificity : 0.7573   
## Pos Pred Value : 0.4417   
## Neg Pred Value : 0.8427   
## Prevalence : 0.2500   
## Detection Rate : 0.1440   
## Detection Prevalence : 0.3260   
## Balanced Accuracy : 0.6667   
##   
## 'Positive' Class : Yes   
##

# Objective 2: Additional Models

## Complex Model

In an attempt to narrow down the features to just those that predict the outcome the most, feature selection is a useful tool. It should be noted that while it may help, it is typically not a complete solution to this problem.

## 46 x 1 sparse Matrix of class "dgCMatrix"  
## s1  
## (Intercept) -2.391133849  
## bonemedYes .   
## bonemed\_fuYes .   
## bonetreatYes .   
## priorfracYes .   
## age .   
## weight .   
## height .   
## bmi .   
## premenoYes .   
## momfracYes 0.048205932  
## armassistYes 0.147574353  
## smokeYes .   
## rateriskSame .   
## rateriskGreater .   
## fracscore 0.165470088  
## weight:height .   
## bonemedYes:weight .   
## bonemedYes:bmi .   
## bonemed\_fuYes:weight .   
## bonemed\_fuYes:height .   
## bonemed\_fuYes:bmi 0.022291244  
## bonetreatYes:weight .   
## priorfracYes:age .   
## priorfracYes:weight .   
## priorfracYes:bmi 0.021137542  
## momfracNo:poly(age, 2)1 .   
## momfracYes:poly(age, 2)1 .   
## momfracNo:poly(age, 2)2 .   
## momfracYes:poly(age, 2)2 .   
## momfracNo:poly(weight, 2)1 .   
## momfracYes:poly(weight, 2)1 .   
## momfracNo:poly(weight, 2)2 -0.649246914  
## momfracYes:poly(weight, 2)2 -16.973494814  
## momfracNo:poly(bmi, 2)1 .   
## momfracYes:poly(bmi, 2)1 .   
## momfracNo:poly(bmi, 2)2 .   
## momfracYes:poly(bmi, 2)2 .   
## height:smokeYes .   
## weight:rateriskSame .   
## weight:rateriskGreater 0.003999953  
## bonetreatNo:weight:poly(height, 2)1 .   
## bonetreatYes:weight:poly(height, 2)1 -0.054784534  
## bonetreatNo:weight:poly(height, 2)2 .   
## bonetreatYes:weight:poly(height, 2)2 .   
## weight:height:bmi .

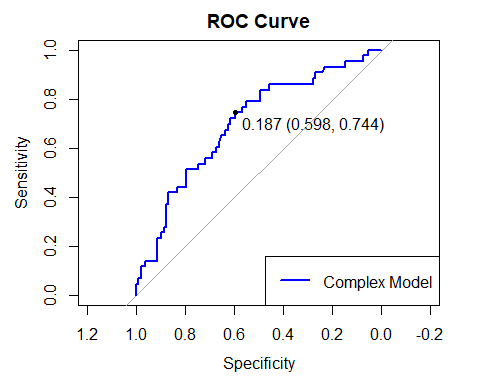
This confirms that at least some of the interaction terms may have a significant contribution to the output. In addition, it supports the idea that there may in fact be a polynomial interaction between momfrac and age, as well as momfrac and weight. The last step of model training for now is to look at validation scores.

## Setting direction: controls < cases

## Max Sensitivity Threshold: -Inf

## Max Specificity Threshold: 0.7073405

## Best Overall Threshold: 0.1867167

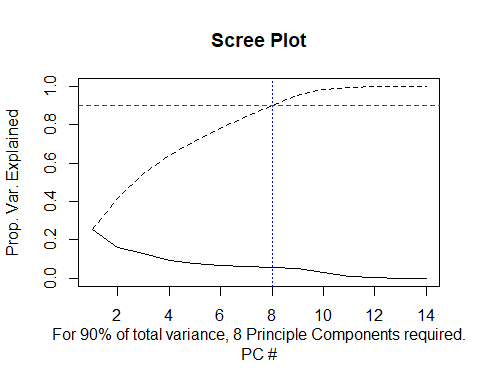


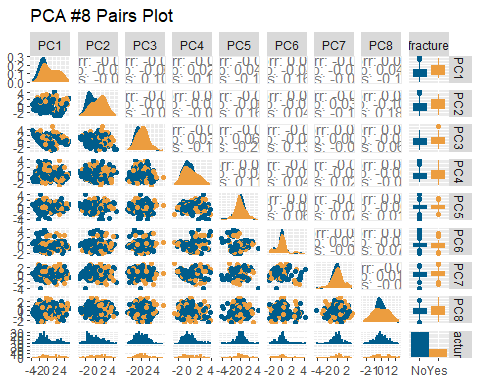
## GLMNET AUROC

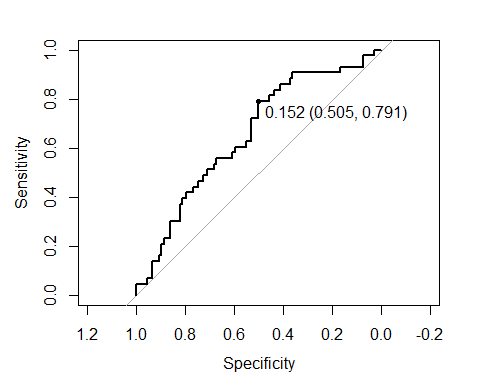
## Area under the curve: 0.7001

## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 64 11  
## Yes 43 32  
##   
## Accuracy : 0.64   
## 95% CI : (0.5577, 0.7167)  
## No Information Rate : 0.7133   
## P-Value [Acc > NIR] : 0.9793   
##   
## Kappa : 0.28   
##   
## Mcnemar's Test P-Value : 2.459e-05   
##   
## Sensitivity : 0.7442   
## Specificity : 0.5981   
## Pos Pred Value : 0.4267   
## Neg Pred Value : 0.8533   
## Prevalence : 0.2867   
## Detection Rate : 0.2133   
## Detection Prevalence : 0.5000   
## Balanced Accuracy : 0.6712   
##   
## 'Positive' Class : Yes   
##

Some of this performance has been lower than hoped

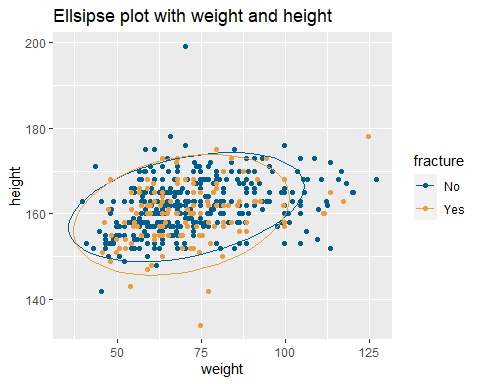
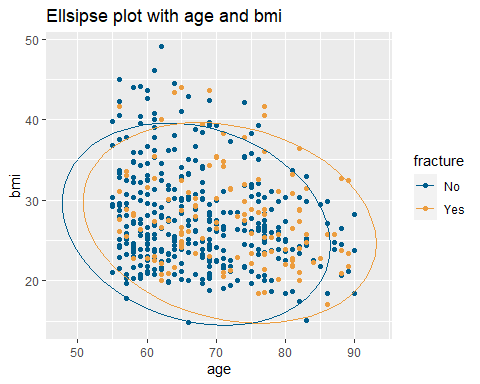
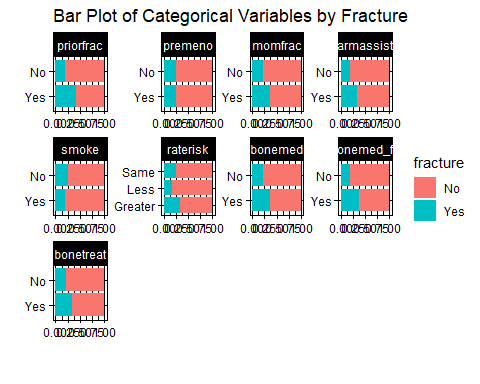
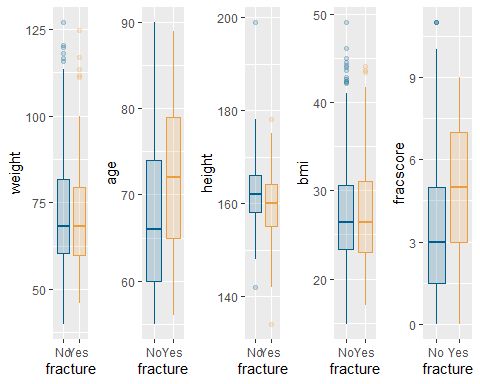






## Confusion Matrix and Statistics  
##   
## Reference  
## Prediction No Yes  
## No 68 19  
## Yes 39 24  
##   
## Accuracy : 0.6133   
## 95% CI : (0.5305, 0.6916)  
## No Information Rate : 0.7133   
## P-Value [Acc > NIR] : 0.9968   
##   
## Kappa : 0.17   
##   
## Mcnemar's Test P-Value : 0.0126   
##   
## Sensitivity : 0.5581   
## Specificity : 0.6355   
## Pos Pred Value : 0.3810   
## Neg Pred Value : 0.7816   
## Prevalence : 0.2867   
## Detection Rate : 0.1600   
## Detection Prevalence : 0.4200   
## Balanced Accuracy : 0.5968   
##   
## 'Positive' Class : Yes   
##

## QDA Model



## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 0.0188 0.8911  
## 498

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 0.9475 0.3308  
## 498

## Levene's Test for Homogeneity of Variance (center = median)  
## Df F value Pr(>F)  
## group 1 0.4352 0.5098  
## 498

## Average Sensitivity: 0.458355

## Average Specificity: 0.8374281

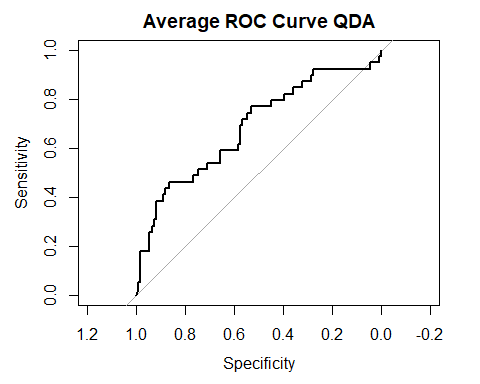
## Average PPV: 0.5384725

## Average NPV: 0.7876493

## Average Accuracy: 0.7251467

## Average Prevalence: 0.2933067

## Average AUROC: 0.7099156



We conducted an assessment of the equal variance assumption to determine whether Linear Discriminant Analysis (LDA) or Quadratic Discriminant Analysis (QDA) is a more suitable fit. Both box plots and the Levene test for the numerical variables indicated that they exhibit similar variances. However, upon examining the proportions within the categorical variable, a noticeable disparity emerged in most categories when tabulated by fracture status. This discrepancy suggests that the linear assumption necessary for LDA might not be satisfied in our case. So, we’ve decided to use the Quadratic Discriminant Analysis (QDA) model since the categorical variable proportions show differences that don’t align with the linear assumption needed for Linear Discriminant Analysis (LDA).

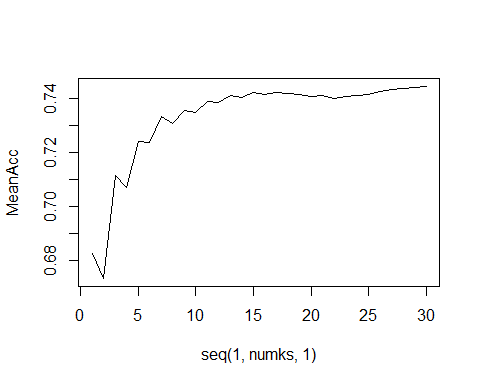
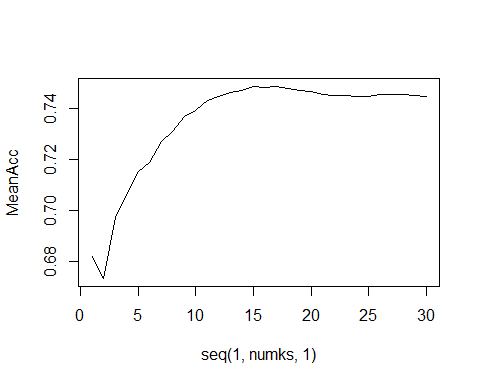
We’ll stick with the same parameters we used in the previous feature selection for the QDA model. However, we couldn’t include two components (“bonetreat” and “raterisk:weight”) due to collinearity problems. After running the model 500 times, we got an average accuracy of 0.72, sensitivity of 0.46, specificity of 0.84, and an average AUROC of 0.71.

## KNN

## Confusion Matrices for various KNN models

## ############ k = 3 ############## #### k = 15 (cont. variables only) #### ####### k = 15 (all variables) ####### #### k = 15 (previous feature selection) ###

## Confusion Matrix and Statistics Confusion Matrix and Statistics Confusion Matrix and Statistics Confusion Matrix and Statistics  
##   
##   
## classifications No Yes classifications No Yes classifications No Yes classifications No Yes  
## No 96 35 No 102 37 No 104 38 No 108 38  
## Yes 14 5 Yes 8 3 Yes 6 2 Yes 2 2  
##   
## Accuracy : 0.6733 Accuracy : 0.7 Accuracy : 0.7067 Accuracy : 0.7333   
## 95% CI : (0.5921, 0.7476) 95% CI : (0.6199, 0.772) 95% CI : (0.6269, 0.7781) 95% CI : (0.6551, 0.8022)  
## No Information Rate : 0.7333 No Information Rate : 0.7333 No Information Rate : 0.7333 No Information Rate : 0.7333   
## P-Value [Acc > NIR] : 0.958023 P-Value [Acc > NIR] : 0.845 P-Value [Acc > NIR] : 0.7982 P-Value [Acc > NIR] : 0.5424   
##   
## Kappa : -0.0027 Kappa : 0.003 Kappa : -0.0061 Kappa : 0.0446   
##   
## Mcnemar's Test P-Value : 0.004275 Mcnemar's Test P-Value : 2.993e-05 Mcnemar's Test P-Value : 2.962e-06 Mcnemar's Test P-Value : 3.13e-08   
##   
## Sensitivity : 0.12500 Sensitivity : 0.07500 Sensitivity : 0.05000 Sensitivity : 0.05000   
## Specificity : 0.87273 Specificity : 0.92727 Specificity : 0.94545 Specificity : 0.98182   
## Pos Pred Value : 0.26316 Pos Pred Value : 0.27273 Pos Pred Value : 0.25000 Pos Pred Value : 0.50000   
## Neg Pred Value : 0.73282 Neg Pred Value : 0.73381 Neg Pred Value : 0.73239 Neg Pred Value : 0.73973   
## Prevalence : 0.26667 Prevalence : 0.26667 Prevalence : 0.26667 Prevalence : 0.26667   
## Detection Rate : 0.03333 Detection Rate : 0.02000 Detection Rate : 0.01333 Detection Rate : 0.01333   
## Detection Prevalence : 0.12667 Detection Prevalence : 0.07333 Detection Prevalence : 0.05333 Detection Prevalence : 0.02667   
## Balanced Accuracy : 0.49886 Balanced Accuracy : 0.50114 Balanced Accuracy : 0.49773 Balanced Accuracy : 0.51591   
##   
## 'Positive' Class : Yes 'Positive' Class : Yes 'Positive' Class : Yes 'Positive' Class : Yes   
##



Let’s run a Cross Validation on the model:

## # Internal CV for continuous variables # ### Internal CV for all variables ### #### Internal CV for custom model ####

## Confusion Matrix and Statistics Confusion Matrix and Statistics Confusion Matrix and Statistics  
##   
##   
## classifications No Yes classifications No Yes classifications No Yes  
## No 359 110 No 365 116 No 359 109  
## Yes 16 15 Yes 10 9 Yes 16 16  
##   
## Accuracy : 0.748 Accuracy : 0.748 Accuracy : 0.75   
## 95% CI : (0.7075, 0.7855) 95% CI : (0.7075, 0.7855) 95% CI : (0.7096, 0.7874)  
## No Information Rate : 0.75 No Information Rate : 0.75 No Information Rate : 0.75   
## P-Value [Acc > NIR] : 0.5649 P-Value [Acc > NIR] : 0.5649 P-Value [Acc > NIR] : 0.524   
##   
## Kappa : 0.1032 Kappa : 0.0632 Kappa : 0.1135   
##   
## Mcnemar's Test P-Value : <2e-16 Mcnemar's Test P-Value : <2e-16 Mcnemar's Test P-Value : <2e-16   
##   
## Sensitivity : 0.1200 Sensitivity : 0.0720 Sensitivity : 0.1280   
## Specificity : 0.9573 Specificity : 0.9733 Specificity : 0.9573   
## Pos Pred Value : 0.4839 Pos Pred Value : 0.4737 Pos Pred Value : 0.5000   
## Neg Pred Value : 0.7655 Neg Pred Value : 0.7588 Neg Pred Value : 0.7671   
## Prevalence : 0.2500 Prevalence : 0.2500 Prevalence : 0.2500   
## Detection Rate : 0.0300 Detection Rate : 0.0180 Detection Rate : 0.0320   
## Detection Prevalence : 0.0620 Detection Prevalence : 0.0380 Detection Prevalence : 0.0640   
## Balanced Accuracy : 0.5387 Balanced Accuracy : 0.5227 Balanced Accuracy : 0.5427   
##   
## 'Positive' Class : Yes 'Positive' Class : Yes 'Positive' Class : Yes   
##

Final KNN Model:

## Setting direction: controls < cases

## KNN #1 AUROC

## Area under the curve: 0.775

## Setting direction: controls < cases

## KNN #2 AUROC

## Area under the curve: 0.7173

## Setting direction: controls < cases

## KNN #3 AUROC

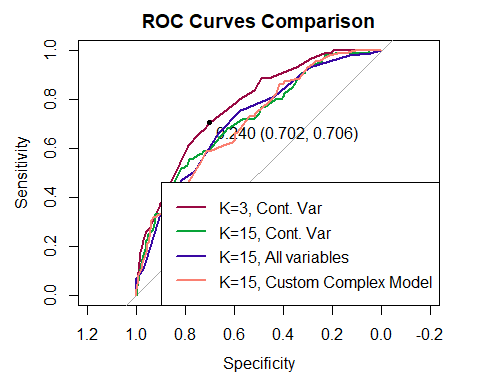
## Area under the curve: 0.7167

## Setting direction: controls < cases

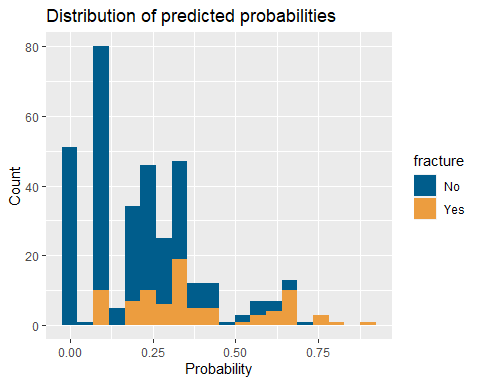
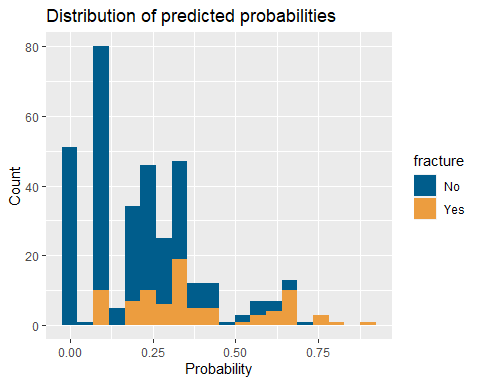
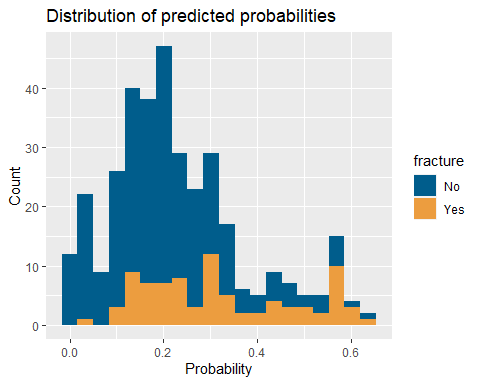
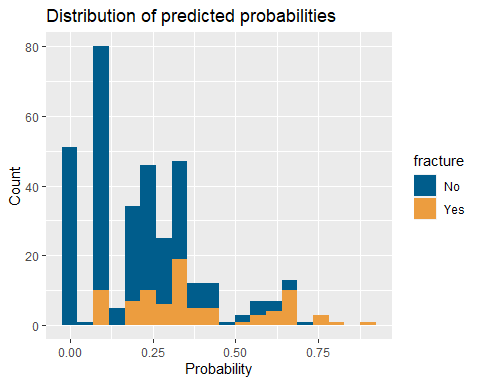
## KNN #4 AUROC

## Area under the curve: 0.7083

## Note: KNN # corresponds with legend in the following chart.



Inspect Probabilities

Source: [https://shihchingfu.github.io/knn-caret-example/#](https://shihchingfu.github.io/knn-caret-example/) 

Fine tuning KNN threshholds

Source: [https://shihchingfu.github.io/knn-caret-example/#](https://shihchingfu.github.io/knn-caret-example/)

## Confusion Matrices for various KNN models (Tuned Thresholds)  
## ############ k = 3 ############## #### k = 15 (cont. variables only) #### ####### k = 15 (all variables) ####### #### k = 15 (previous feature selection)   
## Confusion Matrix and Statistics Confusion Matrix and Statistics Confusion Matrix and Statistics Confusion Matrix and Statistics  
##   
##   
## No Yes No Yes No Yes No Yes  
## No 186 25 No 208 38 No 154 17 No 190 27  
## Yes 79 60 Yes 57 47 Yes 111 68 Yes 75 58  
##   
## Accuracy : 0.7029 Accuracy : 0.7286 Accuracy : 0.6343 Accuracy : 0.7086   
## 95% CI : (0.652, 0.7503) 95% CI : (0.6787, 0.7745) 95% CI : (0.5814, 0.6848) 95% CI : (0.6579, 0.7557)  
## No Information Rate : 0.7571 No Information Rate : 0.7571 No Information Rate : 0.7571 No Information Rate : 0.7571   
## P-Value [Acc > NIR] : 0.9914 P-Value [Acc > NIR] : 0.90353 P-Value [Acc > NIR] : 1 P-Value [Acc > NIR] : 0.984   
##   
## Kappa : 0.3354 Kappa : 0.314 Kappa : 0.2771 Kappa : 0.3351   
##   
## Mcnemar's Test P-Value : 2.024e-07 Mcnemar's Test P-Value : 0.06478 Mcnemar's Test P-Value : <2e-16 Mcnemar's Test P-Value : 3.26e-06   
##   
## Sensitivity : 0.7059 Sensitivity : 0.5529 Sensitivity : 0.8000 Sensitivity : 0.6824   
## Specificity : 0.7019 Specificity : 0.7849 Specificity : 0.5811 Specificity : 0.7170   
## Pos Pred Value : 0.4317 Pos Pred Value : 0.4519 Pos Pred Value : 0.3799 Pos Pred Value : 0.4361   
## Neg Pred Value : 0.8815 Neg Pred Value : 0.8455 Neg Pred Value : 0.9006 Neg Pred Value : 0.8756   
## Prevalence : 0.2429 Prevalence : 0.2429 Prevalence : 0.2429 Prevalence : 0.2429   
## Detection Rate : 0.1714 Detection Rate : 0.1343 Detection Rate : 0.1943 Detection Rate : 0.1657   
## Detection Prevalence : 0.3971 Detection Prevalence : 0.2971 Detection Prevalence : 0.5114 Detection Prevalence : 0.3800   
## Balanced Accuracy : 0.7039 Balanced Accuracy : 0.6689 Balanced Accuracy : 0.6906 Balanced Accuracy : 0.6997   
##   
## 'Positive' Class : Yes 'Positive' Class : Yes 'Positive' Class : Yes 'Positive' Class : Yes   
##

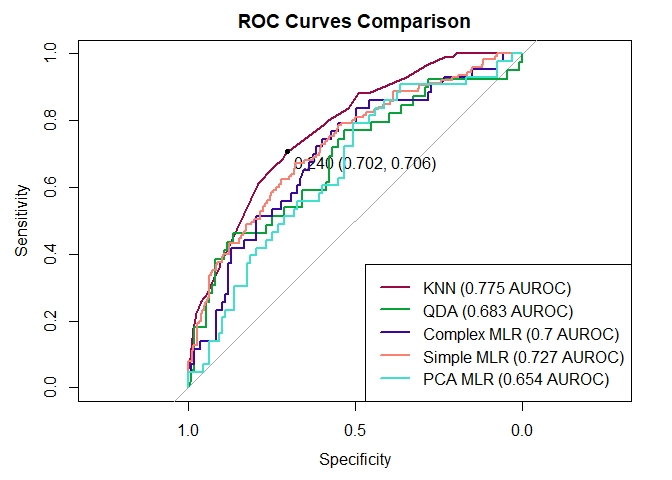
# Model Scores

## ############ KNN ############## ############ QDA ############## ######### Complex Model #########

## Confusion Matrix and Statistics Confusion Matrix and Statistics Confusion Matrix and Statistics  
##   
## Reference Reference  
## No Yes Prediction No Yes Prediction No Yes  
## No 186 25 No 78 18 No 64 11  
## Yes 79 60 Yes 33 21 Yes 43 32  
##   
## Accuracy : 0.7029 Accuracy : 0.66 Accuracy : 0.64   
## 95% CI : (0.652, 0.7503) 95% CI : (0.5783, 0.7353) 95% CI : (0.5577, 0.7167)  
## No Information Rate : 0.7571 No Information Rate : 0.74 No Information Rate : 0.7133   
## P-Value [Acc > NIR] : 0.9914 P-Value [Acc > NIR] : 0.98844 P-Value [Acc > NIR] : 0.9793   
##   
## Kappa : 0.3354 Kappa : 0.2144 Kappa : 0.28   
##   
## Mcnemar's Test P-Value : 2.024e-07 Mcnemar's Test P-Value : 0.04995 Mcnemar's Test P-Value : 2.459e-05   
##   
## Sensitivity : 0.7059 Sensitivity : 0.5385 Sensitivity : 0.7442   
## Specificity : 0.7019 Specificity : 0.7027 Specificity : 0.5981   
## Pos Pred Value : 0.4317 Pos Pred Value : 0.3889 Pos Pred Value : 0.4267   
## Neg Pred Value : 0.8815 Neg Pred Value : 0.8125 Neg Pred Value : 0.8533   
## Prevalence : 0.2429 Prevalence : 0.2600 Prevalence : 0.2867   
## Detection Rate : 0.1714 Detection Rate : 0.1400 Detection Rate : 0.2133   
## Detection Prevalence : 0.3971 Detection Prevalence : 0.3600 Detection Prevalence : 0.5000   
## Balanced Accuracy : 0.7039 Balanced Accuracy : 0.6206 Balanced Accuracy : 0.6712   
##   
## 'Positive' Class : Yes 'Positive' Class : Yes 'Positive' Class : Yes

## ########### Simple ############# ############ PCA ##############

## Confusion Matrix and Statistics Confusion Matrix and Statistics  
##   
## Reference  
## No Yes Prediction No Yes  
## No 213 35 No 68 19  
## Yes 162 90 Yes 39 24  
##   
## Accuracy : 0.606 Accuracy : 0.6133   
## 95% CI : (0.5616, 0.6491) 95% CI : (0.5305, 0.6916)  
## No Information Rate : 0.75 No Information Rate : 0.7133   
## P-Value [Acc > NIR] : 1 P-Value [Acc > NIR] : 0.9968   
##   
## Kappa : 0.2151 Kappa : 0.17   
##   
## Mcnemar's Test P-Value : <2e-16 Mcnemar's Test P-Value : 0.0126   
##   
## Sensitivity : 0.7200 Sensitivity : 0.5581   
## Specificity : 0.5680 Specificity : 0.6355   
## Pos Pred Value : 0.3571 Pos Pred Value : 0.3810   
## Neg Pred Value : 0.8589 Neg Pred Value : 0.7816   
## Prevalence : 0.2500 Prevalence : 0.2867   
## Detection Rate : 0.1800 Detection Rate : 0.1600   
## Detection Prevalence : 0.5040 Detection Prevalence : 0.4200   
## Balanced Accuracy : 0.6440 Balanced Accuracy : 0.5968   
##   
## 'Positive' Class : Yes 'Positive' Class : Yes   
##



# Conclusion

We have found that the best combination of scores can be found within our KNN model. No model performs the best with every single metric here, but KNN did lead on both Negative Predictive Value and AUROC scores, while also scoring near the top in every metric used in this analysis. Its strong performance at all decision thresholds is also evident in the above ROC Curve comparison.