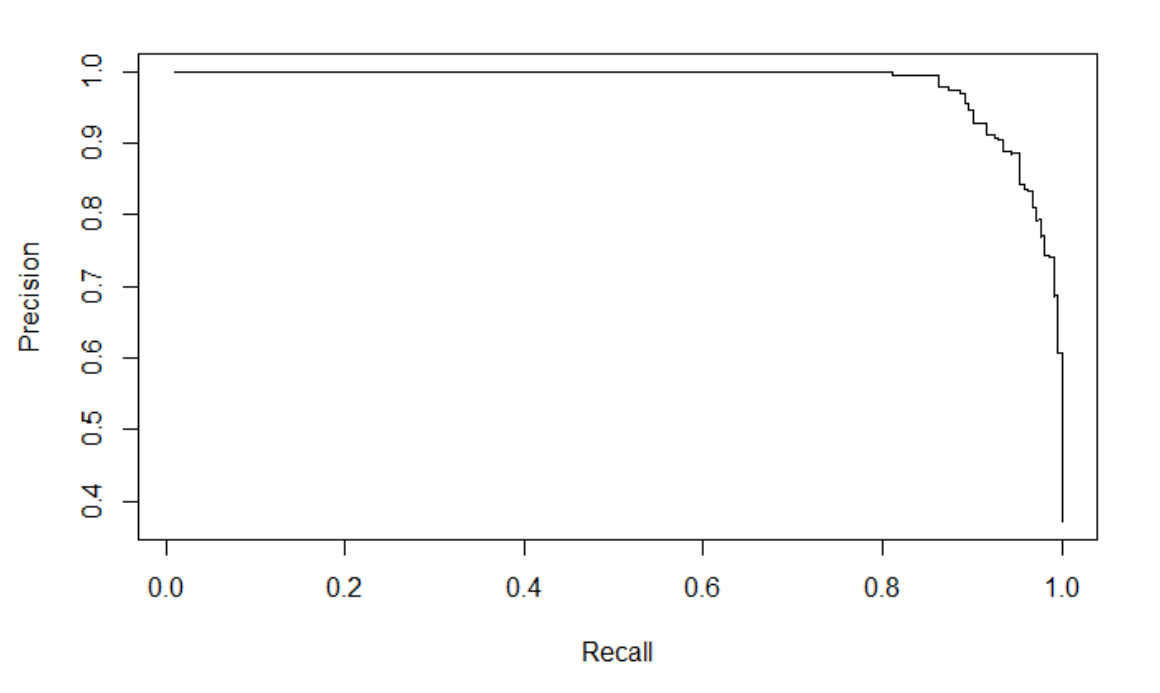
ROC Curve for Reduced Logistic Model

We can see significant lift on the reduced model that accounts for Area mean, texture mean, smoothness mean, and concave points mean.



Running a LASSO and Stepwise Feature Selection.

Looking at the output the StepAIC() seems to likely overfit the data, so despite having a lower AIC score of 50, the LASSO derived model will likely be the better one on newer data.

**Lasso Model**

|  |
| --- |
| > summary(main.lasso.glm)  Call:  glm(formula = diagnosis ~ concavity\_mean + concave\_points\_mean +  fractal\_dimension\_mean + radius\_se + smoothness\_se + compactness\_se +  symmetry\_se + fractal\_dimension\_se + radius\_worst + texture\_worst +  smoothness\_worst + concavity\_worst + concave\_points\_worst +  symmetry\_worst, family = binomial(link = "logit"), data = bc.clean)  Deviance Residuals:  Min 1Q Median 3Q Max  -1.6432 -0.0231 -0.0011 0.0001 3.4988  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -49.94186 14.89788 -3.352 0.000801 \*\*\*  concavity\_mean 12.70977 38.43786 0.331 0.740903  concave\_points\_mean 26.32819 70.41186 0.374 0.708466  fractal\_dimension\_mean -44.16790 144.43201 -0.306 0.759754  radius\_se 15.63875 4.93316 3.170 0.001524 \*\*  smoothness\_se 273.77291 203.10733 1.348 0.177683  compactness\_se -95.33134 60.50492 -1.576 0.115119  symmetry\_se -73.18646 122.55678 -0.597 0.550398  fractal\_dimension\_se -270.98565 545.40470 -0.497 0.619293  radius\_worst 1.25629 0.42384 2.964 0.003036 \*\*  texture\_worst 0.37798 0.08829 4.281 1.86e-05 \*\*\*  smoothness\_worst 31.84886 36.50939 0.872 0.383019  concavity\_worst 8.12142 10.47793 0.775 0.438282  concave\_points\_worst 30.35111 27.64683 1.098 0.272285  symmetry\_worst 23.54037 15.20640 1.548 0.121609  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 751.440 on 568 degrees of freedom  Residual deviance: 55.675 on 554 degrees of freedom  **AIC: 85.675**  Number of Fisher Scoring iterations: 10  **Stepwise Selection Model**  > summary(main.glm.step)  Call:  glm(formula = diagnosis ~ radius\_mean + texture\_mean + area\_mean +  smoothness\_mean + compactness\_mean + concavity\_mean + concave\_points\_mean +  symmetry\_mean + fractal\_dimension\_mean + perimeter\_se + area\_se +  smoothness\_se + compactness\_se + concavity\_se + concave\_points\_se +  symmetry\_se + fractal\_dimension\_se + radius\_worst + texture\_worst +  perimeter\_worst + area\_worst + concavity\_worst + symmetry\_worst +  fractal\_dimension\_worst, family = binomial(link = "logit"),  data = bc.clean)  Deviance Residuals:  Min 1Q Median 3Q Max  -0.003832 0.000000 0.000000 0.000000 0.004291  Coefficients:  Estimate Std. Error z value Pr(>|z|)  (Intercept) -5.914e+03 2.619e+05 -0.023 0.982  radius\_mean -6.630e+03 1.150e+05 -0.058 0.954  texture\_mean 1.913e+02 1.345e+03 0.142 0.887  area\_mean 6.077e+01 1.079e+03 0.056 0.955  smoothness\_mean 3.914e+04 2.517e+05 0.155 0.876  compactness\_mean -8.621e+04 9.326e+05 -0.092 0.926  concavity\_mean 2.852e+04 2.402e+05 0.119 0.905  concave\_points\_mean 5.886e+04 1.544e+06 0.038 0.970  symmetry\_mean -1.964e+04 1.347e+05 -0.146 0.884  fractal\_dimension\_mean 1.626e+05 1.120e+06 0.145 0.885  perimeter\_se -1.253e+03 1.822e+04 -0.069 0.945  area\_se 1.562e+02 2.259e+03 0.069 0.945  smoothness\_se -9.793e+04 1.472e+06 -0.067 0.947  compactness\_se 9.217e+04 7.142e+05 0.129 0.897  concavity\_se -8.131e+04 1.097e+06 -0.074 0.941  concave\_points\_se 4.398e+05 6.736e+06 0.065 0.948  symmetry\_se -1.038e+05 2.160e+06 -0.048 0.962  fractal\_dimension\_se -1.092e+06 1.065e+07 -0.103 0.918  radius\_worst 2.226e+03 2.134e+04 0.104 0.917  texture\_worst 7.269e+01 3.150e+03 0.023 0.982  perimeter\_worst 1.267e+02 1.355e+03 0.093 0.926  area\_worst -1.626e+01 1.165e+02 -0.140 0.889  concavity\_worst 6.737e+03 1.051e+05 0.064 0.949  symmetry\_worst 2.201e+04 3.283e+05 0.067 0.947  fractal\_dimension\_worst 5.899e+04 1.032e+06 0.057 0.954  (Dispersion parameter for binomial family taken to be 1)  Null deviance: 7.5144e+02 on 568 degrees of freedom  Residual deviance: 1.6713e-04 on 544 degrees of freedom  **AIC: 50**  Number of Fisher Scoring iterations: 25 |
|  |
| |  | | --- | | > | |

**Code: ROC Curve for Reduced Logistic Model**

#ROC CURVE TO ASSESS

library(ROCR)

bc\_lasso\_pred <- predict(redux.glm, newx = bc.clean, type = "response")

bc.lasso.pred <- prediction(bc\_lasso\_pred, bc.clean$diagnosis)

bc.lasso.perf <- performance(bc.lasso.pred, measure = "prec", x.measure = "rec")

plot(bc.lasso.perf)

**Code: LASSO Model Selection**

library(glmnet)

#NOTE: GLMNET requires dataframe to be converted to matrix

bc\_lasso\_mat <- model.matrix(diagnosis ~ ., bc.clean)[,-1]

bc.lasso.glm <- glmnet(bc\_lasso\_mat, bc.clean$diagnosis, family = "binomial" )

bc.lasso.cv <- cv.glmnet(bc\_lasso\_mat, bc.clean$diagnosis, family = "binomial")

bc\_lambda\_lasso <- bc.lasso.cv$lambda.min

bc\_lambda\_lasso

# Output the final coefficients from GLMNET LASSO

predict(bc.lasso.cv, type = "coefficients", s = bc\_lambda\_lasso )

**Code: AIC Stepwise Feature Selection**

library(MASS)

main.glm.step <- stepAIC(

main.glm, trace = 0, family = binomial(link = "logit"), direction = "both", test="Chisq"

)

plot(main.glm.step)

summary(main.glm.step)