Misc results, datathon 2022

Andrew Zimolzak

2022-04-20

Misc. results (too detailed)

The analysis is pretty sensitive to the phrasing of the descriptors. Limiting to "unpleasant" results in far fewer notes than "pleasant." Also, making that change results in more negativity for: *men, non-Latinos, and black patients,* unlike the prior word list when it was just the substring "pleasant."

In these tables, an upward trend in ratio means positive correlation of underprivileged group status and negative descriptors. In other words, ratios are $r = u/p$, where u is the count from a group hypothesized to be underprivileged, and p is the group hypothesized to be privileged.

Sex vs. negative descriptors

```
table(joined$Sex, joined$negativity) -> x
round((x[1,] / x[2,]), 3) -> f.m.ratio  # adhoc
rbind(x, f.m.ratio) %>% kable()
```

	0	1	2	3	4	5	6	8	13	24
Female	8506.000	213.000	36.000	9.000	4.000	4	2.0	0	0	0
Male	4515.000	128.000	46.000	16.000	15.000	2	5.0	3	1	1
f.m.ratio	1.884	1.664	0.783	0.562	0.267	2	0.4	0	0	0

```
women <- x[1,]  # adhoc
totals <- x[1,] + x[2,]  # adhoc
prop.trend.test(women, totals)
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  women out of totals ,
##  using scores: 1 2 3 4 5 6 7 8 9 10
## X-squared = 45.015, df = 1, p-value = 1.956e-11
```

Ethnicity vs. negative descriptors

```
table(joined$Ethnic.Group, joined$negativity)  -> x
round((x[1,] / x[2,]), 3) -> latino.non.ratio  # adhoc
rbind(x, latino.non.ratio) %>% kable()
```

	0	1	2	3	4	5	6	8	13	24
Hispanic or Latino	3032.000	117.000	17.000	3.000	5.000	0	1.000	2	0	0
Not Hispanic or Latino	9989.000	224.000	65.000	22.000	14.000	6	6.000	1	1	1
latino.non.ratio	0.304	0.522	0.262	0.136	0.357	0	0.167	2	0	0

```
latino <- x[1,]  # adhoc
totals <- x[1,] + x[2,]  # adhoc
prop.trend.test(latino, totals)
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  latino out of totals ,
##  using scores: 1 2 3 4 5 6 7 8 9 10
## X-squared = 1.1807, df = 1, p-value = 0.2772
```

Race vs. negative descriptors

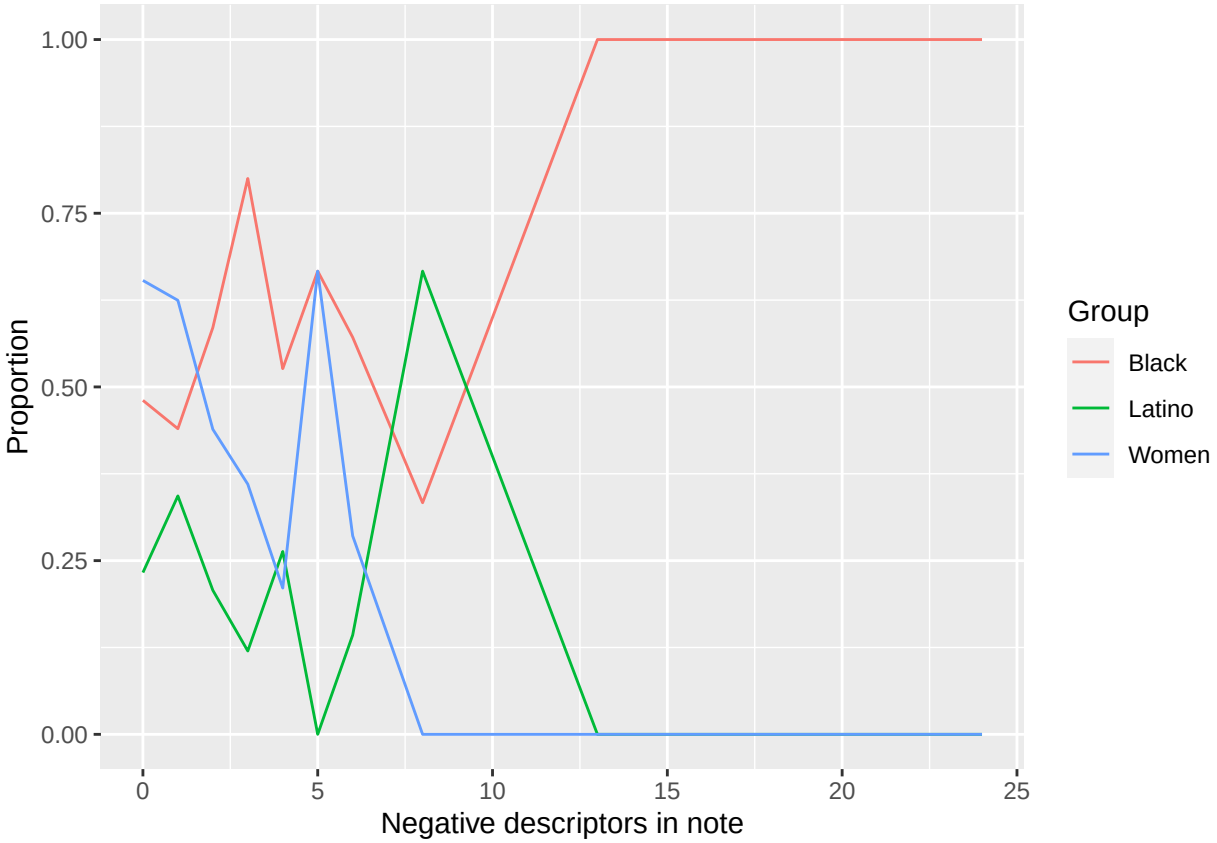
```
table(joined$Race, joined$negativity)  -> x
round((x['Black or African American',] / x['White or Caucasian',]), 2) -> b.w.ratio
rbind(x, b.w.ratio) %>% kable()
```

	0	1	2	3	4	5	6	8	13	24
American Indian or Alaska Native	1.00	0.00	0.00	0	0.00	0	0.00	0.0	0	0
Asian	314.00	1.00	1.00	0	0.00	0	0.00	0.0	0	0
Black or African American	6259.00	150.00	48.00	20	10.00	4	4.00	1.0	1	1
Native Hawaiian or Other Pacific Islander	15.00	0.00	0.00	0	0.00	0	0.00	0.0	0	0
Unable to Determine	17.00	0.00	0.00	0	0.00	0	0.00	0.0	0	0
White or Caucasian	6415.00	190.00	33.00	5	9.00	2	3.00	2.0	0	0
b.w.ratio	0.98	0.79	1.45	4	1.11	2	1.33	0.5	Inf	Inf

```
black <- x['Black or African American',]
totals <- colSums(x)
prop.trend.test(black, totals)
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  black out of totals ,
##  using scores: 1 2 3 4 5 6 7 8 9 10
## X-squared = 4.4787, df = 1, p-value = 0.03432
```

Single proportion line plot



Combined race/ethnicity vs. negative descriptors

Without a "cap" on negativity

	0	1	2	3	4	5	6	8	13	24
American Indian or Alaska Native	1.000	0.000	0.000	0.00	0.000	0.000	0.000	0.000	0	0
Asian	314.000	1.000	1.000	0.00	0.000	0.000	0.000	0.000	0	0
Black or African American	6259.000	150.000	48.000	20.00	10.000	4.000	4.000	1.000	1	1
Latino	3032.000	117.000	17.000	3.00	5.000	0.000	1.000	2.000	0	0
White or Caucasian	3415.000	73.000	16.000	2.00	4.000	2.000	2.000	0.000	0	0
Total	13021.000	341.000	82.000	25.00	19.000	6.000	7.000	3.000	1	1
p_asian	0.024	0.003	0.012	0.00	0.000	0.000	0.000	0.000	0	0
p_black	0.481	0.440	0.585	0.80	0.526	0.667	0.571	0.333	1	1
p_latino	0.233	0.343	0.207	0.12	0.263	0.000	0.143	0.667	0	0
p_white	0.262	0.214	0.195	0.08	0.211	0.333	0.286	0.000	0	0

```
prop.trend.test(x['White or Caucasian',], Total)
```

```
##
```

```
##  Chi-squared Test for Trend in Proportions
##
## data:  x["White or Caucasian", ] out of Total ,
##  using scores: 1 2 3 4 5 6 7 8 9 10
## X-squared = 7.1101, df = 1, p-value = 0.007665
```

```
prop.trend.test(x['Latino',], Total)
```

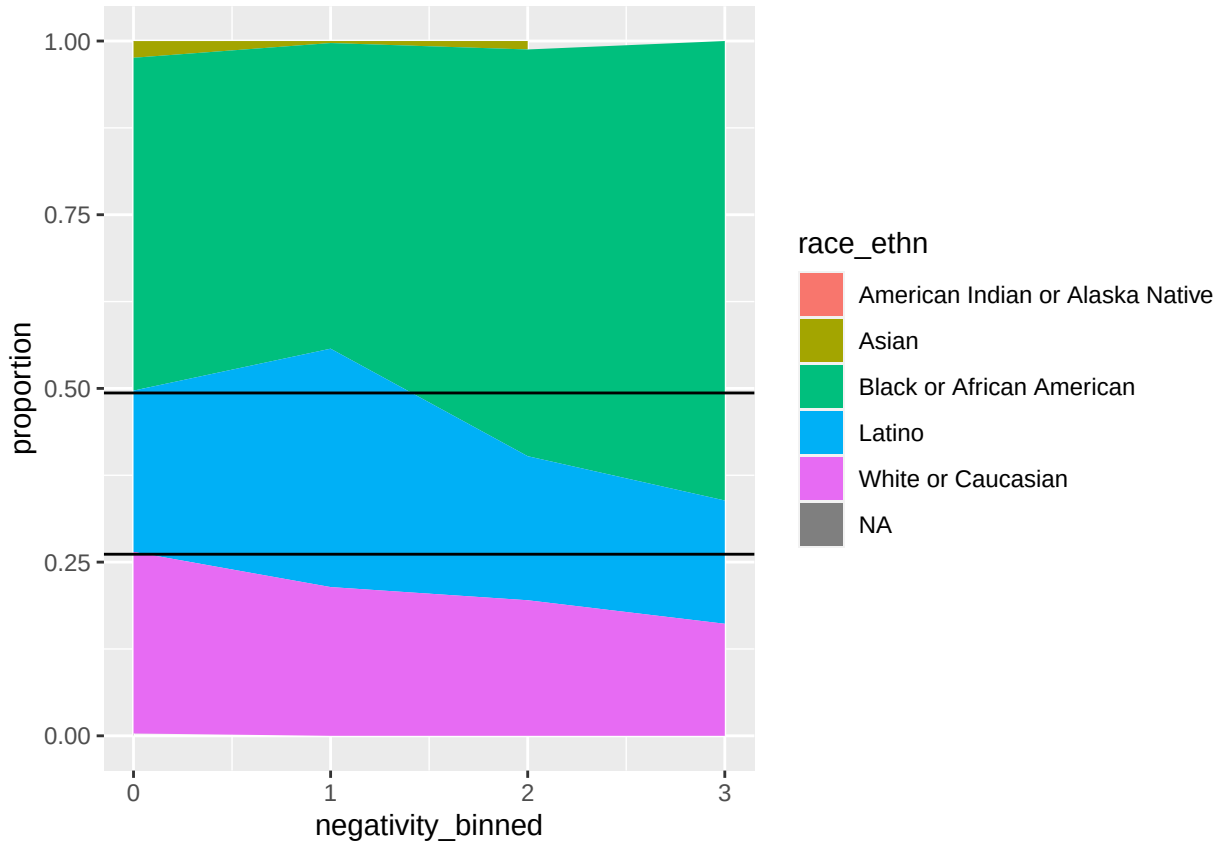
```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  x["Latino", ] out of Total ,
##  using scores: 1 2 3 4 5 6 7 8 9 10
## X-squared = 1.1807, df = 1, p-value = 0.2772
```

```
prop.trend.test(x['Black or African American',], Total)
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  x["Black or African American", ] out of Total ,
##  using scores: 1 2 3 4 5 6 7 8 9 10
## X-squared = 4.4787, df = 1, p-value = 0.03432
```

With a "cap" on negativity

	0	1	2	3
American Indian or Alaska Native	1.000	0.000	0.000	0.000
Asian	314.000	1.000	1.000	0.000
Black or African American	6259.000	150.000	48.000	41.000
Latino	3032.000	117.000	17.000	11.000
White or Caucasian	3415.000	73.000	16.000	10.000
Total	13021.000	341.000	82.000	62.000
p_asian	0.024	0.003	0.012	0.000
p_black	0.481	0.440	0.585	0.661
p_latino	0.233	0.343	0.207	0.177
p_white	0.262	0.214	0.195	0.161




```
prop.trend.test(x['White or Caucasian',], Total)
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  x["White or Caucasian", ] out of Total ,
##  using scores: 1 2 3 4
## X-squared = 8.7441, df = 1, p-value = 0.003106
```

```
prop.trend.test(x['Latino',], Total)
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  x["Latino", ] out of Total ,
##  using scores: 1 2 3 4
## X-squared = 2.1477, df = 1, p-value = 0.1428
```

```
prop.trend.test(x['Black or African American',], Total)
```

```
##
##  Chi-squared Test for Trend in Proportions
##
## data:  x["Black or African American", ] out of Total ,
##  using scores: 1 2 3 4
## X-squared = 4.4897, df = 1, p-value = 0.0341
```

Logistic

Big model

```
my_model <- glm(negativity_any ~ ., data = logit_me, family = "binomial")
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
summary(my_model)
```

```
##
## Call:
## glm(formula = negativity_any ~ ., family = "binomial", data = logit_me)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0838  -0.3276  -0.2640  -0.1690   3.3800
##
## Coefficients:
##                                          Estimate Std. Error z value Pr(>|z|)
## (Intercept)                            -6.009e+10  3.805e+11  -0.158 0.874498
## SexMale                                 1.094e+00  1.961e-01   5.578 2.44e-08 ***
## race_ethnAsian                          6.009e+10  3.805e+11   0.158 0.874498
## race_ethnBlack or African American      6.009e+10  3.805e+11   0.158 0.874498
## race_ethnLatino                         6.009e+10  3.805e+11   0.158 0.874498
## race_ethnWhite or Caucasian             6.009e+10  3.805e+11   0.158 0.874498
## Employment.StatusFull Time              1.264e+00  5.176e-01   2.443 0.014569 *
## Employment.StatusNot Employed           8.738e-01  5.329e-01   1.640 0.101100
## Employment.StatusPart Time              1.840e+00  7.706e-01   2.388 0.016929 *
## Employment.StatusRetired                1.547e+00  5.099e-01   3.034 0.002412 **
## Employment.StatusStudent - Full time    1.679e-01  1.127e+00   0.149 0.881583
## Employment.StatusUnknown               -2.021e+01  2.045e+05   0.000 0.999921
## interpreterY                            2.618e+01  8.286e+04   0.000 0.999748
## LanguageFarsi, Persian                  6.009e+10  3.805e+11   0.158 0.874498
## LanguageLaotian                        -5.025e+01  3.449e+05   0.000 0.999884
## LanguageSpanish                        -2.524e+01  8.286e+04   0.000 0.999757
## LanguageUnknown                        -2.307e+01  3.251e+05   0.000 0.999943
## LanguageVietnamese                     -4.867e+01  8.524e+04  -0.001 0.999544
## Marital.StatusLegally Separated        -4.783e+00  7.350e-01  -6.507 7.67e-11 ***
## Marital.StatusLife Partner             -2.692e+00  7.464e-01  -3.607 0.000309 ***
## Marital.StatusMarried                  -3.916e+00  6.300e-01  -6.216 5.10e-10 ***
## Marital.StatusSingle                   -3.504e+00  6.281e-01  -5.579 2.42e-08 ***
## Marital.StatusUnknown                  -2.685e+01  2.045e+05   0.000 0.999895
## Marital.StatusWidow/Widower            -2.104e+00  5.694e-01  -3.695 0.000219 ***
## mcaidTRUE                              -2.371e+01  7.676e+04   0.000 0.999754
## mcareTRUE                              -1.031e+00  2.716e-01  -3.797 0.000147 ***
## Financial.ClassMedicaid                -2.382e+01  3.569e+05   0.000 0.999947
## Financial.ClassMedicaid Mgd Care       -3.672e-02  3.879e-01  -0.095 0.924568
## Financial.ClassMedicare                 1.847e+00  4.113e-01   4.492 7.07e-06 ***
## Financial.ClassMedicare Mgd Care        1.549e+00  4.592e-01   3.374 0.000741 ***
## Financial.ClassSelf-Pay                 1.501e+00  4.542e-01   3.305 0.000951 ***
## Age                                    -2.481e-02  1.295e-02  -1.915 0.055493 .
```

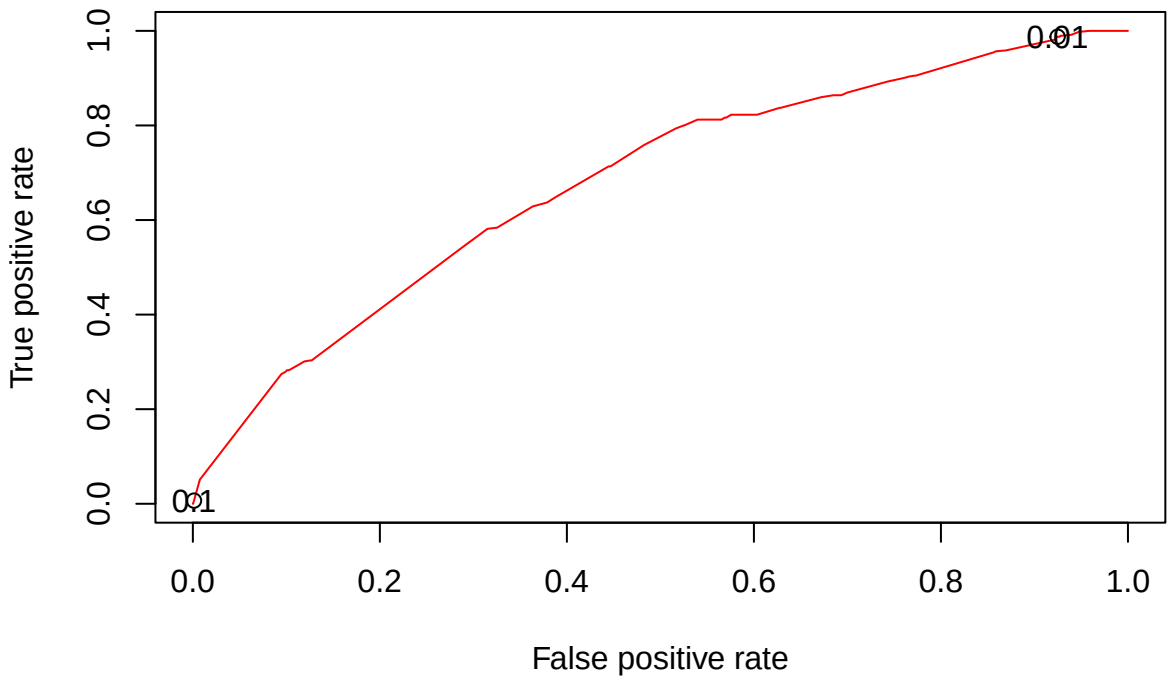
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4179.3  on 13505  degrees of freedom
## Residual deviance: 3978.4  on 13474  degrees of freedom
##   (44 observations deleted due to missingness)
## AIC: 4042.4
##
## Number of Fisher Scoring iterations: 25
```

ROC curve

```
rocr_pred = prediction(my_model$fitted.values, my_model$y)
rocr_perf <- performance(rocr_pred, measure = "tpr", x.measure = "fpr")
auc = performance(rocr_pred, measure = "auc")
auc@y.values
```

```
## [[1]]
## [1] 0.6807869
```

```
plot(rocr_perf, col=rainbow(10), print.cutoffs.at=c(0.01, 0.1))
```



Little model

```
little_model <- glm(negativity_any ~ Sex + race_ethn + Age, data = logit_me, family = "binomial")
summary(little_model)
```

```
##
```

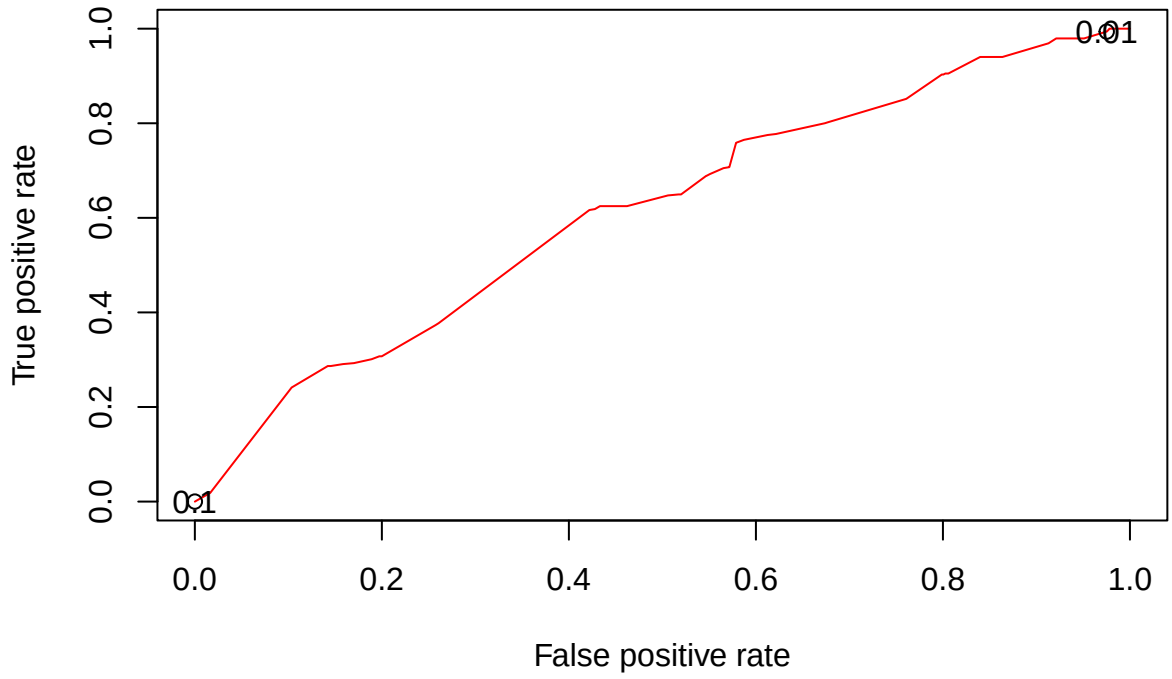
```
## Call:
## glm(formula = negativity_any ~ Sex + race_ethn + Age, family = "binomial",
##     data = logit_me)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.3701  -0.3002  -0.2573  -0.2276   3.1639
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -13.678025 196.967948  -0.069    0.945
## SexMale                            0.391663   0.094909   4.127 3.68e-05 ***
## race_ethnAsian                     6.826930 196.968980   0.035    0.972
## race_ethnBlack or African American   8.858191 196.967703   0.045    0.964
## race_ethnLatino                    9.003980 196.967712   0.046    0.964
## race_ethnWhite or Caucasian        8.249610 196.967713   0.042    0.967
## Age                                0.022636   0.004117   5.499 3.83e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 4179.3  on 13505  degrees of freedom
## Residual deviance: 4102.5  on 13499  degrees of freedom
##   (44 observations deleted due to missingness)
## AIC: 4116.5
##
## Number of Fisher Scoring iterations: 10
```

ROC curve

```
rocr_pred = prediction(little_model$fitted.values, little_model$y)
rocr_perf <- performance(rocr_pred, measure = "tpr", x.measure = "fpr")
auc = performance(rocr_pred, measure = "auc")
auc@y.values
```

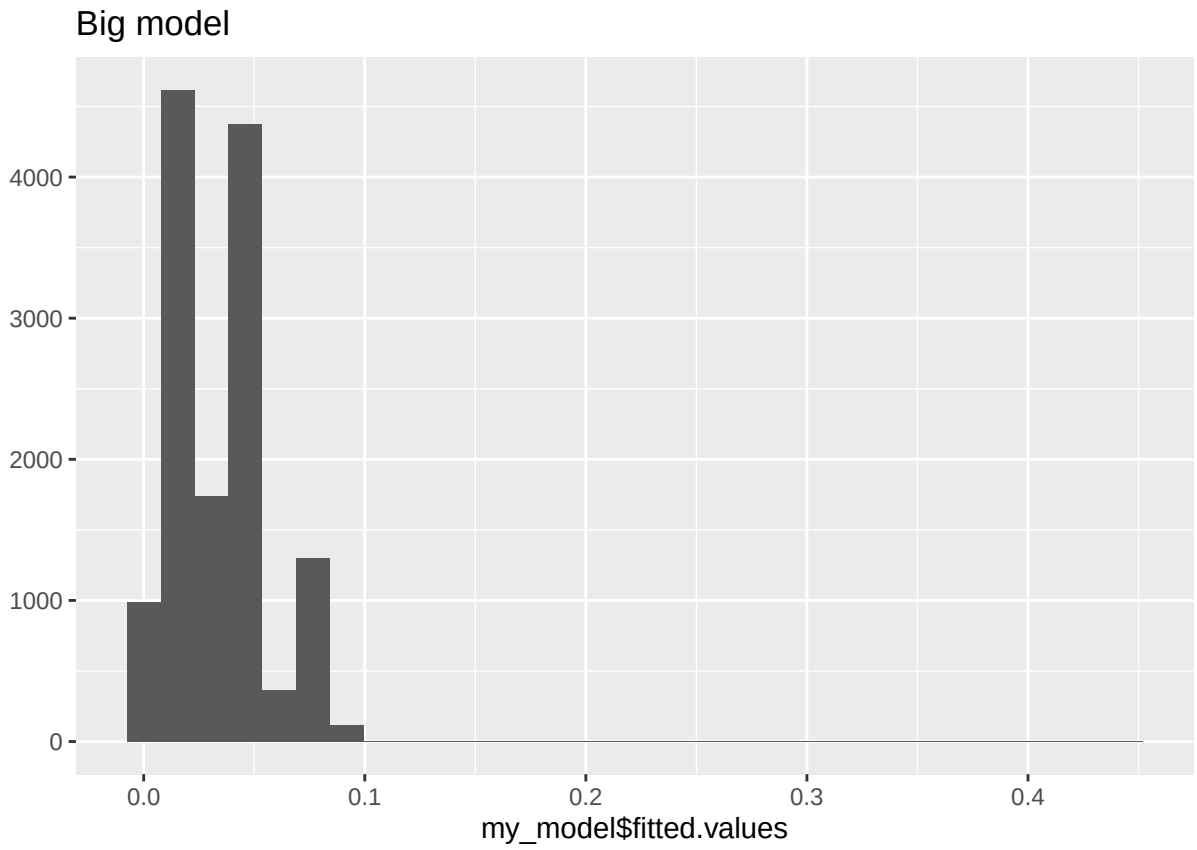
```
## [[1]]
## [1] 0.6145736
```

```
plot(rocr_perf, col=rainbow(10), print.cutoffs.at=c(0.01, 0.1))
```

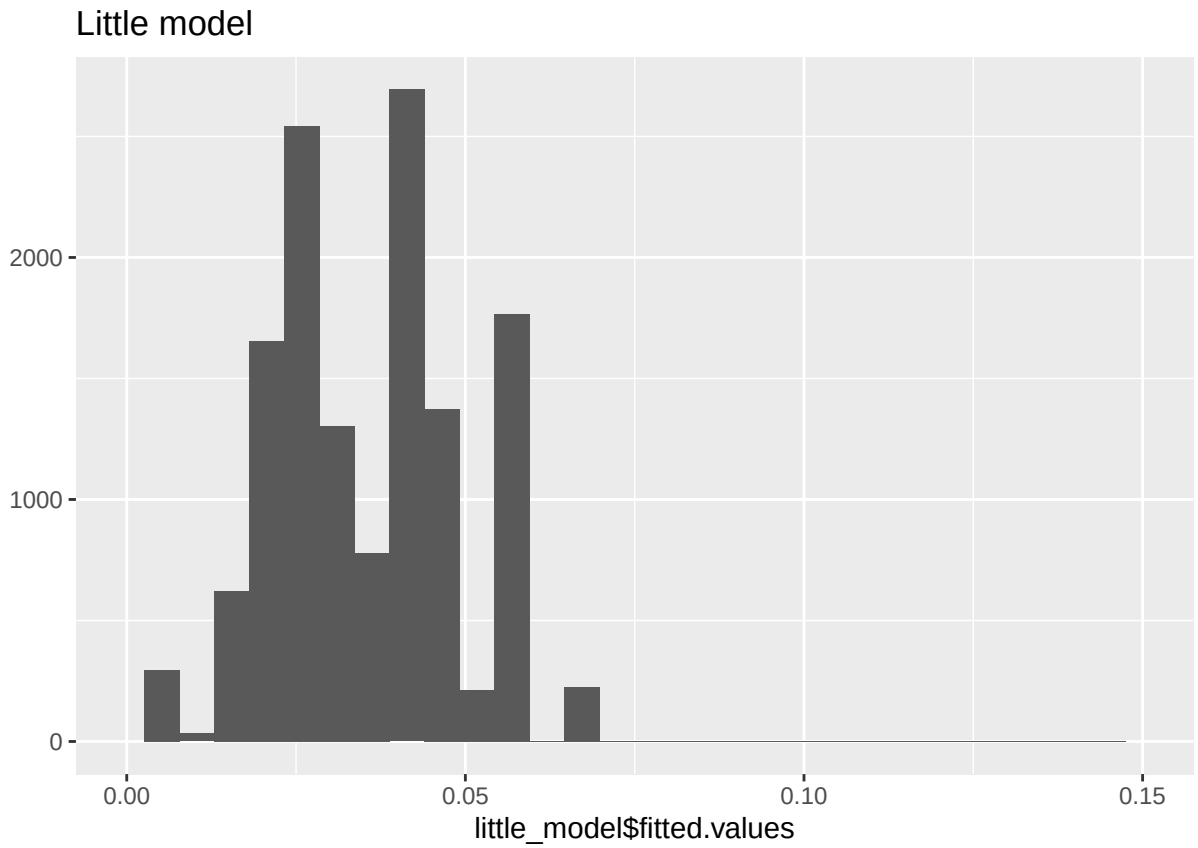


Predicted probabilities, big vs. little

```
qplot(my_model$fitted.values) + labs(title = 'Big model')
```



```
qplot(little_model$fitted.values) + labs(title = 'Little model') + xlim(0, 0.15)
```



I get the feeling it hates unbalanced classes.

Try column plot of coefficients

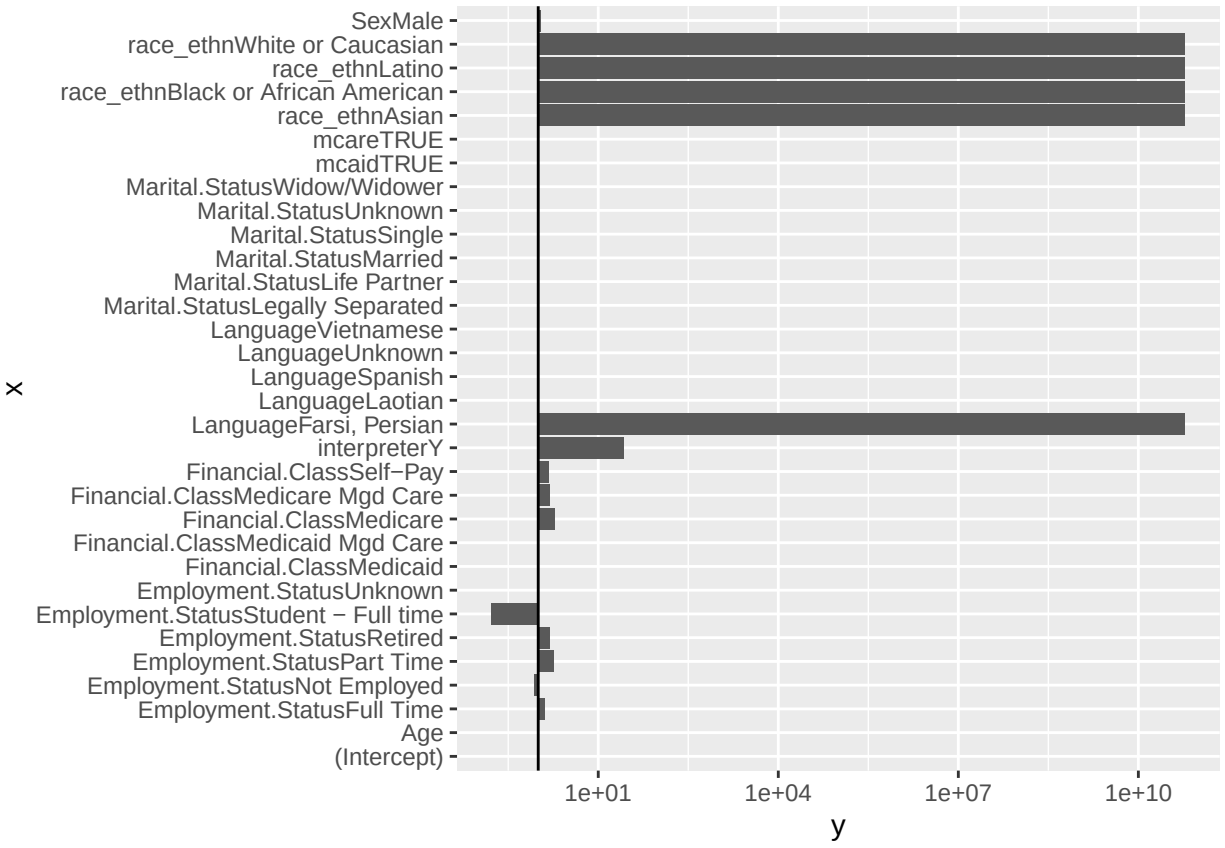
```
df = data.frame(
  y=my_model$coefficients,
  x=names(my_model$coefficients)
)

ggplot(df, aes(x=y, y=x) ) + geom_col() + scale_x_log10() +
  geom_vline(xintercept = 1)
```

```
## Warning in self$trans$transform(x): NaNs produced

## Warning: Transformation introduced infinite values in continuous x-axis

## Warning: Removed 17 rows containing missing values (position_stack).
```



To do, or not

- Limit only to those discharged from ED to home.
- More rigorous handling of those with multiple demographic entries.
- Get complete note data from IT.
- Show length distribution of notes.