Ludovic Njiosseu

A2.4 Write one paragraph interpreting some of the coefficients. If the model did not fit some coefficients, explain why in your write up.

Levels in the Race, age, party, and ideology variables were all significant predictors. For example, racehispanic has an estimate of -2.3503186762 with a p value < .001, which is significant. It shows that as compared to black race(reference), the log odds for voting for Obama will be 2.35 times less for hispanic voters. The log odds for voting for Obama will be 2.74 times less in other races and in white voters the log odds for voting for Obama will be 3.07 times less. All of these values are statistically significant.

The model did not fit coefficients for state_contestedness. The specific error related to this was 'not defined because of singularities' which usually indicates a variable was perfectly correlated with another variable. In this case, because variables were all categorical, it refers to the state of individual respondents; since each respondent was only from one state, and all individuals from that state would have the same state_contestedness, the state and state_contestedness variables were perfectly correlated.

A3.5 Report these values and any notable differences between the metrics corresponding to the predictions in steps 2 and 3 in your writeup.

	At .5	**At .7**
Accuracy	0.858	0.829
Precision	0.861	0.917
Recall	0.895	0.771

A4.3 Report the accuracy of your classifier in your writeup.
84.15%

A4.5 Briefly describe the histograms in your writeup

The histograms show an overwhelming probability of voters would vote for a major party. There is a bimodal distribution of votes among those who vote for a major party candidate. Voters are either highly likely to vote for Obama (with probabilities > .75) or very unlikely (with probabilities near 0, indicating in this case a high likelihood for voting for McCain). The count for those with a high probability of voting for Obama appears larger than the count for those with a low probability.

A4.6 What is the advantage of generating pr_obama and pr_mccain using the procedure in steps 2 and 3?
This procedure yields output that adds up to 1, accounting for the relative probability of each person's voting. Since the likelihood of anyone voting for a third party is very small, this procedure may help identify the individuals most likely to make a rare choice.

As Obama and McCain are logically nested within Major party candidate, having this procedure also seems more methodologically correct. If you use two logistic regression models, one for PR Obama and one for PR McCain, you have less clarity on what the 'Other' category contains -- it contains both the other candidate and any third party. Those probabilities may not sum to 1, so interpreting them is more difficult.

B2.4 Give a precise statement interpreting one of these coefficients in your writeup.
monthAugust is .364; holding all else constant, the log odds of finding a weapon are .364 times greater if the stop happened in August than in December.

B3.1 If your model were used to predict the ex-ante probability that this person were carrying a weapon, what would this probability be?

14.6%
Had it been a man, the probability of him carrying a gun is 14.6%

What if this person were a woman, everything else being equal?
Answer :12.5%
Had it been a woman, the probability of her carrying a gun is 12.5%

B3.2 Compute the AUC of this model on all data from 2009, using the ROCR package (as in lecture). Report this number in your writeup.

81.2%

B3.3 Computing the proportion of pairs where your model predicts that the true example is more likely to find a weapon than the false example. Confirm that your answer is approximately equal to the answer computed in Question B3.2, and report this number in your write up.
78.9%
Yes, these are approximately equal.

B4
Write at least one paragraph explaining what you did and what you found.

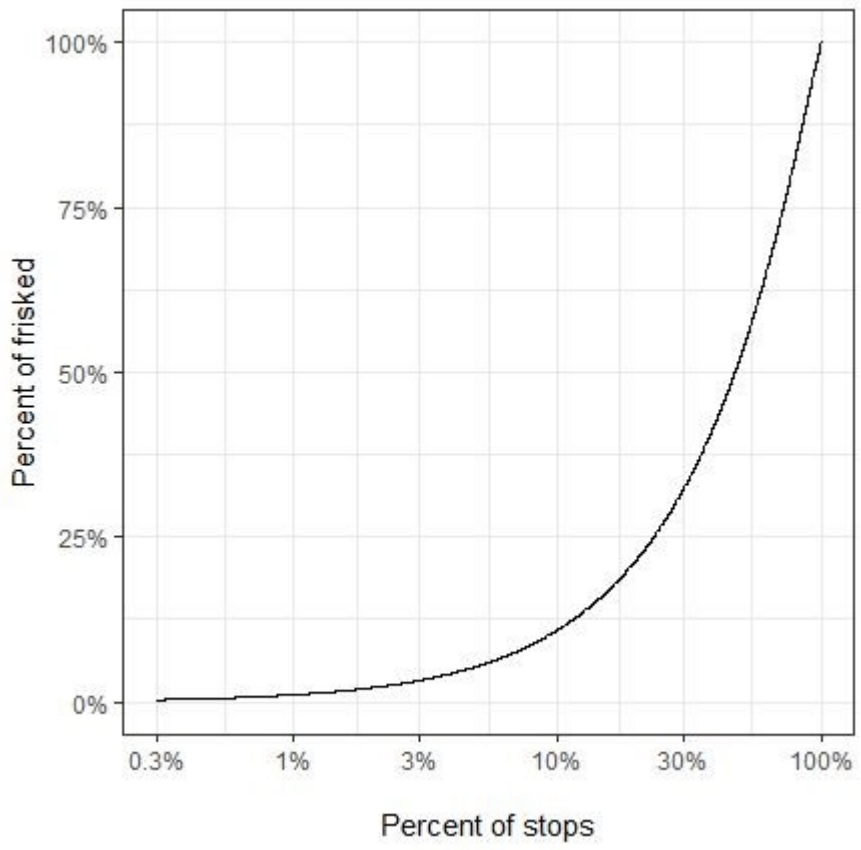
The decisions made in this section were as follows:

- Chose the target variable as whether the suspect was frisked or not. “frisked”
- Subset data to those having “suspected.crime==cpw” and took years 2008, 2012, 2016
- Chose a train/test split; chose 50% of data for test and 50% for train
- Model chosen = logistic regression
- Chose a set of predictors as suspect.race', 'suspect.age', 'suspect.build',
  'suspect.sex', 'suspect.height', 'suspect.weight',
  'precinct', 'inside', 'location.housing',
  'observation.period', 'additional.report',
  'additional.investigation', 'additional.proximity',
  'additional.evasive', 'additional.associating',
  'additional.direction', 'additional.highcrime',
  'additional.time', 'additional.sights', 'additional.other',
  'radio.run', 'day', 'month', 'time.period' and year

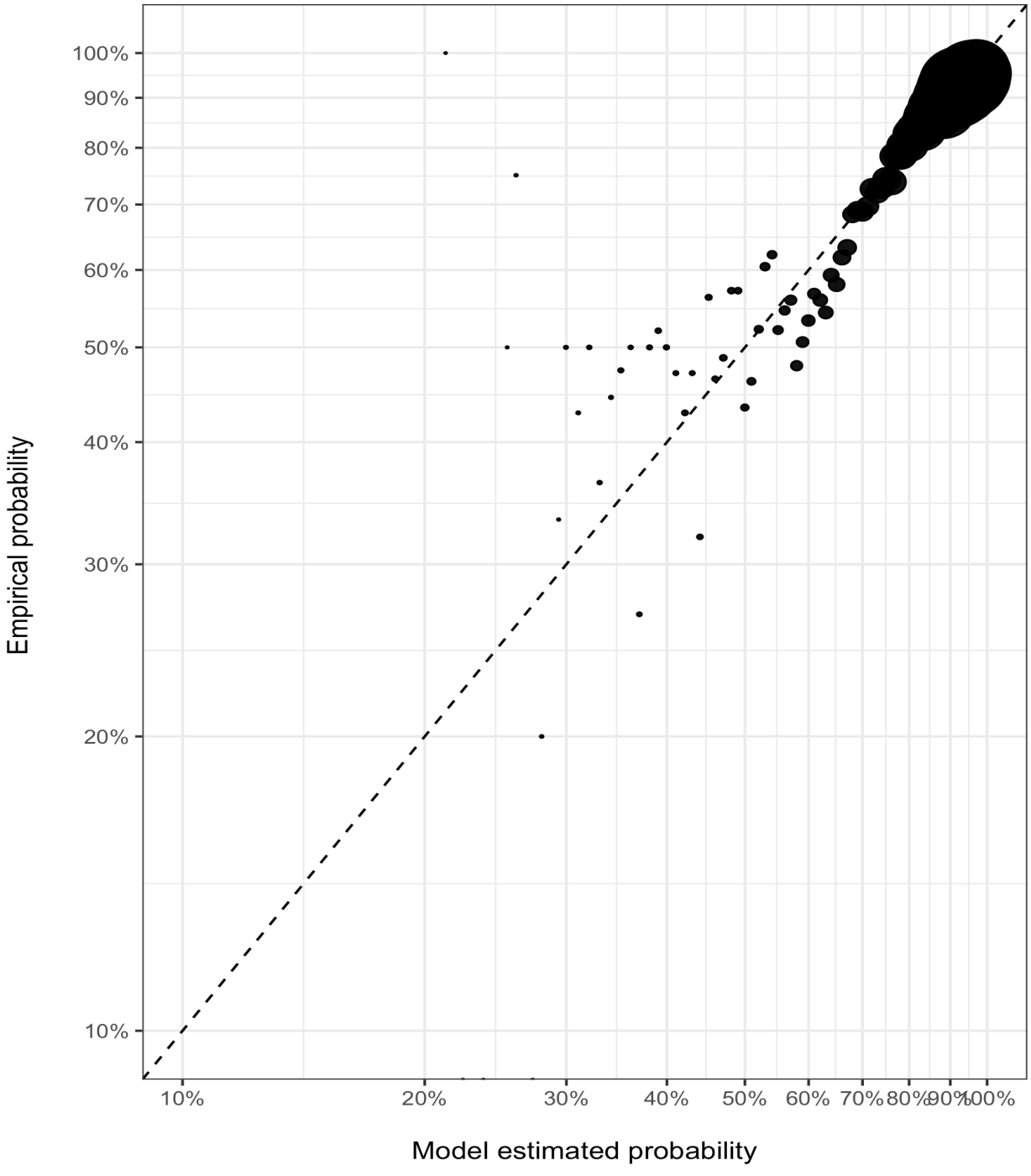
Based on the values calculated, we generated model performance plots as follows:

1. Recall at k% 2.

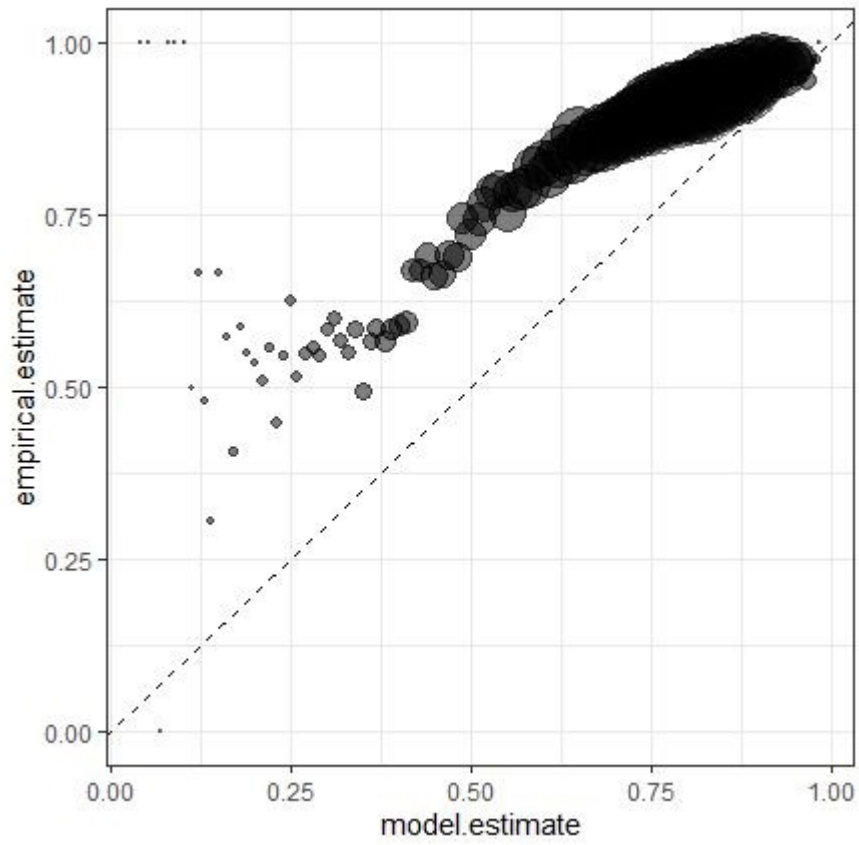
Calibration plot:



The curve does not seem to be “outward pushed” which shows the prediction is quite skewed to produce false positive. The prime reason being the data in which there is a larger proportion of people who were not frisked as compared to those that were. Additionally, it may also be because the features selected in the model are not sensitive to determining whether the suspect was frisked.



Same plot with different scale to see better:



The plot was rescaled to see the distribution more clearly. All the observations are lying above the y=x line showing that the proportion of times people were actually frisked is greater than what our model predicted for this data. The only time the model approached real value was when the empirical estimate was 100%. This plot reiterates the higher number of false positives present in the data.