A Logistic Regression Classifier for Suicidal Reddit Posts

Linda Gai

10/31/2016

```
##
## Please cite as:

##  Hlavac, Marek (2015). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2. http://CRAN.R-project.org/package=stargazer
```

Introduction

Suicide is the second leading cause of the death worldwide among people ages 15-29 (WHO). In particular, young people often the use internet to obtain mental health support (Horgan 2009). In this project, we examine the expression of suicidal ideation on Reddit, a popular online message board aggregator that hosts discussion on user-created topics in smaller sub-boards (known as "subreddits"). Notably, some subreddits are used to provide and ask for mental health support, and many users share deeply personal information, including thoughts of suicide.

Here, we develop a logistic regression model that scores text posts from mental health subreddits for the presence of suicidal ideation. Previous work on suicidal ideation among Reddit users incorrectly assumed that popular mental health subreddits other than r/SuicideWatch, a subreddit dedicated to helping suicidal users, do not contain suicidal posts (de Choudhury 2015). As such, developing reliable identification methods for suicidal ideation in text posts is an important step to improve future work on admissions of suicidal ideation on the internet. Our work suggests that suicidal ideation is associated with text written at a higher grade level, higher counts of the words "die", among other words, and lower counts of second-person pronouns.

Methods

Previous work in suicidal ideation on Reddit by de Choudhury (2015) identified several characteristics of the posts that users who eventually post in SuicideWatch, a proxy measure for suicidal ideation. Here, we first seek to confirm that the same patterns are upheld in users that make suicidal posts and comments to mental health subreddits generally. Secondly, we use the variables that appear to be predictive to create multiple logistic regression models using the training set, and select the best model using backwards stepwise selection. Finally, we calculate the sensitivity and specificity of the best logistic regression model on both the training and test sets, based on a random sample of posts made in June 2016 obtained from r/depression.

Data Collection

Data was collected in three stages. First, we obtained 95 submissions, randomly sampled from a set of the largest mental health subreddits excluding r/depression, using the PRAW python package, which contains a function for randomly sampling submissions from subreddits, as well as a function for randomly sampling a subreddit from a given list of subreddits (scraped by Lacey). The posts were selected by first randomly selecting a subreddit from a manually-created list of the 20 largest mental health subreddits, excluding r/depression, then randomly sampling a post from that subreddit. Each of these was scored for suicidal ideation (by Lacey and Linda). Second, we scraped all submissions and comments from r/depression made between June 1-30, 2016, as well as all replies to submissions made between June 1-30, 2016, using a Python

subreddit archiver, subredditarchive.py (scraped by Linda). From this dataset, a second set of submissions and comments was scored for suicidal ideation (by Lacey and Linda).

Because there were relatively few suicidal posts in the first two training sets (66 out of over 700), a phrase-matching classification algorithm (developed by Lacey and coded by Lacey and Linda) was written to enrich the training sets with additional suicidal posts from the June 2016 dataset. The phrase-matching algorithm works by searching posts for matches to a dictionary phrases that indicated suicidal ideation in the posts in the first two training sets. The phrase-matching algorithm had very high specificity when applied to the June 2016 dataset, identifying only 1 false positive out of 385 submissions labeled as suicidal (work done by Lacey). Moreover, this consitutes of the 4000 submission June dataset, suggesting the number of suicidal posts in r/depression is substantial (work done by Lacey). However, the sensitivity of this algorithm is unknown.

To create the final dataset, we combined the hand-scored submissions from multiple mental health subreddits and the hand-scored submissions and comments from r/depression in June for which the suicidal scores matched for both Lacey and Linda, with the suicidal posts identified by phrase-matching algorithm from the June 2016 r/depression posts. The final dataset contained 611 suicidal posts and 757 non-suicidal posts. 75% of the dataset was then randomly sampled into a training set used to create the model, while the other 25% was reserved as a test set.

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. Exploratory data analysis was used to perform quality control on the data and determine the predictor variables for the the logistic regression model. Several R packages were used to measure aspects of the text posts during exploratory analysis, though only one was ultimately useful for the final model.

To score the posts for readability, we used the korpus package to score each post for Flesch-Kincaid Grade Level, a standard measure of how difficult a passage is to read in English. Scores range from -3.40 and above, with the score roughly corresponding to the US grade level in difficult. For example, text with a score of "13" should be readable by a first-year college student, while text with a score of "1" would be readable by a first-grade student. To avoid negative values, linear transformation of the readability, age, was then used in lieu of the grade-level, using the korpus package.

Statistical Modeling

Here, we use logistic regression as the analysis framework. Although there are several other binary classification algorithms, such as support vector machines, discriminant analysis, and random forests that would be potentially useful for determining more complex relationships between predictors and the outcome, they suffer some drawbacks that make them unsuitable for this project. First, these algorithms are more difficult to interpret (Hastie et al 2001), and thus the insights gleaned from such a model may be of limited public health use. Furthermore, it has been argued that, in practice, the performance of logistic regression to more complex methods like random forests is comparable (Hand 2006, Ruiz et al 2007). As such, logistic regression was ultimately chosen as it is a relatively easy-to-understand framework.

The predictive variables were chosen on the basis of the exploratory analysis, prior knowledge of characteristics of suicidal posts identified by de Choudhury (2015), and the drop-in-deviance test. Coefficients were estimated using iteratively reweighted least squares.

Reproducibility

All analyses performed in this manuscript are reproduced in the R markdown file Code_Markdown.Rmd. To reproduce the results exactly, the analysis must be performed on the cached data, as many of the text posts were scored by hand by human raters.

Results

Several patterns seemed to emerge from the exploratory analysis. Previous work on Reddit mental health boards showed that text posts from suicidal users has decreased readability and higher word counts (de Choudhury 2015). Both patterns are replicated in our data (Fig.1). Furthermore, the sentiment of suicidal posts seemed to be much more negative, and the magnitude of the score indicated that suicidal posts were longer (Fig. 1). There were also word choice differences for suicidal and non-suicidal posts. For example, while "die" and "kill" were much more frequently used in suicidal text, non-suicidal text was more likely to contain second-person pronouns like "you" or "your". (To make the predictors appear linearly-related to the log odds, all predictors were log-transformed.)

To start with, we created a full model using all the predictors on a training set containing 75% of the posts, and used backwards stepwise selection and the drop-in-deviance test to choose the best model (Fig. 2). The final model is

$$\begin{aligned} logit(\pi) = \beta_0 + \beta_1 * age + \beta_2 * want + \beta_3 * person + \\ \beta_4 * positive + \beta_5 * die + \beta_6 * anymore + B_7 * life + B_8 * fucking + \\ \beta_9 * secondpronouns + \beta_1 0 * job \end{aligned}$$

where π is the probability of suicidal ideation in the text; $logit(\pi)$ is the log-odds of a text containing suicidal ideation; β_0 is the intercept; $\beta_1, \beta_2 - \beta_8$ are the effects on the log-odds of probability of suicidal ideation in the text associated with a change of 1 unit in the log-counts of the words "want", "person", "positive", "die", "anymore", "life", and "fucking", respectively, assuming the log-counts of the other words are held constant; β_9 is the the effect on the log-odds of probability of suicidal ideation in the text associated with a change of 1 unit in the log-counts of the second-person pronouns; and β_{10} is the effects on the log-odds of probability of suicidal ideation in the text associated with a change of 1 unit in the log-counts of the word "job" or strings containing "employ".

The Flesch-Kincaid Readability Score (which has been linearly transformed to age in the model) was significant (p = 0.006699). A change of one unit in the log-readability score would correspond to a change of 0.5870 (95% CI = 0.172,1.02) in the log-odds of probability of suicidal ideation. That is, for each additional grade-level year in the reading difficulty of a post, assuming all other variables in the model are held constant, we would expect that the odds of suicidal ideation in that post increases by a factor of (exp(0.5870)) = 1.80.

The association between suicidal ideation and the log-counts of the words "want" (p < 2e-16), "die" (p = 1.61e-15), "anymore" (p < 4.15e-06), and second-person pronouns (p = 6.08e-05) is highly significant. For example, an additional unit increase in the log-count of the word "want" in the text would correspond to an increase of 1.6681 (95% CI = (1.28,2.07)) in the log-odds of suicidal ideation in the text, assuming all other variables are held constant, while a unit increase in the log-count in the second-person pronouns would result in a decrease of -0.4770 (95% CI = 0.172,1.02) in the log-odds probability of suicidal ideation. For details, see Table 1.

To assess the model fit and detect influential points, we performed several diagnostic tests. First, we checked for collinearity by calculating the variance inflation factor (VIF) of each of the predictors, and removed predictors from the full model that had VIF >2.5 (UCLA). We also checked for influential points using deviance residuals and hat diagonals. 4 influential points, for which at least one predictor was not significant without it in the dataset, were removed prior to the final model selection. Finally, we assessed the degree to data met the model assumptions by determining which of the predictors were linearly-related to the log-odds of the probability of the text containing suicidal ideation, and removed one predictive variable, the log-counts of the word "kill", that was obviously not linear. For details, see the RMarkdown code.

Finally, to test the validity of the model, we compared its sensitivity and specificity on the training set (n = 1024 posts) and test set (n = 344 posts) from r/depression. The ROC and AUC (Training = 0.8135, Test = 0.8063) were comparable for both sets, indicating the model validity is probably accurate. (Fig.3)

Discussion and Conclusion

Our analysis is limited by several factors. Most significantly, most of the suicidal posts in the dataset were not randomly sampled and may not be representative of the population of suicidal text posts. Because the training set of the randomly sampled hand-classified text was extremely unbalanced, with only 65 suicidal posts and ~700 non-suicidal, the phrase-matching algorithm was used to enrich the set of suicidal texts. This means that the occurrence of the words used in the phrase-matching classifier are overrepresented in the suicidal posts in the training set relative to actual suicidal posts on Reddit, implying that the significance of the coefficients for the words "want","die","anymore", and "life", which are are parts of the phrases used in the phrase-matching algorithm, is perhaps much lower. However, "fucking", job-related words, "positive", "person", and second-person pronouns, remains potentially good predictors, as the phrase-matching algorithm did not contain phrases containing these words. A direction for future work, then, might be to use alternative logistic regression methods that are especially well-suited to classifying unbalanced with rare outcomes (King and Zheng 2009).

Furthermore, three of the the predictors, the log- word counts for 'fucking', 'die', and 'anymore', were not linearly-related to the log-odds of the probability of the text being suicidal. However, they were were retained because the model's predictive power was greatly reduced without them (Drop-in-Deviance = 318.26, Δ df = 4, P($\chi^2 \geq 2.2e-16$), with the AUC for the training set being 0.6842 under the reduced the model. A consequence of including predictor variables in the model that don't satisfy the model assumptions, the model estimates for the coefficient are probably further biased (Department of Statistics, Pennsylvania State University).

Thirdly, we used only the main body of the text of the post to predict suicidal ideation. Information from other variables, like the title of the submission or thread, the number of comments, and the number of upvotes and downvotes may also be viable predictors for a suicide detection text classifier, though adding such predictors may limit the usefulness of such a model to Reddit or similar message boards. Furthermore, our model assumes fixed effects of each of the predictors, whereas a mixed or random effects model may have also been effective. Also, backwards subset selection is not guaranteed to find the optimal model, though it is relatively computationally efficient compared to best subsets selection (Hastie et al 2001). Lastly, we did not check for interactions between the predictors, which may have provided further insight into the composition of suicidal text, as well.

This algorithm can be used to build a large dataset of text posts, classified by the presence of suicidal ideation. Such a dataset would be useful to would be useful to mental health researchers, and could be used answer questions relevant to suicide prevention, such as predicting whether someone will make a suicidal post based on their post history, and to measure the effectiveness of public health interventions on reducing suicidal ideation. The study of internet-based mental health support communities has only just begun, and presents a rich area of future research and opportunities to improve support for people with mental illnesses.

Figures

Table 1

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Nov 01, 2016 - 11:25:05

```
## Waiting for profiling to be done...
```

% Table created by stargazer v.5.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu
% Date and time: Tue, Nov 01, 2016 - 11:25:05

Tables of coefficient estimates, significance, and confidence intervals.

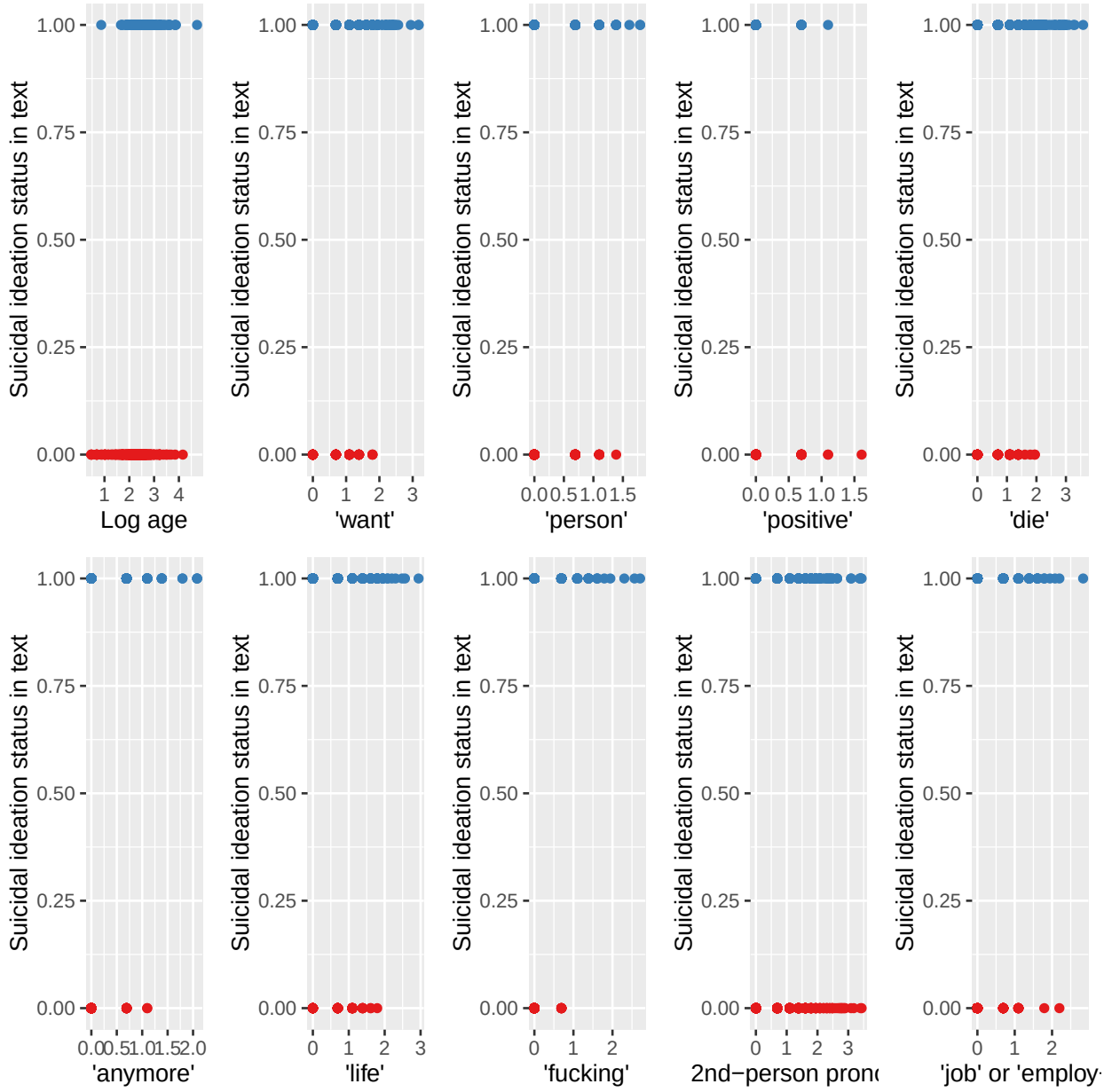

Figure 1: **Logistic regression model predictive variables.** The y-axis is the suicidal ideation status of the text, with 1 being suicidal and 0 being non-suicidal. The x-axes are the log-readability score linearly transformed into age, or the log-counts of the predictive word.

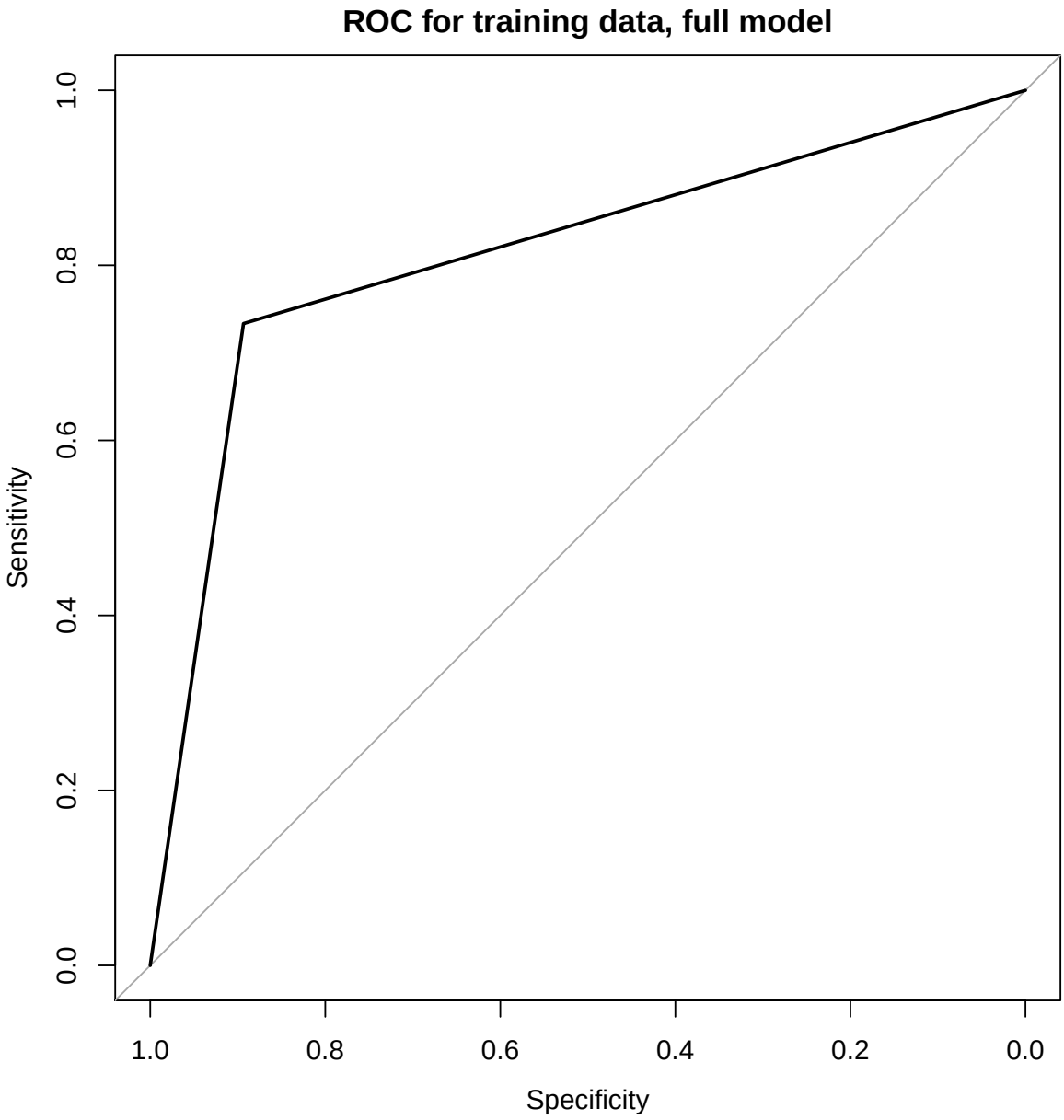

Figure 2: **ROC curves for the full model on the training set.** The ROC is clearly much lower for the reduced model, while it is about the same for the training and test sets for the full model.

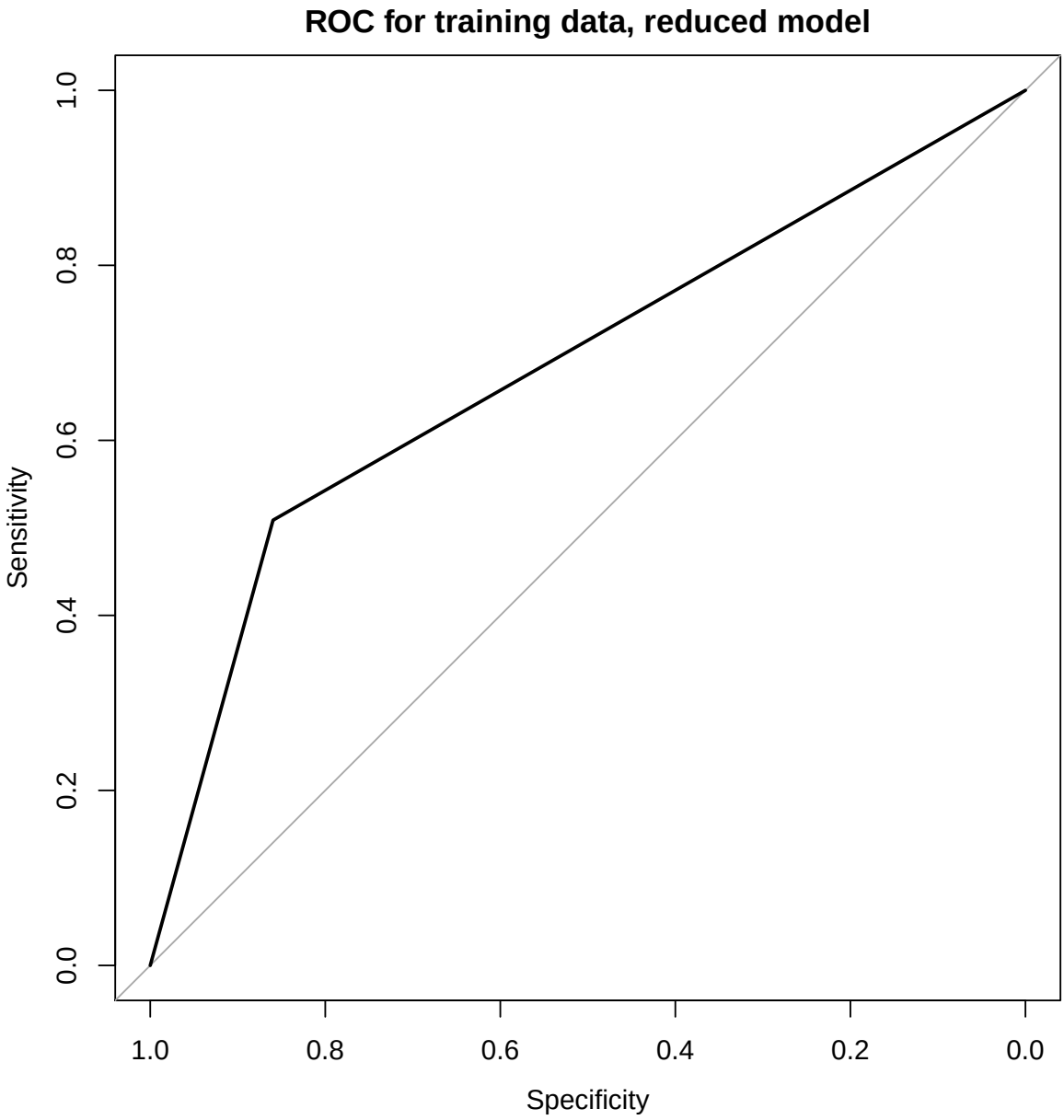

Figure 3: **ROC curves for the reduced model on the training set. The ROC is clearly much lower for the reduced model, while it is about the same for the training and test sets for the full model.

Table 1:

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-2.905	0.524	-5.544	0.00000
age	0.587	0.216	2.711	0.007
want_word	1.668	0.200	8.329	0
person_word	-0.829	0.354	-2.343	0.019
positive_word	-2.507	0.760	-3.298	0.001
die_word	1.278	0.160	7.968	0
anymore_word	2.039	0.443	4.604	0.00000
life_word	1.098	0.222	4.936	0.00000
fucking_word	2.298	0.613	3.748	0.0002
sec_pronouns	-0.477	0.119	-4.009	0.0001
job_words	-0.521	0.324	-1.609	0.108

Table 2:

	2.5 %	97.5 %
(Intercept)	-3.967	-1.911
age	0.172	1.022
want_word	1.283	2.069
person_word	-1.529	-0.139
positive_word	-4.131	-1.109
die_word	0.969	1.599
anymore_word	1.207	2.951
life_word	0.667	1.540
fucking_word	1.177	3.600
sec_pronouns	-0.715	-0.248
job_words	-1.157	0.116

Works Cited

De Choudhury, Munmun, et al. "Discovering shifts to suicidal ideation from mental health content in social media." Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems. ACM, 2016.

King, Gary, and Langche Zeng. "Logistic regression in rare events data." Political analysis 9.2 (2001): 137-163.

Horgan, A., and John Sweeney. "Young students' use of the Internet for mental health information and support." Journal of psychiatric and mental health nursing 17.2 (2010): 117-123.

Hand, David J. "Classifier technology and the illusion of progress." Statistical science 21.1 (2006): 1-14.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction: With 200 Full-color Illustrations. New York: Springer, 2001. Print.

McCullough, Peter, and John A. Nelder. "Generalized linear models." (1989).

Pennsylvania State University, Department of Statistics. "7.2.1 - Model Diagnostics." STAT 504: Analysis of Discrete Data. Pennsylvania State University, Department of Statistics. Web.

UCLA."Chapter 3 Logistic Regression Diagnostics." Institute for Digital Research and Education, UCLA. Web. 31 Oct. 2016.

Ruiz, Anne, and Nathalie Villa. "Storms prediction: Logistic regression vs random forest for unbalanced data." Case Studies in Business, Industry and Government Statistics 1.2 (2007): 91-101.

WHO. "Suicide Data." World Health Organization. World Health Organization, n.d. Web. 28 Oct. 2016.