2018-01-01

Estimating The Optimal Cutoff Point For Logistic Regression

Zheng Zhang
University of Texas at El Paso, zhaaeeng@gmail.com

ESTIMATING THE OPTIMAL CUTOFF POINT FOR LOGISTIC REGRESSION

ZHENG ZHANG

Master's Program in Mathematical Sciences

APPROVED:

Xiaogang Su, Ph.D., Chair.

Feixue Xie, Ph.D.

Michael Pokojovy, Ph.D.

Charles Ambler, Ph.D.
Dean of the Graduate School

Dedication

to my

MOTHER and FATHER

with love

ESTIMATING THE OPTIMAL CUTOFF POINT FOR LOGISTIC REGRESSION

by

ZHENG ZHANG

THESIS

Presented to the Faculty of the Graduate School of

The University of Texas at El Paso

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE

Department of Mathematical Sciences

THE UNIVERSITY OF TEXAS AT EL PASO

August 2018

Acknowledgements

I would like to express my deep-felt gratitude to my advisor, Dr. Xiaogang Su, of the Mathmatical Science Department at The University of Texas at El Paso, for his advice, encouragement, and patience.

I also want to thank the rest two members of my committee, Dr. Feixue Xie of the College of Business and Dr. Michael Pokojovy of the Mathematical Science Department, both at The University of Texas at El Paso. Their suggestions and comments were precious to this work.

In addition, I wish to thank all professors and staff in the University of Texas at El Paso for all their hard work and dedication, providing me the means to complete my degree.

At last, I must thank my family. Thank you so much for your continuing support.

NOTE: This thesis was submitted to my Supervising Committee on the June 12, 2018.

Abstract

Binary classification is one of the main themes of supervised learning. This research is concerned about determining the optimal cutoff point for the continuous-scaled outcomes (e.g., predicted probabilities) resulting from a classifier such as logistic regression. We make note of the fact that the cutoff point obtained from various methods is a statistic, which can be unstable with substantial variation. Nevertheless, due partly to complexity involved in estimating the cutpoint, there has been no formal study on the variance or standard error of the estimated cutoff point.

In this thesis, a bootstrap aggregation method is put forward to estimate the optimal cutoff point c. In our approach, the out-of-bag samples facilitate a natural way of performing cross validation when computing the predicted probabilities. The ensemble learning helps reduce the variation in the estimated cutoff point. Furthermore, we are able to compute the standard error of the estimated cutoff point conveniently via the infinitesimal jackknife method, a by-product of the bootstrap aggregation method without adding much to the computational cost. Accordingly, valid confidence intervals for c can be constructed.

Extensive simulation studies are conducted to investigate the performance of the proposed method and gain experience with its usage. Throughout the research, our focus is restricted to logistic regression. While the bootstrap aggregation method yields valid and promising results in general, a number of interesting observations and useful conclusions are drawn from the simulated experiments. For an empirical illustration, we applied our proposed method to a viral news detection study. Logistic regression analysis results in meaningful findings and accurate predictions with an appropriate choice of the cutoff point.

Table of Contents

List of Tables

List of Figures

Chapter 1

Background

1.1 Cutoff Point for Classification Problems

In the field of data mining and machine learning, classification is one of the main themes of supervised learning where the response or target variable is categorical with two or more levels. Throughout this thesis, we shall focus on binary classification problems only. See, e.g., [1] for how to extend binary classifiers to multiclass problems. A *binary classifier* aims to classify each observation into one of the two categories. Nevertheless, the outputs from most classifiers present themselves as the predicted probabilities. A cutoff point (or cutpoint in short) or threshold value is applied to these probabilities to finally allocate an observation to a particular class.

Take the widely used binary linear classifier, logistic regression, as an example. For binary responses 1 and 0, after obtaining the estimated probability that an observation belongs to 1 or 0, one needs to pick up a decimal number c between 0 and 1 so that the observation is assigned to class 1 if the calculated probability is greater than c and 0 otherwise.

1.2 Determination of Cutoff point c

Determining the cutoff point is clearly important in decision-making process for many application fields. In economic and business applications, we aim at predicting a phenomenon correctly so that we can control the risk or seek to identify new potential business opportunities. For example, we want to predict whether or not a new piece of online news will

become popular or make a hit with readers before or just right after its release. Accordingly, we can sell the advertisement positions at a higher price (for website owners) or make our advertisement campaigns more effective with lower cost for advertisers.

In many cases, from the perspective of optimal Bayes classifier, $c = 0.5$ is a common choice for the cutoff point, often referred to as the majority voting classification rule. Nevertheless, this choice can be quite erroneous depending on the performance criterion and/or the data sources.

Any stochastic decision making process is subject to two types of errors, false positive (FP) and false negative (FN). If we regard "popular" as positive, the false positive scenario means a piece of news is not popular but we assign it to the popular class. One the other hand, a false negative error occurs where a piece of popular news is wrongly assigned to the non-popular class.

Numerous procedures have been developed to determine the optimal cutoff point c by controlling both types of errors; see, e.g., [10]. Among these, the ROC (Receiver Operating Characteristic) curve is a graphical method of displaying the accuracy of a diagnostic test result by taking all distinctive cutpoints into account. The Youden Index (see [18]) is another approach that considers both the true positive rate (also known as sensitivity, Se) and true negative rate (also known as specificity, Sp).

1.3 Problem Statement

While finding the optimal cutoff points have been extensively discussed, there are two remaining problems. Firstly, the cutoff point c is a statistic estimated from data. This estimator can be very unstable with large variation. Hence, variance reduction techniques can be helpful in improving the precision of the cutpoint estimate. Secondly, partly owing to complexity involved in estimating the cutpoint, there has not been any formal study on its variance or standard errors. As a result, no statistical inference is available for the estimated cutpoint in various methods. Ideally, it is hoped to construct a confidence

interval for the estimated cutoff point.

In this thesis research, bootstrap aggregation will be considered for variance reduction in estimating the optimal cutoff point c. In our proposed ensemble learning method, the out-of-bag (OOB) samples facilitate a natural way of performing cross validation in computing the predicted probabilities. Moreover, a confidence interval for c can be conveniently constructed via the infinitesimal jackknife (IJ; see [5]) method. IJ is computed as a by-product of the bootstrap aggregation method and thus adds nothing much to the computational cost.

1.4 Organizations of Following Chapters

The remainder of the thesis is organized as follows. In the next chapter, we review existing methods on finding optimal cutoff points for classifiers and introduce the study on on news popularity, whose aim is to detect "viral" online news with top popularity. In Chapter 3, our proposed methods are discussed in detail. Specifically, a bootstrap aggregation method is proposed to reduce the variability in the estimated cutpoint c, as well as the way of obtaining its standard error (SE) via Infinitesimal Jackknife. Chapter 4 contains simulation experiments designed to investigate performances of the proposed methods in different scenarios, and assess the validity of the SE formula. In Chapter 5, we illustrate our approach with the real data set about online news popularity. Charter 6 ends the thesis with conclusions and future research work.

Chapter 2

Literature Review

In this chapter, we provide a literature review on the best cutoff points and their applications.

2.1 Optimal Cutoff points Selection Methods

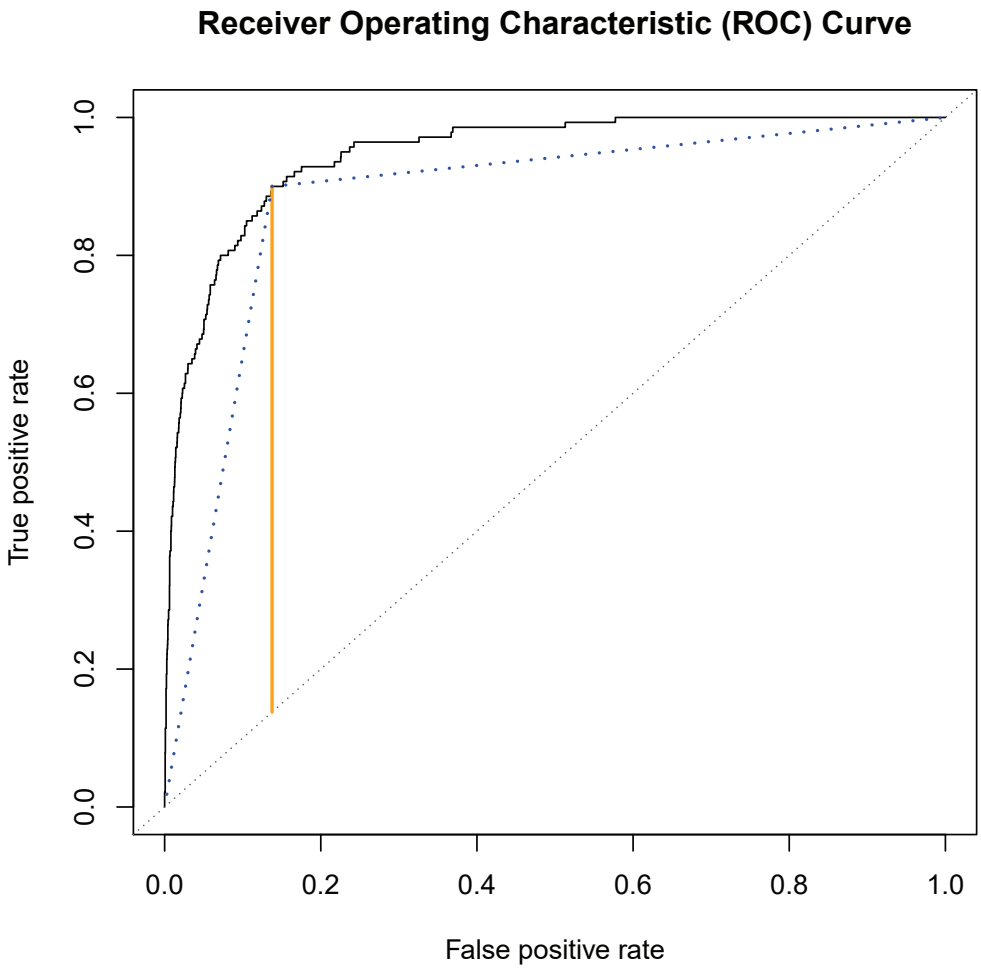
To decrease the possibility of making wrong decisions, determining a cutpoint for a quantitative variable is a common problem, and has indeed been an active area of study. To find an optimal cutoff point on a continuous region, many studies (see [11, 9]) concentrate on receiver operating characteristic curve, which is a graph that measures the diagnostic ability of a binary classifier globally.

The ROC curve is created by plotting and connecting all corresponding true positive rate (TPR) against the false positive rate (FPR). The true positive rate is also known as sensitivity. The false positive rate is obtained by subtracting specificity from 1.

The area under the curve (AUC), ranging from 0.5 to 1, is a measure of the usefulness of the classifier. The larger the AUC, the better classifier we have.

Based on the ultimate goal, several cretiria, such as "Youden Index" (see [18]) and proportion of correctly classified cases (see [6]) have been proposed in existed studies. Youden Index is defined as $Sensitivity + Specificity - 1$, a measurement considering both true positive rate and true negative rate. In Figure 2.1, this index is quivalent to the orange vertical segment, making the area under the curve subtended by a single operating point. The prediciton accuracy (or efficiency) is obtained by $p_+ \cdot Sensitivity + (1 - p_+) \cdot Specificity$, where p_+ is the probability of positive observations.

Figure 2.1: Receiver Operating Characteristic Curve



Other cretiria are also introduced by assuming certain values, or defining a linear combination or function of both sensitivity and specificity. Moreover, ROC curve allows decision-makers to select optimal cutpoints, based on the prevalence and the relative risks and benefits of the possible decisions.

2.2 R Packages of Best Cutpoints

Important contributions to this issue have been made by the R (see [12]) packages. For example, `pROC` (see [13]) provides a method based on Youden Index. `PresenceAbsence` (see [7]) combines 12 methods including user defined required specificity. `OptimalCutpoints` (see [10]) introduces 34 strategies for selecting optimal cutpoints. It incorporates criteria costs, prevalence, and test accuracy.

As proposed by [10], there are several ways of determining c, depending on the analytic purpose, including: 1) Sensitivity and specificity measures: maximize sensitivity; Youden Index, maximize prediction accuracy...; 2) Predictive values: maximize negative predictive value; Equal positive and negative predictive value...; 3) Diagnostic likelihood ratios: positive likelihood ratio...; 4) Cost-benefit analysis of diagnosis: minimizing misclassification cost; 5) Minimum p value (which is based on a two-sample χ^2 test resulting from dichotomization); and 6) Prevalence-based: closest value to observed prevalence, among others. These 34 methods fit most demands of cutoff points determination.

Nevertheless, there are two remaining problems. To start with, we find that The cutoff point, c, a statistic estimated from data, is unstable. For instance, we take two runs from the same simulation: the first c we obtain is 0.9725, while the second one is 0.7475. In addition, for the random variable c, there is lack of statistical inference. Previous studies calculate the confidence intervals of AUC (see [8, 10]) and of Youden Index (see [14, 19]). However, no research takes the confidence interval for c into account. These drawbacks motivate us to study ways of how to obtain a more stable optimal cutoff point and construct confidence intervals.

2.3 Innovations of Our Method

In our study, we use bootstrap aggregation to increase stability in estimating c. Meanwhile, we take out-of-bag samples to provide internal cross validation (CV) for bootstrap. Besides, we apply Infinitesimal Jackknife to make statistical inference on c.

Chapter 3

Bootstrap Aggregation

Ensemble learning is a machine learning approach in which multiple learners are trained to solve the same problem and the results are then integrated. This paradigm is exemplified by boosting, bagging, and random forest for predictive modeling purposes. An ensemble model often results in a much improved prediction performance compared to a single model. One key to the success of ensemble learning is variance reduction.

Considering this advantage, it is natural to expect that ensemble learning might help with the cutoff point estimation. This motivates us to develop a bootstrap aggregating method to estimate c. Our proposed method has the following attractive considerations. First of all, the bootstrap resampling method facilitates a convenient configuration where model training and prediction can be done separately in estimating the optimal cutoff point c. This is because of the fact that each bootstrap sample has a corresponding independent out-of-bag (OOB) sample. Moreover, the estimated c on basis of ensemble learning is expected to be more precise with reduced variation. Besides, reckoning that any estimated c is a statistic itself, it is statistically desirable that a standard error can be obtained as a reliability measure for the estimator. We will present the proposed method with greater details in ensuing sections.

3.1 Bootstrap and Out-of-bag Samples

Bootstrap is a random sampling method by regarding the observed sample as the "population". It mimics the sampling distribution with the bootstrap distribution, allowing us to assign measures of accuracy (e.g., bias, variance, confidence intervals, among some other

measures) to sample estimates (see [4]).

Let $\mathcal{D}$ denote the sample data, which is of size n. While many variants are available, the standard version of bootstrap is to randomly select n observations with replacement from $\mathcal{D}$. Clearly there are $B = n^n$ possible choices of distinct bootstrap samples in total. Whenever a bootstrap sample is taken, some individuals in the original sample data may not be selected at any time. We can use these left-out examples (expected with a proportion of 37%) to form accurate estimates for several quantities (see [2]). The set of the left-out observations is termed as the out-of-bag (OOB) sample. Meanwhile, the selected ones are called as in-bag samples.

3.2 The Algorithm

Let $\theta(F)$ denote the parameter of interest. The natural plug-in estimator is a functional statistics of form $\theta(\hat{F})$, where $\hat{F}$ denotes the empirical cumulative distribution function (ECDF) of data $\mathcal{D}$ by assigning weight $1/n$ to each observation. In general, the bootstrap aggregation method for estimation works as follows. A total of B bootstrap samples are taken. From each bootstrap sample $\mathcal{D}_b$, a reweighted ECDF $\hat{F}_b$ can be obtained, which amounts to assigning weight p_{bi} to each observation in $\mathcal{D}$. The bootstrap weights p_{bi} satisfies $0 \leq p_{bi} \leq 1$ and $\sum_{i=1}^{n} p_{bi} = 1$ for any $b = 1, \ldots, B$. Accordingly, B bootstrap estimates $\{\theta(\hat{F}_b) : b = 1, \ldots, B\}$ can be computed. The aggregated estimator is simply given as their average

$$\theta(\hat{F}_\cdot) = \frac{1}{B} \sum_{b=1}^{B} \theta(\hat{F}_b).$$

Applying the similar procedure, we can estimate the optimal cutoff point c as follows:

- Take B bootstrap samples;
- Let $\mathcal{D}_b$ denote the b-th bootstrap sample and $\mathcal{D}'_b$ be the b-th out-of-bag (OOB) sample, for $b = 1, \ldots, B$;

- Train the logistic model with $\mathcal{D}_b$ and obtain the fitted probabilities $\hat{\pi}_i$ for $i \in \mathcal{D}_b$. Optionally, one may apply the fitted model to the OOB sample $\mathcal{D}'_b$ and obtain the predicted probabilities $\hat{\pi}_{i'}$ for $i' \in \mathcal{D}'_b$.

- Based on the fitted or predicted probabilities, obtain the ROC curve and the optimal $\hat{c}_b$ for the b bootstrap sample;

- The final optimal threshold $\hat{c}$ is the average over B bootstrap sample: $\hat{c} = \sum_{b=1}^{B} \hat{c}_b/B$.

In the above algorithm, it is worth noting that, with the OOB option, the OOB sample $\mathcal{D}'_b$ provides the convenience for conducting cross-validation so that each cutpoint estimate $\hat{c}_b$ is based on predicted probabilities instead of fitted probabilities. Besides, the aggregated estimator $\theta(\hat{F}_\cdot)$ has a smaller variance than $\theta(\hat{F})$ in general due to the effect of averaging. We expect to see this attractive feature preserved in estimating the cutoff point as well.

3.3 Estimating $\mathbf{var}(\hat{c})$

The variance and SE for $\hat{c}$ can be obtained via the Infinitesimal Jackknife method (see [5, 17, 15]). The IJ method is equivalent to the nonparametric delta method as well as the influence function approach; see Efron (1982).

Without the OOB option, the IJ estimate of variance of $\hat{c}$ for n observations is given by:

$$\hat{V} = \sum_{i=1}^{n} \bar{Z}_i^2, \tag{3.1}$$

where $\bar{Z}_i = \sum_{b=1}^{B} Z_{bi}/B$; $Z_{bi} = (N_{bi} - 1)(\hat{c}_b - \hat{c})$ with N_{bi} being the number of times that the i-th observation appears in the b-th bootstrap resample and satisfying $\sum_{i=1}^{n} N_{bi} = n$. That is, the quantity $\bar{Z}_i$ is the bootstrap covariance between N_{bi} and $\hat{c}_b$. We note that, strictly speaking, the above SE formula is not directly applicable when the OOB option is used.

3.3.1 Bias Correction

In practice, $\hat{V}$ is biased upwards due to additional Monte Carlo noise, especially when B is small or moderate (see [15]). Therefore, we consider a bias correction procedure. Following the step proposed by [5], a bias-corrected version is given below:

$$\hat{V}_{unbiased} = \hat{V} - \frac{1}{B^2}\sum_{i=1}^{n}\sum_{b=1}^{B}(Z_{bi} - \bar{Z}_i)^2. \tag{3.2}$$

By further assuming approximate independence of N_{bi} and $\hat{c}$ in Z_{bi} (see [17]), another bias-corrected version is given by:

$$\hat{V}_{unbiased} = \hat{V} - \frac{n-1}{B^2}\sum_{b=1}^{B}(\hat{c}_b - \hat{c})^2. \tag{3.3}$$

The validity of the above-mentioned SE formulas will be assessed by simulation in Chapter 4. The SE given in (3.3) is particularly recommended because of its simpler computation than (3.2).

3.4 Pros and Cons

In this last section, we summarize some key features of our method and discuss its pros and cons.

Bootstrapping makes it straightforward to derive estimates of SE and confidence intervals. It is more precise than the standard intervals that are based on sample variance and normality assumptions (see [3]). The ensemble learning enjoys a variance reduction in estimating c.

Essentially, IJ applies the nonparametric delta method (see [5]). In this method, the number B of bootstrap samples needs to be large to have valid SE estimates according to [5]. We will investigate on this via simularion in Section 4. The bias-corrected SE formulas in (3.2) and (3.3) generally lead to similar results, both with superior performance to the uncorrected version (3.1) (see [15]). IJ is computed as a by-product of the ensemble learning, without adding much to the computational burden.

The out-of-bag examples can be used to provide internal cross-validations. In our bootstrapping method, we make the use of the OOB flexible. In other words, whether to use OOB or not would be optional. After all, the cross validation is uncommon in the current practice of estimating c. Besides, use of the OOB sample would entail development of new SE formula, which is beyond the scope of this thesis research. We will investigate the proposed method using either in-bag or out-of-bag samples for the calculation and compare their performances in the subsequent chapters.

As for cons of our approach, the whole procedure is computationally intensive. For this reason, we will not investigate the empirical size, power, and coverage based on the estimated SE since the involved computation would be overwhelmingly formidable. Besides, the asymptotic normality of the estimated cutoff via bootstrap aggregation entails further exploration.

Chapter 4

Simulation Studies

In this chapter we will run several simulations to verify our method. Various scenarios will be set to make comparations. Additionaly, the validity of Standard Error formula will also be assessed during the process.

4.1 Data Generations

In this section, the following regression model is introduced:

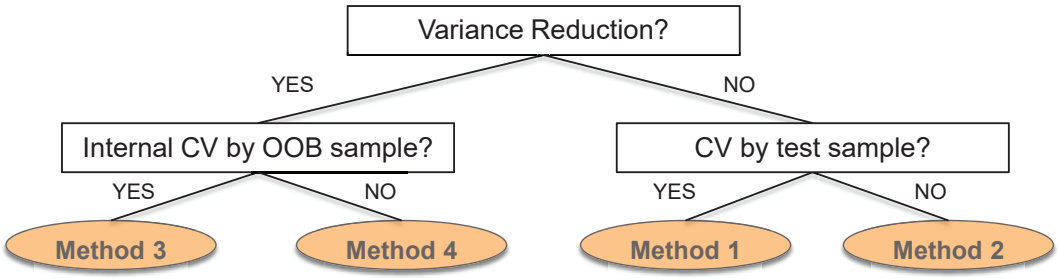
$$\text{logit}\{\Pr(y = 1)\} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \tag{4.1}$$

The data are generated with the following scheme. At first, we simulate $x_j \sim \text{unif}[0, 1]$ for $j = 1, \ldots, 5$ independently. Then we compute $\pi = \Pr(y = 1|\mathbf{x}) = \text{expit}(\mathbf{x}^T\boldsymbol{\beta})$. and simulate $y \sim \text{Bernoulli}(\pi)$. After those procedures, we repeat the above procedure for n times.

4.2 Simulation Settings

We initialize different settings by controlling the following variables: 1)Sample size: small: $n = 100$, medium: $n = 500$, and large: $n = 2000$. 2) Signal strength: stronger: $\boldsymbol{\beta}_s = (3,\ -3,\ 3,\ -3,\ 3)^T$, and weaker: $\boldsymbol{\beta}_w = (-0.5, 0.6, -0.5, 0.6, 0.5)^T$. 3) Number of bootstrap samples, B. Each setting will be run for 200 times.

Figure 4.1: Simulation Methods



4.3 Simulation Methods

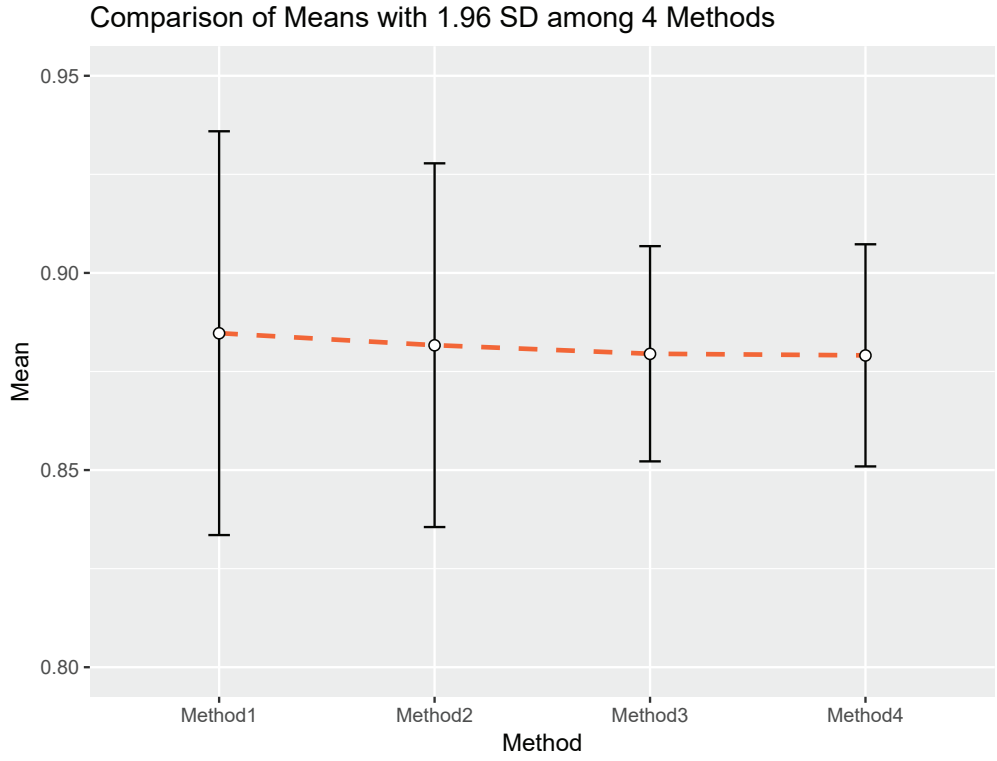
We consider four methods to implement the simulations (see Figure 4.1). Each method will give us a distribution of cutoff points. Method 1 uses the test sample to find the cutoff points. While the rest methods, reuse the training sample again. For Method 2, we simply run the optimal cutpoints codes without adding bootstrap method. For Method 3 and Method 4, we apply the bootstrap samples to make the resampling process. The difference between Method 3 and Method 4 is that the former method uses out-of-bag samples but the latter one does not.

4.4 Performances of Each Method

In this section, we set the model as large sample size: $n = 2000$, stronger signal $\boldsymbol{\beta}_s$, $B = 5000$). Based on the Youden Index criterion, we find the cutoff points of each run, and calculate the means and standard deviations (SDs).

Several findings can be marked from the two comparison figures of all the methods. 1) 0.5, the default value, is covered by the 95% confidence inteval of $\hat{c}$ in none of these methods. 2)Bootstrap aggregations can reduce the variation. 3) The averages do not vary a lot from method to method. 4) Cross validation does not affect the results remarkably.

Figure 4.2: Comparisons of Method Performances



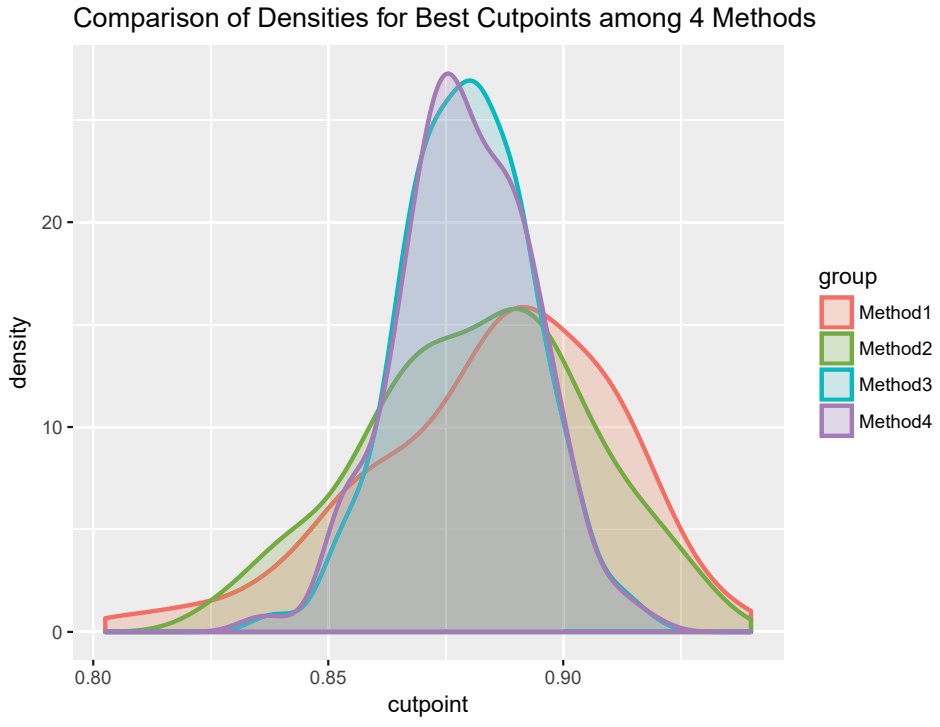
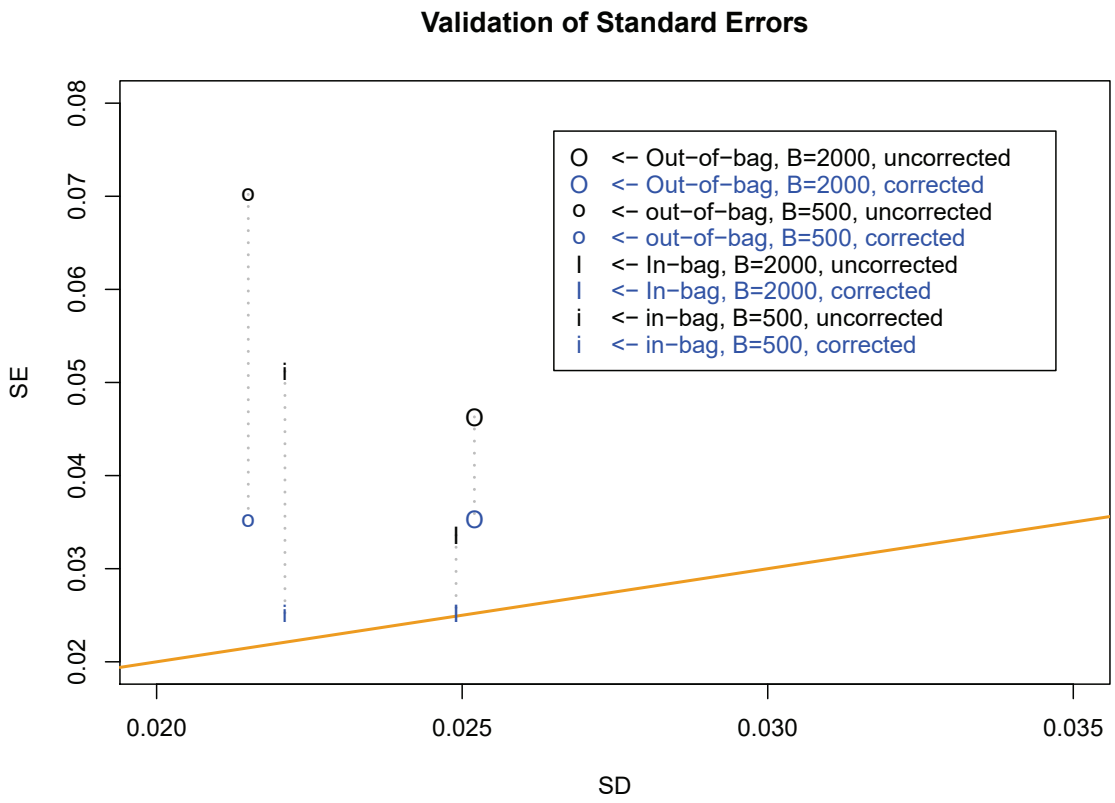


Figure 4.3: SE Validation (One Case)



4.5 Comparisons of Different Settings

Table 4.1 shows the results in different scenarios.

We notice that the simulations fail to converge for small n with stronger signal $\boldsymbol{\beta}_s$ (shown as N/A in the table) and small B (in most cases, $B < 300$). Besides, standard deviation becomes smaller for larger sample size. When the signal strengths are different, the means of the cutpoints change a lot but the SDs are not affected remarkably.

Table 4.1: Comparason of Different Simulation Settings

	Settings		Weaker Signal $\boldsymbol{\beta}_w$		Stronger Signal $\boldsymbol{\beta}_s$	
Bootstrap	Method	Sample	Mean	SD	Mean	SD
$B = 500$	Method 1	$n = 100$	0.5320	0.1151	N/A	N/A
		$n = 500$	0.5257	0.0490	0.8844	0.0428
		$n = 2000$	0.5246	0.0264	0.8853	0.0270
	Method 2	$n = 100$	0.5186	0.0748	N/A	N/A
		$n = 500$	0.5223	0.0415	0.8779	0.0373
		$n = 2000$	0.5212	0.0238	0.8849	0.0252
	Method 3	$n = 100$	0.5267	0.0603	N/A	N/A
		$n = 500$	0.5244	0.0687	0.8775	0.0215
		$n = 2000$	0.5223	0.0151	0.8815	0.0149
	Method 4	$n = 100$	0.5246	0.0516	N/A	N/A
		$n = 500$	0.5241	0.0261	0.8736	0.0222
		$n = 2000$	0.5222	0.0147	0.8811	0.0156
$B = 2000$	Method 1	$n = 100$	0.5287	0.1145	N/A	N/A
		$n = 500$	0.5240	0.0473	0.8760	0.0505
		$n = 2000$	0.5227	0.0279	0.8835	0.0271
	Method 2	$n = 100$	0.5181	0.0736	N/A	N/A
		$n = 500$	0.5248	0.0373	0.8780	0.0422
		$n = 2000$	0.5217	0.0236	0.8813	0.0244
	Method 3	$n = 100$	0.5267	0.0603	N/A	N/A
		$n = 500$	0.5244	0.0687	0.8752	0.0252
		$n = 2000$	0.5223	0.0153	0.8804	0.0134
	Method 4	$n = 100$	0.5253	0.0475	N/A	N/A
		$n = 500$	0.5269	0.0240	0.8716	0.0249
		$n = 2000$	0.5221	0.0148	0.8799	0.0139

4.6 SE Validations

As we propose in the last chapter, our goal is to estimate the standand error by using bootstrap method. The validation in this section helps us to justify if the SE successfully captures the variations, which we can obtain by the standard deviations when implementing simulations. We take $n = 500$ and Stronger signal $\boldsymbol{\beta}_s$ as an example (Note: same conclusions can be drawn from other settings). Figure 4.3 shows eight points with a reference line $SE = SD$. First, we find that the bias-correction is not optional but necessary. We can see the corrected points (in blue color) "drag" the uncorrected ones closer to the orange reference line. Second, the out-of-bag sample does not work successfully. We can see that the labels "O" and "o" are still far away from the reference line. Third, from the in-bag cases, we conclude that larger bootstrap samples give us more precise result, but the smaller ones are acceptable as well.

4.7 Remarks of Simulations

This Chapter explains the features of our method through simulations. Firstly, we discover that in-bag sample works successfully; while out-of-bag samples does not lead to the results. Secondly, Bias correction is not optional but necessary. Without bias correction, the standard error will be overly estimated. Lastly, to obtain a stable SE, we suggest that B, the bootstrap samples, should be as large as 2000.

Chapter 5

Data Examples

For further illustration, in this chapter, we apply the proposed methods to a real data set. We demonstrate how the bootstrap aggregated cutpoint with stability enhancement and confidence interval can be useful in real world applications. At the same time, we elaborate some important specifics of viral news detection via analytical methods.

5.1 Data Description

We obtain the Mashable Online News Popularity data from the UCI Machine Learning Repository (`http://archive.ics.uci.edu/ml/datasets.html`). Figure 5.1 indicates that each webpage allows the news to be shared through different social media such as Facebook and Twitter. The data set has the dimension of 39,644 $\times$ 61, meaning 39,644 pieces of online news and 61 variables. The original target variable is the number of shares of the news since its release. Besides the target and two other non-predictive variables, the data set has 58 predictors or attributes, among which 14 are categorical and 44 are continuous.

The 58 predictors capture the following eight categories of characteristics of each news piece: number of words in titles, contents; number of non-word elements on the page (links, videos...); channels (world, business...); shares of keywords (max, avg, min); time (Monday to Sunday, weekend); closeness of top 5 topics estimated by Latent Dirichlet Allocation (LDA) model; subjectivity; and polarity (positive/negative).

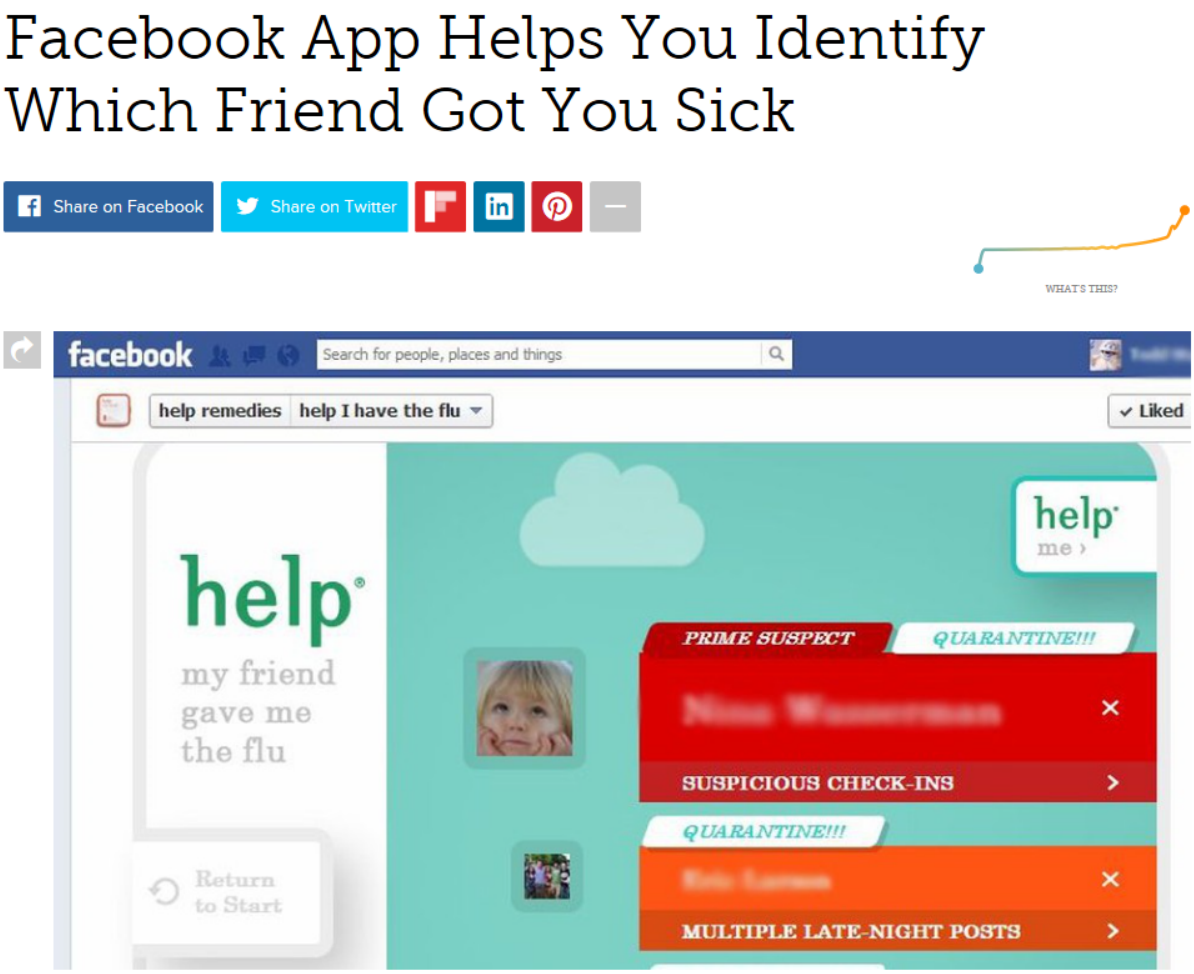


Figure 5.1: Screenshot of a Piece of News on Mashable

5.2 Data Preparation

In this study, we shall focus only on news that had been released for more than 1 year upon data collection. Specifically, a subset of the news data was extracted by restricting to the time period from Jan 7, 2013 to Jan 7, 2014. This one-year restriction ensures a long enough time period for readers to respond but also takes into account the lifespan of news stories.

We formally define *viral news* as a news piece that makes it to the top 5% list in terms of the total number of sharings since its release. This induces a 0-1 binary target variable y, with value 1 indicating the viral news status and 0 otherwise.

We start the analysis with extensive exploratory data analysis. There is no missing value in this data set. We removed observations that are posted less than 1 year since we assume that prolonged time has little effect on its sharing after being released for more than 365 days. We verified this assumption partially by checking the sharing frequencies beyond one year. Next, we remove a few highly linearly correlated predictors (i.e., with multicollinearity), including: Sunday, Thursday, and LDA02. Most of them are dummy variables introduced to account for some categorical variables. After that, we marked the TOP 5% popular news (or more than 11,000 shares) as Viral News (1) and Non-viral News (0) for the remaining ones.

Finally, we ended up with a data frame of dimension $18,511 \times 55$. To proceed, the data set is partitioned into the training set, the validation set, and the test set with a ratio about 1:1:1. The proportions of viral news in these three data sets are 5.3%, 4.8%, and 5.0%, respectively.

Besides descriptive statistics and graphical summaries, we used the variable importance ranking feature in random forest for variable screening purpose. The results are shown in the top panel of Figure 5.2. We can see that different variables show quite heterogeneous predictive power in viral news detection. The bottom panel shows a zoomed-in version of the top 10 important variables. It can be seen that the keywords, topics, subjectivity are

among the most important factors. We, nevertheless, did not remove any variable at this stage for the subsequent logistic regression analysis, since $p = 58$ is not too formidable given the power of regularization.

5.3 Fitting Logistic Regression Model

Next, we fit logistic regression models with the training sample. For variable selection purposes, we apply three different selecting methods: Stepwise Regression, MIC (Minimizing approximated Information Criterion; see [16]) and Adaptive LASSO (see [20]). BIC and approximated BIC were used in stepwise selection and MIC. Ten-fold cross-validation was used to determine the best tuning parameter in adaptive LASSO.

Then the fitted results were applied to the validation set. The ROC curves were obtained based on the predicted probabilities; see Figure 5.2. It can be seen that the best logistic model from each method yields quite similar performance. Nevertheless, our final model is the one with the largest AUC or C-statistic, which is the best adaptive LASSO model with AUC 67.82%.

Finally, we refit the best logistic model by pooling the training data and the validation data together. The model fit is summarized in Table 5.1. It can be seen that a total of six variables are present in the final logistic model. Among the six variables, `avg_positive_polarity` and `abs_title_sentiment_polarity` are just under 0.1 significant level. Then we calculate the odds ratio and corresponding 95% confidence interval of each interpreter, as is shown in Table 5.2. Because the intervals of `avg_positive_polarity` and `abs_title_sentiment_polarity` cover 1, we do not consider these two variables are good interpreters. Therefore, we make implications from the rest four variables.

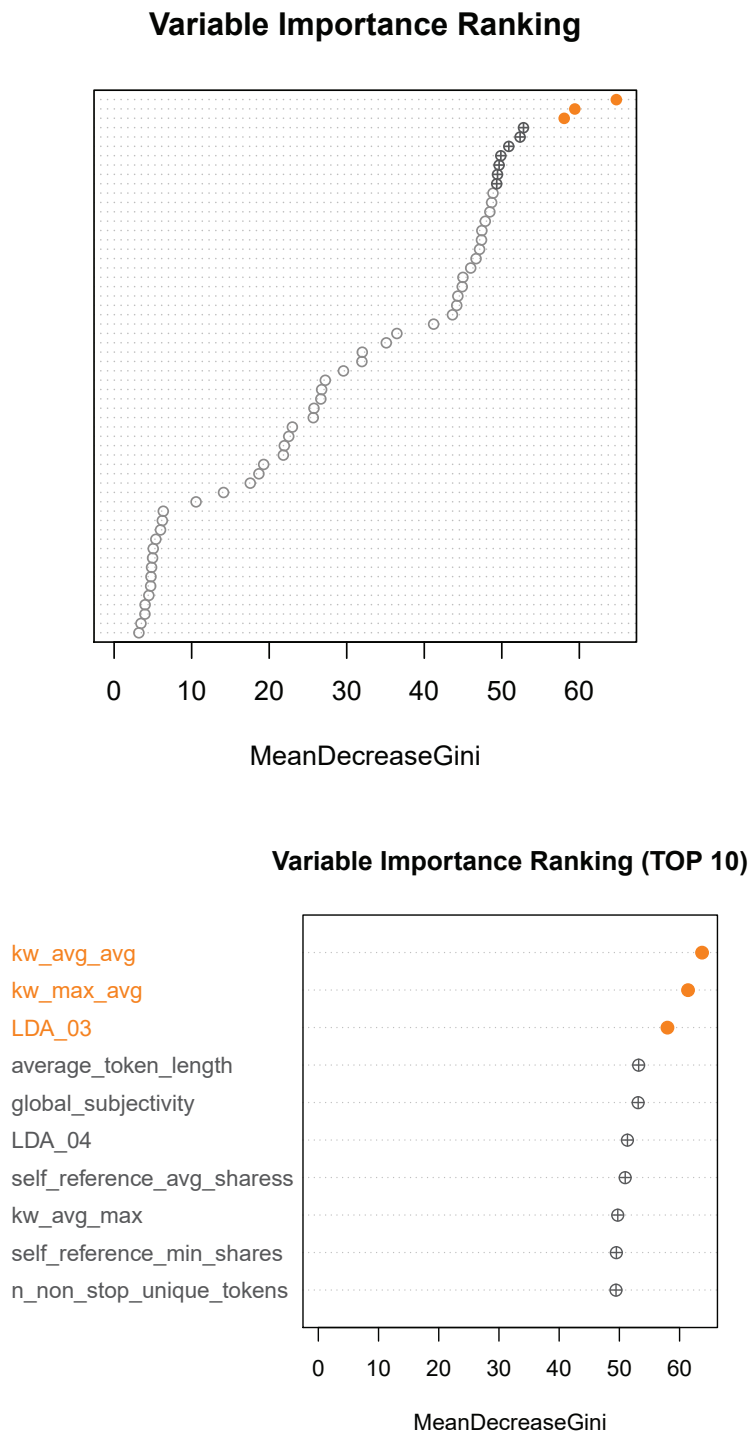


Figure 5.2: Variable Importance Ranking from Random Forests

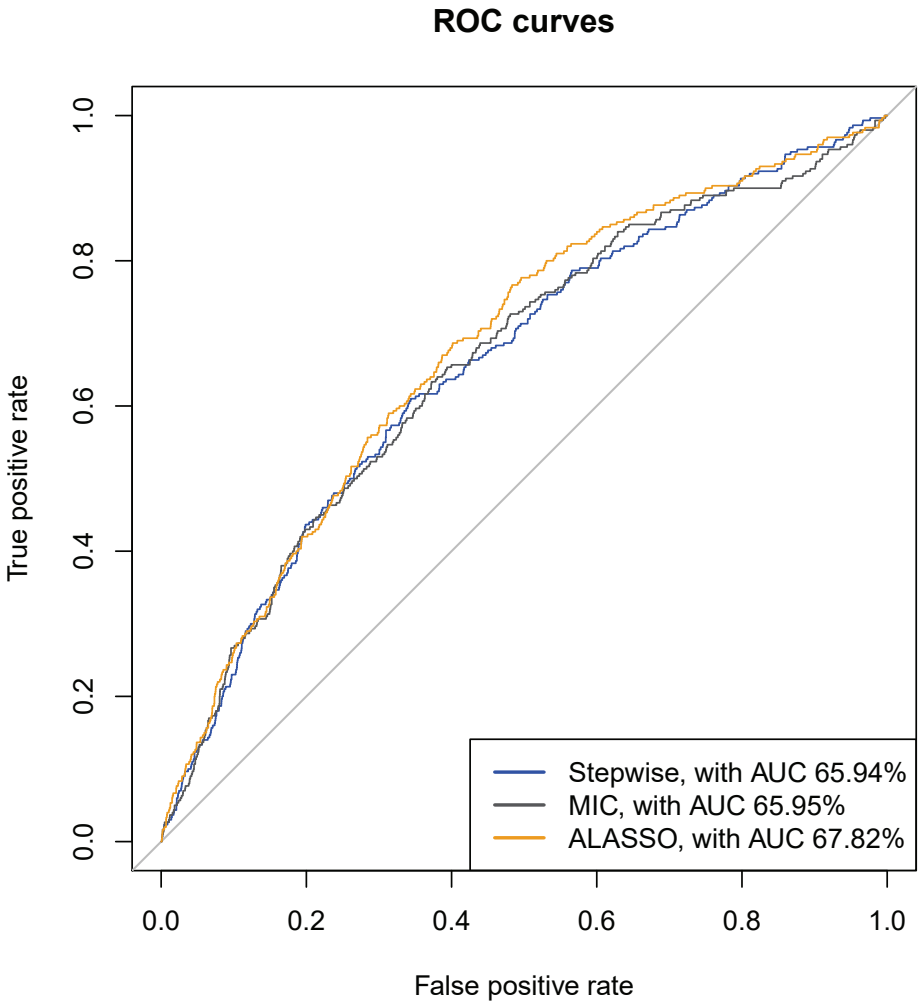


Figure 5.3: ROC Curves for 3 Models

Table 5.1: Final Logistic Regression Model (Sig. Level: ***< 0.001; **< 0.01; .< 0.1)

Variable	Estimate	Std. Error	z value	$\Pr(>\|z\|)$	Sig. Level
(Intercept)	$-5.143e+00$	$2.014e-01$	-25.542	$< 2e-16$	***
kw_avg_avg	$1.620e-04$	$1.948e-05$	8.314	$< 2e-16$	***
LDA_03	$9.777e-01$	$1.107e-01$	8.835	$< 2e-16$	***
global_subjectivity	$1.856e+00$	$4.260e-01$	4.357	$1.32e-05$	***
avg_positive_polarity	$7.399e-01$	$4.268e-01$	1.734	0.08298	.
avg_negative_polarity	$-7.753e-01$	$2.625e-01$	-2.954	0.00314	**
abs_title_sentiment_polarity	$2.462e-01$	$1.371e-01$	1.795	0.07260	.

Table 5.2: Odds Ratio of Each Interpreter

Variable	Odds Ratio	95% Confidence Interval
kw_avg_avg	1.00016	(1.00012, 1.00020)
LDA_03	2.65825	(2.13995, 3.30208)
global_subjectivity	6.39913	(2.77645, 14.74864)
avg_positive_polarity	2.09565	(0.90793, 4.83709)
avg_negative_polarity	0.46056	(0.27535, 0.77035)
abs_title_sentiment_polarity	1.27911	(0.97768, 1.67348)

5.4 Tips for Editors - Remarks of Interpretations

The final logistic regression model is quite meaningful to the editors by informing them how to make the news more viral and providing tips on what should be emphasized and what needs to be avoided in writing a news story.

The variable `LDA_03` corresponds to Topic 03, a factor that enhances the popularity as indicated by the positive sign of the estimated coefficient. From the database, this topic may be summarized as "Internet Plus" for its coverage, whose contents may be extended to Internet-related news such as iPhone, viral video, and teenagers' sending messages. For

example, the following viral news pieces are typical examples of Topic 3:

- "Leaked: More Low-Cost iPhone Photos" (843,300 Shares)
- "Viral Video Shows the Extent of U.S. Wealth Inequality" (617,900 Shares)
- "87% of American Teenagers Send Text Messages Each Month" (139,600 Shares)

Polarity is another crucial determinant of viral news. From the results, we find that a piece of news can earn more shares if it becomes more negative ("worse").

- "MySpace Tom's Twitter Feed Will Make You Feel Worse" (227,300 Shares)

Subjectivity is also an important factor. For example, the following news made use of subjective words like "never" in the title.

- "Your Childhood Will Never Be the Same After These 12 Mashups" (233,400 Shares)

5.5 Predictions and Diagnostics

The raw output from the logistic model, i.e., the predicted probability, can only yield ambiguous answers in the decision making. A cutoff point is needed. Another important concern that we should keep in mind is that sensitivity is of major concern for viral news detection. This is because predictions of viral news are very important not only to editors but also to advertisers. If marketing departments can predict which news will be viral and target such news-releasing webpages accordingly, they can make their advertising campaigns more effective with lower cost.

5.5.1 Cutoff Points Found by Standard Method

By applying the Youden Index and Prediction Accuracy criteria on the test set, we obtained two choices of the cutoff point: $c = 0.0474$ and $c = 0.9823$; see the ROC curve in Figure 5.3 for illustration. Next, we applied our proposed bootstrap aggregation method to estimate

Table 5.3: Cutpoints and Corresponding Confidence Intervals

	Youden Index	Prediction Accuracy
Cutpoints From Standard Ways	$c = 0.0474$	$c = 0.9823$
Cutpoints Obtained via Bootstrap	$c = 0.0449$	$c = 0.7207$
95% CI Constructed by IJ	$(0.0382, 0.0516)$	$(0.4789, 0.9625)$

the cutpoint with the same criteria, which led to $c = 0.0449$ and $c = 0.7207$, respectively. It is worth noting that the estimates based on the Youden Index are quite close while the estimates based on minimum classification error show a substantial difference. The additional advantage of our method is the availability of the confidence interval for the estimated cutoff point c. As shown in Table 5.3 , the 95% bootstrap confidence intervals constructed with IJ cover the cutpoint estimate obtained by the standard ways for Youden Index, but not for Prediction Accuracy. Also, we note that the standard error for c with the prediction accuracy criteria is much larger than that with Youden Index. These results may imply that prediction accuracy for imbalanced data such as viral news detection may not be a reliable criterion. As known to us in this case, a nearly 95% prediction accuracy can be reached by simply assigning all observations to non-viral news, which amounts to setting $c = 0$.

5.5.2 Cutpoints and Corresponding Confidence Intervals Obtained by Proposed Method

To further investigate and compare the performance of the estimated cutoff points, we applied the fitted final logistic model to the test set and obtained a number of performance measures with each estimated cutpoint. For the comparison purpose, we also include the natural choice of $c = 0.5$ induced from the Bayes rule.

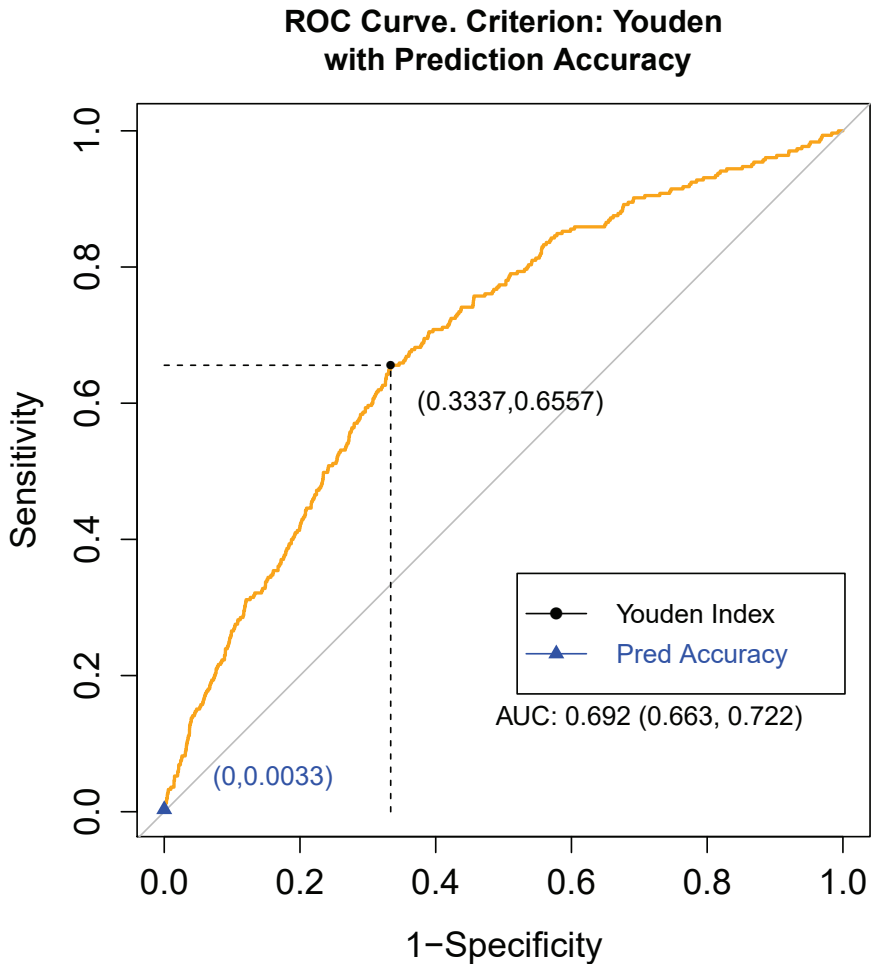


Figure 5.4: Cutpoints Obtained by 2 Methods

Table 5.4: Predicted Outcomes of 4 Cutpoints.

	Youden Index (Standard) $c = 0.0474$		Default (Bayes Rule) $c = 0.5$		Youden Index (Bootstrap with IJ) $c = 0.0449$		Prediction Accuracy (Bootstrap with IJ) $c = 0.7207$	
Predicted	0	1	0	1	0	1	0	1
Observed 0	3890	1948	5833	5	3696	2142	5836	2
Observed 1	105	200	303	2	98	207	304	1
Sensitivity	0.6557		0.0066		<u>0.6787</u>		0.0033	
Specificity	0.6663		0.9991		0.6331		<u>0.9997</u>	
Youden Index	<u>0.3220</u>		0.0057		0.3118		0.0030	
Prediction Accuracy	65.93%		94.99%		63.54%		<u>95.02%</u>	

5.5.3 Comparison among 4 Different Cutpoints

Table 5.4 presents a summary of the results, where the sensitivity, specificity, Youden index, and prediction accuracy are computed for each cutpoint choice. The top winner for each measure is underlined. It can be seen that using $c = 0.5$ blindly can be senseless. The cutpoint estimates based on prediction accuracy do well in terms of specificity, but all have very poor sensitivity. The two choices with Youden Index are quite close, though the standard method yields a slightly higher Youden Index on the test data. The Youden Index based cutpoint sacrifices some in the specificity to boost its sensitivity in predicting viral news. Because of the aforementioned reasons, detecting viral news is more vital than finding common news.

5.6 Summary of Empirical Studies

Detecting viral news is vital in this application. Hence more attention should be paid to sensitivity in viral news detection. The logistic regression analysis gives us quite meaningful interpretations and promising prediction results with an appropriate choice of the cutpoint

estimate. We showed that the proposed bootstrap aggregation method is quite reliable for estimating the optimal cutoff point estimation and leads to convenient confidence intervals without added computational cost.

Concerning tips on how to write potential viral news, Editors may want to select fancy topics which are related to modern lifestyles, for example, by focusing more on "Internet +". Besides, they may not want to be too neutral or objective when writing news. Polar and subjective words can attract more people to read and share.

Chapter 6

Concluding Remarks

6.1 Significance of the Results

Now that we have shown the simulated and real applications of our method, several advantages need to be highlighted. First, Bootstrap aggregation indeed enhances the stability of the cutpoints calculation. Second, as a by-product of bootstrap, we compute SE effectively by applying Infinitesimal Jackknife.

Moreover, we also learn some experiences during these applications: 1) Bias-correction is necessary, 2) B should be large, 3) OOB does not work for SE, 4) It is vital to detect viral news, and 5) Our method is appropriate for viral news detection.

6.2 Future Directions

The problem now shifts to improving the methodology to apply OOB for SE computing. When dealing with data sets via statistical learning methods, it is important to provide cross validation. Thus, the exploration of using OOB is a worthwhile endeavor.

Further, future studies may also focus on generalizing the method to other classifiers and multi-class classification problems, and seeking other interpreters/preditors to make better prediction for viral news.

References

[1] Allwein, E. L., Schapire, R. E., Singer, Y., and Kaelbling, P. (2000). Reducing multiclass to binary: A unifying approach for margin classifiers. Journal of Machine Learning Research, 1:113–141.

[2] Breiman L. Out-of-bag estimation[J]. 1996.

[3] Efron, B., and Tibshirani, R. (1986). Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. Statistical science, 54-75.

[4] Efron, B., and Tibshirani, R. J. (1994). An introduction to the bootstrap. CRC press.

[5] Efron B. (2014) Estimation and accuracy after model selection (with discussion). Journal of the American Statistical Association,109: 991–1007.

[6] Feinstein, S. H. (1975). The accuracy of diver sound localization by pointing. Undersea biomedical research, 2(3), 173-184.

[7] Freeman, E. A., and Moisen, G. (2008). PresenceAbsence: An R package for presence absence analysis. Journal of Statistical Software. 23 (11): 31 p.

[8] Hanley, J.A. and NcNeil, B.J. (1982). 'The Meaning and Use of the Area under a Receiver Operating Characteristic(ROC) Curve.' Radiology, Vol 148, 29-36.

[9] Lahiri, K. , and Yang, L. (2018). Confidence bands for ROC curves with serially dependent data. Journal of Business & Economic Statistics, 36(1), 115-130.

[10] Lopez-Raton, M., Rodriguez-Alvarez, M. X., Cadarso-Suarez, C., and Gude-Sampedro, F. (2014). OptimalCutpoints: an R package for selecting optimal cutpoints in diagnostic tests. Journal of Statistical Software, 61(8), 1-36.

[11] Metz CE (1978). Basic Principles of ROC Analysis." Seminars in Nuclear Medicine, 8, 183-298.

[12] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2017. URL `https://www.R-project.org/`.

[13] Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J. C., and Mller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC bioinformatics, 12(1), 77.

[14] Schisterman, E. F., and Perkins, N. (2007). Confidence intervals for the Youden index and corresponding optimal cut-point. Communications in StatisticsSimulation and Computation, 36(3), 549-563.

[15] Su, X., Pena, A. T., Liu, L., and Levine, R. A. (2017). Random forests of interaction trees for estimating individualized treatment effects in randomized trials. Statistics in medicine.

[16] Su, X. (2018). R Package glmMIC: Sparse estimation of GLM via MIC. Available at GitHub. To install, devtools::install_github("xgsu/glmMIC").

[17] Wager, S., Hastie, T., and Efron, B. (2014). Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. The Journal of Machine Learning Research, 15(1), 1625-1651.

[18] Youden, W. J. (1950). Index for rating diagnostic tests. Cancer, 3(1), 32-35.

[19] Zhou, H. (2011). Statistical inferences for the youden index.

[20] Zou, H. (2006). The adaptive lasso and its oracle properties. Journal of the American statistical association, 101(476), 1418-1429.

Curriculum Vitae

Zheng Zhang was born and raised in Beijing, China, He is the first son of Boli Zhang and Shurong Zheng. He graduated from the High School Affiliated to Beijing Normal University (China) in the summer of 2006. He then went to Capital University of Economics and Business in China and obtained his bachelor degree in economics in 2010. In 2013, he graduated from Tohoku University in Japan with an master degree in environmental studies (concerntration: environmental economics). From 2014 to 2016, he worked as an analyst in Enterprise Risk Management Department at Deloitte China in Beijing. In the fall of 2016, Zheng entered the University of Texas at El Paso. While pursuing his master degree in Statistics, he worked as a Teaching Assistant.

Contact Information: zhaaeeng@gmail.com