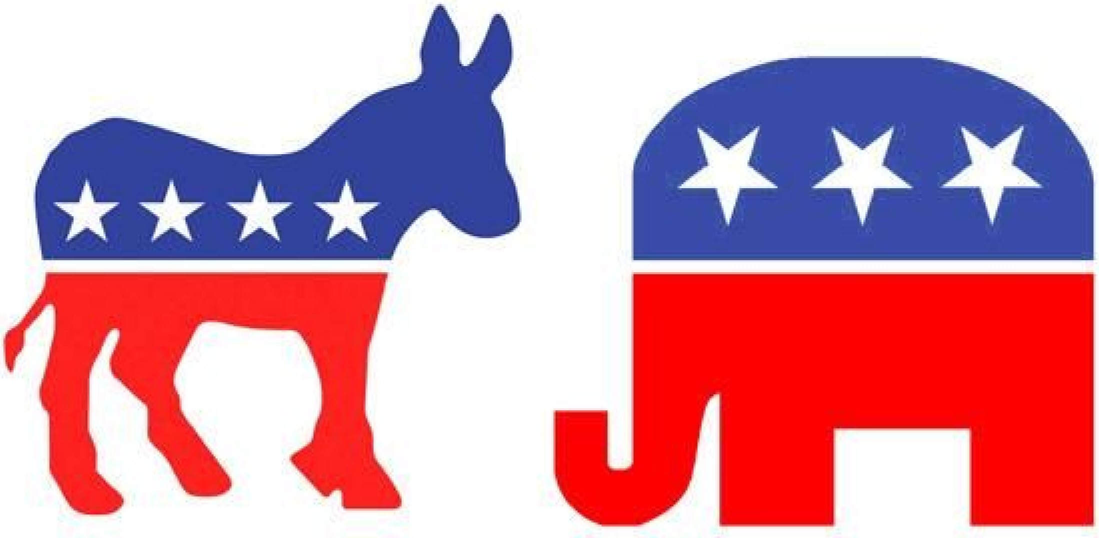


Instructor: Caitlin Steiner

Zhiwei Zhang (zz3px)

Jingyi Sun (js6sm)

Introduction:

Data: endorsement-june-30.csv

Data overview:

The endorsement dataset gives information about candidates' campaign performances in primary through June 30th from each year in 1980 - 2012. It evaluates 109 candidates' performances in terms of their weighted endorsement points, percentage of weighted endorsement points, money raised, percentage of money raised, percentage of primary vote won and primary result.

Background information (United States Presidential Primary):

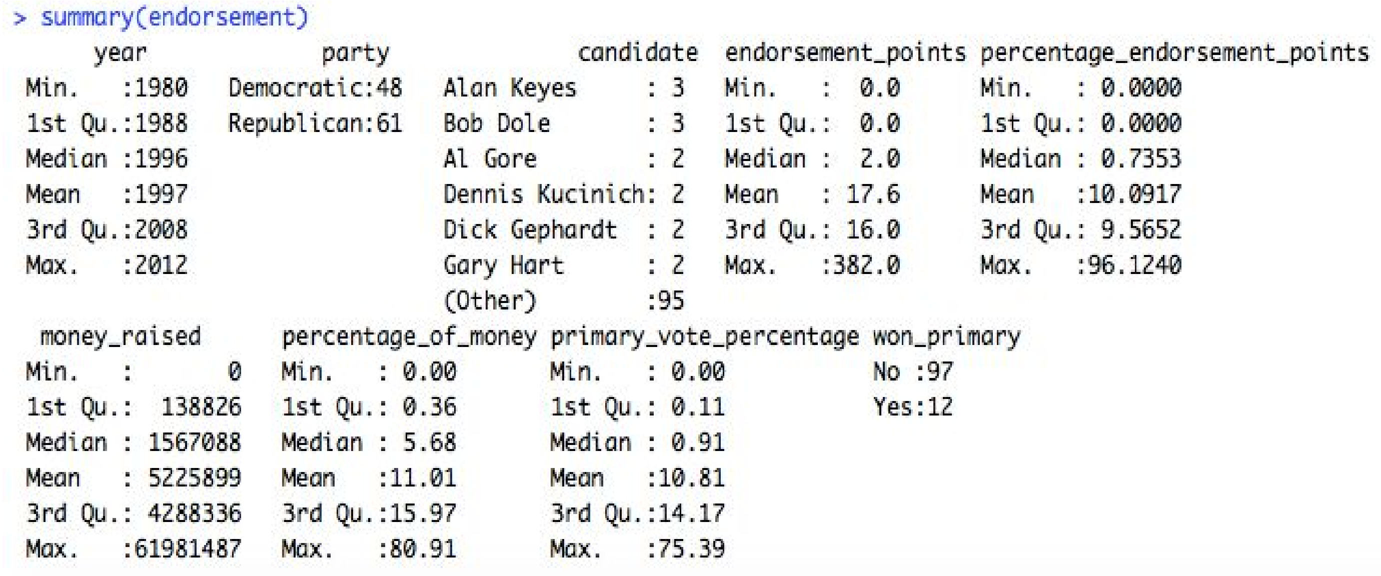
Prior to a general election, there is a selection process to determine which candidate will appear on the ballot for a given political party in the nationwide general election. One way for choosing delegates for to the national convention is primary. In a primary election, registered voters may participate in choosing the candidate for the party's nomination by voting through secret ballot, as in a general election.

Variables:

Predictors that are used to start the model include all the quantitative variables:

1. Endorsement_point: Weighted endorsements through June 30th before primary
2. Percentage_endorsement_points: Percentage of total weighted endorsement points for the candidate's political party through June 30th of the year before the primary
3. Money_raised: Money raised through June 30th of the year before the primary
4. Percentage_of_money: Percentage of total money raised by the candidate's political party through June 30th of the year before the primary
5. Primary_vote_percentage: Percentage of votes won in the primary

Summary Statistics:



```
> summary(endorsement)
      year              party                 candidate   endorsement_points percentage_endorsement_points
 Min.   :1980   Democratic:48   Alan Keyes      : 3   Min.   :  0.0      Min.   : 0.0000
 1st Qu.:1988   Republican:61   Bob Dole        : 3   1st Qu.:  0.0      1st Qu.: 0.0000
 Median :1996                   Al Gore         : 2   Median :  2.0      Median : 0.7353
 Mean   :1997                   Dennis Kucinich: 2   Mean   : 17.6      Mean   :10.0917
 3rd Qu.:2008                   Dick Gephardt   : 2   3rd Qu.: 16.0      3rd Qu.: 9.5652
 Max.   :2012                   Gary Hart       : 2   Max.   :382.0      Max.   :96.1240
                                (Other)         :95
  money_raised      percentage_of_money primary_vote_percentage won_primary
 Min.   :       0   Min.   : 0.00       Min.   : 0.00           No :97
 1st Qu.:  138826   1st Qu.: 0.36       1st Qu.: 0.11           Yes:12
 Median : 1567088   Median : 5.68       Median : 0.91
 Mean   : 5225899   Mean   :11.01       Mean   :10.81
 3rd Qu.: 4288336   3rd Qu.:15.97       3rd Qu.:14.17
 Max.   :61981487   Max.   :80.91       Max.   :75.39
```

Important variables explanations:

Endorsement_point: The value of political endorsements varies, depending on whom they are from, when they are given, and other factors, so the endorsement point used here is an attempt to quantify the importance of endorsements by weighting each one according to the position held by the endorser: 10 points for each governor, 5 points for each senator and 1 point for each representative. The team is interested in how early endorsements were correlated with the success candidates achieved in primary.

Money_raised: Before primary, candidates raise funds for the upcoming primary elections and attempt to garner support of political leaders and donors, the party establishment. The team is interested in how the primary results were affected by the money raised by the candidates before primary.

Questions of Interests:

1. The team is interested in figuring out which variables are especially crucial in predicting the outcome of the primary. In this question, the team will adopt the logistic regression in order to get the predictors.

2. Building on to the model established, the team want to test which predictors have stronger weights in determining the outcome of the primary. Also, the team will examine closely to find out which method is better in testing the model.
3. Lastly, the team is interested in predicting the probability that candidates fall into two classes of the binary response as a function of the explanatory variables. In order to approach that, the team is going to use classification methods, Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and classification tree to predict whether a candidate will win the primary on the basis of their previous performances.

Methods:

Logistic Regression:

The team considers the problem of predicting whether the candidate will win the primary using multiple quantitative predictors. The team first has a logistic regression model that uses all the quantitative variables to predict the probability of winning. Then the team needs to decide on important variables by t-test for individual slope.

H0: There is no relationship between won_primary and the tested variable

H1: There is some relationship between won_primary and the tested variable

Once the coefficients have been estimated, the team can use that to make predictions for the testing data set and compare the results to get error rates for using this method.

Discriminant Analysis:

Both LDA and QDA can be derived from simple probabilistic models which model the class conditional distribution of the data for each class k. Predictions can be obtained by using Bayes' rule and select the class k which maximizes this conditional probability.

- **Linear Discriminant Analysis (LDA):**

  LDA classifier results from assuming that the observations within each class come from a multivariate normal distribution with a class-specific mean vector and a common variance if there are more than 1 predictor.

- **Quadratic Discriminant Analysis (QDA)**:

  QDA classifier results from assuming that the observation within each class came from a normal distribution with a class-specific mean vector, μk, and specific covariance-variance matrix, Σ, and plugging estimates for these parameters into the Bayes classifier δk(x).

Confusion Matrix (Table 2):

The team cares about how well the three classifiers did in classifying the data and how accurate they were in making prediction. A confusion matrix, shown for the data set in Table 2, is a great way to display the information. Elements on the diagonal are those correctly predicted and off-diagonal elements are misclassified by the method.

Five-fold cross validation(Table 4):

The team uses the five-fold cross validation method to estimate the expected prediction error for each method. The team splits set of observations into two parts: the training set is used to train the classifier and the testing set is used to estimate the error rate. The team randomly divides the data set into 5 folds of approximately equal size, one fold would be validation set. The team fits classification method on the K-1 folds, then predicts yhat for validation point and calculate test error, such as MSE. The whole process is repeated K times and the test rate is averaged to give the expected prediction error.

Classification Tree (Figure 2):

A classification tree is very similar to a regression tree, except that it is classification tree used to predict a qualitative response rather than a quantitative one. For a classification tree, the team predicts that each observation belongs to the most commonly occurring class of training observations in the region to which it belongs. In interpreting the results of a classification tree, the team is interested in the class proportions among the training observations that fall into that region.

Bagging, random forests and Boosting (Table 5) :

The team also uses bagging, random forests, and boosting after classification tree in order to compensate for tree's prediction accuracy. Each of these approaches involves producing multiple trees which are then combined to yield a single consensus prediction.

Results:

Logistics regression:

The team decides to take out primary_vote_percentage from the logistic model because it might not be of interest to the team and R cannot converge the algorithm and keep returning warning message when the team includes that variable. The team first has a logistic regression model that uses all the quantitative variables except for primary_vote_percentage. Then the team needs to decide on important variables using t-test for individual slope. Since endorsement points and money raised yield high p-value, 0.923 and 0.981, the team further simplifies the model and keeps only two predictor with relatively smaller p-values, percentage_endorsement_points and percentage_of_money. The team doesn't make a separate testing data set because within 109 observations, there are only 12 yes and 97 no. In order to avoid having randomly chosen testing test that does not include yes in response variable, the team randomly chooses 30 observations as training data and test on the entire endorse data.

	Estimates	P-value
Intercept	-4.34421	5.51e-08
Percentage_endorsement_points	0.07860	0.00121
Percentage_of_money	0.04320	0.10144

Table 1: Coefficients and p-values from logistic regression

Confusion Matrix		Predicted Class		
		Negative	Positive	Total
True Class	Negative	96	1	97
	Positive	4	8	12
	Total	100	9	109

Table 2.1 Confusion Matrix from Logistic Regression

Linear Discriminant Analysis (LDA):

The team also performs LDA on our training data. In R, the team fits an LDA model using the lda() function on response variables and the two predictors the team chooses to include. The LDA output indicates 86.7% of the training observations correspond to not winning the primary and 13.3% correspond to winning the primary. The average of each predictor within each class suggests that there is a tendency for candidate with higher percentage_endorsement_points and higher percentage of money to win the primary.

Confusion Matrix		Predicted Class		
		Negative	Positive	Total
True Class	Negative	92	5	97
	Positive	4	8	12
	Total	96	13	109

Table 2.2 Confusion Matrix from LDA

Quadratic Discriminant Analysis (QDA):

QDA provides an alternative approach.

Confusion Matrix		Predicted Class		
		Negative	Positive	Total
True Class	Negative	84	13	97
	Positive	4	8	12
	Total	88	21	109

Table 2.3 Confusion Matrix from QDA

Classification Error, Sensitivity & Specificity:

Using confusion matrix given above, the team can come up with classification error, sensitivity and specificity for each classifier. Classification error tells us how often it is wrong and use (1 – accuracy) to get classification error in the table; sensitivity tells us when it is actually yes, how often it predicts yes and specificity tells us when it is actually no, how often it predicts no.

	Accuracy	Classification Error	Sensitivity	Specificity
Logistic	0.954128	0.045872	0.666667	0.989691
LDA	0.917431	0.082569	0.666667	0.948454
QDA	0.844037	0.155963	0.666667	0.865980

Table 3 summary of classification error, sensitivity and specificity

ROC Curve (figure 1):

ROC curve summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the true positive value against the false positive value when varying the threshold for assigning the candidates to different classes.

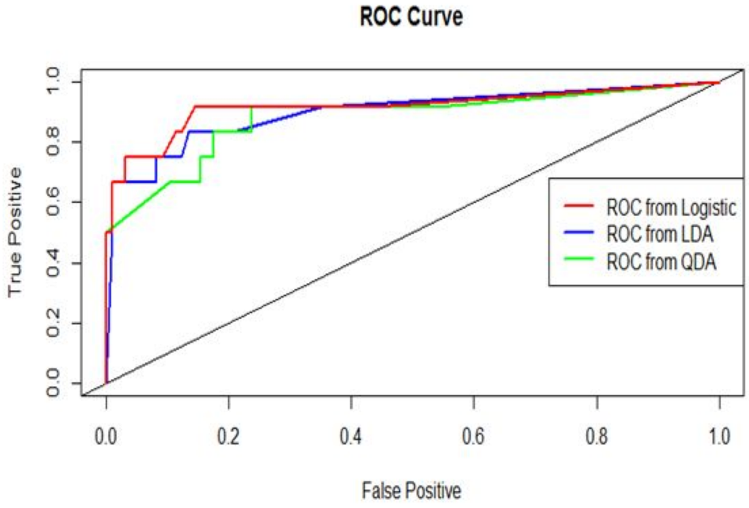


Figure1 : ROC Curve

Five-fold cross validation:

The table below gives expected prediction error which is the measured squared error for each of the three methods using the five-fold cross validation. The smaller the error, the closer the predicted value to the true value. Simplified logistic regression yields the smallest expected prediction error, 5.38%.

	Logistic (simplified)	LDA	QDA
Average Exp pred error	5.383512%	48.74971%	47.51204%

Table 4: The expected prediction error using cross validation method

Classification Tree:

A classification tree is used to predict the qualitative response, so in this case, the team uses classification trees to predict whether a candidate will win the primary based on his or her endorsement points and money raised. The team is using two impurity function: gini and cross-entropy. It turns out that the gini index and the cross-entropy are quite similar numerically.

	Tree	Bagging	Forest	Boosting
Estimators of Prediction Error	0.06422018	0.10091743	0.07339450	0.05504587

Table 5: error rate from Tree, bagging, random forest and boosting

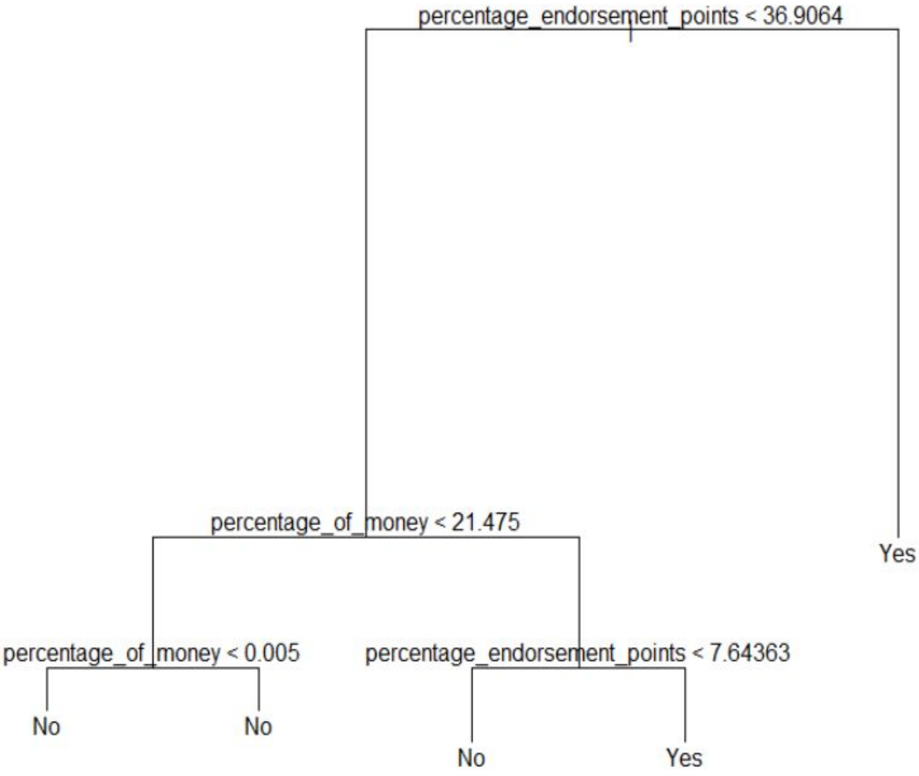

percentage_endorsement_points < 36.9064
percentage_of_money < 21.475
Yes
percentage_of_money < 0.005
percentage_endorsement_points < 7.64363
No
No
No
Yes

Figure 2.1 Classification Tree

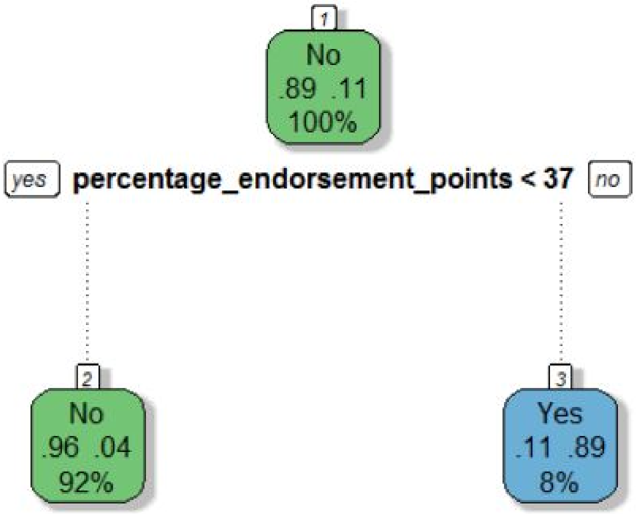


Figure 2.2: simplified plot

Discussion:

Interpretation of logistic regression:

The p-values associated with percentage_endorsement_points and percentage_of_money are relatively small, indicating that each of these variables is associated with the response variable. The positive coefficients for these two variables indicate that both have positive effects on the response variable, increasing percentage_endorsement_points and percentage_of_money will be associated with increasing p(winning the primary). The smaller p-value from percentage_endorsement_points than from percentage_of_money indicates that percentage_of_points is more significant and the team is more confidence about the existence of relationship between percentage_endorsement_points with the response variable.

For the data, estimated coefficients predict the probability of winning the primary using the 2 predictors. For example, estimated coefficient 0.07860 suggests that increasing percentage_endorsement_points by one unit changes the log odds by 0.07860. Once the coefficients have been estimated in the model, it is easier to make predictions by computing the probability of winning the primary for any given percentage_endorsement_points and percentage_of_money.

Interpretation of confusion matrix:

The accuracy tells that overall, how often is the classifier correct, which gives the proportion of total number of predictions that were made correctly. Overall, all these classifiers perform well in terms of making prediction about qualitative response about at 95.4%, 91.7% and 84.4%. Sensitivity and specificity characterize the performance of the classifiers. The higher the rate, the better performance of the method.

- Sensitivity is the percentage of candidates who won the primary that are correctly identified. For example, when LDA predicts a candidate won, there is 66.7% chance that it is correct. Logistic regression, LDA and QDA yield the same rate of 66.7%.
- The specificity is the percentage of candidates who didn’t win the primary that are correctly identified. For example, when QDA predicts a candidate will not win, there is a 86.6% chance that the candidate won’t win. All the three methods perform well in terms

of classifying candidates who will not win the primary, each gives specificity of 98.9%, 94.8% and 86.6%.

All these three classifiers yield higher specificity than sensitivity which means they do much better in classifying candidates who won the primary than classifying those who didn't win the primary.

Interpretation of expected prediction error from cross-validation :

Cross-validation is primarily a way of measuring the predictive performance of statistical model. Ideally, the team would want MSE resulted from the five-fold cross validation method to be small and the team can compare the three models to see which one yields the lowest error. In the data set, the team has response variable = 1 if the candidate wins the primary, response variable = 0 if not. The team predicts whether winning the primary would happen with a threshold of 0.5. Logistic regression yields the smallest mean squared error, 5.3835%; LDA and QDA both have error rates higher than that. If the smaller the Mean Squared Error, the closer the fit is to the true value, the team can conclude that logistic regression gives more accurate prediction than the other two methods.

Interpretation of ROC Curve:

The ROC curves are not smooth which generally means that the model can only provide discrete predictions, rather than continuous. The reason why it occurs is that the data only has 109 observations. The ROC curve can be remedied by adding more samples to our dataset and having more features in the model. The ROC curve hugs the top left corner, indicating a high sensitivity and a low specificity and the closer the curve follows the left-hand border, the larger the area under the ROC curve, the more accurate the model. ROC curve from logistic has larger area under, which means it does a better job the other two methods.

Interpretation of Tree Classification:

Tree-based methods are useful in interpretation but it is not very competitive in terms of prediction accuracy. Hence the team uses bagging, random forests and boosting. Figure 2 from the result section shows a classification tree fit to our endorse data.

It consists of a series of splitting rules, starting at the top of the tree. The top split assigns observations having percentage_endorsement_points < 36.9064% to the left branch and the rest to the right branch. For candidates with percentage_endorsement_points greater than 36.9064%, they will win the primary. The left branch is further divided by percentage_of_money. Candidates with percentage_of_money < 21.475% will not win the primary, and with percentage_of_money > 21.475%, the group will be divided based on percentage_endorsement_points again. Overall, the classification tree stratifies the candidates into three regions of predictor space: candidates who have percentage_endorsement_points >36.9064%, candidates who have percentage_endorsement_points <36.9064% and who have percentage_of_money < 21.475%, and candidates who have percentage_of_money > 21.475% and percentage_endorsement_points <7.65363.

Summary:

1. Based on the logistic regression, the team determines that percentage_money_rasied and percentate_endorsement_points are two variables significant in predicting the outcome of the primary. From the logistic regression, it's easy to see that these two variables yield lower p-value.

   Final model based on logistic regression:

   P(won_primary) =

   e^(-4.34421+0.07860X1 + 0.04320X2) / (1 + e^ (-4.34421+0.07860 X1 + 0.04320 X2)

   Where X1 = percentage_endorsement_points and X2 = percentage_of_money.
2. The smaller p-value from percentage_endorsement_points than from percentage_of_money indicates that percentage_of_points is more significant and the team is more confidence about the existence of relationship between percentage_endorsement_points with the response variable. If p-value alone doesn't indicate the size of the effect, the team decides to take a closer look at the regression coefficient which is a measure of how strongly each predictor influences the response variable Increasing percentage_endorsement_points by one unit changes the log of probability of winning by 0.07860; increasing percentage_of_money by one unit changes the log of probability of winning by 0.0432.

   Conclusion above is also supported by the classification tree. One might interpret the classification tree (figure 2) as follows: percentage_endorsement_points is a more important factor in determining if a candidate will win the primary, and candidate with less endorsement points is less likely to win the primary than candidate with more endorsement points. Given that a candidate has more endorsement points, the money the candidate raised play little role in deciding if he or she will win the primary. But among candidates with percentage_endorsement_points less than 36.9064%, money raised does affect the probability of winning, and candidates who have more money raised before the primary have higher probability of winning.

3. Based on expected prediction error from five-fold cross-validation and what ROC curves implied, simplified logistic regression yields the smallest expected prediction error, 5.38% which means predictions from logistic regression are closer to the true value than predictions from the other two methods. Moreover, ROC curve from logistic has larger area under, which means it does a better job the other two methods.