Discriminant Analysis

Sravani Vadlamani

Agenda

- Logistic Regression vs LDA
- Linear Discriminant Analysis
- Multiple Linear Discriminant Analysis
- Quadratic Discriminant Analysis

Linear Discriminant Analysis (LDA)

Logistic Regression vs LDA

- In logistic regression we model the conditional distribution of the response variable, given the predictor(s) X i.e., $Pr(Y=k \mid X = x)$
- In LDA we model the distribution of the predictors separately in each of the response classes Y ($Pr(X|Y=k)$ and then use Bayes theorem to invert the probabilities to estimate the conditional distribution.
- Reasons to prefer LDA over Logistic regression
  - Parameter estimates from logistic regression are surprisingly unstable when there is a substantial separation between classes. Discriminant analysis approaches do not have this issue
  - Logistic regression estimates are less accurate when the predictors are approximately normally distributed, and the sample size (n) is small

Classification with Bayes' Theorem

- Assume a qualitative variable Y can take on K possible distinct and unordered values
- π_k denotes the prior probability that a randomly chosen observation comes from the kth class.
- The density function of X for an observation that comes from the kth class is given by
- $f_k(x) = Pr(X = x \mid Y = k)$
- $f_k(x)$ is relatively large if there is a high probability that an observation from the kth class features $X = x$.
- Conversely, $f_k(x)$ is small if it is unlikely that an observation from the kth class features $X = x$.

Logistic Regression vs LDA

- In accordance to Bayes theorem, the posterior probability that an observation $X = x$ belongs to the kth class is denoted as p_k and is given by
- $p_k(x) = Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{j=1}^{K} \pi_j f_j(x)}$
- Unlike logistic regression, where we directly calculate the posterior probability, we plug in estimates of $\pi_k \; and \; f_k(x)$ into the above equation.
- π_k - Computed as the fraction of training observations that belong to the kth class.
- $f_k(x)$ - estimation is challenging and requires some assumptions
- A classifier that classifies an observation x to the class for which $p_k(x)$ is the largest will have the lowest possible error rate.
- LDA, QDA and naive Bayes classifiers use different estimates for $f_k(x)$ which we will discuss next

LDA with One Predictor

- When there is only one predictor, we assume that $f_k(x)$ follows a normal or gaussian distribution which takes the form

$$f_k(x) = \frac{1}{\sqrt{2\pi\sigma_k}} exp\left(-\frac{1}{2\sigma_k^2}\left((x-\mu_k)^2\right)\right)$$

- μ_k is the mean for the kth class
- σ_k^2 is the variance for the kth class. LDA assumes that the variance across all K classes is the same and is denoted as σ^2 for simplicity.
- The posterior probability can now be denoted as

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma}} \exp(\frac{1}{2\sigma^2}(x-\mu_k)^2)}{\sum_{j=1}^{K} \pi_j \frac{1}{\sqrt{2\pi\sigma}} \exp(\frac{1}{2\sigma^2}(x-\mu_j)^2)}$$

- π_k denotes the prior probability that an observation belongs to the kth class.

LDA with One Predictor

- Taking the log of the previous equation $p_k(x)$ and rearranging the terms, we obtain the following which allows us to classify an observation by taking the class that yields the largest value.

$$\delta_k(x) = x.\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

- The above discriminant function is linear in terms of x and hence the name of the classifier is LDA
- LDA uses the following estimates for μ_k and σ^2 to estimate the discriminants

LDA with One Predictor

- LDA uses the following estimates for μ_k and σ^2 to estimate the discriminants

$$\widehat{\mu_k} = \frac{1}{n_k}\sum_{i:y_i=k} x_i$$

$$\hat{\sigma}^2 = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i}(x_i - \hat{\mu}_k)^2$$

- n is the total number of training observations, n_k is the number of training observations in class k
- $\widehat{\mu_k}$ is the average of all the training observations in class k
- $\hat{\sigma}^2$ is the weighted average of the sample variance for all the k classes

LDA with One Predictor - Summary

- In summary, LDA assumes that the observations from each class follow a normal distribution with a class specific mean vector and constant variance across the classes to build a Bayes' theorem-based classifier.

Multiple Predictor LDA

- Assumes $X = (X_1, X_2, ...X_p)$ follows a multivariate normal or multivariate Gaussian distribution with a class-specific mean vector and a common covariance matrix.
- Multivariate normal distribution implies that each predictor follows a one-dimensional normal distribution with some correlation between the predictors. The bell shape of the normal distribution will be distorted if the predictors are highly correlated.

Multiple Predictor LDA

- The multivariate Gaussian density function is given by

$$f(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

- x is a p dimensional random variable that follows multivariate Gaussian distribution with the notation $X \sim N(\mu, \Sigma)$.
- μ is the mean of x which is a vector with p components
- $Cov(X = \Sigma)$ is the p^* p covariance matrix of x

Multiple Predictor LDA

- The combination of the above density function with Bayes' theorem yields the following discriminant function, the largest value of which gets the highest probability classification.

$$\delta_k(x) = x^T\Sigma^{-1}\mu_k - \frac{1}{2}\mu_k^T\Sigma^{-1}\mu_k + \log(\pi_k)$$

- $The\ Bayes\ decision\ boundaries\ are\ defined\ by\ the\ values\ for\ which\ \delta_j(x) = \delta_k(x)$. Since both the classes are assumed to have the same number of training observations, $\log(\pi_k)$ terms cancel out.

Quadratic Discriminant Analysis

- An alternative approach to LDA
- Same assumptions as LDA regarding the observations from each class following a Gaussian/Normal distribution.
- Difference is in the covariance estimation
  - QDA assumes each class has its own covariance matrix.
- This results in assuming that an observation from the k th class follows a distribution of the form $X \sim N(\mu_k, \Sigma_k)$ where Σ_k is the covariance matrix for k th class

Quadratic Discriminant Analysis

- The Bayes classifier that assigns an observation X = x to the class with the largest value for

$$\delta_k(x) = -\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}\ (x-\mu_k)\ \text{-}\ -\frac{1}{2}\log(\Sigma_k) + \log(\pi_k)$$

$$\delta_k(x) = -\frac{1}{2}x^T\Sigma_k^{-1}x + x^T\Sigma_k^{-1}\mu_k\ -\frac{1}{2}\mu_k^T\Sigma_k^{-1}\mu_k - \frac{1}{2}\log(\Sigma_k) + \log(\pi_k)$$

- The above discriminant function is quadratic in terms of x and hence the name of the classifier is QDA

Common Covariance Matrix or Not

- Why does common covariance matrix in LDA vs. class specific covariance matrix in QDA matter?
  - Bias-Variance trade off
- With p predictors, we need $\frac{p(p+1)}{2}$ parameters in LDA to estimate the covariance matrix.
- QDA estimates a separate covariance matrix for each class requiring $\frac{Kp(p+1)}{2}$ parameters.
- By assuming a shared covariance matrix, we only estimate Kp parameters in LDA. Hence, LDA is much less flexible than QDA and has substantially low variance. This can lead to improved prediction performance. However, if the common variance assumption is highly off LDA can suffer from high bias.

Comparison of Methods

- Generally, LDA is better than QDA if there are relatively few training observations and so reducing variance is relevant.
- If the training set is very large that the variance of classifier is not an issue or if the assumption of common covariance matrix is unrealistic, QDA can be a better choice.
- LDA and logistic regression work well when the decision boundary is linear
- QDA gives better results when the decision boundary is moderately non-linear
- K-nearest neighbors (KNN) is a non-parametric approach and outperforms LDA and logistic regression when the decision-boundary is highly non-linear

Performance Metrics

- Confusion Matrix
- Sensitivity
- Specificity
- Precision
- Accuracy
- ROC Curve

Confusion Matrix

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I Error)
Test Negative	False Negative (Type II Error)	True Negative

- A confusion matrix is used to compare the model predictions to the actual (or true) values.
- Sensitivity and Specificity characterize the performance of a classifier
- Sensitivity measures the ability of a classifier to identify positive results
- Specificity measures the ability of a classifier to identify negative results

Metrics

	Condition Positive	Condition Negative
Test Positive	True Positive	False Positive (Type I Error)
Test Negative	False Negative (Type II Error)	True Negative

- Sensitivity =Recall = $P(Test+ \mid Condition+) = \frac{TP}{TP+FN}$
- Specificity = $P(Test- \mid Condition-) = \frac{TN}{FP+TN}$
- False Negative Rate (β) = $P(Test- \mid Condition+) = \frac{FN}{TP+FN}$
- False Positive Rate (α) = $P(Test+ \mid Condition-) = \frac{FP}{FP+TN}$
- Precision = $\frac{TP}{TP+FP}$
- Sensitivity = 1 – False Negative Rate = Power
- Specificity = 1 – False Positive Rate
- F1 Score = $\frac{2*Precision*Recall}{Precision+Recall}$
- Accuracy = $\frac{Correct\ Predictions}{Total\ Predictions}$

Metrics

- Accuracy
  - % of correct predictions
  - One value for the entire model
- Prediction
  - Exactness of the model
  - Each class/label has a value
- Recall
  - Completeness of model
  - Correctly detected over total observations
  - Each class/label has a value
- F1 Score
  - Combines precision and recall (Harmonic mean of precision & recall)
  - Each class/label has a value

ROC Curves

- A curve for simultaneously displaying the positive and negative error types for various thresholds
- ROC stands for “Receiver Operating Characteristics” and derives its name from communications theory (historical)
- The x-axis is False Positive Rate (=1-specificity)
- The y-axis is True Positive Rate (= Sensitivity)
- Overall performance is given by the Area Under the Curve, denoted as AUC.
- Larger the AUC, better the classifier i.e., the curve is closer to the top left corner
- Let us now look at an example of how a ROC Curve is plotted for various thresholds. *(The slides for the example are borrowed from Colin Rundel & Mine Cetinkaya-Rundel)*

Back to Spam

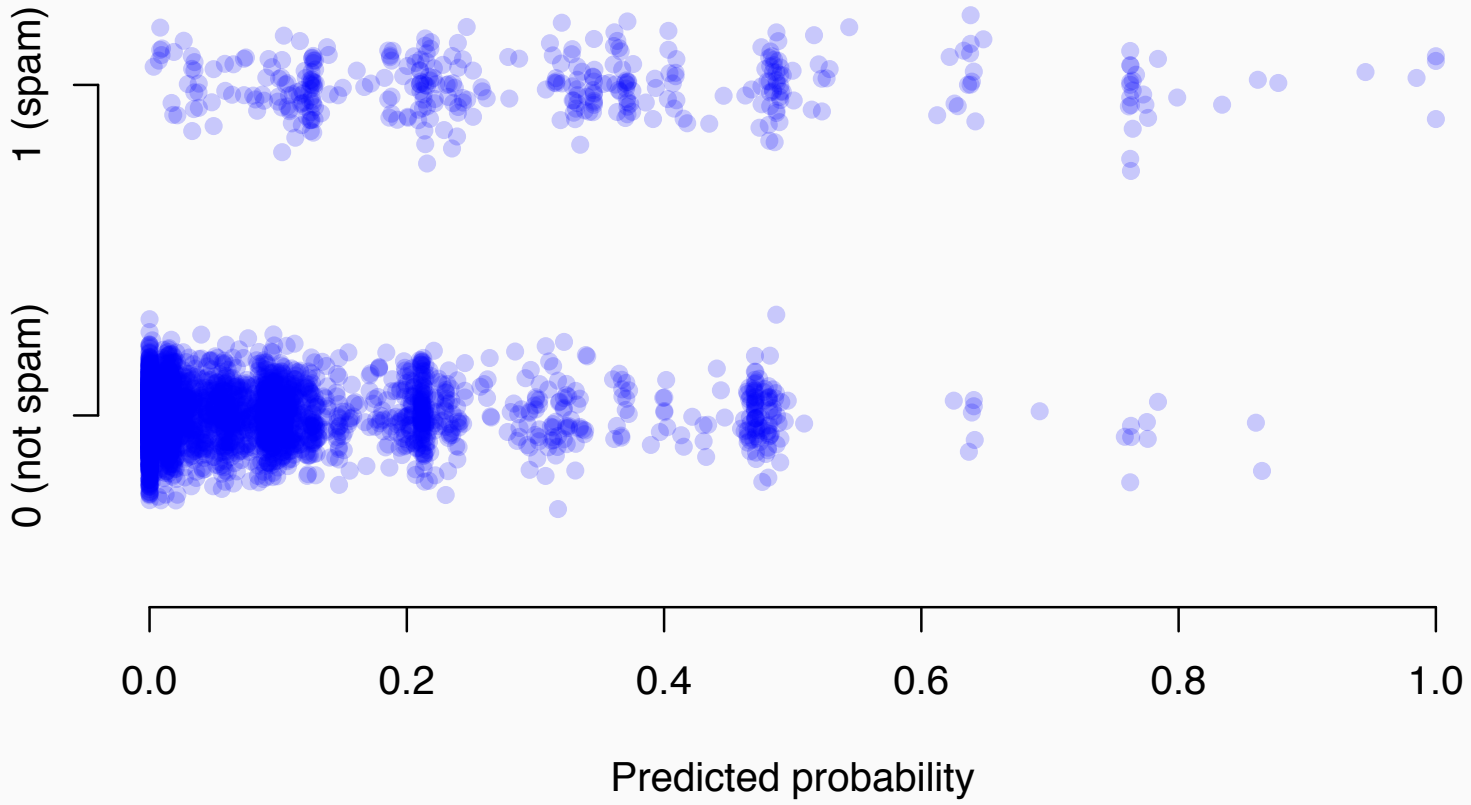
we will now examine a data set of emails where we are interested in identifying spam messages. We will examine several different logistic regression models, however these models only predict the probability an incoming message is spam. If we were designing a spam filter this would only be half of the battle, we also need to design a decision rule about which emails get flagged as spam (e.g. what probability should we use as out cutoff?)

Back to Spam

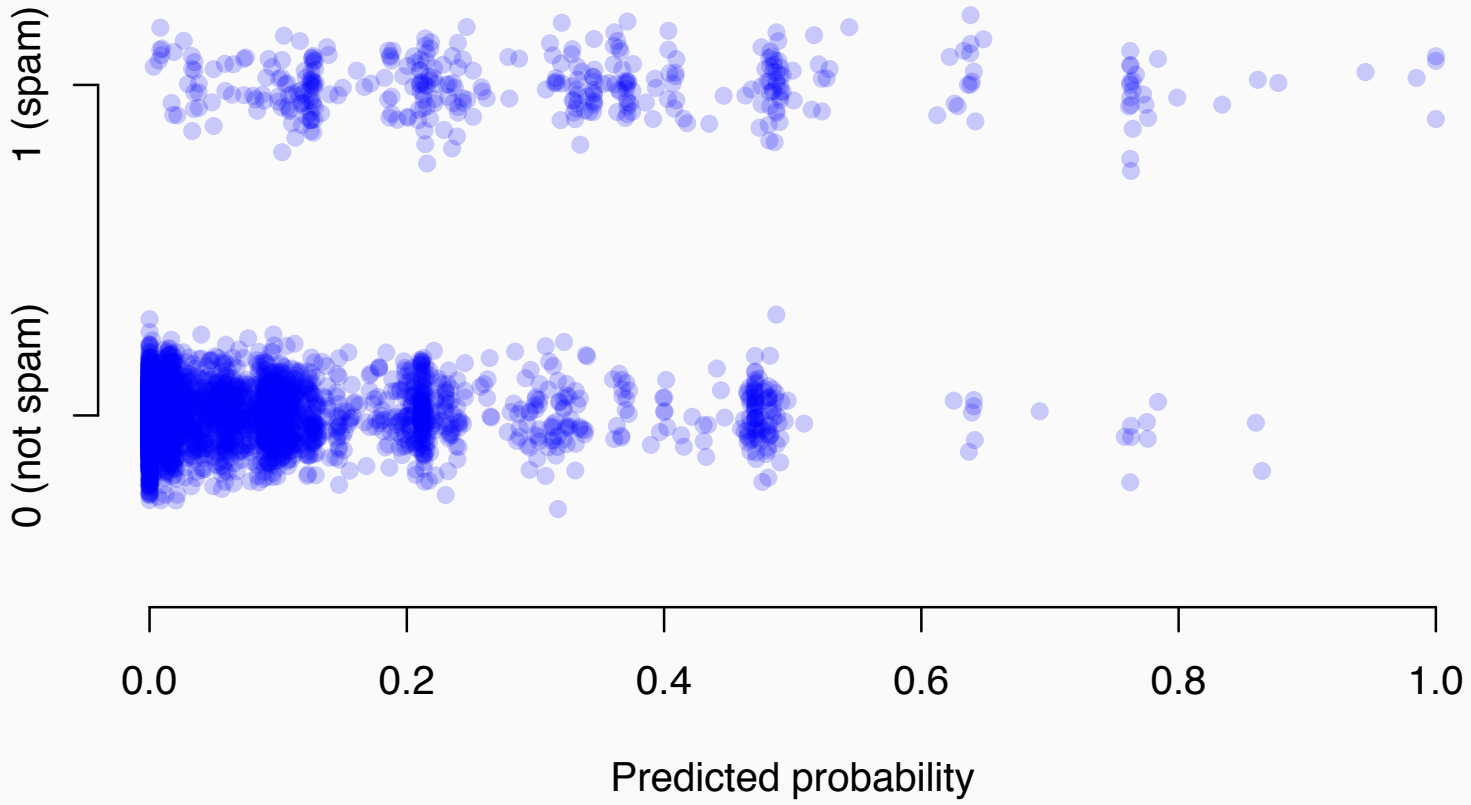
we will now examine a data set of emails where we are interested in identifying spam messages. We will examine several different logistic regression models, however these models only predict the probability an incoming message is spam. If we were designing a spam filter this would only be half of the battle, we also need to design a decision rule about which emails get flagged as spam (e.g. what probability should we use as out cutoff?)

While not the only possible solution, we will consider a simple approach where we choose a single threshold probability and any email that exceeds that probability is flagged as spam.

Picking a threshold

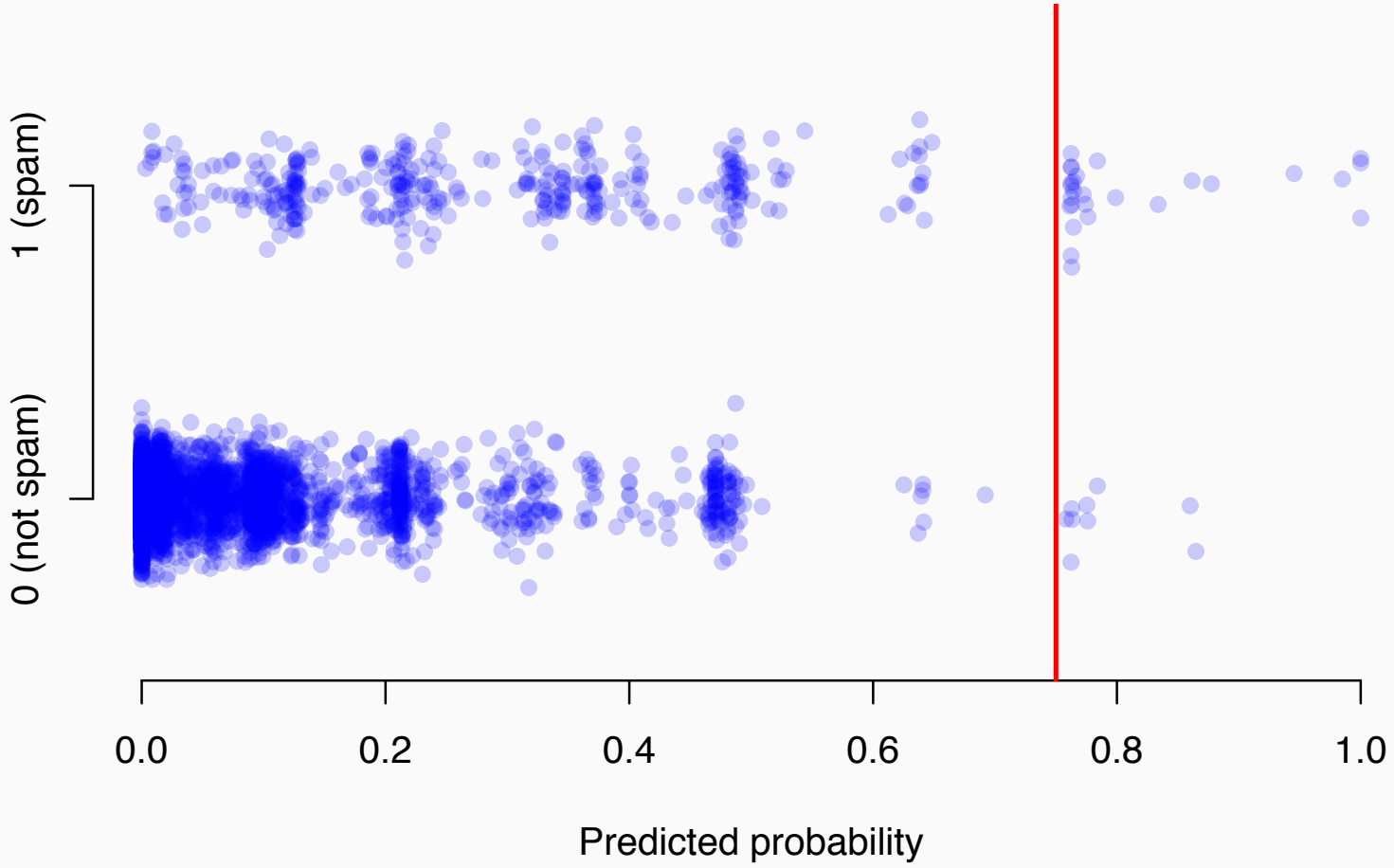


Picking a threshold



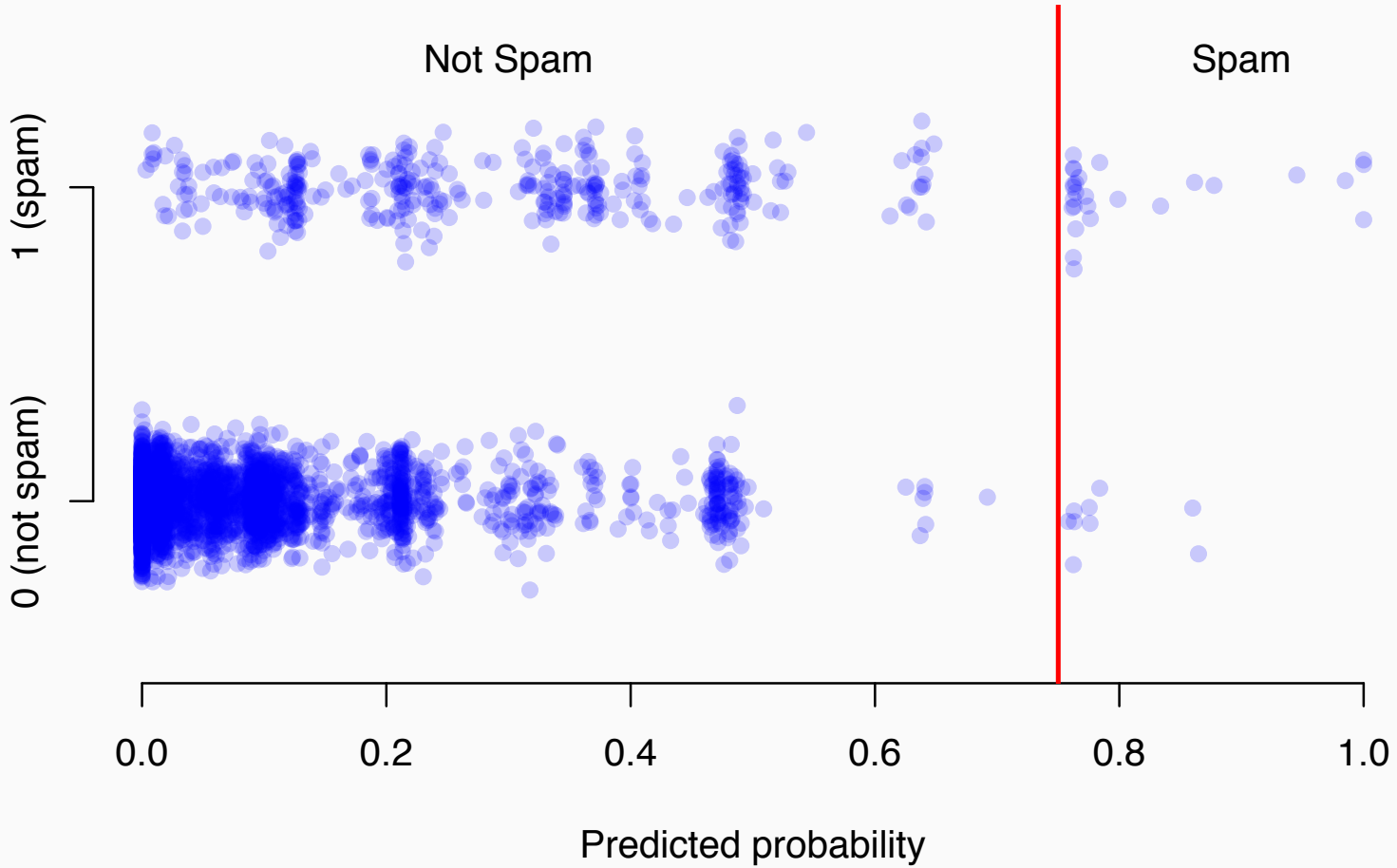
Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



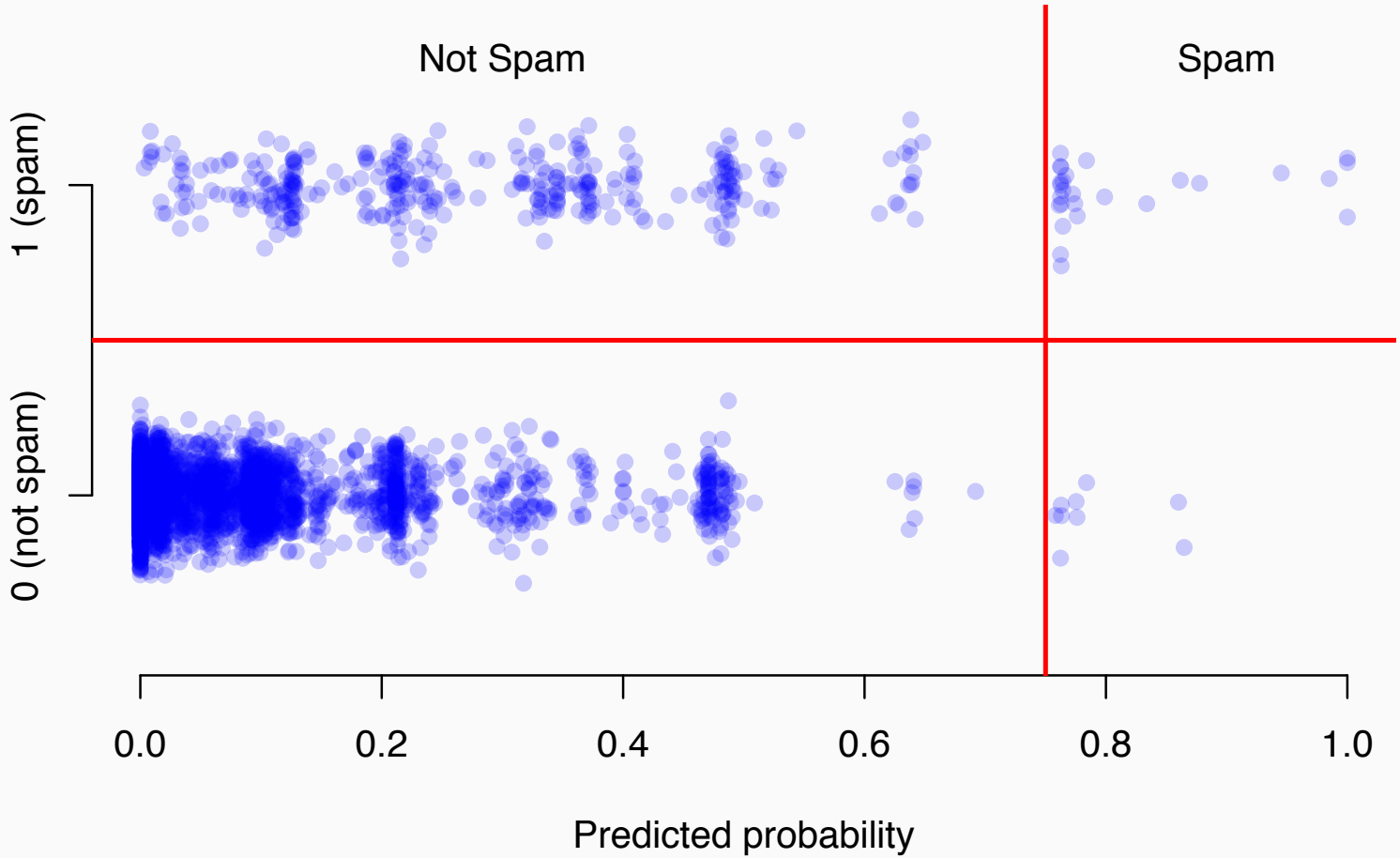
Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



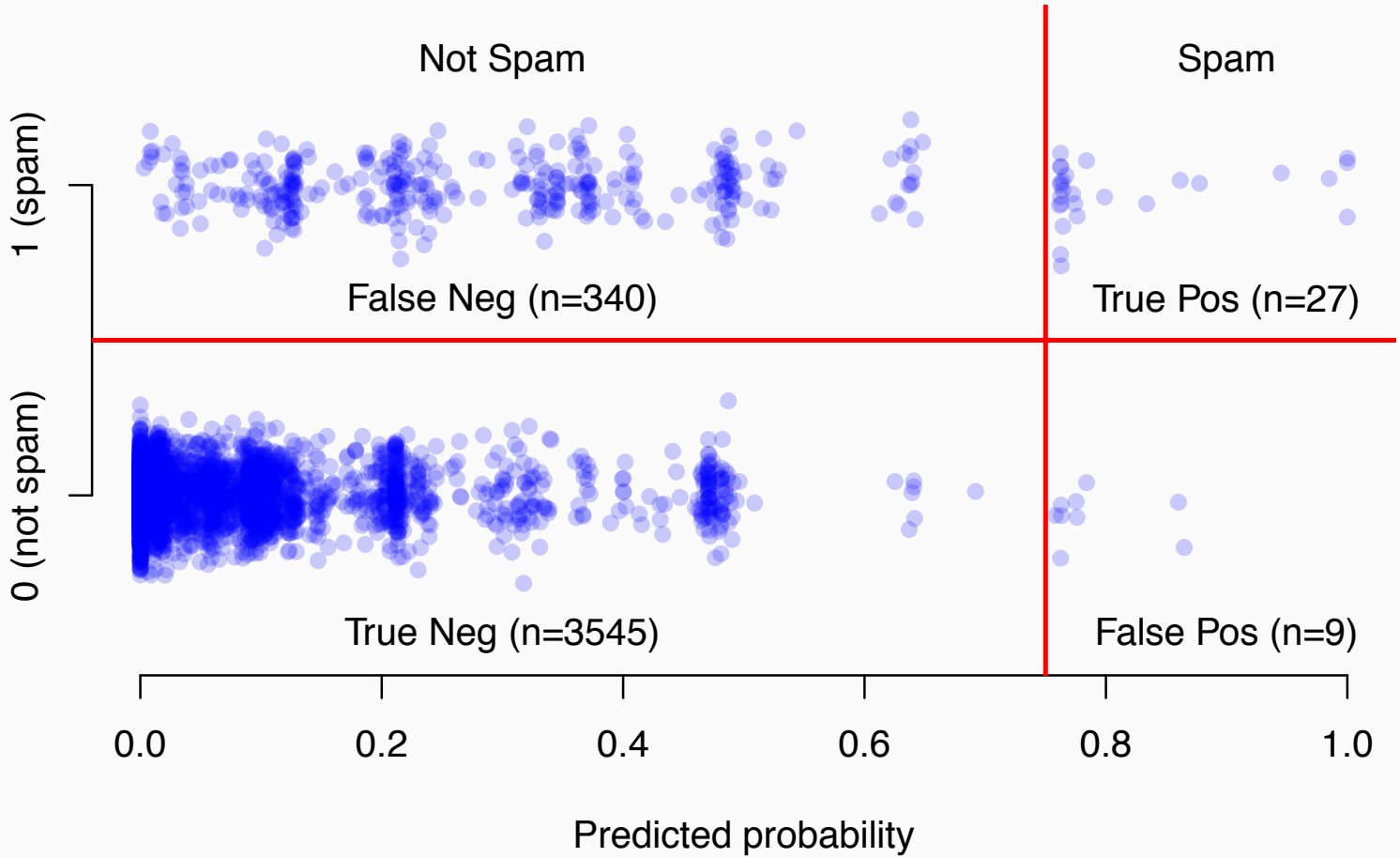
Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



Lets see what happens if we pick our threshold to be *0.75*.

Picking a threshold



Lets see what happens if we pick our threshold to be *0.75*.

Consequences of picking a threshold

For our data set picking a threshold of 0.75 gives us the following results:

$$FN = 340 \qquad TP = 27$$

$$TN = 3545 \qquad FP = 9$$

Consequences of picking a threshold

For our data set picking a threshold of 0.75 gives us the following results:

$$FN = 340 \qquad TP = 27$$

$$TN = 3545 \qquad FP = 9$$

What are the sensitivity and specificity for this particular decision rule?

Consequences of picking a threshold

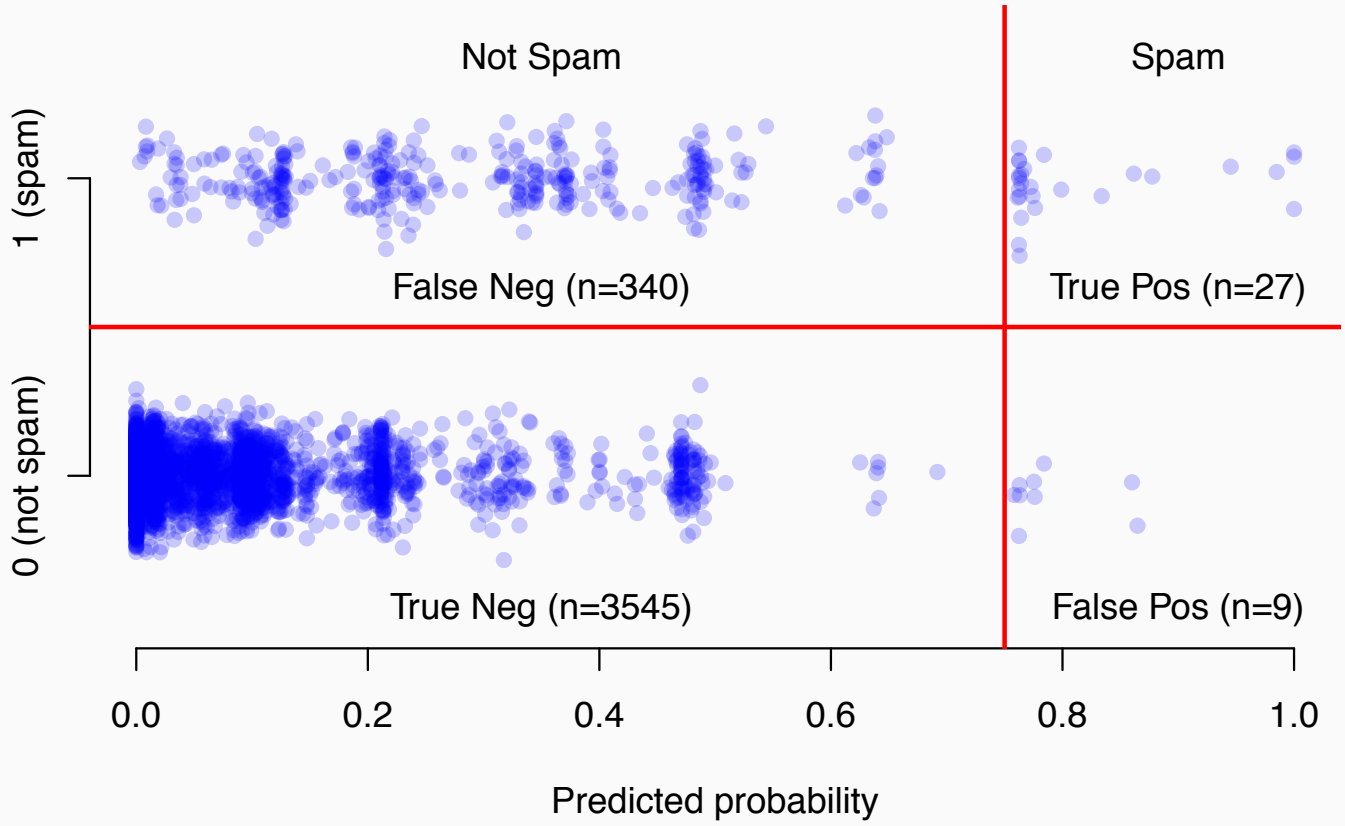
For our data set picking a threshold of 0.75 gives us the following results:

$$FN = 340 \qquad TP = 27$$
$$TN = 3545 \qquad FP = 9$$

What are the sensitivity and specificity for this particular decision rule?

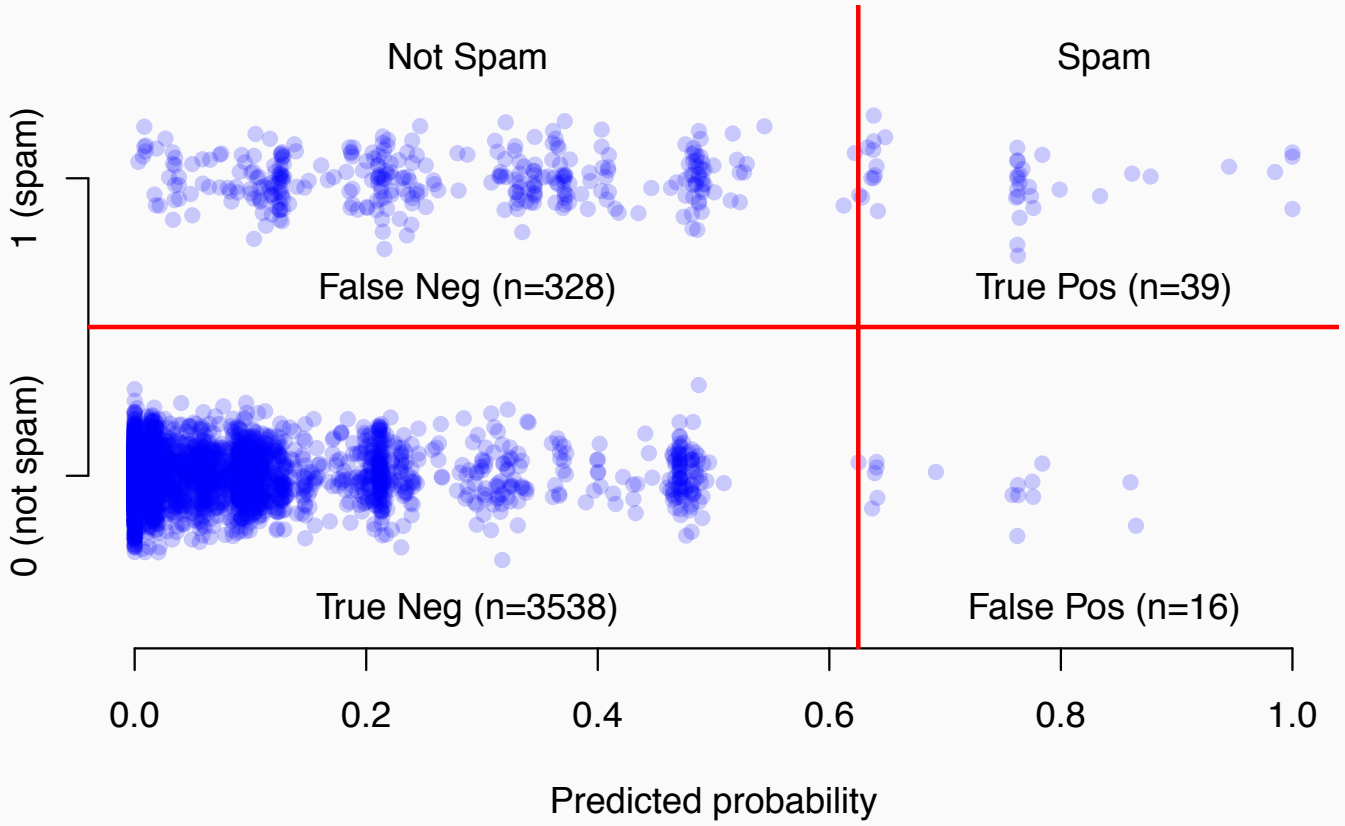
$$Sensitivity = TP/(TP + FN) = 27/(27 + 340) = 0.073$$
$$Specificity = TN/(FP + TN) = 3545/(9 + 3545) = 0.997$$

Trying other thresholds



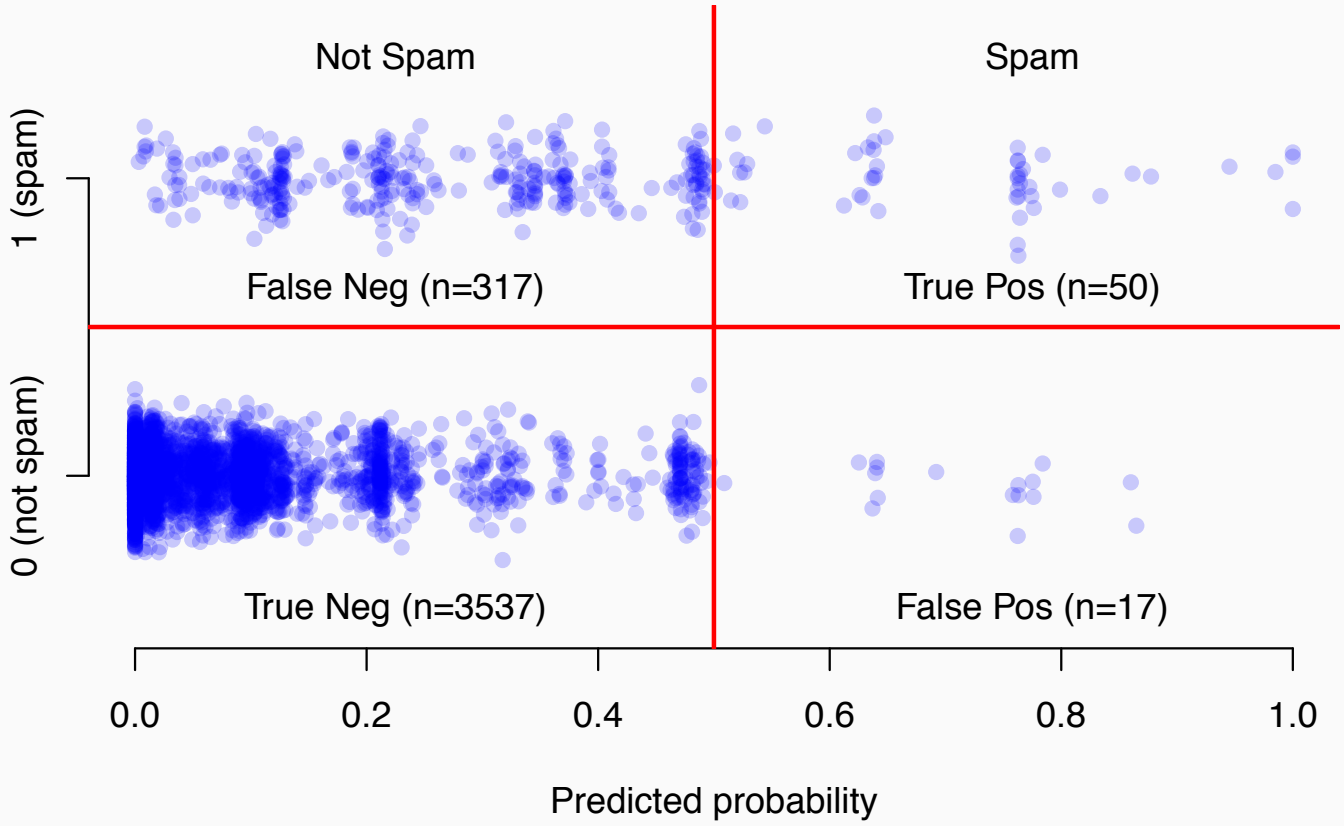
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074				
Specificity	0.997				

Trying other thresholds



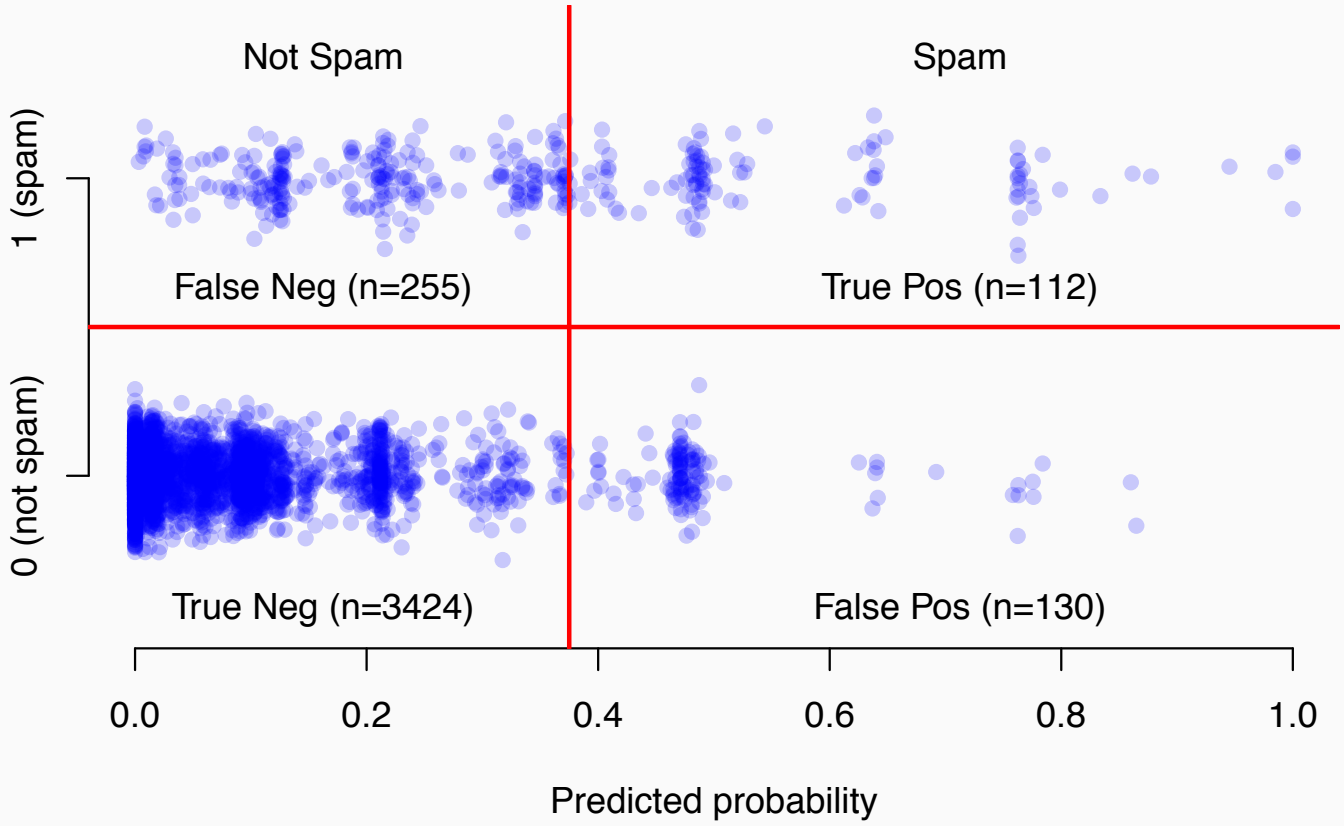
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106			
Specificity	0.997	0.995			

Trying other thresholds



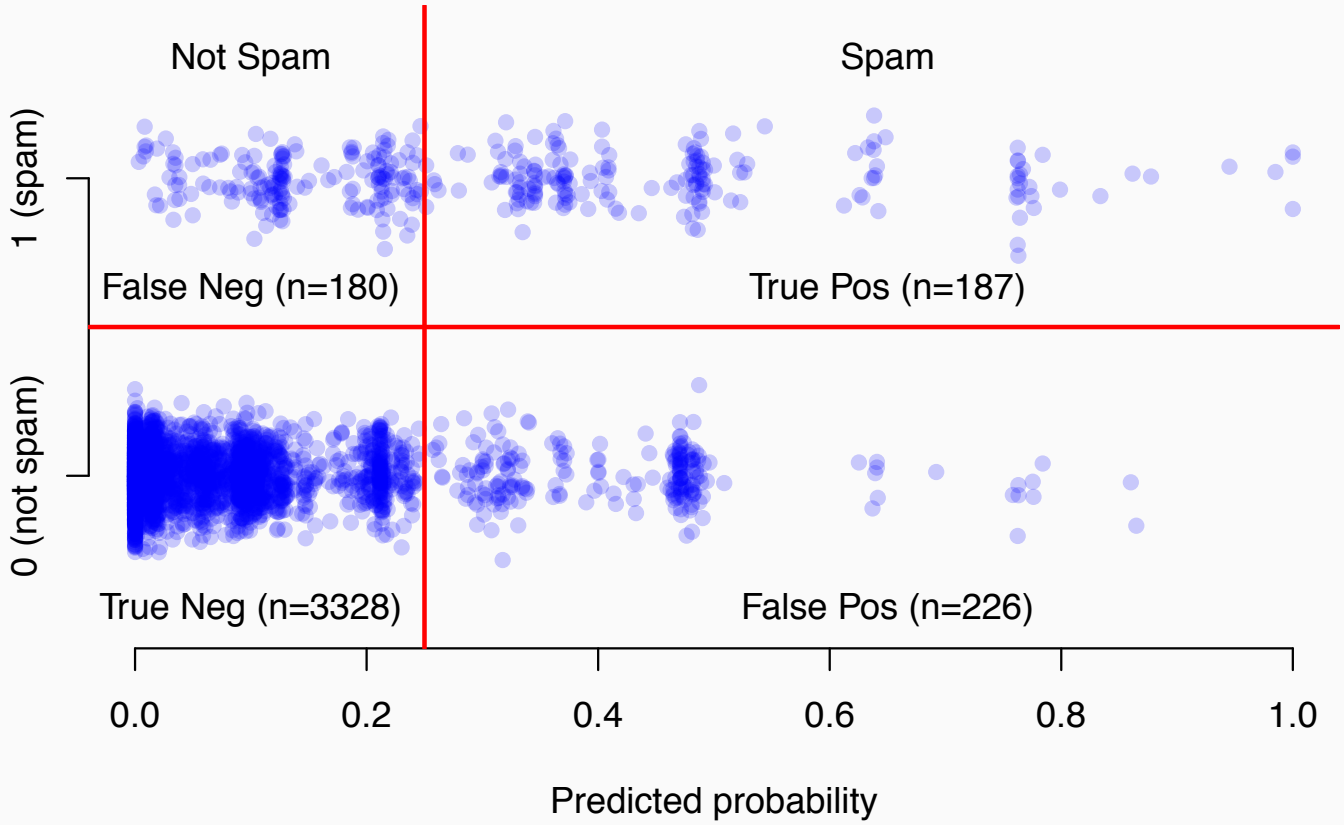
Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136		
Specificity	0.997	0.995	0.995		

Trying other thresholds



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	
Specificity	0.997	0.995	0.995	0.963	

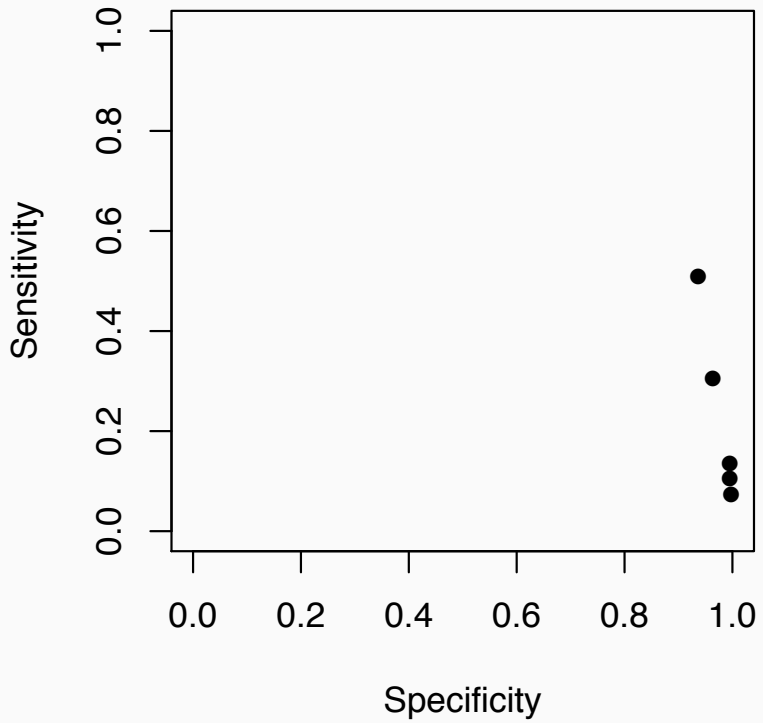
Trying other thresholds



Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

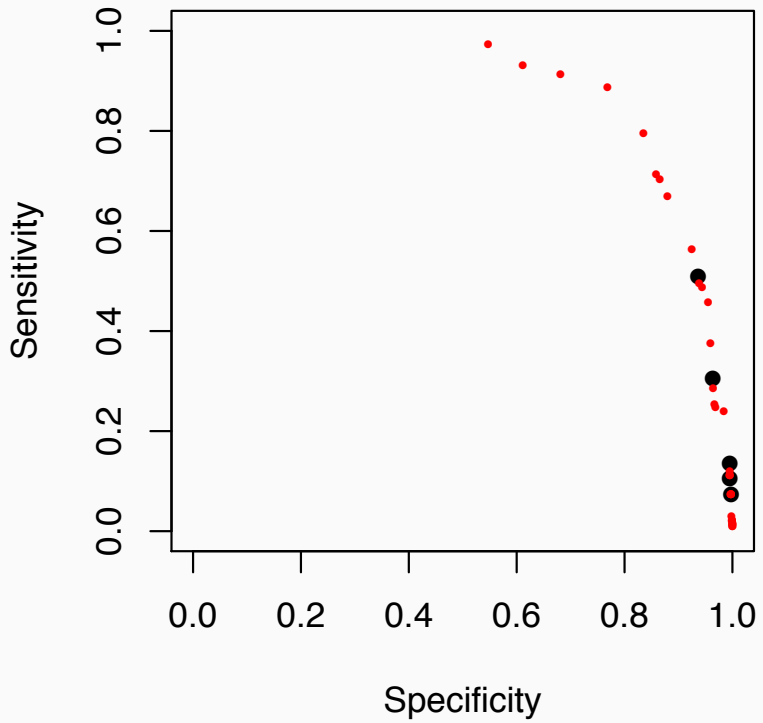
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



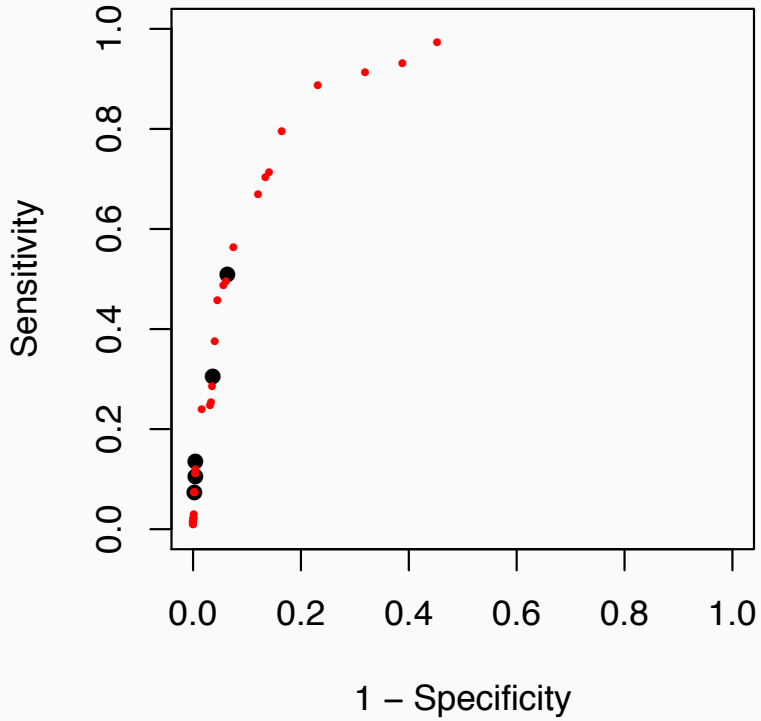
Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936

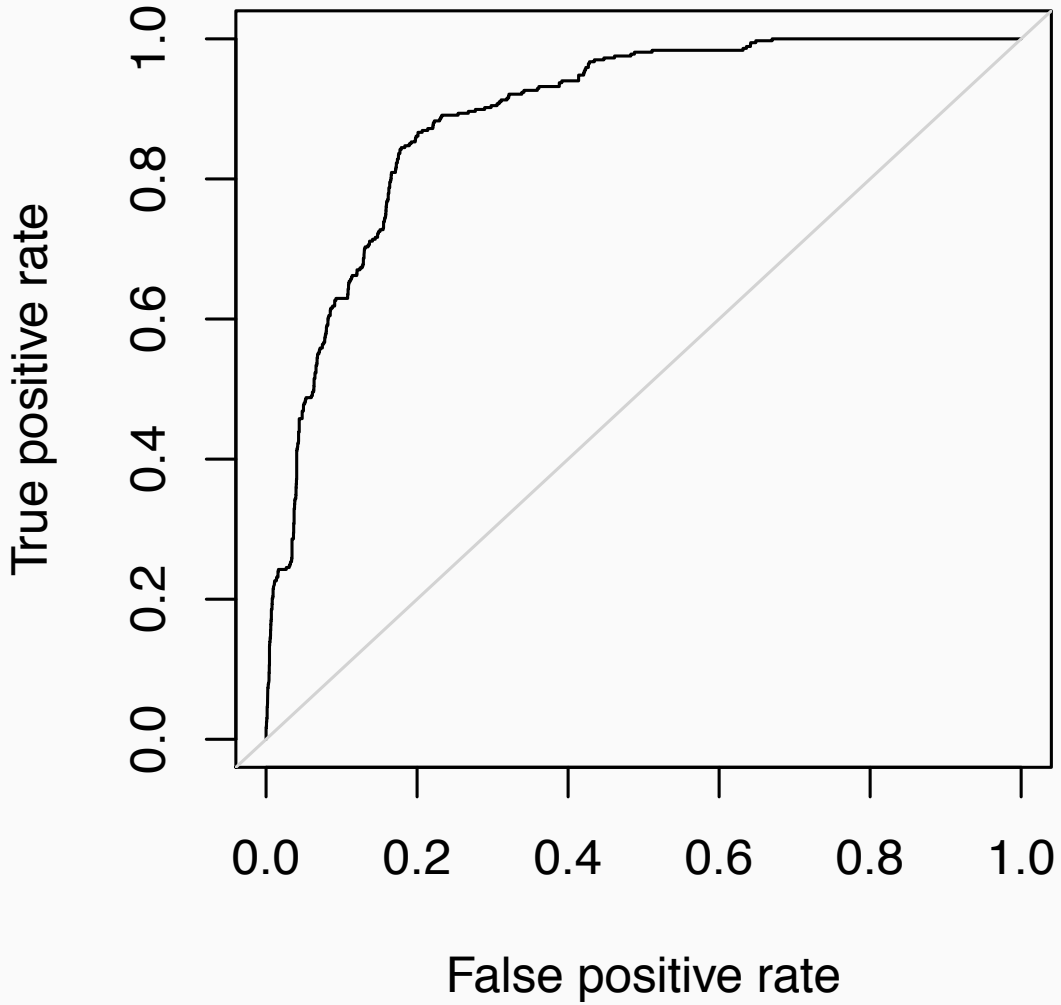


Relationship between Sensitivity and Specificity

Threshold	0.75	0.625	0.5	0.375	0.25
Sensitivity	0.074	0.106	0.136	0.305	0.510
Specificity	0.997	0.995	0.995	0.963	0.936



Receiver operating characteristic (ROC) curve



Receiver operating characteristic (ROC) curve (cont.)

Why do we care about ROC curves?

- Shows the trade off in sensitivity and specificity for all possible thresholds.
- Straight forward to compare performance vs. chance.
- Can use the area under the curve (AUC) as an assessment of the predictive ability of a model.