Advanced Statistics
DS2003 (BDS-4A)
Lecture 27

Instructor: Dr. Syed Mohammad Irteza

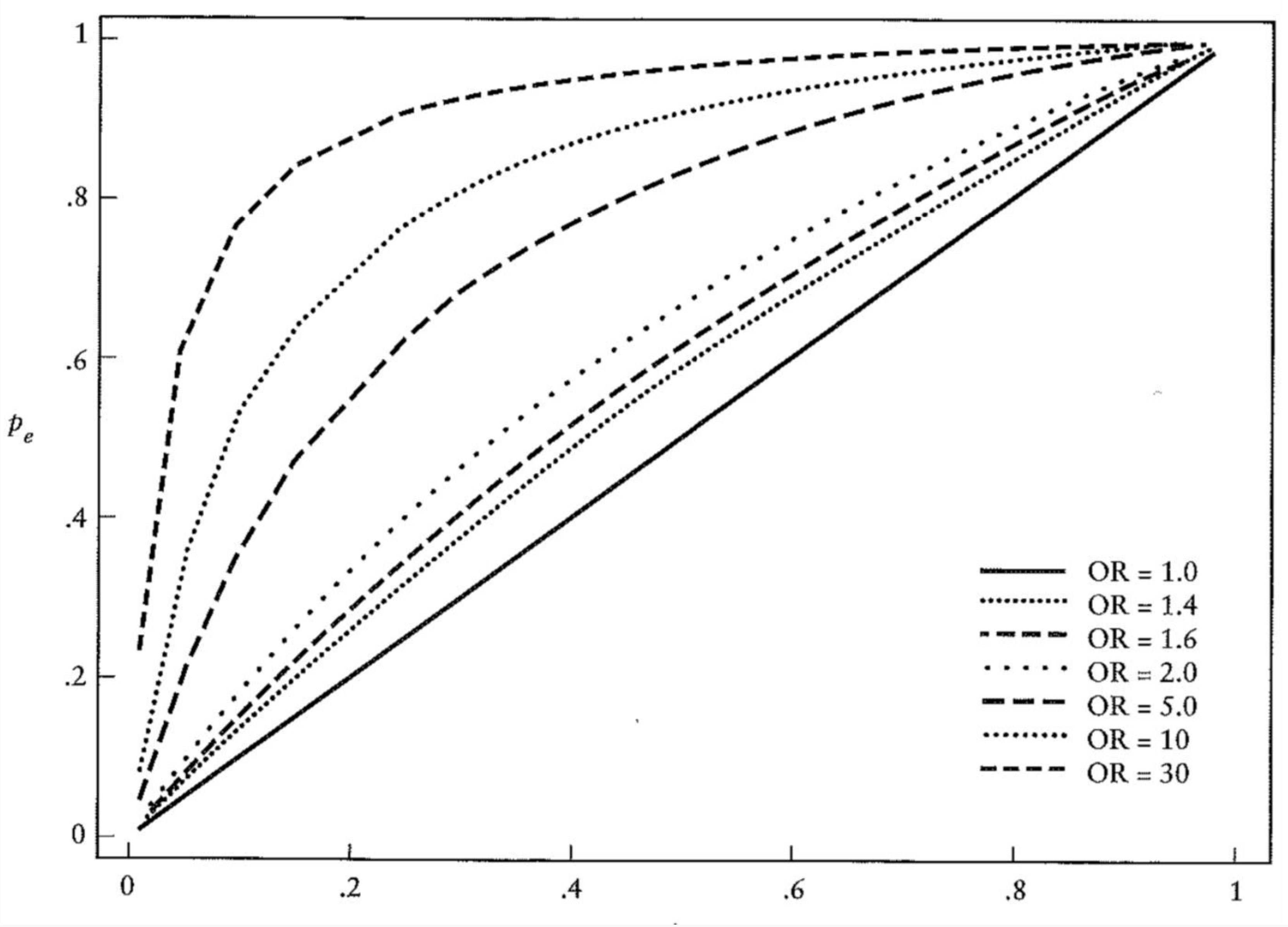
Assistant Professor, Department of Computer Science, FAST

31 May, 2022

Previous Lecture

- Logit Function
- Confidence Interval using PE (point estimates) and SE (standard errors)
- Concept of Odds Ratio and Log Odds Ratio
- Example of Bird-keeping and Lung Cancer
  - Discarding the predictors who have p-values of less than 0.05
  - Calculating the odds ratio on the basis of the slope values for (BKBird, YR)
  - Interpreting the OR (odds ratio) value, and how it is different from the RR (relative risk)
    - We can compute RR if we have more information provided
  - Bird OR Curve

OR Curves



(An old) Example - House

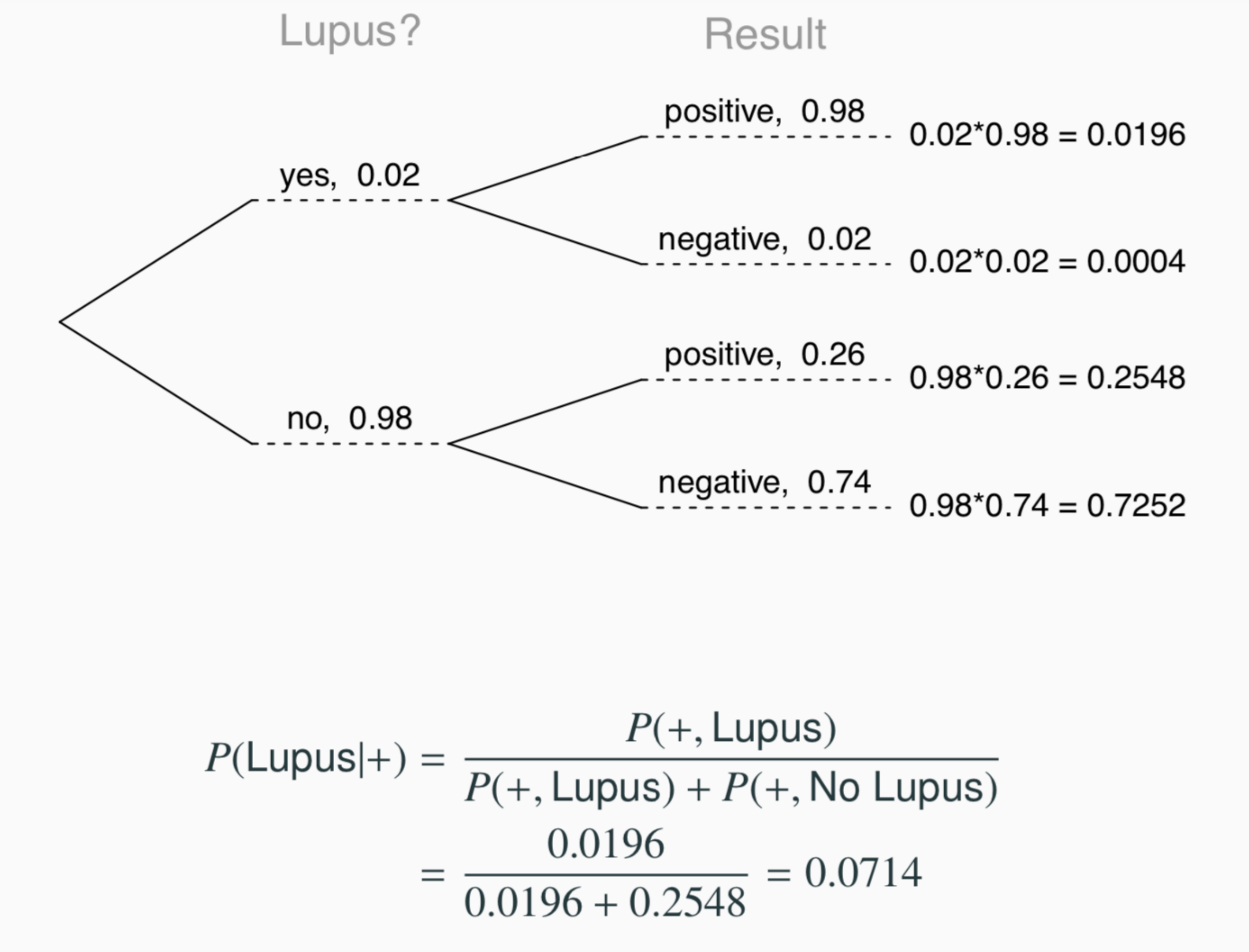
If you've ever watched the TV show House on Fox, you know that Dr. House regularly states, "It's never lupus."

Lupus is a medical phenomenon where antibodies that are supposed to attack foreign cells to prevent infections instead see plasma proteins as foreign bodies, leading to a high risk of blood clotting. It is believed that 2% of the population suffer from this disease.

The test for lupus is very accurate if the person actually has lupus, however is very inaccurate if the person does not. More specifically, the test is 98% accurate if a person actually has the disease. The test is 74% accurate if a person does not have the disease.

Is Dr. House correct even if someone tests positive for Lupus?

(An old) Example - House



$$P(\text{Lupus}|+) = \frac{P(+, \text{Lupus})}{P(+, \text{Lupus}) + P(+, \text{No Lupus})}$$

$$= \frac{0.0196}{0.0196 + 0.2548} = 0.0714$$

Testing for lupus

It turns out that testing for Lupus is actually quite complicated, a diagnosis usually relies on the outcome of multiple tests, often including: a complete blood count, an erythrocyte sedimentation rate, a kidney and liver assessment, a urinalysis, and or an antinuclear antibody (ANA) test.

It is important to think about what is involved in each of these tests (e.g. deciding if complete blood count is high or low) and how each of the individual tests and related decisions plays a role in the overall decision of diagnosing a patient with lupus.

Testing for lupus

At some level we can view a diagnosis as a binary decision (lupus or no lupus) that involves the complex integration of various explanatory variables.

The example does not give us any information about how a diagnosis is made, but what it does give us is just as important - the sensitivity and the specificity of the test. These values are critical for our understanding of what a positive or negative test result actually means.

Sensitivity and Specificity

Sensitivity - measures a tests ability to identify positive results.

$$P(\text{Test} + \mid \text{Conditon} +) = P(+|\text{lupus}) = 0.98$$

Specificity - measures a tests ability to identify negative results.

$$P(\text{Test} - \mid \text{Condition} -) = P(-|\text{no lupus}) = 0.74$$

Sensitivity and Specificity

Sensitivity - measures a tests ability to identify positive results.

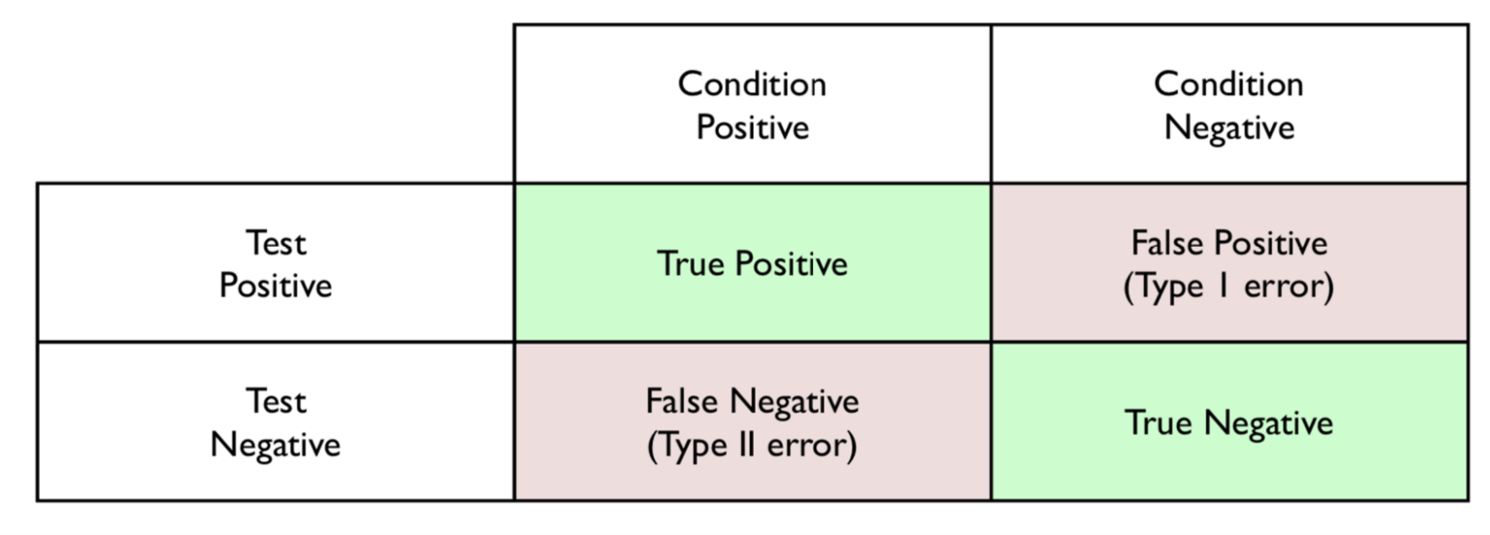
P(Test + | Conditon +) = P(+|lupus) = 0.98

Specificity - measures a tests ability to identify negative results.

P(Test − | Condition −) = P(−|no lupus) = 0.74

It is illustrative to think about the extreme cases - what is the sensitivity and specificity of a test that always returns a positive result? What about a test that always returns a negative result?

Sensitivity and Specificity (cont.)



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

$$\textit{Sensitivity} = P(\text{Test } + \mid \text{Condition } +) = TP/(TP + FN)$$

$$\textit{Specificity} = P(\text{Test } - \mid \text{Condition } -) = TN/(FP + TN)$$

$$\textit{False negative rate } (\beta) = P(\text{Test } - \mid \text{Condition } +) = FN/(TP + FN)$$

$$\textit{False positive rate } (\alpha) = P(\text{Test } + \mid \text{Condition } -) = FP/(FP + TN)$$

$$\textit{Sensitivity} = 1 - \textit{False negative rate} = \text{Power}$$

$$\textit{Specificity} = 1 - \textit{False positive rate}$$

So what?

Clearly it is important to know the Sensitivity and Specificity of test (and or the false positive and false negative rates). Along with the incidence of the disease (e.g. P(lupus)) these values are necessary to calculate important quantities like P(lupus|+).

Additionally, our brief foray into power analysis before the first midterm should also give you an idea about the trade offs that are inherent in minimizing false positive and false negative rates (increasing power required either increasing α or n).

How should we use this information when we are trying to come up with a decision?

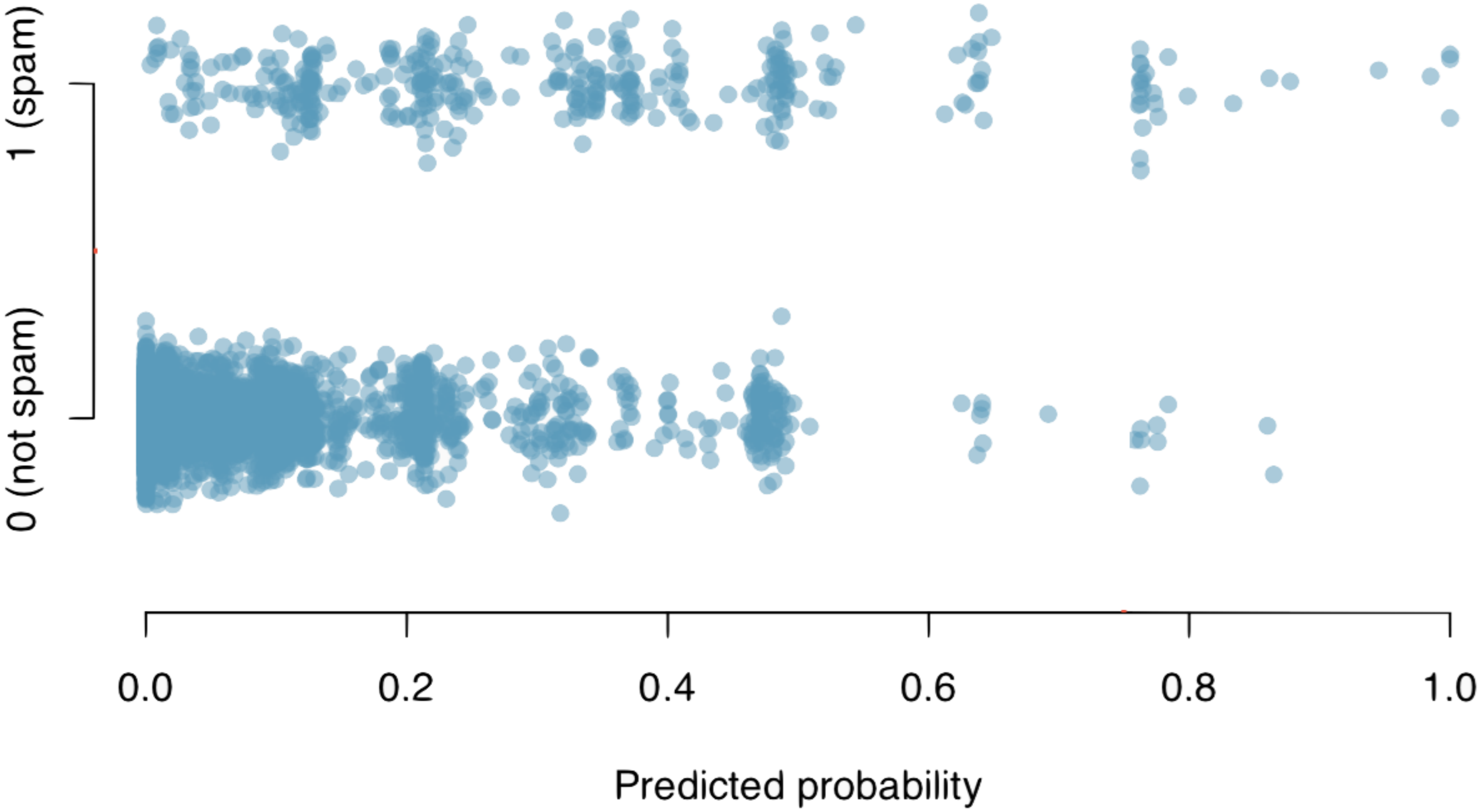
Back to Spam

In lab this week, we examined a data set of emails where we were interesting in identifying the spam messages. We examined different logistic regression models to evaluate how different predictors influenced the probability of a message being spam.

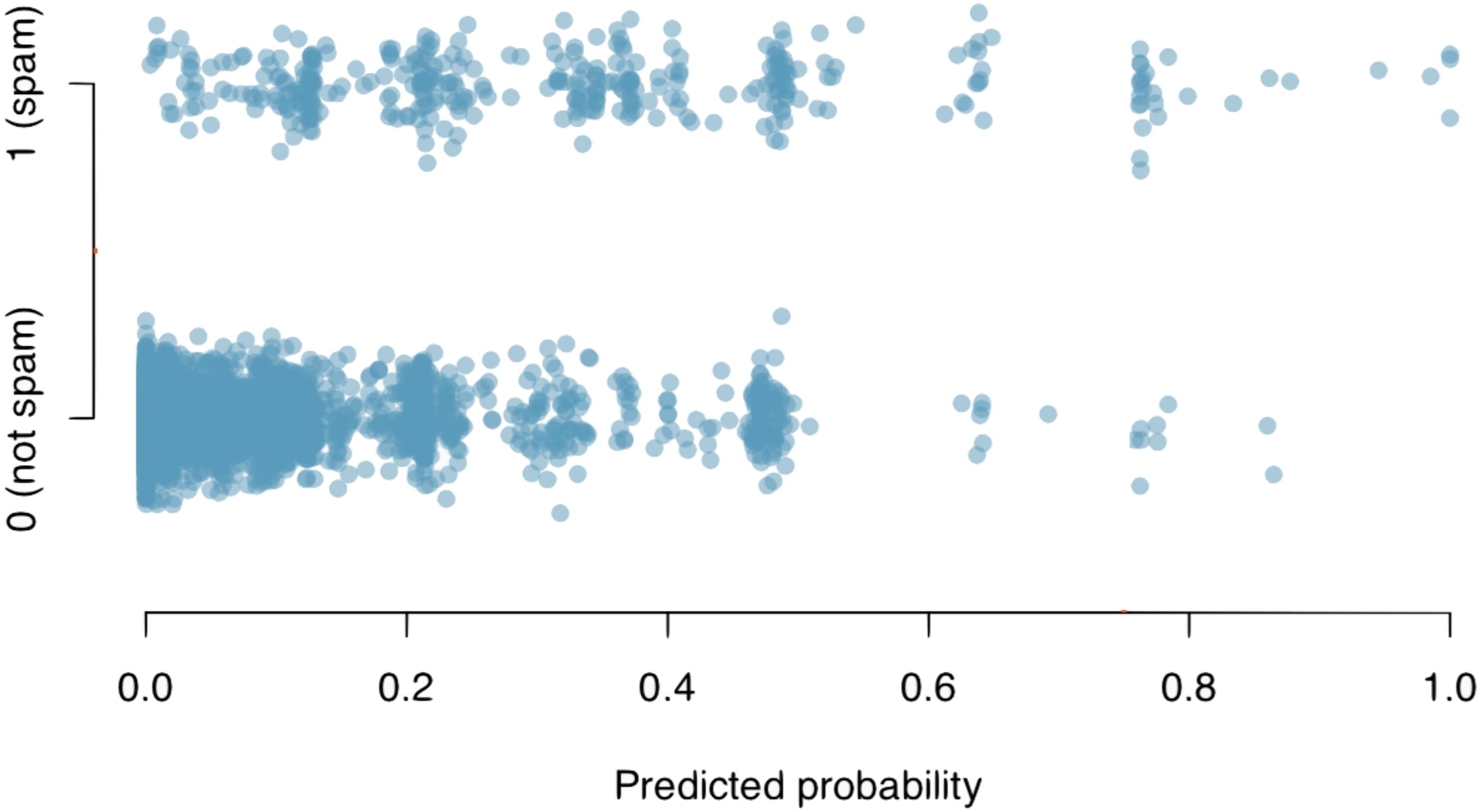
These models can also be used to assign probabilities to incoming messages (this is equivalent to prediction in the case of SLR / MLR). However, if we were designing a spam filter this would only be half of the battle, we would also need to use these probabilities to make a decision about which emails get flagged as spam.

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

Picking a threshold

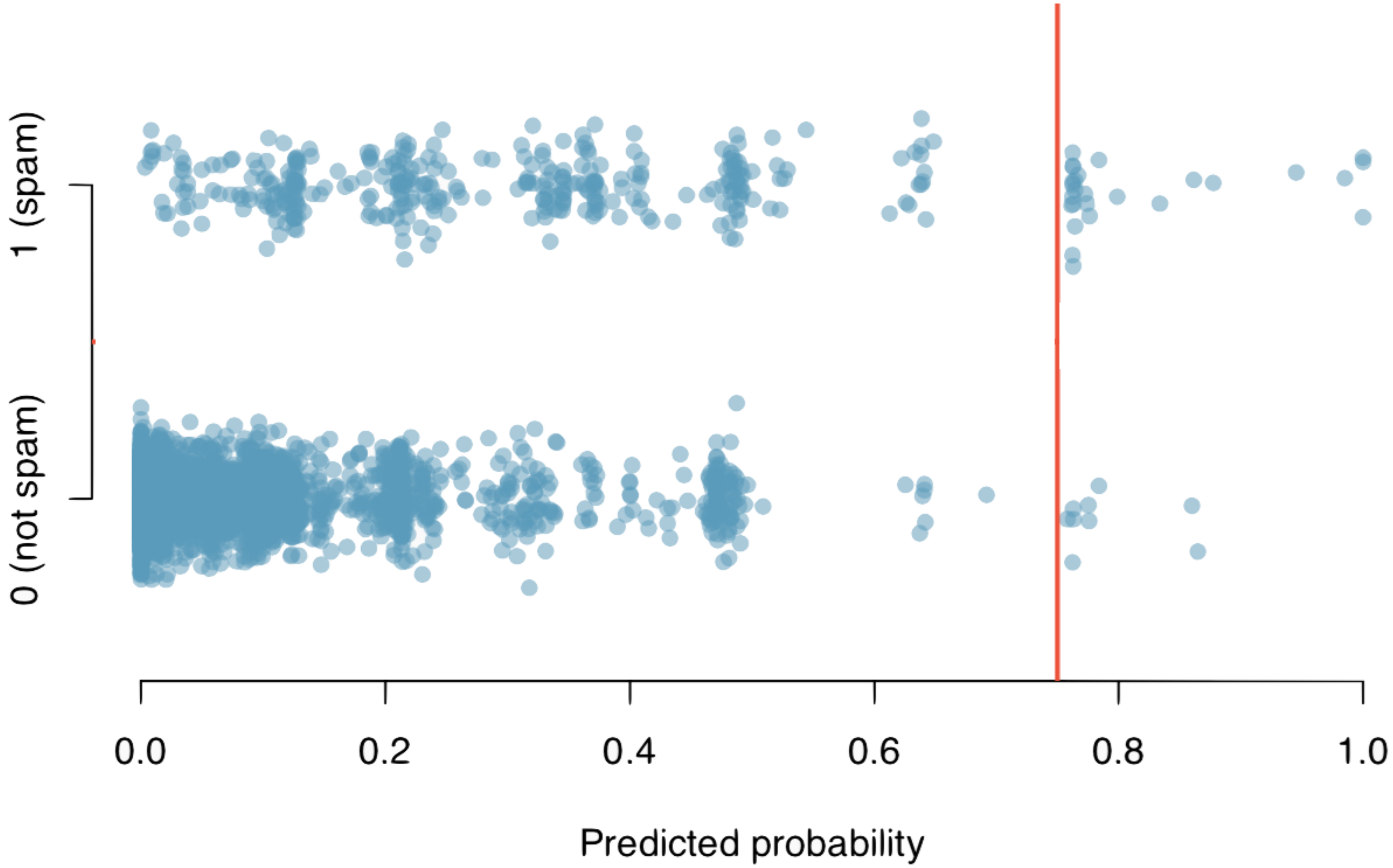


Picking a threshold



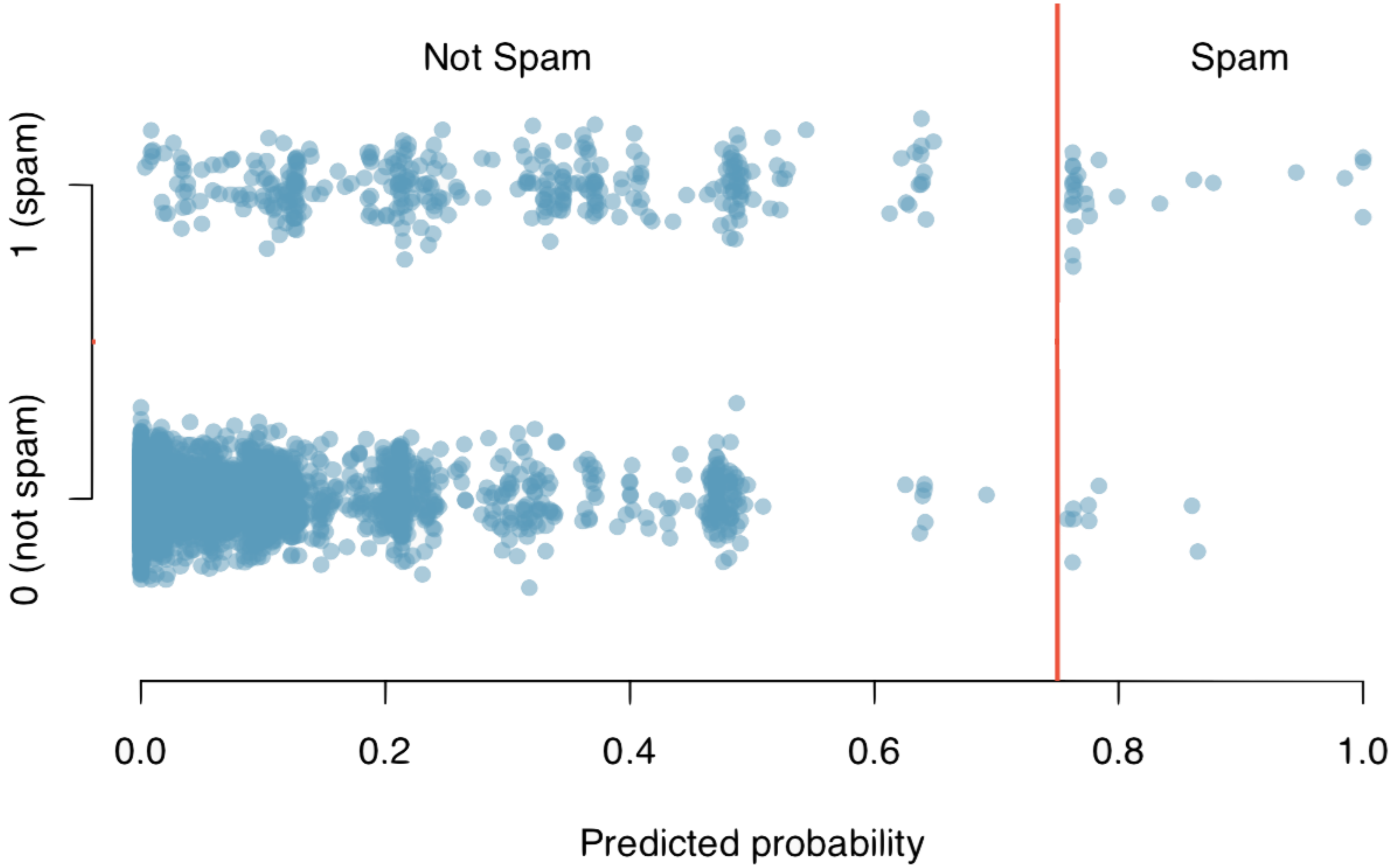
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



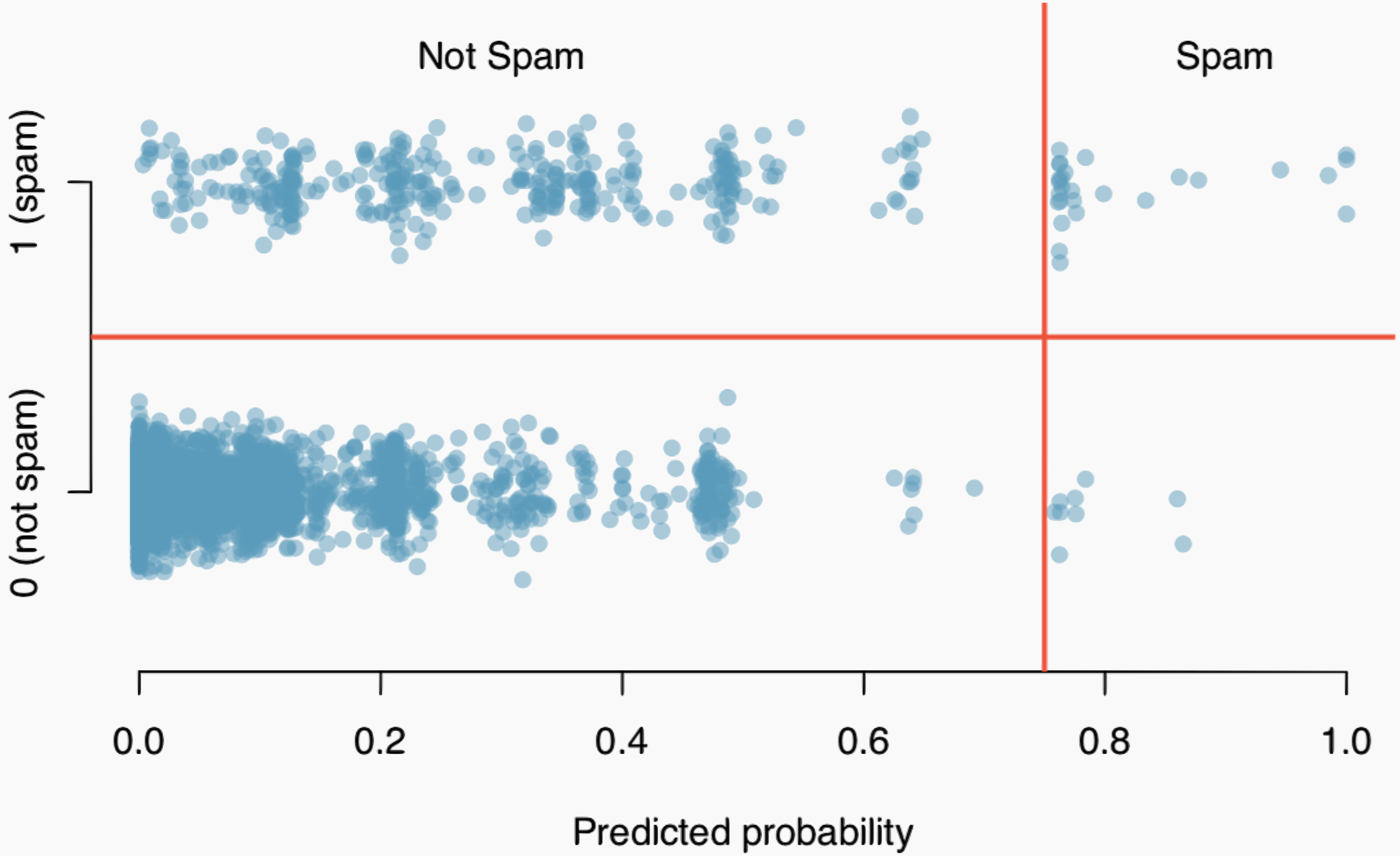
Lets see what happens if we pick our threshold to be 0.75.

Picking a threshold



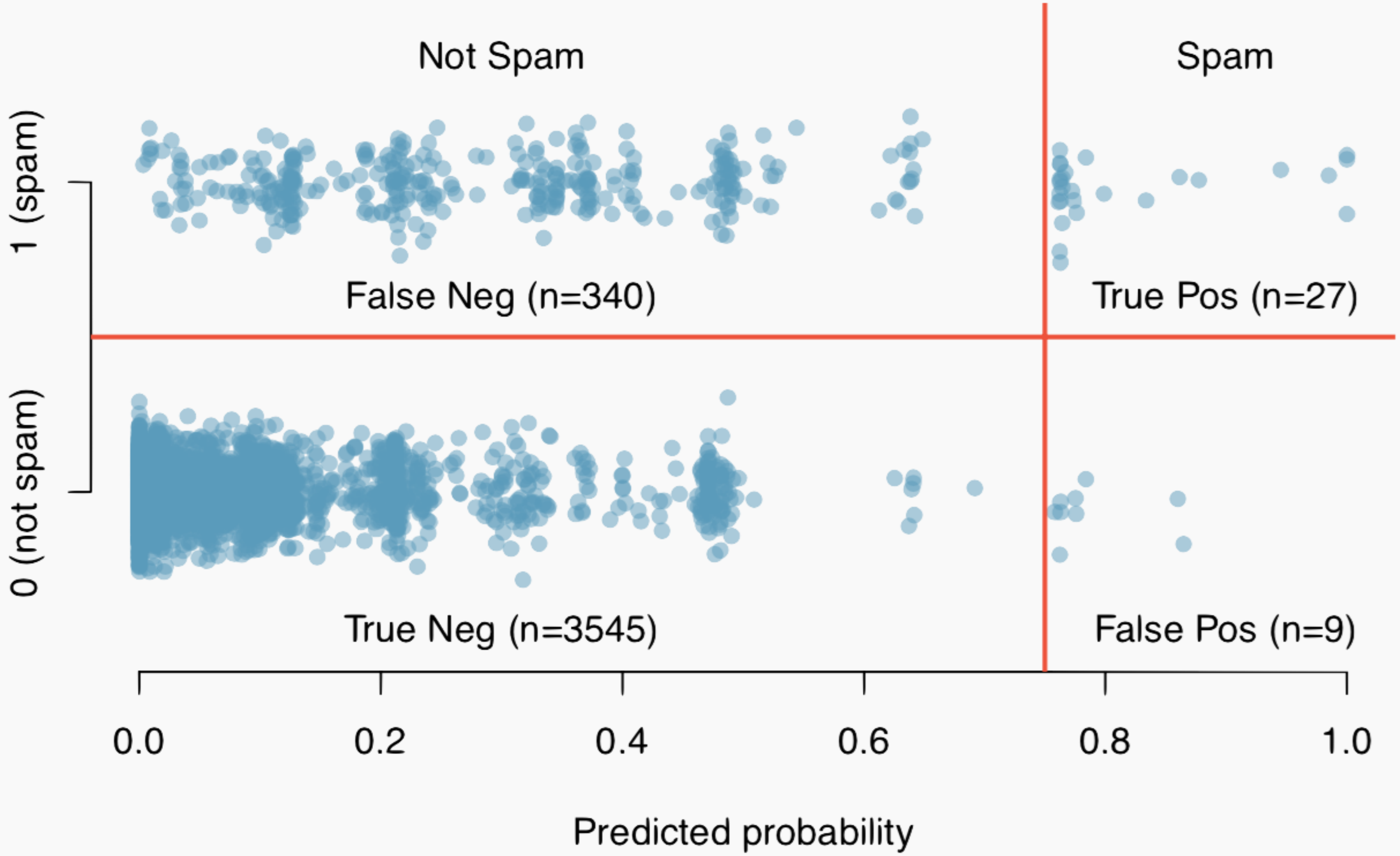
Lets see what happens if we pick our threshold to be 0.75.

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	TP = 27
TN = 3545	FP = 9

Consequences of picking a threshold

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

FN = 340	TP = 27
TN = 3545	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:

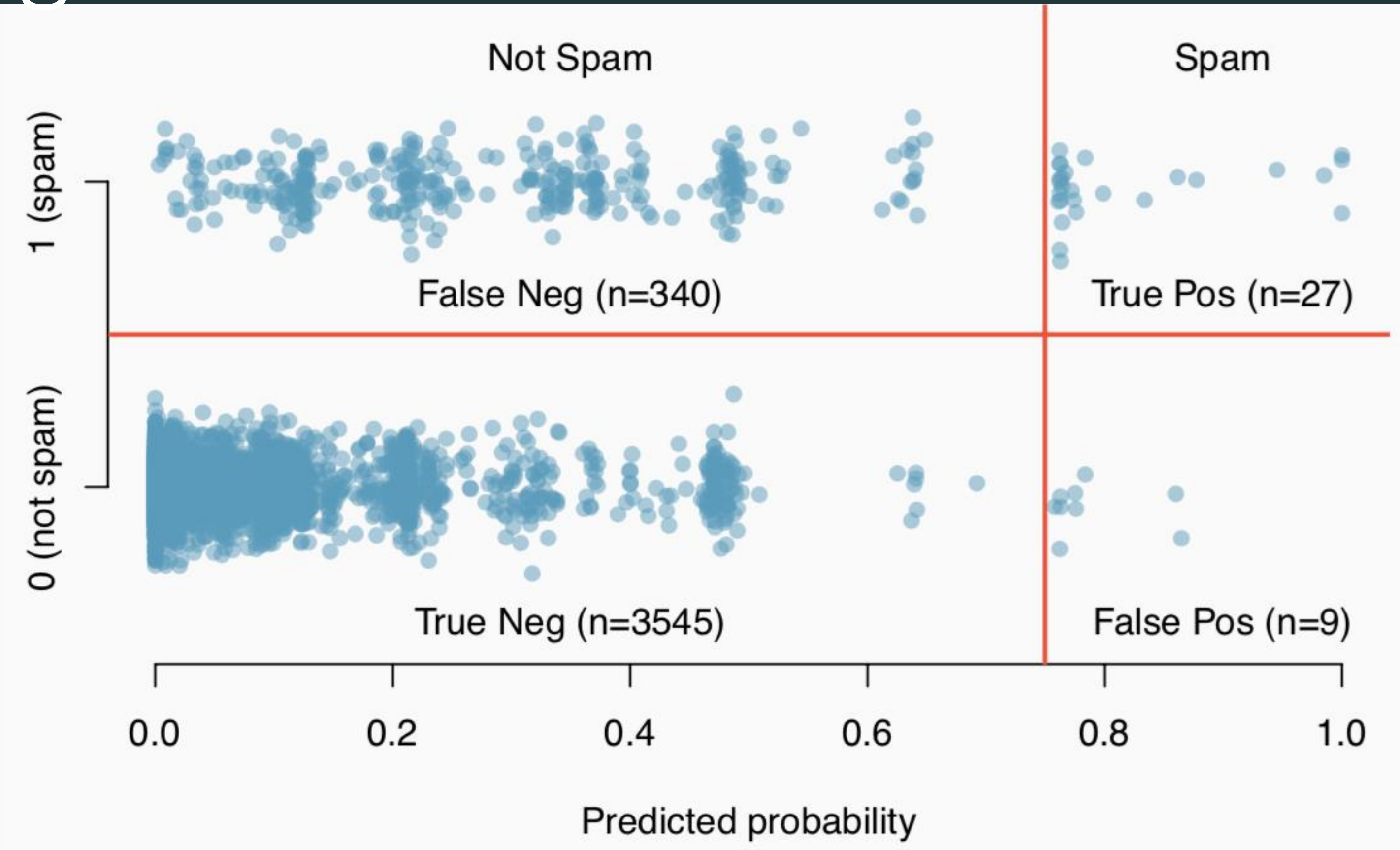
FN = 340	TP = 27
TN = 3545	FP = 9

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				

Useful Links & Resources

- **Reference**:
  - openintro.org/os (Chapter 9, Section 9.5)
- **Helpful Links:**
  - https://www.youtube.com/watch?v=59y8cOwb-xs