Introduction to Logistic Regression

April 22, 2020

Logistic Regression

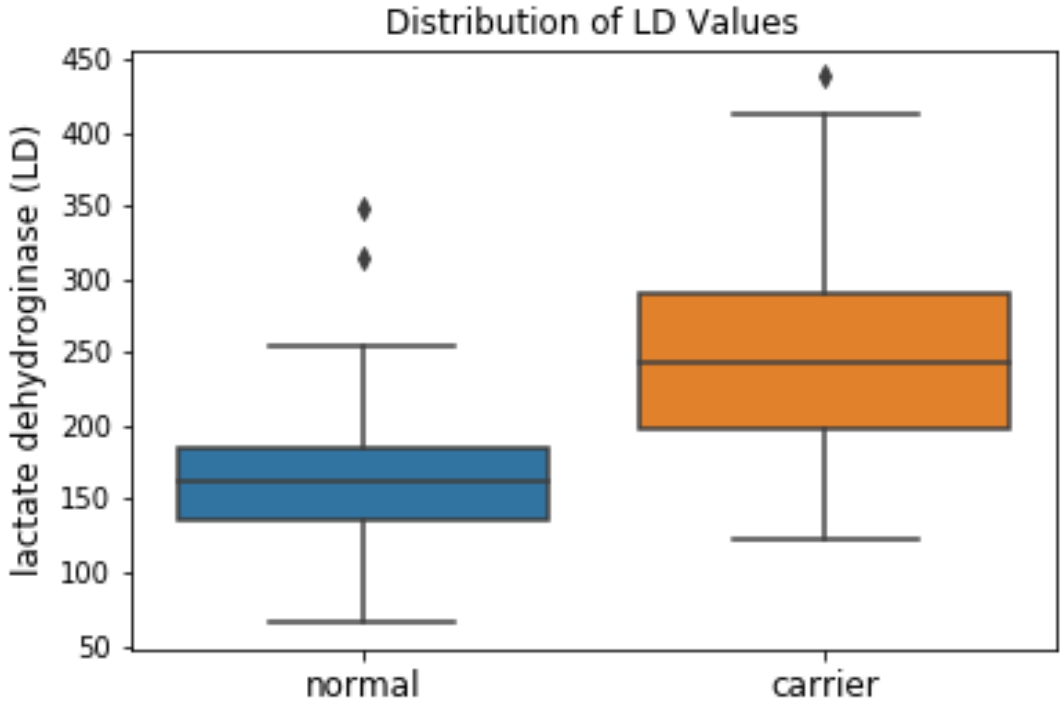
Example: Duchenne Muscular Dystrophy (DMD)

- Genetically-transmitted disease
- Passed from a mother to her children
- Female offspring suffer no apparent symptoms, but male offspring with the disease die at a young age.
- Female carriers tend to exhibit elevated levels of certain serum enzymes or proteins.

Let's say we want to build a model which takes as input the serum levels and outputs our prediction about whether or not a female is a carrier.

Logistic Regression

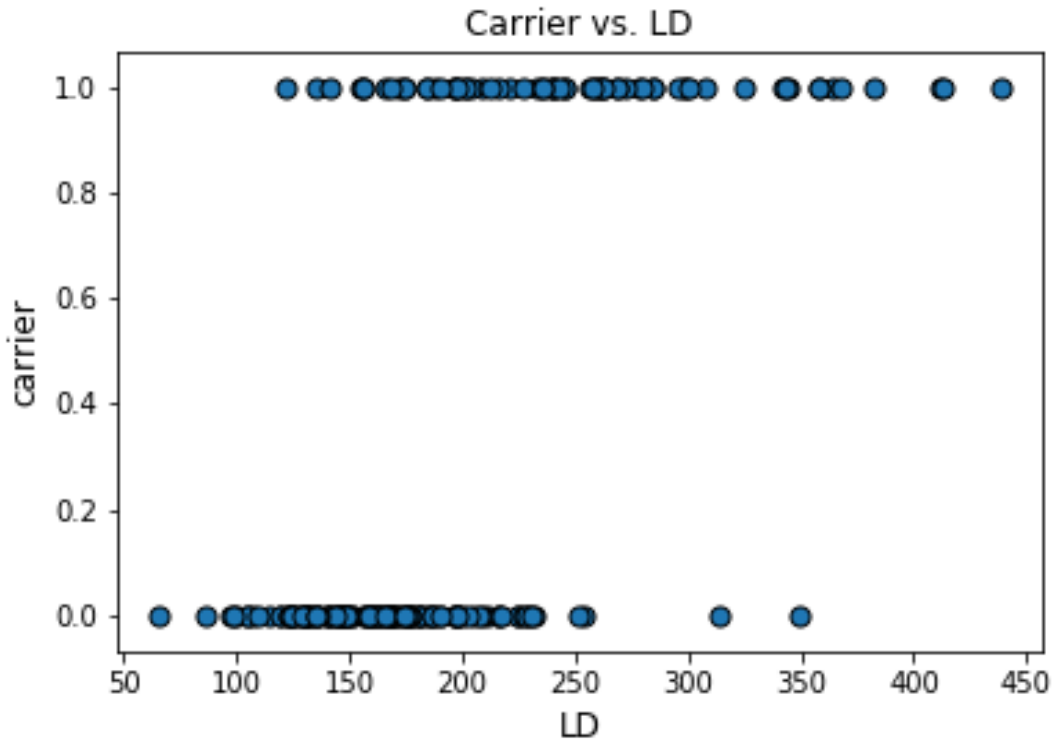
We can start be doing some exploratory analysis.



What we can see is that in our datset, carriers tend to have higher LD values. However, there is some overlap between carriers and non-carriers in the middle values. There is not a single cutoff we can use to classify a person as a carrier or non-carrier.

Logistic Regression

Here is another view of the data, using a scatterplot



Here, we have encoded whether a person is a carrier or not using a numeric 0/1 value. A value of 1 indicates that a person is a carrier.

Logistic Regression

Since there is overlap between carriers and non-carriers, we would probably be best off to not just make a simple prediction of carrier/non-carrier, but instead predict the likelihood or probability that a female is a carrier.

From what we have seen in the plots, females with higher LD values look more likely to be carriers that those with lower values.

Between 175 and 225, it is not clear if a person is a carrier or not, so we might be best off assigning a probability close to 0.5 for people in that range.

Logistic Regression

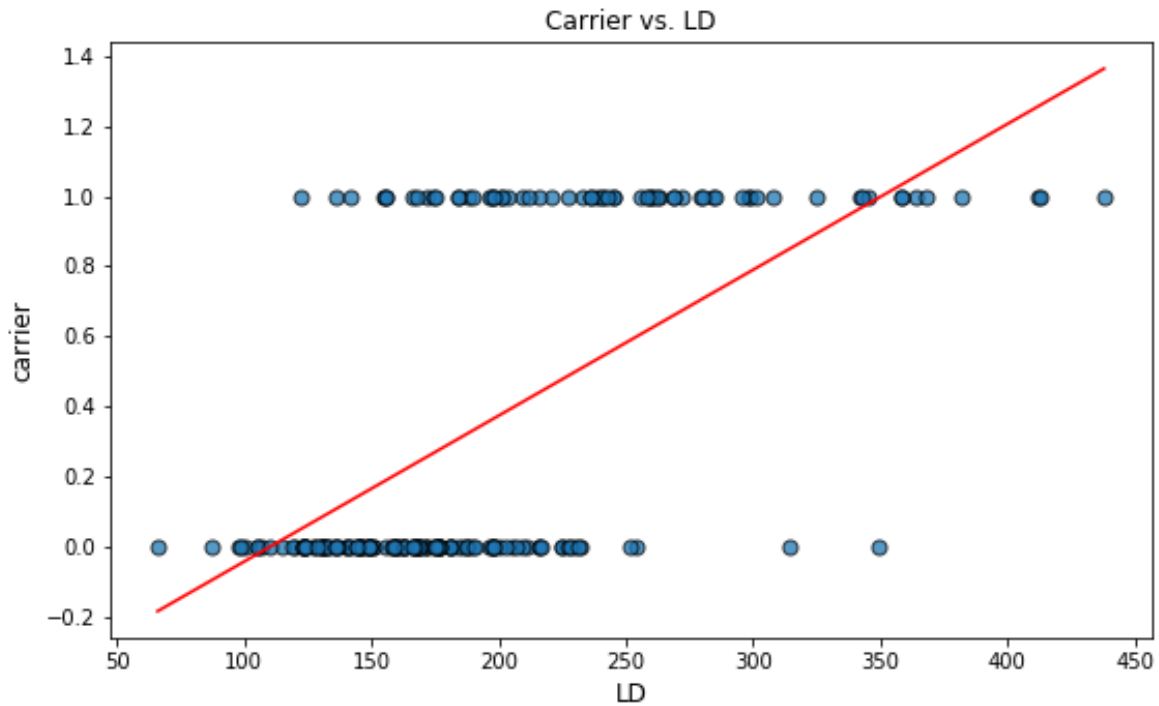
So how do we create our model? We can try using a linear regression model.

Logistic Regression

So how do we create our model? We can try using a linear regression model.

A linear regression model produces the following result:

$$\mathrm{P}(\text{carrier}) = 0.00423 \cdot (\text{LD}) - 0.4757$$



Logistic Regression

This approach has a big problem: probabilities are values between 0 and 1, but this equation has guarantee of outputting values between 0 and 1.

In fact, we can see that for some values, we get predictions less than 0 or greater than 1.

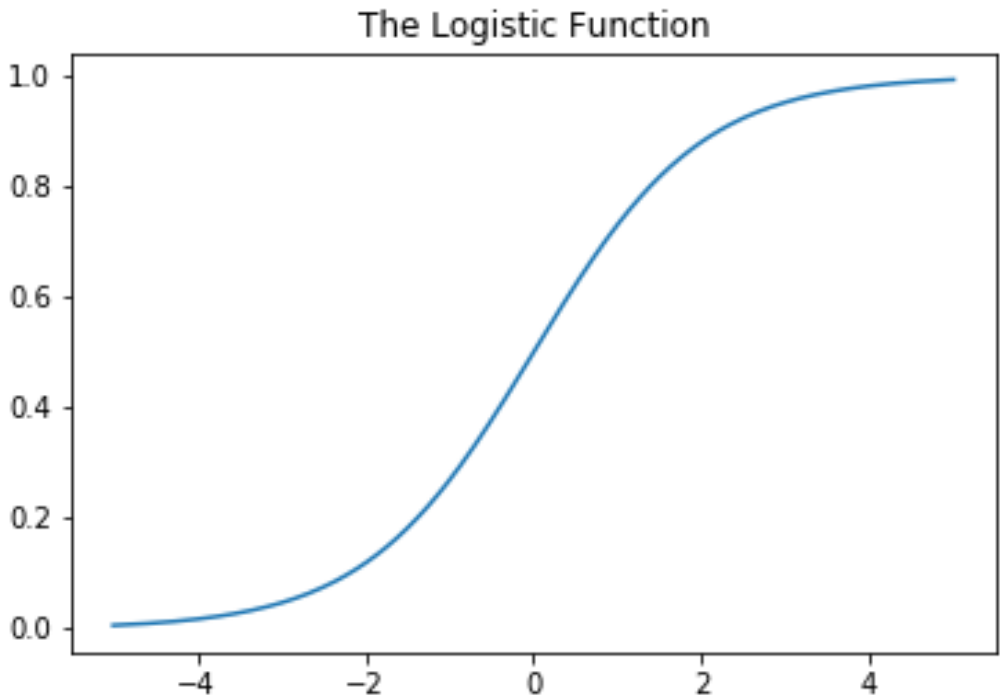
Another problem is that it assumes a fixed change in LD will have a fixed effect on the probability. That is, a change from 60 to 70 will have the same impact as a change from 250 to 260.

Logistic Regression

One possible solution to this problem is to "squash" our output between 0 and 1.

A common way to do this is to pass the output from a linear model into the **logistic function**:

$$l(x) = \frac{1}{1 + e^x}$$



Logistic Regression

This means that instead of our model looking like

$$\mathrm{P}(\text{carrier}) = \beta_1 \cdot (\mathrm{LD}) + \beta_0$$

it will look like

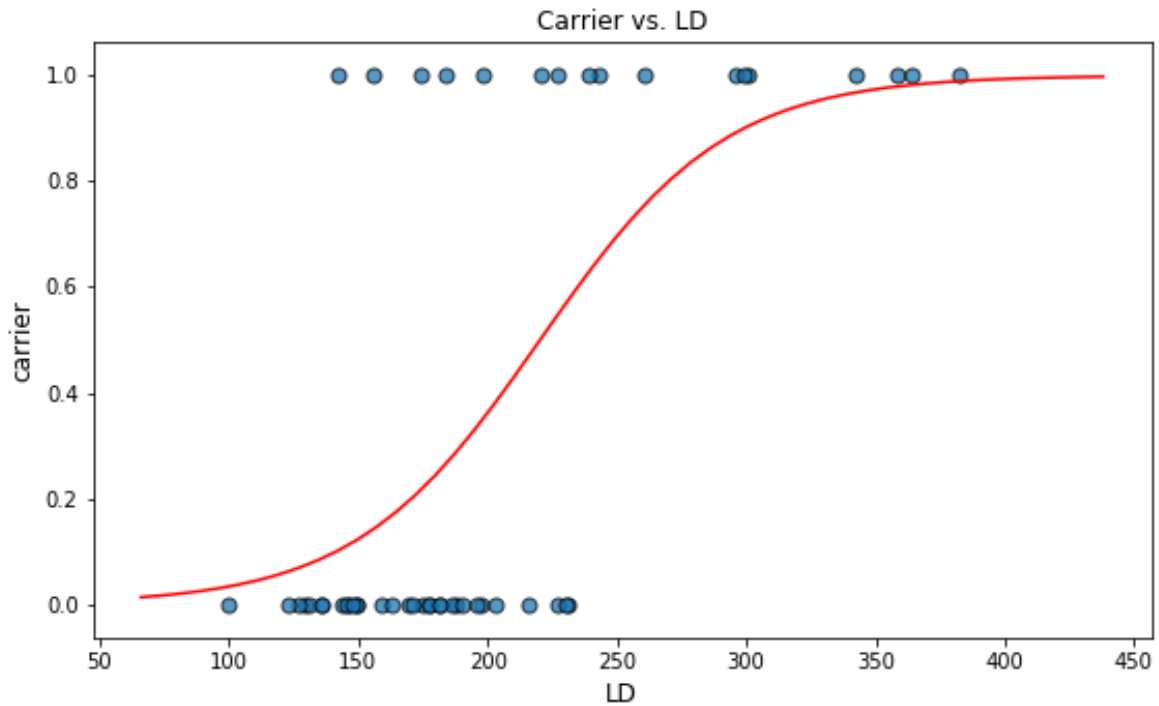
$$\mathrm{P}(\text{carrier}) = \frac{1}{1 + e^{-(\beta_1 \cdot (\mathrm{LD}) + \beta_0)}}$$

Logistic Regression

Fitting this model, we obtain

$$\mathrm{P}(\text{carrier}) = \frac{1}{1 + e^{-(0.0279 \cdot (\mathrm{LD}) - 6.1492)}}$$

And here are those probabilities plotted against the test set:



Logistic Regression

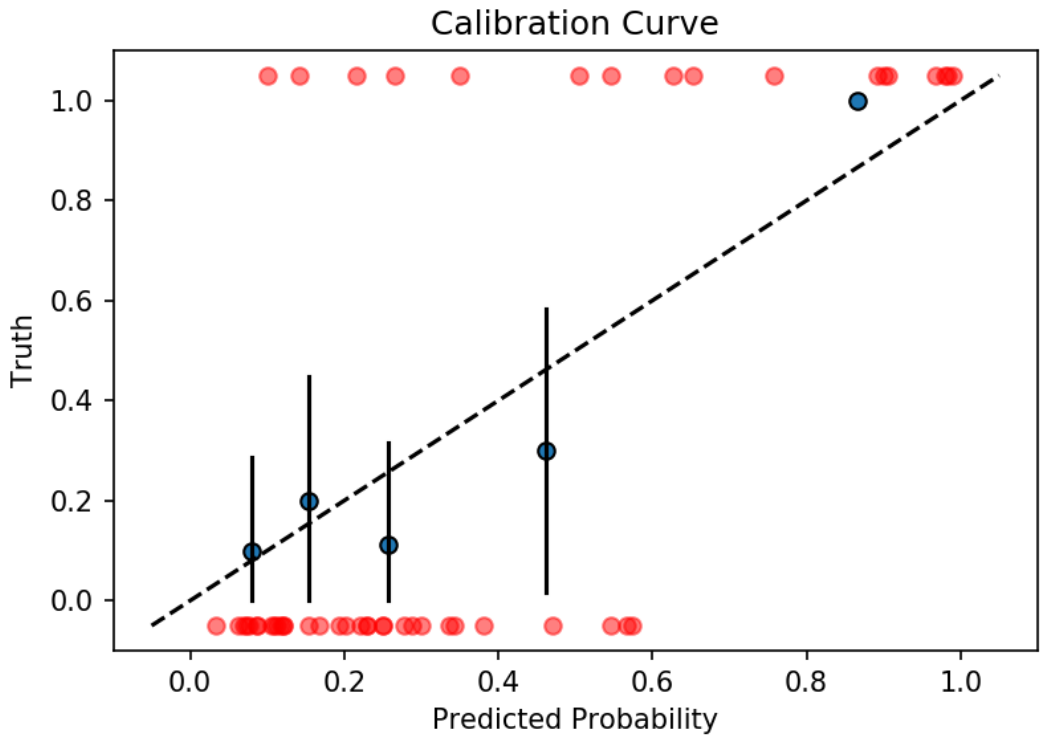
How do we assess how good our model is?
One option is to determine how "calibrated" our model is. That is, if we look at people who we say have a 25% probability of being a carrier, we want about 25% of them to be carriers.

Using this idea, we can build a **calibration curve**. This curve by constructed by:

1. Binning the data into groups based on predicted probabilities.
2. Computing the average predicted probability for each group.
3. Computing the observed proportion for each group, along with a confidence interval (usually a 95% confidence interval).
4. Plot the observed probabilities and confidence intervals against the average probabilities for each group.

Ideally, the plotted points should fall close to the line $y = x$ (because predicted probabilities should be similar to observed probabilities).

Logistic Regression

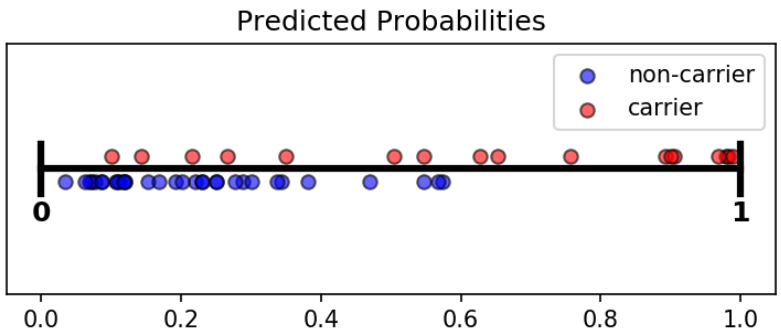


The model is decently well calibrated. All of our confidence intervals intersect the line $y = x$, but the second and third point estimates are a bit low. Also, we have a smallish sample size, so we need to be cautious.

Logistic Regression

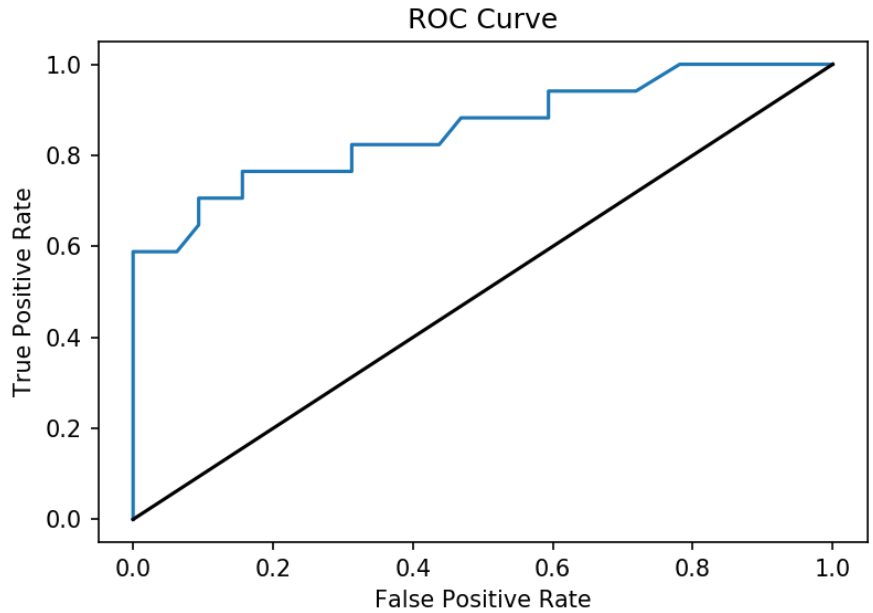
Another way we could look at performance is to calculate how well our model discriminates between positive classes and negative classes.

That is, does the model tend to assign higher probabilities to positive observations compared to negative observations.



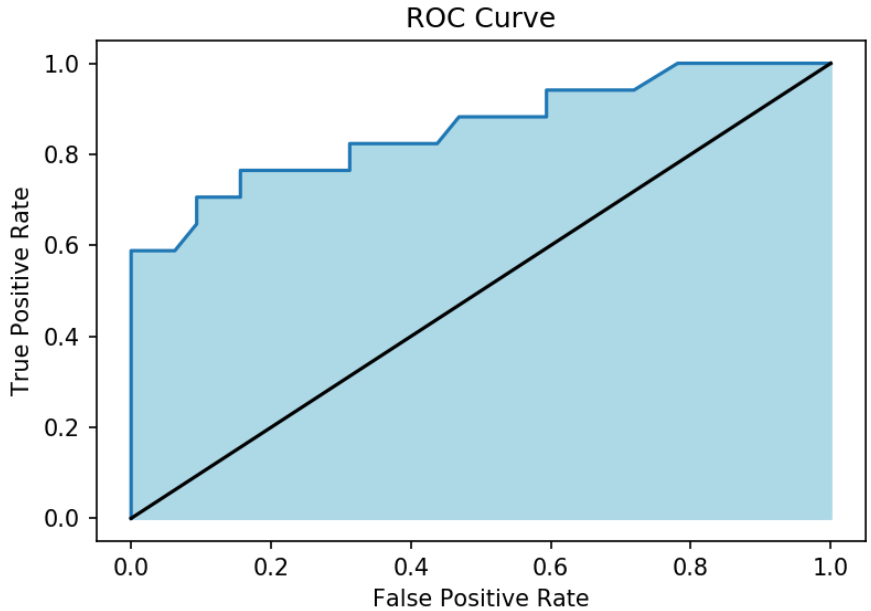
Logistic Regression

Closely related to the plot of the predicted probabilities is the Receiver Operator Characteristic Curve (ROC).



This curve shows the tradeoff between correctly classifying those who are carriers vs. incorrectly classifying those who are not carriers as we adjust the threshold for how sure we have to be to predict that a person is a carrier (see demonstration).

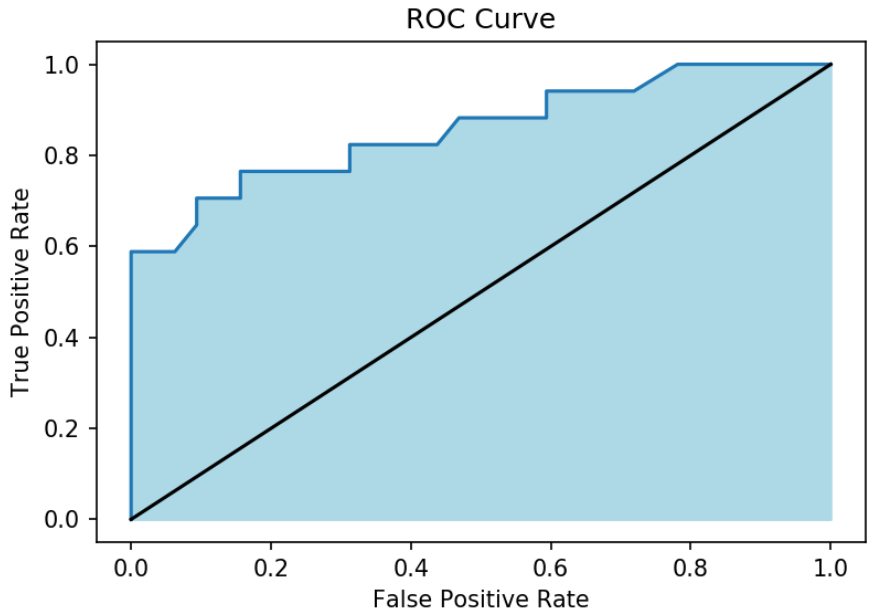
Logistic Regression



The Area under the Receiver Operater Characteristic Curve (ROC AUC) is a useful metric to see how well the model discriminates between the positive and negative classes.

This area is the same as the probability that, given a random positive and a random negative observation, the model assigns a higher probability to the positive observation.

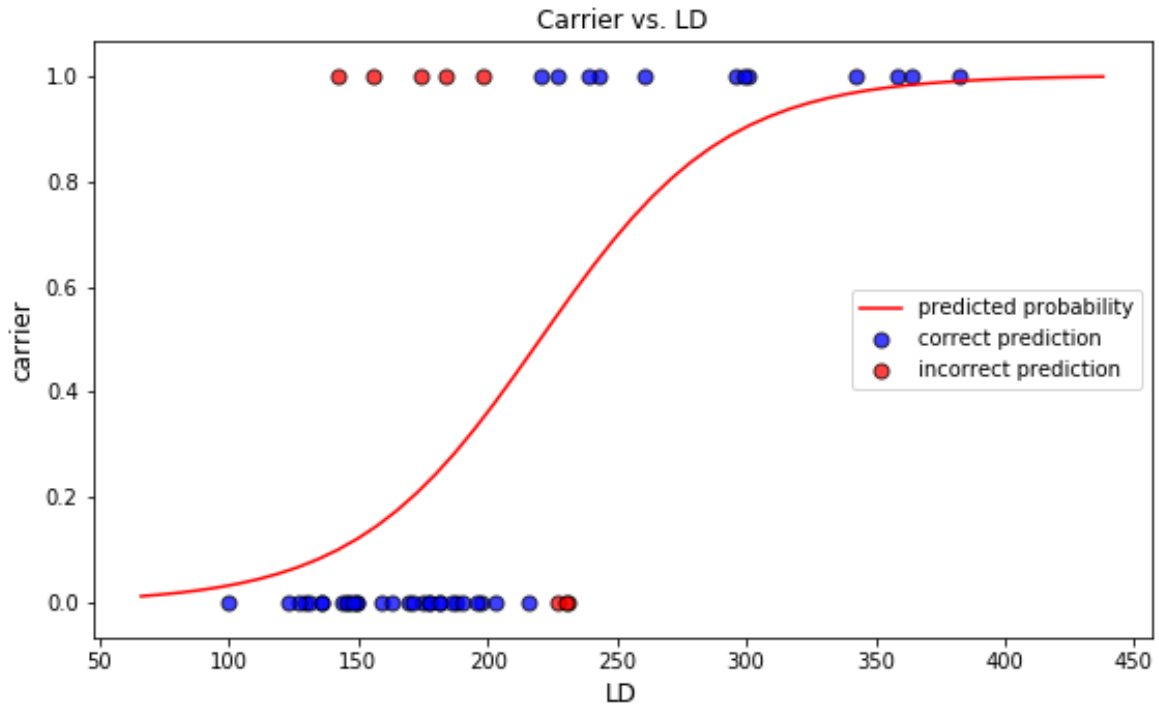
Logistic Regression



Here, the area under the curve is 0.8539.

Logistic Regression

The last thing to do is to assign predictions based on probabilities. The normal convention would be to predict that anyone with a predicted probability greater than 0.5 is a carrier and everyone else as a non-carrier. If we do this, we get the following results on the test set:



Logistic Regression

There are two type of wrong predictions:

Logistic Regression

There are two type of wrong predictions:

- Non-carriers who were predicted as being carriers (**False Positives**)

Logistic Regression

There are two type of wrong predictions:

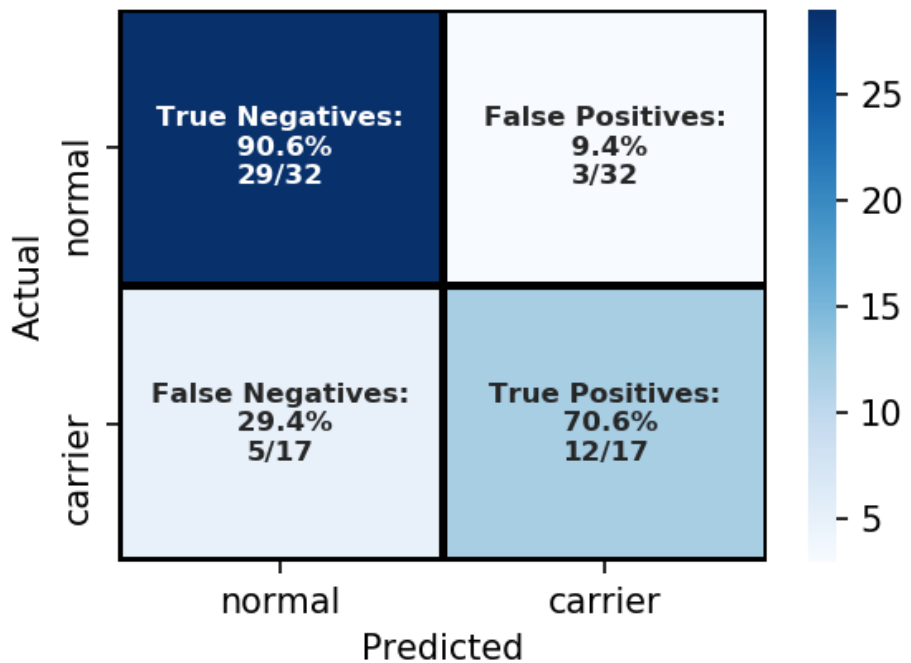
- Non-carriers who were predicted as being carriers (**False Positives**)
- Carriers who were predicted as not being carriers (**False Negatives**)

Logistic Regression

There are two type of wrong predictions:

- Non-carriers who were predicted as being carriers (**False Positives**)
- Carriers who were predicted as not being carriers (**False Negatives**)

We can organize all of the cases into a **confusion matrix**:



Logistic Regression

How might we try and improve our model?

We have three other serum values in our datset, so we could add those in as additional predictor variables. This will turn our model into one that looks like

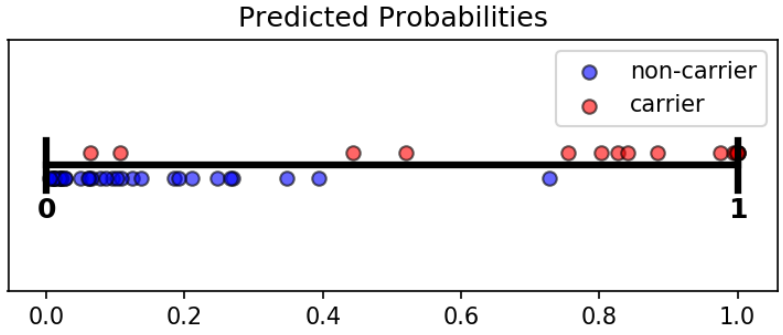
$$P(carrier) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \cdot (\mathrm{CK}) + \beta_2 \cdot (\mathrm{H}) + \beta_3 \cdot (\mathrm{PK}) + \beta_4 \cdot (\mathrm{LD}))}}$$

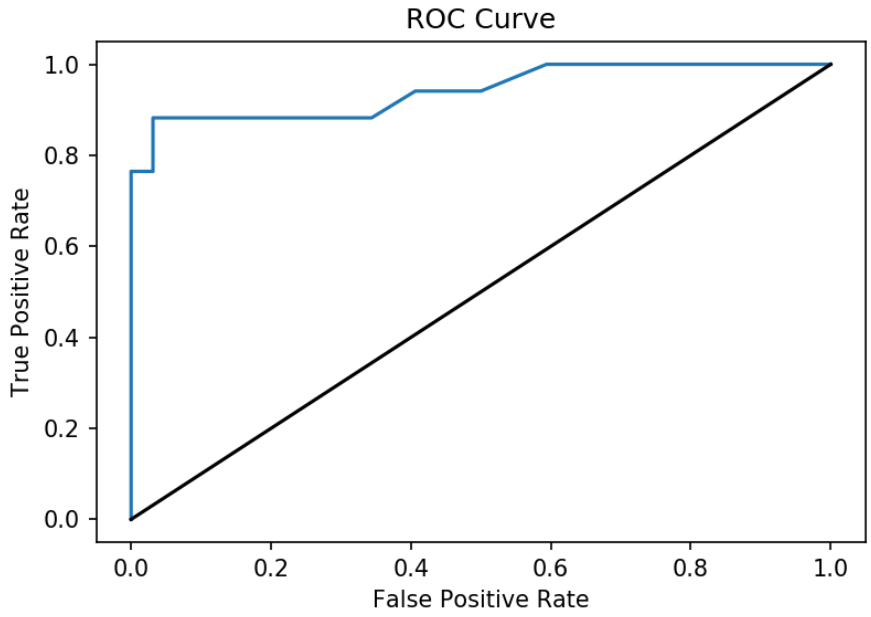
Logistic Regression

After fitting the model, we get

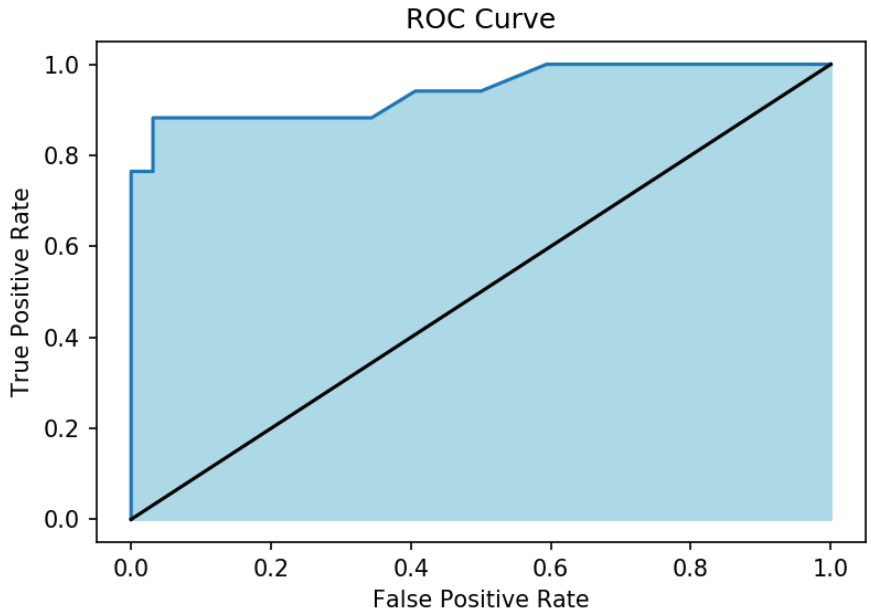
$$P(carrier) = \frac{1}{1 + e^{-(-17.2254+0.0446(\mathrm{CK})+0.1163(\mathrm{H})+0.0875(\mathrm{PK})+0.0122(\mathrm{LD}))}}$$

Logistic Regression





Logistic Regression



This time, the area under the curve is 0.9430.

Logistic Regression

Using this model, we see a higher percentage of correct predictions.