QUICK QUESTION 7 (1/1 point)

In the previous video, we used CART and Random Forest to predict sentiment. Let's see how well logistic regression does. Build a logistic regression model (using the training set) to predict "Negative" using all of the independent variables. You may get a warning messge after building your model - don't worry (we explain what it means in the explanation).

Now, make predictions using the logistic regression model:

predictions = predict(tweetLog, newdata=testSparse, type="response")

where "tweetLog" should be the name of your logistic regression model. You might also get a warning message after this command, but don't worry - it is due to the same problem as the previous warning message.

Build a confusion matrix (with a threshold of 0.5) and compute the accuracy of the model. What is the accuracy?

0.8056338

0.8056338

Answer: 0.8197183

EXPLANATION

You can build a logistic regression model in R by using the command:

tweetLog = glm(Negative ~ ., data=trainSparse, family="binomial")

Then you can make predictions and build a confusion matrix with the following commands:

predictLog = predict(tweetLog, newdata=testSparse, type="response")

table(testSparse$Negative, predictLog >= 0.5)

The accuracy is (254+37)/(254+46+18+37) = 0.8197183, which is worse than the baseline. If you were to compute the accuracy on the training set instead, you would see that the model does really well on the training set - this is an example of over-fitting. The model fits the training set really well, but does not perform well on the test set. A logistic regression model with a large number of variables is particularly at risk for overfitting.

Note that you might have gotten a different answer than us, because the glm function struggles with this many variables. The warning messages that you might have seen in this problem have to do with the number of variables, and the fact that the model is overfitting to the training set. We'll discuss this in more detail in the Homework Assignment.

Is this worse or better than the baseline model accuracy of 84.5%? Think about the properties of logistic regression that might make this the case!

Check Save Hide Answer *You have used 1 of 5 submissions*

EdX is a non-profit created by founding partners Harvard and MIT whose mission is to bring the best of higher education to students of all ages anywhere in the world, wherever there is Internet access. EdX's free online MOOCs are interactive and subjects include computer science, public health, and artificial intelligence.