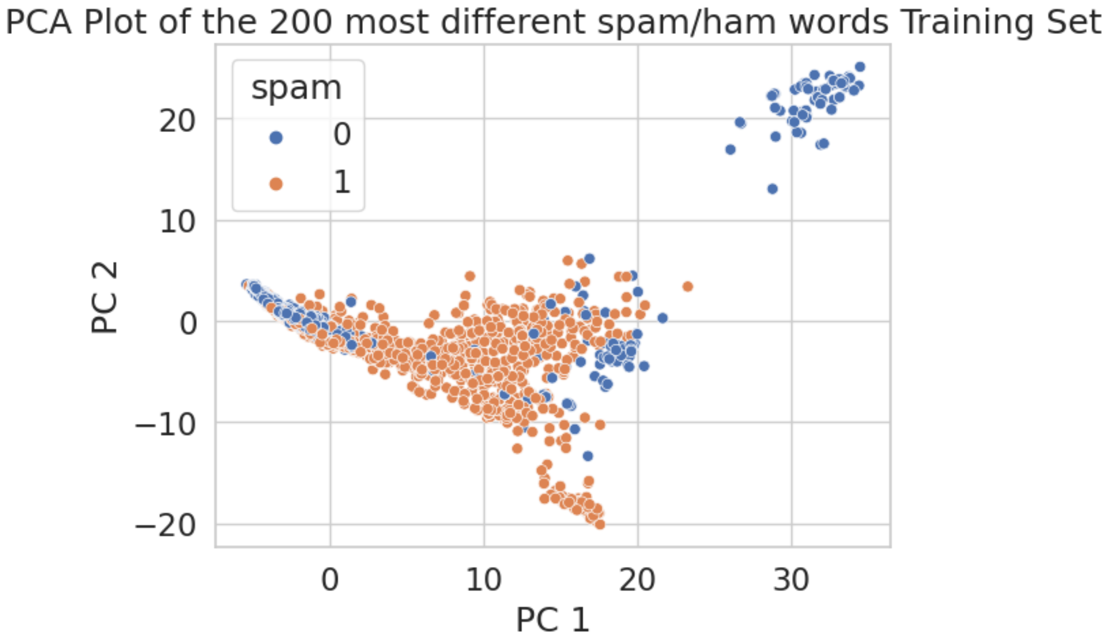
0.1 Question 1a

Generate your visualization in the cell below.

```
In [41]: #### Do PCA with 2 components ###
         from sklearn.decomposition import PCA
         from sklearn.preprocessing import StandardScaler
         scaler = StandardScaler()
         scaled_data = scaler.fit_transform(words_in_texts(words, train['email']))
         pca = PCA(n_components=2)
         principal_components = pca.fit_transform(scaled_data)

         #### Make Plot ####
         sns.scatterplot(data=pd.DataFrame(principal_components).join(train[['spam']]), x=0, y=1, hue='
         plt.title('PCA Plot of the 200 most different spam/ham words Training Set')
         plt.xlabel('PC 1')
         plt.ylabel('PC 2');
```



0.2 Question 1b

In two to three sentences, describe what you plotted and its implications with respect to your features.

I found the 200 words with the largest absolute differnece in the proportion of spam and ham emails (e.g. if it made up 2% of all words in spame emails, and 3% of all words in ham emails, it would get a differnece value of 1%) and conducted a PCA to reduce dimensionality to 2. Plotting this with data colored by spam/ham class/ I can see that there seems to be some clustering of ham emails, especially around the top right of the graph. This indicates that these 200 features will be useful in classification.

1 Question 4

Describe the process of improving your model. You should use at least 2-3 sentences each to address the following questions:

1. How did you find better features for your model?
2. What did you try that worked or didn't work?
3. What was surprising in your search for good features?

1. I looked for features that showed a difference between spam and ham emails. For some features I looked at summary statistics like the mean for each group, and for others like the words I calculated the difference in the proportion of words in each class that the words made up. **2.** I found that some features that I had expected to help improve the model were not that effective. These were mostly the numerical summary statistics like number of words. **3.** I was surprised by the clustering of emails based on my PCA of the 200 most differnt words.

2 Question 5: ROC Curve

In most cases, we won't be able to get 0 false positives and 0 false negatives, so we have to compromise. For example, in the case of cancer screenings, false negatives are comparatively worse than false positives — a false negative means that a patient might not discover that they have cancer until it's too late. In contrast, a patient can receive another screening for a false positive.

Recall that logistic regression calculates the probability that an example belongs to a particular class. To classify an example, we say that an email is spam if our classifier gives it ≥ 0.5 probability of being spam. However, **we can adjust that cutoff threshold**. We can say that an email is spam only if our classifier gives it ≥ 0.7 probability of being spam, for example. This is how we can trade off false positives and false negatives.

The Receiver Operating Characteristic (ROC) curve shows this trade-off for each possible cutoff probability. In the cell below, plot an ROC curve for your final classifier (the one you use to make predictions for Gradescope) on the training data. Lecture 23 may be helpful.

Hint: You'll want to use the `.predict_proba` method (documentation) for your classifier instead of `.predict` to get probabilities instead of binary predictions.

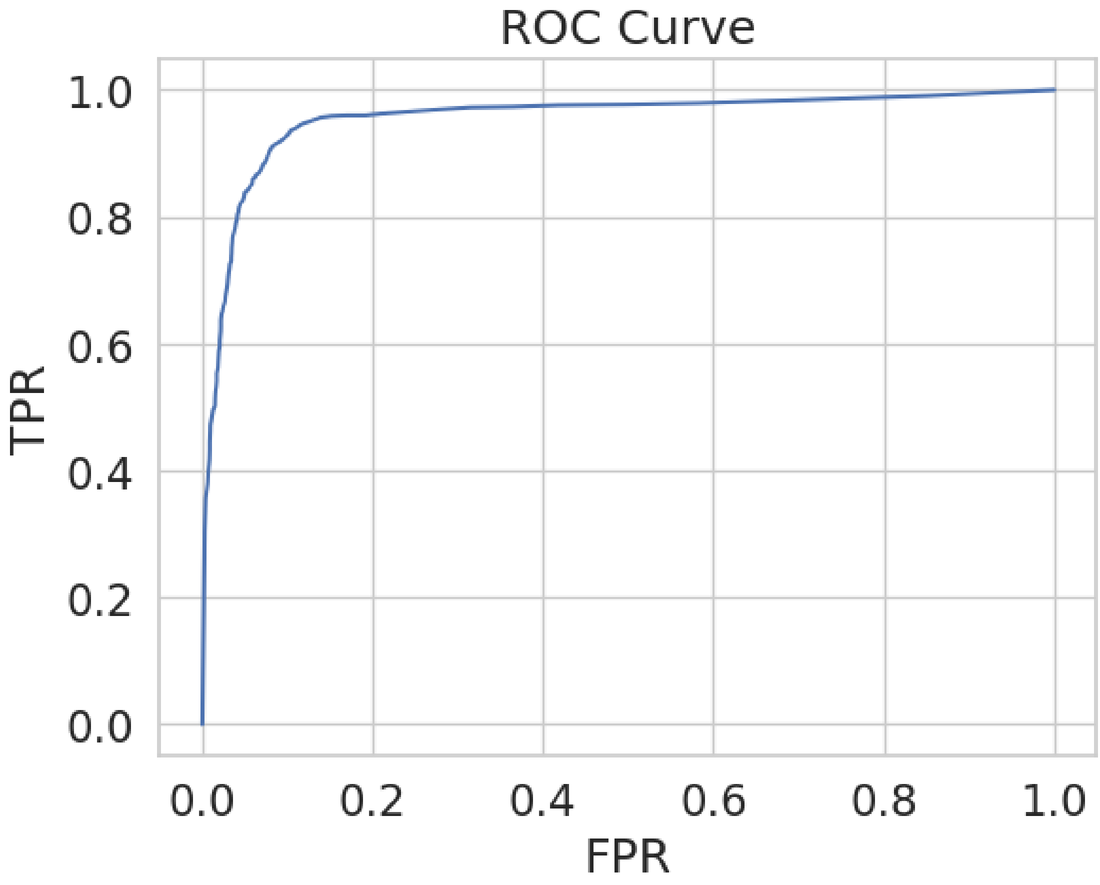
```
In [56]: def predict_threshold(my_model, X, T):
             prob_one = my_model.predict_proba(X)[:, 1]
             return (prob_one >= T).astype(int)

         def tpr_threshold(X, Y, T): # Same as recall
             Y_hat = predict_threshold(my_model, X, T)
             return np.sum((Y_hat == 1) & (Y == 1)) / np.sum(Y == 1)

         def fpr_threshold(X, Y, T):
             Y_hat = predict_threshold(my_model, X, T)
             return np.sum((Y_hat == 1) & (Y == 0)) / np.sum(Y == 0)

         thresholds = np.linspace(0, 1, 100)
         tprs = [tpr_threshold(X, train['spam'], t) for t in thresholds]
         fprs = [fpr_threshold(X, train['spam'], t) for t in thresholds]

         plt.plot(fprs, tprs)
         plt.xlabel("FPR")
         plt.ylabel("TPR")
         plt.title("ROC Curve");
```


ROC Curve
1.0
0.8
0.6
0.4
0.2
0.0
TPR
0.0
0.2
0.4
0.6
0.8
1.0
FPR

2.0.1 Question 6a

Pick at least **one** of the emails provided above to comment on. How would you classify the email (e.g., spam or ham), and does this align with the classification provided in the training data? What could be a reason someone would disagree with *your* classification of the email? In 2-3 sentences, explain your perspective and potential reasons for disagreement.

EXAMPLE 3: The data classifies this email as *ham.* I would classify this email as *spam.* **MY REASON:** The email contains potentially unsolicited promotion for a product. **REASON FOR DISAGREEMENT** The email if from a partner of Ryanair, who the recipient likeley has a flight booked with. The ability to make cheap international phone calls may be very relevant to someone who might be flying overseas and thus be relevant to their life an not a nuisance.

2.0.2 Question 6b

As data scientists, we sometimes take the data to be a fixed “ground truth,” establishing the “correct” classification of emails. However, as you might have seen above, some emails can be ambiguous; people may disagree about whether an email is actually spam or ham. How does the ambiguity in our labeled data (spam or ham) affect our understanding of the model’s predictions and the way we measure/evaluate our model’s performance?

PREDICTIONS: The ambiguity in the labeled data means that we have to understand that our model can only be acurate to the standard set by the original classification of the emails. Specifically, we are trying our best to train our model to approximate the classifications of the original people who labled the data, and thus that we are carrying this standard forward onto new data. **PERFORMANCE:** Knowing this, when we evaluate model perfomance, we need to keep in mind that false positives and negatives are defined in terms of the standard of the orignal data. Thus, our model might be working better or worse than its statistics show when measured against out own standards.

Part ii In 2-3 sentences, explain why you think the feature you chose to remove changed how your email was classified.

The email I chose contained only the word bank out of all of the words used to change the model, so that word had a very large sway over the classification of that particular email. Therefore, when that one word was removed from the model, the probability that that particular email was spam drastiaclly changed, and as a result its classification changed.

Part i In this context, do you think you could easily find a feature that could change an email's classification as you did in part a)? Why or why not?

The method I used (isolating an email which only had one of the features in the model) would be very impractcal as more and more features are added to the model for the simple reason that most messages will likeley contain more than one word out of a set of 1000. Aditionally, each feature likely will exert less influence on its own, so emails will have to be closer to the decision boundry for a single word to change their classification.

Part ii Would you expect this new model to be more or less interpretable than `simple_model`?

Note: A model is considered interpretable if you can easily understand the reasoning behind its predictions and classifications. For example, the model we saw in part a), `simple_model`, is considered interpretable as we can identify which features contribute to an email's classification.

The new model would be *less interpretable* than the old model because it will be much harder to understand why each individual email was classified the way it was. This is because there would be three orders of magnitude more words to consider when making a descision, and it may not be entirely clear which ones are factoring into a given decision.

2.0.3 Question 7c

Now, imagine you're a data scientist at Meta, developing a text classification model to decide whether to remove certain posts / comments on Facebook. In particular, you're primarily working on moderating the following categories of content: * Hate speech * Misinformation * Violence and incitement

Pick one of these types of content to focus on (or if you have another type you'd like to focus on, feel free to comment on that!). What content would fall under the category you've chosen? Refer to Facebook's Community Standards, which outline what is and isn't allowed on Facebook.

Hate Speech: Breifly paraprased: Attacks against an individual or individuals based on protected characteristics such as race, gender, or sexuality.

2.0.4 Question 7d

What are the stakes of misclassifying a post in the context of a social media platform? Comment on what a false positive or false negative means for the category of content you've chosen (hate speech, misinformation, or violence and incitement).

FALSE POSITIVES
INDIVIDUAL: A user posting non-hateful content will have their post removed. Other users will not be able to see their post. Their account may be at risk for punative action.
PLATFORM: If false positives occur most frequentlty in discussion of controversial (but non-hateful) political topics, a chilling effect could occur on the platform

FALSE NEGATIVES:
INDIVIDUAL: A user posting hateful content will be allowed to continue harrasing other users. People will continue to see their posts, and may feel othered, alienated, or otherwise hurt by the posters words. They may choose to leave the platform.
PLATFORM: If false negatives occur too frequently, Meta could become a breeding ground for hateful ideoloies, and/or minority groups might be forces off of the platform.

2.0.5 Question 7e

As a data scientist, why might having an interpretable model be useful when moderating content online?

As mentioned in 7d, there are real consequences to misclassification. Those who have a post incorrectly taken down, or those who are hurt by a post that wasn;t taken down, are likeley to want an explanation (that they can understand / makes sense) about how the descision was made. Aditionally, those who have their posts correctly flagged as hate speech might also demand an explanation as to why their posts ahve been restricted. Further, all users of the site will likeley perfer to feel like the site's moderation "makes sense."