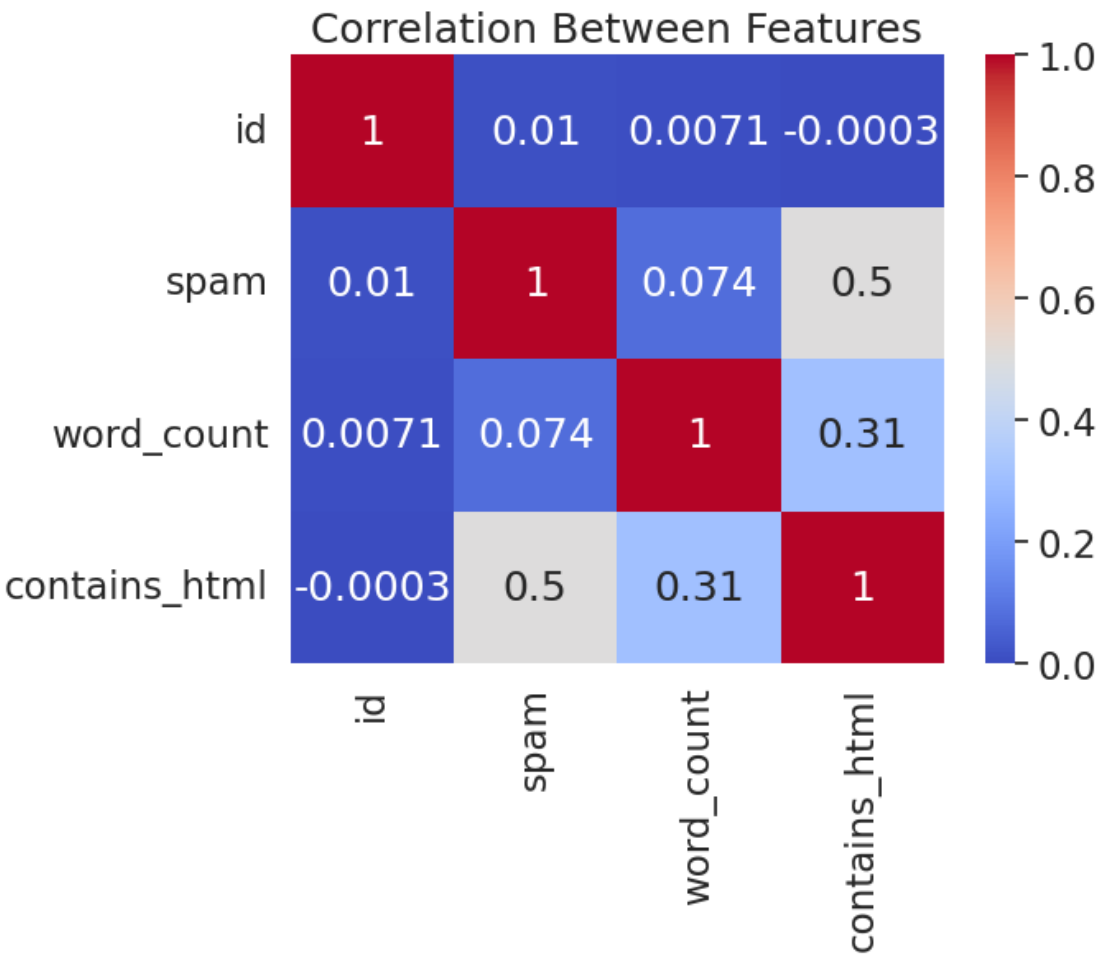
0.1 Question 1a

Generate your visualization in the cell below.

```
In [11]: import seaborn as sns
         import matplotlib.pyplot as plt
         train['word_count'] = train['email'].apply(lambda x: len(x.split()))
         train['contains_html'] = train['email'].apply(lambda x: 1 if '<html>' in x.lower() else 0)
         numeric_train = train.select_dtypes(include=['number'])
         corr_matrix = numeric_train.corr()
         sns.heatmap(corr_matrix, annot=True, cmap="coolwarm")
         plt.title("Correlation Between Features")
         plt.show()
```



0.2 Question 1b

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

The heatmap above shows the correlation between the numeric features in the data set. The contains_html feature has a moderate positive correlation (0.5) with spam, which means that emails containing HTML tags are more likely to be spam. On the other hand, word_count has a weak correlation (0.074) with spam, indicating it might be less useful as a predictor. The id feature has no significant correlation with spam or other features so it should also be excluded from the model.

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. Reviewing the correlation heat map helps give ideas of what variables are going to be important with identifying spam emails. Also, thinking in terms of real life and what types of emails are more likely to be spam is helpful. Things like punctuation, capital letters, and URLs are often signs of a spam email. Key words like "free" or "win" also help the models ability to distinguish between spam and ham.
2. Adding features like the ratio of capital letters and the presence of specific key words helped significantly to boost the accuracy. However, some features like number of lines in the email text did not help in improving the model much. Adjusting the key words could also perhaps continue to increase the accuracy by finding more specific spam words.
3. What was surprising was how strong certain spam keywords influence the predictions. Features like "contains_url" and "contains_html" had more significant impact that initially thought. Punctuation count also had less of an impact than I thought, ideally because I thought that spam emails used a ton of exclamation marks.

2 Question 5: ROC Curve

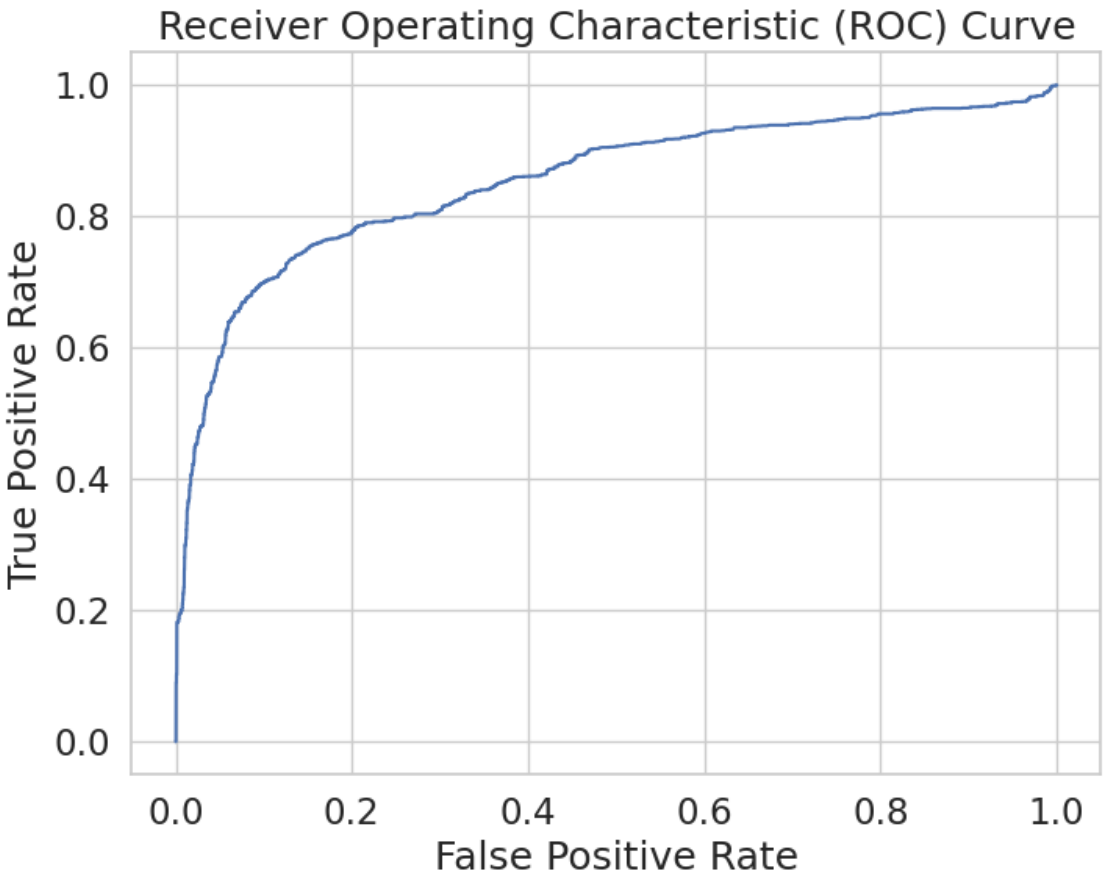
In most cases, we won't be able to get 0 false positives and 0 false negatives, so we have to compromise. For example, in the case of cancer screenings, false negatives are comparatively worse than false positives — a false negative means that a patient might not discover that they have cancer until 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 [19]: from sklearn.metrics import roc_curve, roc_auc_score
         import matplotlib.pyplot as plt
         y_scores = model.predict_proba(X_train)[:, 1]
         fpr, tpr, thresholds = roc_curve(Y_train, y_scores)
         auc_score = roc_auc_score(Y_train, y_scores)
         plt.figure(figsize=(8, 6))
         plt.plot(fpr, tpr)
         plt.xlabel("False Positive Rate")
         plt.ylabel("True Positive Rate")
         plt.title("Receiver Operating Characteristic (ROC) Curve")
         plt.show()
```


Receiver Operating Characteristic (ROC) Curve
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0

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.

I would classify example 1 as ham because it appears to be a personal email with not spam like content. They mention family events and personal things. The email indicates overall personal communication rather than spam. This aligns with the classification in the training data. However it was marked as spam and I could see why given the spelling errors and the incorrect indentations.

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?

The data was likely generated from various email sources, either manually labeled or categorized using automation. If humans labeled it then it was subject to some possible forms of bias. Also, assumptions about what defines spam, like promotional language or certain key words, could lead to inconsistent labeling, especially when contextual factors are involved. For example formal emails from mailing lists could be labeled as ham when in reality they are spam. Since the model performance is based on accuracy metrics, then biases and assumptions that are incorrect could influence the performance.

Part ii Please provide below the index of the email that you removed (`email_idx`). Additionally, in 2-3 sentences, explain why you think the feature you chose to remove changed how your email was classified.

The email index: 7477. The feature I chose to remove was "prescription". The word is strongly associated with spam emails and especially in the context of online pharmacies. Removing the feature reduced the models ability to classify the email as spam from 70.16% to ham at 23.41%.

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?

In the new context, I do not think I would easily find a feature, it would be much harder to find one single feature to change the email classification. Being able to identify important features is more complex as the relationship between the features and classification is less clear due to many features potentially having importance.

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 much less inerpretable than the simple_model because with a larger number of features, it is more difficult to pinpoint certain features that are responsible for a spam or ham classification.

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.

I picked missinformation since I think it is now more relevant than ever. The information would include content that basically spreads false or misleading information that could harm people or communities. Recently in the past 4 years we saw alot of misinformation regarding vaccines / public health, elections, climate change, and even financial crypto scams. There could also be images, videos, or suspicious links that intentionally mislead people. Under Facebook's community standards, any misinformation that could incite panic or undermine the trust in our democratic systems would be monitored.

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 and false negative means for the category of content you've chosen (hate speech, misinformation, or violence and incitement).

In this scenario, a false positive could lead to censorship, frustrate users, and destroy the trust in the platform, much like Twiter a couple years ago. Flagging legitimate news articles about vaccines as misinformation could discourage important disussions. On the other hand false negatives allows for harmful and incorrect information to circulate and spread online. Platforms like X now do community checks and verify information that goes "viral" in order to prevent misinformation. Overall, both scenarios call for a stronger need for more accurate and balanced moderation.

2.0.5 Question 7e

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

It is crucial to have an interpretable model because it allows data scientists and decision makers to make accurate decisions and understand the why. If it was interpretable I think one would have an easier time spotting bias which in turn could lead to better accuracy and precision. The platform users benefit from this in a very direct way by having a safer and more trustworthy online environment. In todays time where AI is having a greater impact on not only our economy, but platforms like Facebook and Instagram, we need models that we an rely on. Important decisions are being made on the daily basis that have real life consequences on peoples lives. Even future AI regulation is going to be influenced and determined by meaningful and accurate models.