
0.1 Question 1

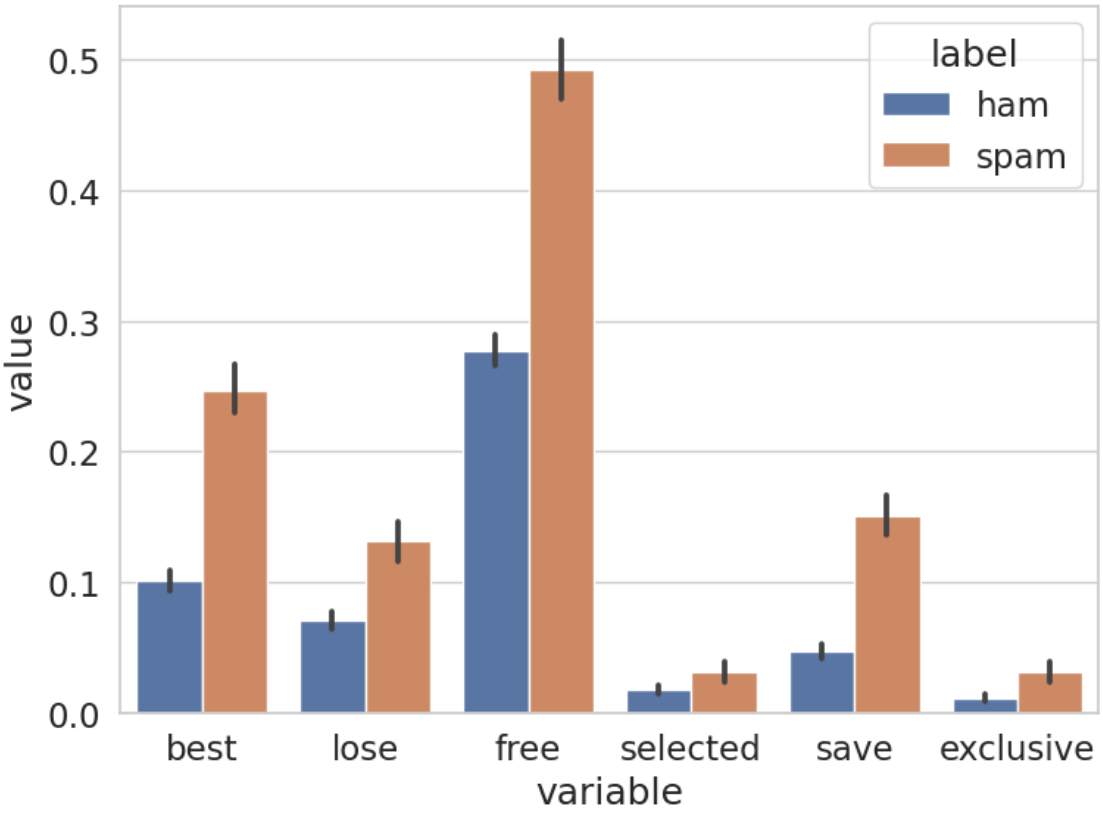
Discuss one attribute or characteristic you notice that is different between the two emails that might relate to the identification of a spam email.

A characteristic that is distinctly different from the spam and the ham email is the formatting. The spam email has a lost of html objects scattered throughout the email, whereas the ham email is very organized. Another detail is that the words used for the spam email use sweeping generalizations that are trying to sell the reader something.

Create your bar chart with the following cell:

```
In [26]: train = train.reset_index(drop=True) # We must do this in order to preserve the ordering of em
         plt.figure(figsize=(8,6))

         texts = [el.lower() for el in train.email.values]
         words = ['best', 'lose', 'free', 'selected', 'save', 'exclusive']

         new_train = train.copy()

         new_train[words] = words_in_texts(words, texts)

         dat = new_train.iloc[:, range(-7, 0)].melt('spam')

         spam_dict = {0: 'ham', 1: 'spam'}

         dat['label'] = dat['spam'].map(spam_dict)

         display(dat)

         sns.barplot(data = dat, x = 'variable', y = 'value', hue = 'label')

         plt.tight_layout()
         plt.show()
```

```
       spam   variable  value label
0         0       best      0   ham
1         0       best      0   ham
2         0       best      0   ham
3         0       best      0   ham
4         0       best      1   ham
…     …         …    …   …
45073     0  exclusive      0   ham
45074     1  exclusive      0  spam
45075     0  exclusive      0   ham
45076     0  exclusive      0   ham
45077     0  exclusive      0   ham

[45078 rows x 4 columns]
```


label
ham
spam
value
0.0
0.1
0.2
0.3
0.4
0.5
best
lose
free
selected
save
exclusive
variable

0.2 Question 6c

Explain your results in Question 6a and Question 6b. How did you know what to assign to `zero_predictor_fp`, `zero_predictor_fn`, `zero_predictor_acc`, and `zero_predictor_recall`?

The accuracy of labeling all of mail as not spam mail is 74%. This is misleading because the number of mail that's spam is not the same as the amount of emails that are not spam. By using the zero-predictor, we are simply measuring the rate of non-spam emails in the inbox.

Since we are using the zero predictor, the accuracy is the rate at which the Y_train == 0.

The zero predictor false positive rate is the rate at which Y_train == 1 and the zero_model predicts 0, which is never.

The zero predictor false negative rate is the rate at which Y_train == 0.

Recall is the true_positive/(true_positive + false_negative), or the rate at which the model correctly predicts 1, or when an email is spam.

0.3 Question 6f

How does the accuracy of the logistic regression classifier `my_model` compare to the accuracy of the zero predictor?

```
In [64]: my_model_acc = np.mean(Y_train_hat == Y_train)
         model_acc_diff = my_model_acc - zero_predictor_acc
         model_acc_diff
```

```
Out[64]: 0.012910954345800585
```

The accuracy is marginally higher than the zero-predictor model. The logitic regression model is 1.2% more accurate.

0.4 Question 6g

Given the word features provided in Question 4, discuss why the logistic regression classifier `my_model` may be performing poorly.

Hint: Think about how prevalent these words are in the email set.

It is likely due to the feature selection. The accuracy is poor because we are underfitting the model. The existing model only has 6 features. Creating and adding more features that can tease-out the details of email classification would make the model more accuracte.

0.5 Question 6h

Would you prefer to use the logistic regression classifier `my_model` or the zero predictor classifier for a spam filter? Why? Describe your reasoning and relate it to at least one of the evaluation metrics you have computed so far.

I would rather use my_model. First, the accuracy is a bit higher. Second, if the rate of spam mail increases into my inbox, the accuracy of the zero-predcitor model will decrease. Finally, using the zero-predictor model also doesn't sound like a smart idea considering we know for a fact that there is spam in the inbox.