Introduction to Data Science

SPAM EMAIL DETECTION

BRIAN D'ALESSANDRO

WHY TEACH EMAIL SPAM?

Fighting email SPAM is an incredibly interesting DS problem, as it encompasses many elements of production grade DS challenges.

- Problem is highly adversarial, so constant updating is necessary
- Problem has high potential for selection bias and negative feedback loops
- Problem is inherently multi-task
- Features are mixed text, categorical and numeric
- Data is very high dimensional, sparse and large

LETS BUILD A DATASET FOR SPAM FILTERING

SPAM – TARGET & SAMPLING

1. What is the Target Variable?

2. What does it look like?

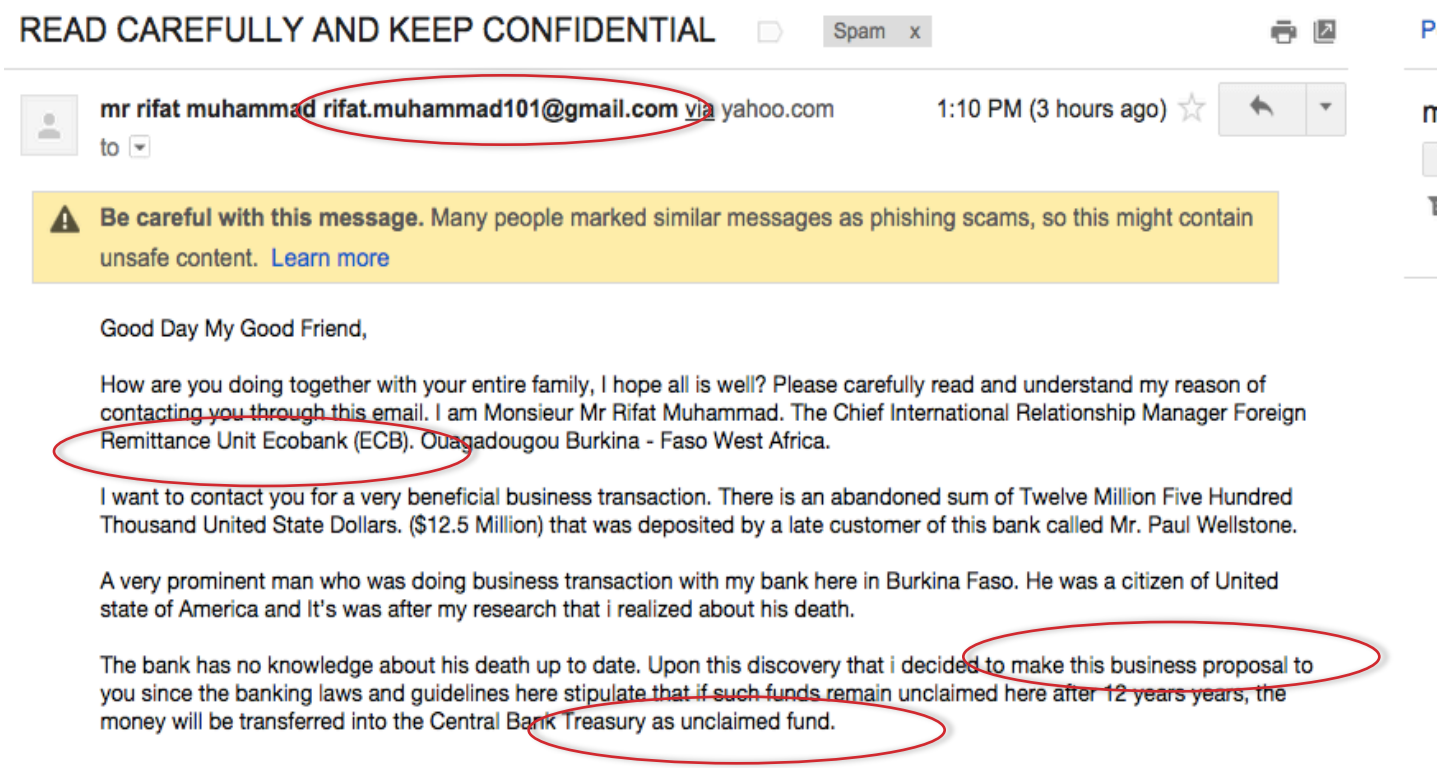
3. How do we know if an email is spam or not?

4. What is the likely distribution of the observed target variable? How should we sample it?

Practical aside: when you start a modeling project from scratch, you generally always have to pull your own data. You should always invest sufficient time to design your data acquisition strategy (answering questions like this). Always document your decisions thoroughly to aid in future peer and risk review.

SPAM – FEATURES

Lets look at an example spam and identify aspects that give it away.
How should we encode these as features?



READ CAREFULLY AND KEEP CONFIDENTIAL Spam x

mr rifat muhammad rifat.muhammad101@gmail.com via yahoo.com 1:10 PM (3 hours ago)

to

Be careful with this message. Many people marked similar messages as phishing scams, so this might contain unsafe content. Learn more

Good Day My Good Friend,

How are you doing together with your entire family, I hope all is well? Please carefully read and understand my reason of contacting you through this email. I am Monsieur Mr Rifat Muhammad. The Chief International Relationship Manager Foreign Remittance Unit Ecobank (ECB). Ouagadougou Burkina - Faso West Africa.

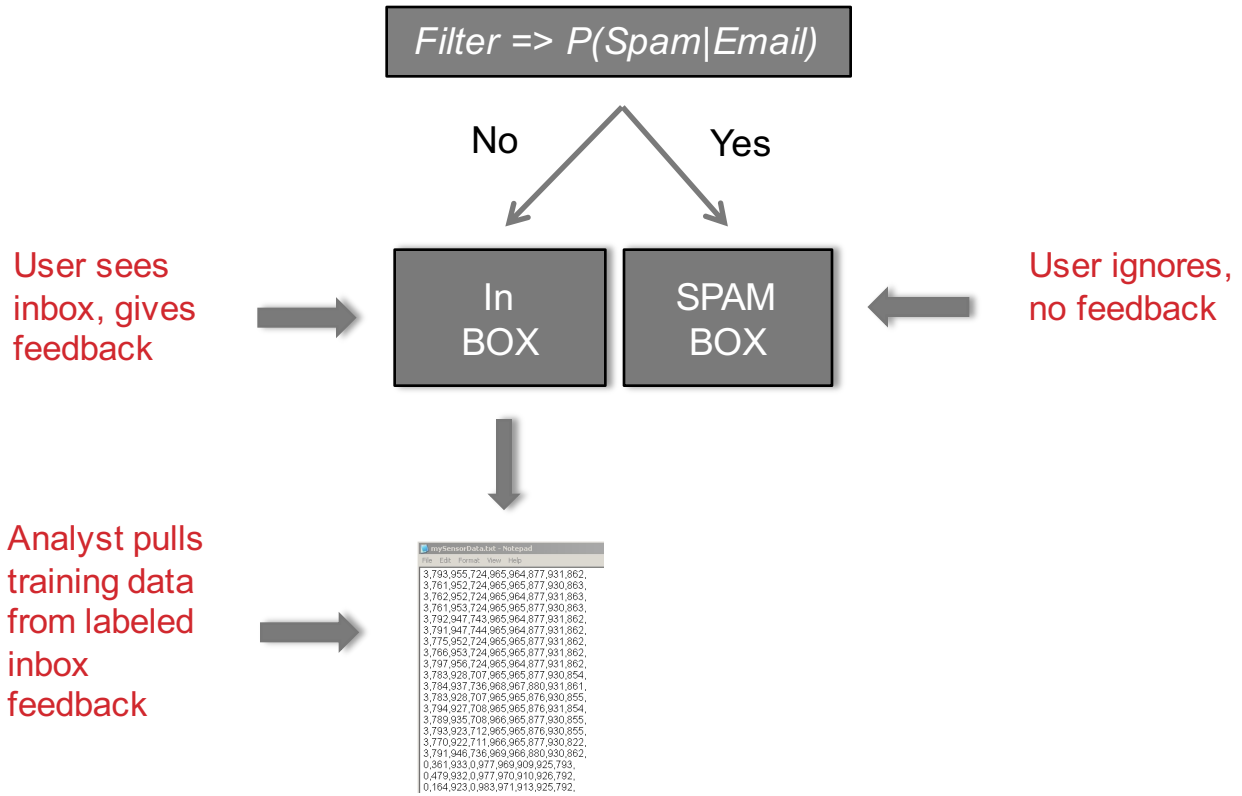
I want to contact you for a very beneficial business transaction. There is an abandoned sum of Twelve Million Five Hundred Thousand United State Dollars. ($12.5 Million) that was deposited by a late customer of this bank called Mr. Paul Wellstone.

A very prominent man who was doing business transaction with my bank here in Burkina Faso. He was a citizen of United state of America and It's was after my research that i realized about his death.

The bank has no knowledge about his death up to date. Upon this discovery that i decided to make this business proposal to you since the banking laws and guidelines here stipulate that if such funds remain unclaimed here after 12 years years, the money will be transferred into the Central Bank Treasury as unclaimed fund.

SPAM – FEEDBACK LOOP

Congrats, you are ready to update your SPAM filter!
Now lets anticipate what could go wrong, including negative feedback loops.



Practical aside: many ML based products involve thinking deeply about the user experience. If you depend on user feedback for labels, how do you ensure you get consistent and thorough labels?

In this case, most of the real spam likely goes unlabeled, and users will inconsistently label SPAM in their inboxes.

As the DS you need to anticipate these problems and design your short and long term strategies to mitigate them.

SPAM – PRACTICAL QUESTIONS

1. If the training data is pulled only from labeled Inbox, is selection bias likely? Why?
2. Should you put some SPAM in the user's inbox so that you can get more training labels?
3. What if user deletes w/o reading but doesn't label spam?
4. How often should we retrain the model?
5. What would happen if you completely retrained model from scratch but didn't include examples from the SPAM box?

Practical aside: Often times, the thought put into questions like these are more important than the time deliberating on what algorithms to use.

Some of these are non-technical, requiring input from product and design colleagues.

SPAM – PRACTICAL QUESTIONS

1. If the training data is pulled only from labeled Inbox, is selection bias likely? Why? Yes. Features that highly predict SPAM will be less represented in the inbox. Predictive features from the first iteration may be less predictive in future iterations, causing model degradation.
2. Should you put some SPAM in the user's inbox so that you can get more training labels? This is the best way to get an unbiased dataset, but too much SPAM makes the service less usable. Best to apply 'active learning' techniques. Can also sample more obvious cases from SPAM box to include in training data.
3. What if user deletes w/o reading but doesn't label spam? This is tricky. Best approach might be to train models with and without these entries labeled as SPAM, and AB test them.
4. How often should we retrain the model? This is an empirical question. If SPAMmers get smarter, more SPAM will end up in Inbox. We can monitor SPAM labeling rates and retrain when they start to trend upwards.
5. What would happen if you completely retrained model from scratch but didn't include examples from the SPAM box? Features that were highly predictive of SPAM, but not observed in Inbox may not get learned in the new model. Thus you end of with a negative feedback loop. One could use the techniques discussed above, or also build the model in an incremental, online fashion. That way learned predictive features that are no longer observe won't change.