Assignment#2. Categorize Mistakes

Jae Dong Hwang

Implement a way to get the raw context for the samples where your model is most-wrong. Recall this includes examples where the true answer was 1, but the model gives very low probabilities, and examples where the true answer was 0, but gives very high probabilities.

Produce a list of the 20 worst false positives made by running logistic regression on the initial train/test split with 10 mutual information features

- False Positive - the true answer was 0, but gives very high probabilities

Probabilities	**Test Raw**
0.9857846245077153	Hello handsome ! Are you finding that job ? Not being lazy ? Working towards getting back that net for mummy ? Where's my boytoy now ? Does he miss me ?
0.9853404694265699	True dear..i sat to pray evening and felt so.so i sms'd you in some time...
0.9853404694265699	Ok. Every night take a warm bath drink a cup of milk and you'll see a work of magic. You still need to loose weight. Just so that you know
0.9853404694265699	Yeah do! Don’t stand to close tho- you’ll catch something!
0.980487581593895	Lovely smell on this bus and it ain't tobacco...
0.980487581593895	Will do. Was exhausted on train this morning. Too much wine and pie. You sleep well too
0.980487581593895	Send this to ur friends and receive something about ur voice..... How is my speaking expression? 1.childish 2.naughty 3.Sentiment 4.rowdy 5.ful of attitude 6.romantic 7.shy 8.Attractive 9.funny <#> .irritating <#> .lovable. reply me..
0.980487581593895	Wen ur lovable bcums angry wid u, dnt take it seriously.. Coz being angry is d most childish n true way of showing deep affection, care n luv!.. kettoda manda... Have nice day da.
0.9798812997702371	Please reserve ticket on saturday eve from chennai to thirunelvali and again from tirunelvali to chennai on sunday eve...i already see in net..no ticket available..i want to book ticket through tackle ..
0.9798812997702371	Doesn't g have class early tomorrow and thus shouldn't be trying to smoke at <#>
0.9798812997702371	Get down in gandhipuram and walk to cross cut road. Right side <#> street road and turn at first right.

Probabilities	Test Raw
0.9798812997702371	Hi. Wk been ok - on hols now! Yes on for a bit of a run. Forgot that i have hairdressers appointment at four so need to get home n shower beforehand. Does that cause prob for u?"
0.9740474449935478	Thank you so much. When we skyped wit kz and sura, we didnt get the pleasure of your company. Hope you are good. We've given you ultimatum oh! We are countin down to aburo. Enjoy! This is the message i sent days ago
0.9740474449935478	Ups which is 3days also, and the shipping company that takes 2wks. The other way is usps which takes a week but when it gets to lag you may have to bribe nipost to get your stuff.
0.9740474449935478	There is os called ubandu which will run without installing in hard disk...you can use that os to copy the important files in system and give it to repair shop..
0.9732465179509926	Yetunde, i'm sorry but moji and i seem too busy to be able to go shopping. Can you just please find some other way to get what you wanted us to get. Please forgive me. You can reply free via yahoo messenger.
0.9732465179509926	Thank you so much. When we skyped wit kz and sura, we didnt get the pleasure of your company. Hope you are good. We've given you ultimatum oh! We are countin down to aburo. Enjoy!
0.9732465179509926	Hey i booked the kb on sat already... what other lessons are we going for ah? Keep your sat night free we need to meet and confirm our lodging
0.9732465179509926	Oh and by the way you do have more food in your fridge! Want to go out for a meal tonight?
0.9698845717052931	That's ok. I popped in to ask bout something and she said you'd been in. Are you around tonght wen this girl comes?

Produce a list of the 20 worst false negatives made by running logistic regression on the initial train/test split with 10 mutual information features

- False Negatives - the true answer was 1, but the model gives very low probabilities

Probabilities	Test Raw
0.017102139176027785	FreeMsg Why haven't you replied to my text? I'm Randy, sexy, female and live local. Luv to hear from u. Netcollex Ltd 08700621170150p per msg reply Stop to end
0.017102139176027785	Hi I'm sue. I am 20 years old and work as a lapdancer. I love sex. Text me live - I'm i my bedroom now. text SUE to 89555. By TextOperator G2 1DA 150ppmsg 18+

Probabilities	Test Raw
0.022260931031803273	Reminder: You have not downloaded the content you have already paid for. Goto http://doit. mymoby. tv/ to collect your content.
0.043793359146198786	08714712388 between 10am-7pm Cost 10p
0.06716028386908282	FreeMsg Hey there darling it's been 3 week's now and no word back! I'd like some fun you up for it still? Tb ok! XxX std chgs to send, £1.50 to rcv
0.06716028386908282	SIX chances to win CASH! From 100 to 20,000 pounds txt> CSH11 and send to 87575. Cost 150p/day, 6days, 16+ TsandCs apply Reply HL 4 info
0.06716028386908282	SMS. ac Sptv: The New Jersey Devils and the Detroit Red Wings play Ice Hockey. Correct or Incorrect? End? Reply END SPTV
0.06716028386908282	As a valued customer, I am pleased to advise you that following recent review of your Mob No. you are awarded with a £1500 Bonus Prize, call 09066364589
0.06716028386908282	Did you hear about the new "Divorce Barbie"? It comes with all of Ken's stuff!
0.06716028386908282	Your free ringtone is waiting to be collected. Simply text the password "MIX" to 85069 to verify. Get Usher and Britney. FML, PO Box 5249, MK17 92H. 450Ppw 16
0.06716028386908282	Will u meet ur dream partner soon? Is ur career off 2 a flyng start? 2 find out free, txt HORO followed by ur star sign, e. g. HORO ARIES
0.06716028386908282	We tried to contact you re your reply to our offer of a Video Handset? 750 anytime networks mins? UNLIMITED TEXT? Camcorder? Reply or call 08000930705 NOW
0.06716028386908282	Hey I am really horny want to chat or see me naked text hot to 69698 text charged at 150pm to unsubscribe text stop 69698
0.06716028386908282	HMV BONUS SPECIAL 500 pounds of genuine HMV vouchers to be won. Just answer 4 easy questions. Play Now! Send HMV to 86688 More info:www.100percent-real.com
0.06716028386908282	SMS. ac Blind Date 4U!: Rodds1 is 21/m from Aberdeen, United Kingdom. Check Him out http://img. sms. ac/W/icmb3cktz8r7!-4 no Blind Dates send HIDE
0.06716028386908282	SMS. ac sun0819 posts HELLO:"You seem cool, wanted to say hi. HI!!!" Stop? Send STOP to 62468
0.06716028386908282	You have an important customer service announcement from PREMIER.
0.06716028386908282	SMSSERVICES. for yourinclusive text credits, pls goto www.comuk.net login= 3qxj9 unsubscribe with STOP, no extra charge. help 08702840625.COMUK. 220-CM2 9AE

Probabilities	Test Raw
0.06716028386908282	Do you want 750 anytime any network mins 150 text and a NEW VIDEO phone for only five pounds per week call 08002888812 or reply for delivery tomorrow
0.06716028386908282	We tried to contact you re our offer of New Video Phone 750 anytime any network mins HALF PRICE Rental camcorder call 08000930705 or reply for delivery Wed

Categorize the false positives into at least 4 categories.

Message Index	Type	Category
0	3	very_long
1	1	has_lower_i
2	3	very_long
3	0	has_exlamations
4	2	has_dots
5	3	very_long
6	2	has_dots
7	2	has_dots
8	2	has_dots
9	3	very_long
10	3	very_long
11	1	has_lower_i
12	0	has_exlamations
13	3	very_long
14	2	has_dots
15	1	has_lower_i
16	0	has_exlamations
17	1	has_lower_i
18	3	very_long
19	3	very_long

Categorize the false negatives into at least 4 categories.

Message Index	Type	Category

Message Index	Type	Category
0	0	has_call_nums
1	1	many_uppers
2	2	has_url
3	3	lengthy_line
4	3	lengthy_line
5	1	many_uppers
6	3	lengthy_line
7	0	has_call_nums
8	3	lengthy_line
9	1	many_uppers
10	3	lengthy_line
11	0	has_call_nums
12	3	lengthy_line
13	1	many_uppers
14	2	has_url
15	3	lengthy_line
16	3	lengthy_line
17	1	many_uppers
18	0	has_call_nums
19	0	has_call_nums

(reference: generated by category_mistakes_report_utils.py)

In no more than 150 words describe the insight you got from this process, including one new heuristic feature you think would reduce the bad false positives, and one that would reduce the bad false negatives.

I categorized the messages above based on the observation, which actually was based on N tokens. I think it helps model to understand the context of sms messages, which make model hard to make mistakes.

In the first problem, feature selection, I learned that the number of features selected would increase accuracy of prediction. Assuming we can utilized the performance optimized library and access to higher volumn of training data, we should be able to increase the accuracy than what currently we are getting in this homework.

And also, if we generate modeling with randomly sorted data and more independent events, I think the performance of model would increase to handle test data incoming.

To categorize the messages, I used heuristic features listed above in table. An here are self-descriptive functions that i used.

```
def many_uppers(line, U=3):
  """Has many upper cases? """
  upper_cnt = 0
  for word in line.split(' '):
      if word.isupper():
          upper_cnt += 1
  if upper_cnt > U:
      return 1
  return 0

def has_call_nums(line):
  """Has 'call' or 'text' followed by numbers"""
  return contains(line, r'(call \d+|text \d+)')

def has_dots(line):
  """Has more than two dots(..)"""
  return contains(line, r'\.\.')

def has_lower_i(line):
  """Has a typo, i, instead of I"""
  return contains(line, ' i ')

def very_long(line):
  """Is message way long?"""
  return lengthy_line(line, 120)

def has_exlamations(line):
  """Has multiple bangs!!"""
  return contains(line, '.*!.*!')
```
