<final project>

Predicting Consumer Disputes

Philipp Grimm

May 19, 2014

GA Data Science

<problem>

[video]

We get a lot of complaints

We cannot investigate all of them

How can we better manage our workflow?

lifecycle of a consumer complaint

1. Consumer submits complaint online

2. CFPB screens complaint and routes it to company

3. Company responds to complaint

4. Consumer reviews response **and has the option to file a dispute**

5. If dispute: CFPB performs its own investigation of complaint

OMG so much text

OMG so much potential for additional workload

</problem>

<data>

What is (are?) the data?

Two data sources:

- Public consumer complaint database (API? woot!)
- Internal complaint data warehouse (SQL? 0.1*woot!)

  (but currently only a small subset (~10k))

```
> str(data)
'data.frame':   233636 obs. of  14 variables:
 $ Complaint.ID        : int  853713 854060 851569 851472 852567 852997 851961
 $ Product             : Factor w/ 8 levels "Bank account or service",..: 5 5
 $ Sub.product         : Factor w/ 28 levels "","(CD) Certificate of deposit",
 $ Issue               : Factor w/ 71 levels "Account opening, closing, or mana
 $ Sub.issue           : Factor w/ 48 levels "","Account status",..: 16 16 1 1
 $ State               : Factor w/ 63 levels "","AA","AE","AK",..: 10 10 29 57
 $ ZIP.code            : int  96150 95826 4210 24014 31027 97062 91902 33972 2
 $ Submitted.via       : Factor w/ 6 levels "Email","Fax",..: 6 6 6 6 6 6 6 6 6
 $ Date.received       : Factor w/ 898 levels "01/01/2012","01/01/2013",..: 40
 $ Date.sent.to.company: Factor w/ 848 levels "","01/01/2013",..: 380 380 380 3
 $ Company             : Factor w/ 1910 levels "(Former)Shapiro, Swertfeger & H
 $ Company.response    : Factor w/ 7 levels "Closed","Closed with explanation"
 $ Timely.response.    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ Consumer.disputed.  : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 2 1 2 2 1 ...
```

</data>

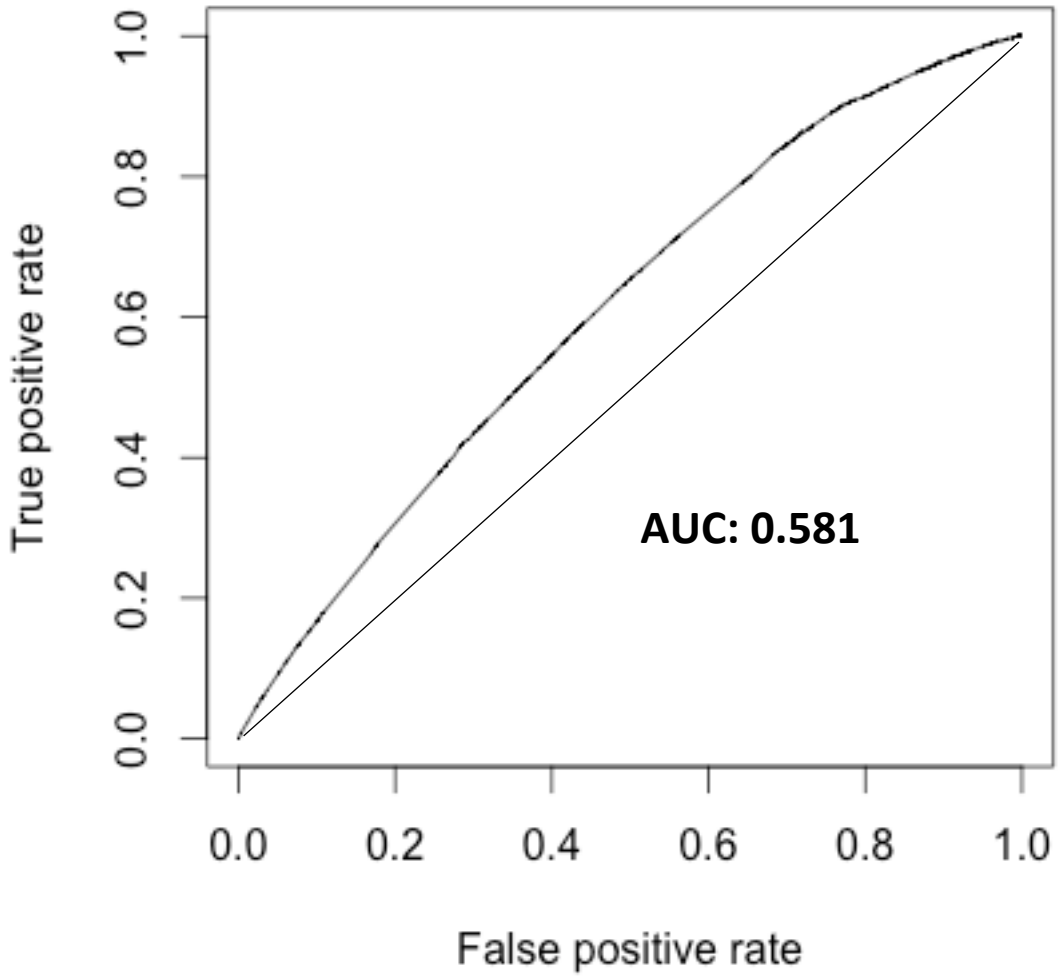
<model>

- Binary classification problem: logistic regression ftw!
- Text mining? Let's call (tm), our friendly neighborhood R package
- I don't have time to babysit my models. Let's do something unsupervised: LDA (topicmodels)

</model>

<findings>

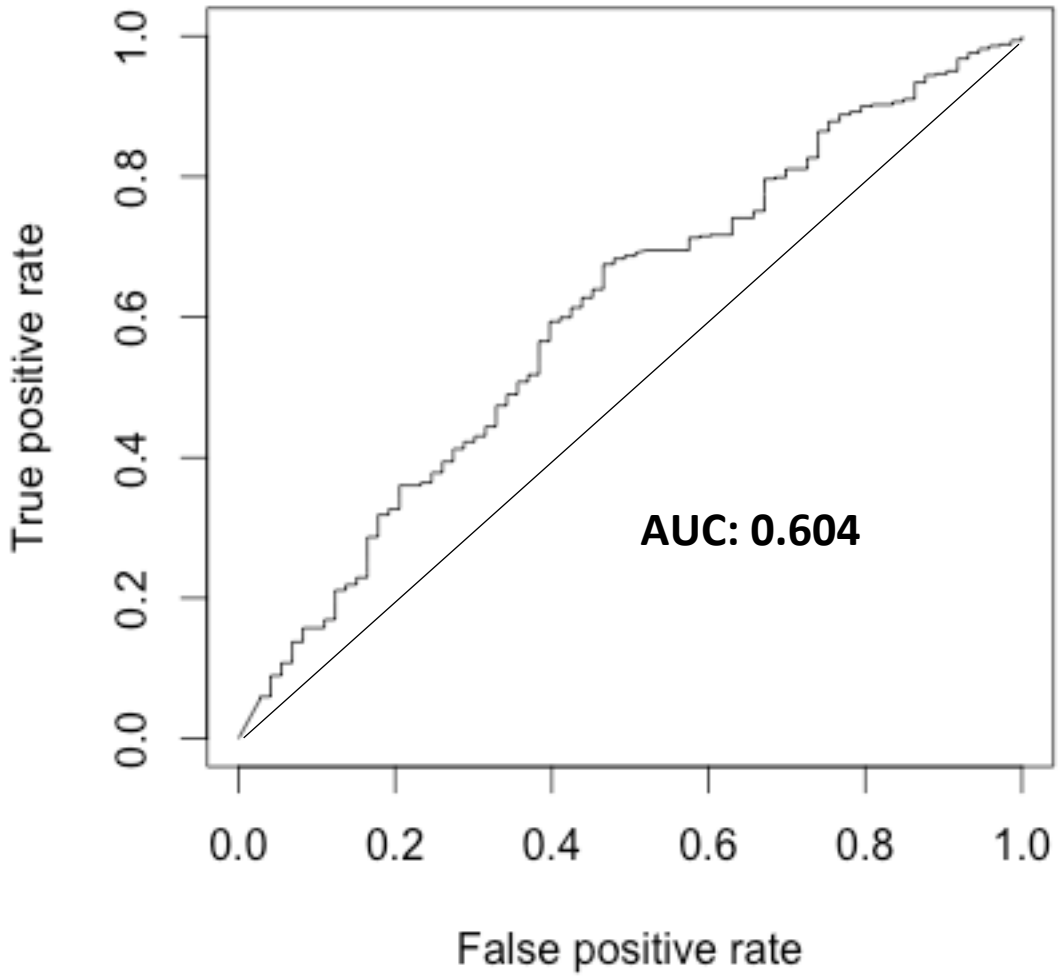
Model: ***Product:Company.response + Issue + Submitted.via***




credit
report
account
inform
equifax
experian
disput
transunion
request
companі
provid
file
address
state
remov
respons
consum
document
complaint
letter
time
date
number
contact
becaus
creditor
pleas
ask
also
will
verifi
sent
receiv
get
whi
ani
never
law
one
delet
mail
need
debt
investig
correct
payment
score
collect
still
paid
information
fcra
issu
agenc
proof
onli
can
show
bank
send
even
record
bureau
phone
incorrect
well
bankruptci
mortgag
wrong
social
dispute
previous
place
confirm
review
done
may
refus
union
without
use
servic
attach
item
due
clear
question
first
fair
list
name
prove
ident
year
regard
capit
requir
bill
must
right
per
bureaus
money
chang
loan
day
feder
data
fail
tri
close
late
legal
pay
accur
fact
now
person
negat
sinc
told
continu
system
reason
amount
result
inaccur
made
like
fraud
want
charg
anoth
act
complet
balanc
alreadi
problem
month
give
them
actual
answer
follow
current
updat
includ
back
said
violat
thank
response
years
financi
statement
furnish
status
claim
respond
secur
take
yet
removed
help
make
two
histori
see
copi
full
mine
issue
check
process
submit
allow
case
valid
call
origin
inquiri
say
know
just
free
last
open
matter
fals
you
cfpb
card
doe
veri
court
direct
past
tell
three
theft
new
owe
resolv

```
> terms(myLDA, 5)
     Topic 1   Topic 2   Topic 3
[1,] "credit"  "account" "report"
[2,] "report"  "debt"    "credit"
[3,] "account" "report"  "inform"
[4,] "score"   "credit"  "equifax"
[5,] "get"     "collect" "disput"
```

Model: ***Base model + DocumentTermMatrix***



</findings>

<next_steps>

Step 1:
Perform text mining on the complete set of complaint narratives

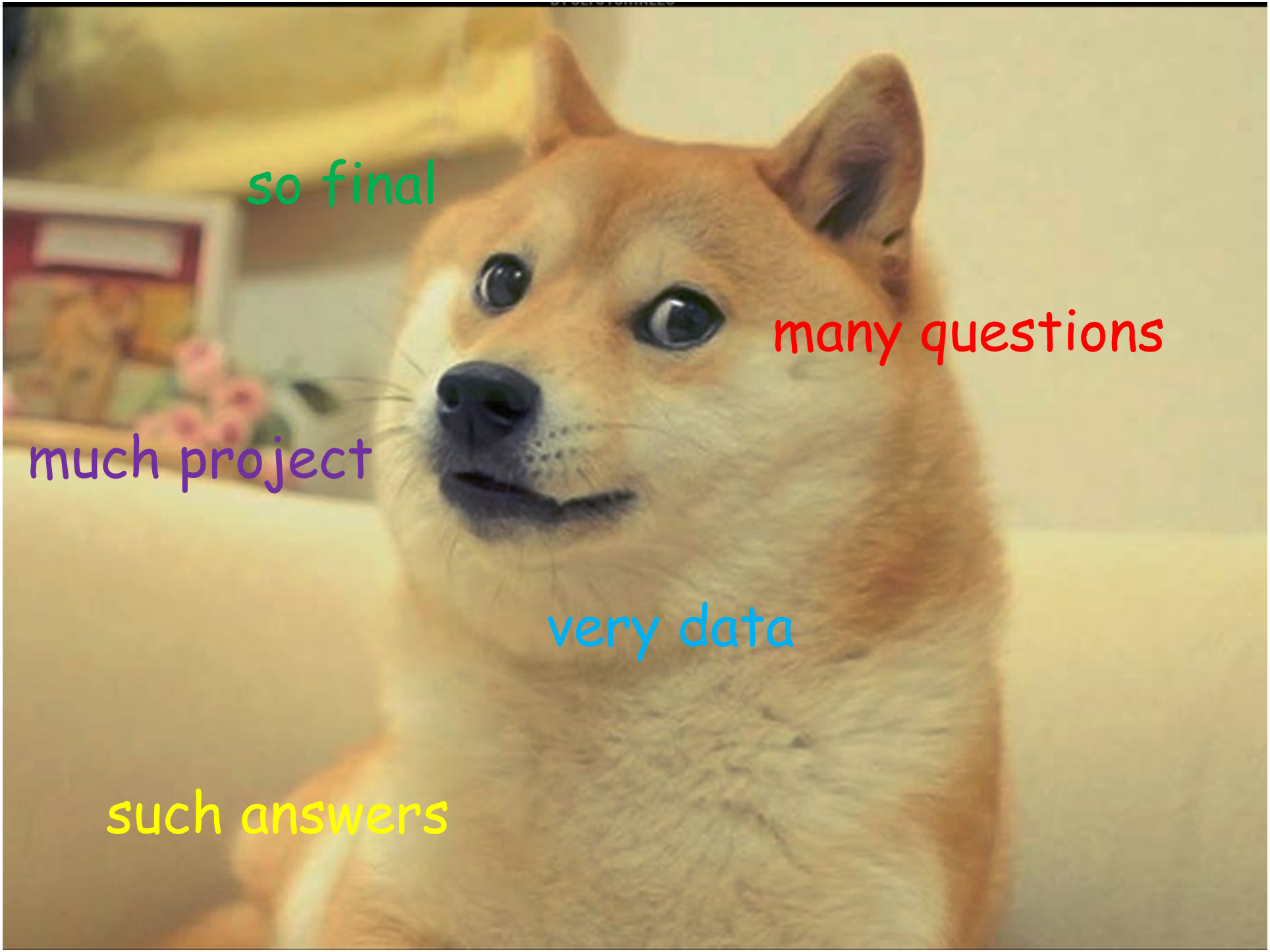
Step 2:
Play around with other classifiers

Step 3:
???

Step 4:
Profit

</next_steps>

</final project>


so final
many questions
much project
very data
such answers