ben

August 6, 2021

[1]:
```
# path to data- first 10,000,000 rows sampled from large reddit dataset
path = '/project/ds5559/r-slash-group8/sample.csv'
```

[2]:
```
import pandas as pd # data management
from pyspark.sql import SparkSession # paralellization
import numpy as np # math
import re # regex
from pyspark.sql.functions import * # pyspark sql functions
from pyspark.sql.types import ArrayType, IntegerType,  StringType,␣
 ↪TimestampType # pyspark sql cast data types
from pyspark.ml import Pipeline # pipeline of ML steps
from pyspark.ml.feature import * # ML features
from pyspark.ml.classification import LogisticRegression # ML regression model
from pyspark.mllib.evaluation import MulticlassMetrics # Model evaluation
```

[3]:
```
# start spark session
spark = SparkSession.builder.getOrCreate()
```

[4]:
```
# read in data
df_full = spark.read.csv(path, inferSchema=True, header = True)
```

[5]:
```
# preview data
df_full.show(5)
```

```
+--------------------------------+----------+----+-----------+---------+------
----+-----------+---------------------+----------------+---------+-------+---
-----------+------+-----+--------+-----------+-----+-----------+--------------
------+-------------+------+----------------+----------+
|                             _c0|created_utc| ups|subreddit_id|  link_id|
name|score_hidden|author_flair_css_class|author_flair_text|subreddit|
id|removal_reason|gilded|downs|archived|      author|score|retrieved_on|
body|distinguished|edited|controversiality| parent_id|
+--------------------------------+----------+----+-----------+---------+------
----+-----------+---------------------+----------------+---------+-------+---
-----------+------+-----+--------+-----------+-----+-----------+--------------
------+-------------+------+----------------+----------+
|                               1| 1430438400|   4|
t5_378oi|t3_34di91|t1_cqug90g|           0|                  NA|
```

```
NA|soccer_jp|cqug90g|            NA|     0|    0|       0|       rx109|    4|
1432703079|                 |         null|  null|            null|      null|
|         |       null|null|        null|     null|      null|
null|                  null|            null|     null|   null|          null|
null| null|    null|        null| null|        null|                null|
null|  null|            null|      null|
|                  "|         NA|   0|           0|t3_34di91|      null|
null|                  null|            null|     null|   null|          null|
null| null|    null|        null| null|        null|                null|
null|  null|            null|      null|
|                              2| 1430438400|   4|
t5_2qo4s|t3_34g8mx|t1_cqug90h|           0|                  Heat|
Heat|      nba|cqug90h|            NA|     0|    0|       0|   WyaOfWade|    4|
1432703079|gg this one's ove…|           NA|     0|               0|
t3_34g8mx|
|                              3| 1430438400|   0|
t5_2cneq|t3_34f7mc|t1_cqug90i|           0|                  NA|
NA| politics|cqug90i|            NA|     0|    0|       0|Wicked_Truth|    0|
1432703079|Are you really im…|           NA|     0|
0|t1_cqufim0|
+------------------------------+----------+----+-----------+---------+------
----+-----------+--------------------+----------------+---------+-------+---
----------+------+-----+--------+-----------+-----+-----------+-------------
------+-------------+------+----------------+----------+
only showing top 5 rows
```

[6]:
```
# remove irrelevant columns
df = df_full.
 ↪drop('_c0','subreddit_id','link_id','name','score_hidden','author_flair_css_class',␣
 ↪\
       'author_flair_text','archived','retrieved_on',␣
 ↪'edited','parent_id','ups', \
      'downs', 'removal_reason', 'distinguished')
# keep score, subreddit, id, created_utc, author, body, gilded, controversiality

df=df.withColumn("score",df.score.cast(IntegerType()))
df=df.withColumn("gilded",df.gilded.cast(IntegerType()))
df=df.withColumn("created_utc",from_unixtime(df.created_utc))
df=df.withColumnRenamed("id", "comment_id")
```

[7]:
```
# preview reduced data
df.show(5)
df.printSchema()
```

```
+------------------+---------+----------+------+-----------+-----+-----------
--------+----------------+
|        created_utc|subreddit|comment_id|gilded|     author|score|
```

```
body|controversiality|
+------------------+---------+----------+------+------------+-----+------------
--------+----------------+
|2015-05-01 00:00:00|soccer_jp|   cqug90g|     0|       rx109|    4|
 |            null|
|               null|     null|      null|  null|        null| null|
null|            null|
|               null|     null|      null|  null|        null| null|
null|            null|
|2015-05-01 00:00:00|      nba|   cqug90h|     0|   WyaOfWade|    4|gg this
one's ove…|               0|
|2015-05-01 00:00:00| politics|   cqug90i|     0|Wicked_Truth|    0|Are you
really im…|               0|
+------------------+---------+----------+------+------------+-----+------------
--------+----------------+
only showing top 5 rows

root
 |-- created_utc: string (nullable = true)
 |-- subreddit: string (nullable = true)
 |-- comment_id: string (nullable = true)
 |-- gilded: integer (nullable = true)
 |-- author: string (nullable = true)
 |-- score: integer (nullable = true)
 |-- body: string (nullable = true)
 |-- controversiality: string (nullable = true)
```

[8]:
```
# filter out rows with missing values in essential columns
df=df.filter(df['body'].isNotNull())
df=df.filter(df['subreddit'].isNotNull())
df=df.filter(df['comment_id'].isNotNull())
```

[9]:
```
# preview clean data
df.show()
```

```
+------------------+-----------------+----------+------+--------------+-----+
--------------------+----------------+
|       created_utc|        subreddit|comment_id|gilded|        author|score|
body|controversiality|
+------------------+-----------------+----------+------+--------------+-----+
--------------------+----------------+
|2015-05-01 00:00:00|         soccer_jp|   cqug90g|     0|          rx109|    4|
 |            null|
|2015-05-01 00:00:00|              nba|   cqug90h|     0|      WyaOfWade|
4|gg this one's ove…|               0|
|2015-05-01 00:00:00|          politics|   cqug90i|     0|   Wicked_Truth|
0|Are you really im…|               0|
```

```
|2015-05-01 00:00:00|         AskReddit|   cqug90j|     0|       jesse9o3|
3|No one has a Euro…|                 0|
|2015-05-01 00:00:00|         AskReddit|   cqug90k|     0| beltfedshooter|
3|"That the kid ""…|                  0|
|2015-05-01 00:00:00|        bloodborne|   cqug90l|     0|     Rubenticus|
1|Haha, i was getti…|                 0|
|2015-05-01 00:00:00|     relationships|   cqug90m|     0|silverraven1189|
6|After reading thi…|              null|
|2015-05-01 00:00:00|   Tennesseetitans|   cqug90n|     0|     Scrubtanic|
2|Let's do this. Se…|                 0|
|2015-05-01 00:00:00|            cigars|   cqug90o|     0|       burnmyiz|
6|You can buy a mys…|                 0|
|2015-05-01 00:00:00|   GlobalOffensive|   cqug90p|     0|      BEE_REAL_|
5|Nihilum and LG ar…|                 0|
|2015-05-01 00:00:00|            eagles|   cqug90q|     0|           SNVG|    4|
F[redacted] that what|                0|
|2015-05-01 00:00:00|         smashbros|   cqug90r|     0|       BiigLord|
1|Don't diss the Gr…|                 0|
|2015-05-01 00:00:00|      makinghiphop|   cqug90s|     0|        KingEze|
1|Your 16 bars seem…|                 0|
|2015-05-01 00:00:00|   GoogleCardboard|   cqug90t|     0|        cyborek|
3|Trinus vr is amaz…|                 0|
|2015-05-01 00:00:00|        offmychest|   cqug90u|     0|     Zekkystyle|
14|It's not your fau…|              null|
|2015-05-01 00:00:00|              nrl|   cqug90v|     0|       jeauxoxo|
3|http://www.reddit…|                 0|
|2015-05-01 00:00:00|elderscrollsonline|   cqug90w|     0| Chainsaw_Ninja|
0|and what about al…|                 0|
|2015-05-01 00:00:00|         childfree|   cqug90x|     0|    tparkelaine|
5|I would be tempte…|              null|
|2015-05-01 00:00:00|       ExNoContact|   cqug90y|     0|         vvvvfl|
1|I can't answer be…|              null|
|2015-05-01 00:00:00|         AskReddit|   cqug90z|     0|  InterimFatGuy|    5|
NSFL|               0|
+-------------------+------------------+----------+------+---------------+-----+
--------------------+----------------+
only showing top 20 rows
```

[10]:
```
# total records remaining in dataset:
df.count()
```

[10]: 10002410

[11]:
```
# choose only the two most popular subreddits
df_reduced = df.filter((df['subreddit'] == 'AskReddit') | (df['subreddit'] ==␣
↪'leagueoflegends'))
```

```
# encode subreddit as a binary variable
df_reduced = df_reduced.withColumn('subreddit_bin',when(df['subreddit'] ==␣
 ↪'AskReddit',0).otherwise(1))
df_reduced.show()
```

```
+------------------+--------------+---------+------+-----------------+-----+
--------------------+--------------------+-------------+
|        created_utc|     subreddit|comment_id|gilded|           author|score|
body|    controversiality|subreddit_bin|
+------------------+--------------+---------+------+-----------------+-----+
--------------------+--------------------+-------------+
|2015-05-01 00:00:00|     AskReddit|   cqug90j|     0|         jesse9o3|
3|No one has a Euro…|                   0|            0|
|2015-05-01 00:00:00|     AskReddit|   cqug90k|     0|   beltfedshooter|
3|"That the kid ""…|                  0|            0|
|2015-05-01 00:00:00|     AskReddit|   cqug90z|     0|    InterimFatGuy|    5|
NSFL|                   0|            0|
|2015-05-01 00:00:01|leagueoflegends|   cqug919|     0|   SenpaiOniichan|
1|well i think new …|               null|            1|
|2015-05-01 00:00:01|     AskReddit|   cqug91c|     0|      JuanTutrego|
1|I'm a guy and I h…|                   0|            0|
|2015-05-01 00:00:01|     AskReddit|   cqug91e|     0|      dcblackbelt|
101|Mid twenties male…|                  0|           0|
|2015-05-01 00:00:02|     AskReddit|   cqug920|     0|    TheDoorsShirt|
1|Fran Drescher lau…|                   0|            0|
|2015-05-01 00:00:02|     AskReddit|   cqug921|     0|   youthfulvictim|
-5|I honestly wouldn…|                  1|            0|
|2015-05-01 00:00:02|     AskReddit|   cqug923|     0|      sonovadoyle|    1|
&amp;lt;/3 |                   0|            0|
|2015-05-01 00:00:02|     AskReddit|   cqug929|     0|         boludo54|
1|no money, just ID…|                   0|            0|
|2015-05-01 00:00:03|     AskReddit|   cqug92o|     0|    mister_sleepy|
1|Smoking tobacco. …|                   0|            0|
|2015-05-01 00:00:03|     AskReddit|   cqug92s|     0|   baconinstitute|    1|
Bootleg Fireworks|                   0|           0|
|2015-05-01 00:00:03|     AskReddit|   cqug92v|     0|      cpmustang90|
1|I'm a member of t…|                   0|            0|
|2015-05-01 00:00:03|     AskReddit|   cqug937|     0|          hamhead|
1|That's one reason…|                   0|            0|
|2015-05-01 00:00:03|     AskReddit|   cqug93k|     0|     mrjosemeehan|
7|Go back to Route …|                   0|            0|
|2015-05-01 00:00:04|     AskReddit|   cqug93o|     0|     Peacehamster|
2|They're oh-so-bra…|                   0|            0|
|2015-05-01 00:00:04|leagueoflegends|   cqug93p|     0|         SureShaw|
1|"Honestly I don't…| while above what…|            1|
|2015-05-01 00:00:04|     AskReddit|   cqug93u|     0|       bunnylumps|
20|"I think the whol…|               null|            0|
```

```
|2015-05-01 00:00:04|      AskReddit|   cqug93v|     0|           DawkTux|
1|F[redacted] pound coins…|                   0|            0|
|2015-05-01 00:00:04|      AskReddit|   cqug940|     0|salamandersnuggles|
1|"Someone was tell…|                null|             0|
+-------------------+---------------+----------+------+------------------+-----+
--------------------+--------------------+-------------+
only showing top 20 rows
```

[12]:
```
# Total records between the two most popular subreddits
df_reduced.count()
```

[12]:
```
950528
```

These two subreddits account for almost 10% of all the data; not a bad sample size!

[13]:
```
# Train/Test split
df_train, df_test = df_reduced.randomSplit([0.75,0.25])
```

[14]:
```
df_train.count()
```

[14]:
```
712005
```

[15]:
```
# Breakdown of test set by subreddit
df_train.groupBy('subreddit_bin').count().show()
```

```
+-------------+------+
|subreddit_bin| count|
+-------------+------+
|            1|145672|
|            0|566333|
+-------------+------+
```

Certainly an imbalance between the two classes.

[16]:
```
df_test.groupBy('subreddit_bin').count().show()
```

```
+-------------+------+
|subreddit_bin| count|
+-------------+------+
|            1| 48782|
|            0|189741|
+-------------+------+
```

The train and test sets show a similar ratio; performance may be improved later on by downsampling the training set.

[17]:
```
# Build up pipeline for modeling

# Separate documents into tokens
tok = Tokenizer(inputCol="body", outputCol="body_tokens")
# Remove stopwords
rem = StopWordsRemover(inputCol="body_tokens", outputCol="tokens_filtered")
# Reduce feature count using hashing function
htf = HashingTF(numFeatures = 4096, inputCol="tokens_filtered", outputCol="tf")
# Vectorize features
w2v = Word2Vec(inputCol="body_tokens", outputCol="w2v")
# Assemble feature column
va = VectorAssembler(inputCols=["tf", "w2v"], outputCol="features")
# Apply logistic regression
lr = LogisticRegression(labelCol='subreddit_bin', featuresCol='features',␣
 ↪maxIter=10, regParam=0.01)

pipeline = Pipeline(stages=[tok, rem, htf, w2v, va, lr])
```

[18]:
```
# Fit model to train data
model = pipeline.fit(df_train)
```

[19]:
```
# Make predictions for test data using model
prediction = model.transform(df_test)
# Check out format of dataframe with predictions
prediction.show()
```

```
+-------------------+--------------+----------+------+--------------+-----+---
-----------------+----------------+-------------+--------------------+----------
----------+--------------------+--------------------+--------------------+------
-------------+--------------------+----------+
|        created_utc|     subreddit|comment_id|gilded|        author|score|
body|controversiality|subreddit_bin|         body_tokens|     tokens_filtered|
tf|                 w2v|            features|       rawPrediction|
probability|prediction|
+-------------------+--------------+----------+------+--------------+-----+---
-----------------+----------------+-------------+--------------------+----------
----------+--------------------+--------------------+--------------------+------
-------------+--------------------+----------+
|2015-05-01 00:00:00|     AskReddit|   cqug90z|     0|  InterimFatGuy|    5|
NSFL|               0|            0|              [nsfl]|              [nsfl]| (
4096,[2955],[1.0])|[0.01905914768576…|(4196,[2955,4096,…|[1.89691694817265..
.|[0.86954218692454…|       0.0|
|2015-05-01 00:00:01|leagueoflegends|   cqug919|     0| SenpaiOniichan|
1|well i think new …|            null|            1|[well, i, think,
…|[well, think, new…|(4096,[31,35,317,…|[-0.0618038852267…|(4196,[31,35,
317,…|[-2.0889314636877…|[0.11017728816876…|       1.0|
|2015-05-01 00:00:02|     AskReddit|   cqug923|     0|    sonovadoyle|    1|
```

```
&lt;/3 |               0|            0|            [&lt;/3]|
[&lt;/3]| (4096,[3775],[1.0])|[0.00736410822719…|(4196,[3775,4096,…|[0.82569
551789326…|[0.69544400120126…|       0.0|
|2015-05-01 00:00:03|      AskReddit|   cqug92s|     0| baconinstitute|    1|
Bootleg Fireworks|               0|            0|[bootleg, fireworks]|[bootleg,
fireworks]|(4096,[170,253],[…|[0.05399652209598…|(4196,[170,253,40…|[3.223
60025170436…|[0.96171280099356…|       0.0|
|2015-05-01 00:00:05|      AskReddit|   cqug94r|     0|  IDUnavailable|
2|"""That's great a…|            null|            0|["""that's, great…|
["""that's, great]|(4096,[1428,3822]…|[-0.0423605451360…|(4196,[1428,3822,..
.|[1.34825458769439…|[0.79384412815095…|       0.0|
|2015-05-01 00:00:05|      AskReddit|   cqug94y|     0|       robondes|   -3|The
implications …|               1|            0|[the,
implication…|[implications, va…|(4096,[511,2596,2…|[-0.0870756695978…|(4
196,[511,2596,2…|[1.64670475263427…|[0.83844519082963…|       0.0|
|2015-05-01 00:00:05|      AskReddit|   cqug956|     0|      [deleted]|    1|
[deleted]|               0|            0|          [[deleted]]|
[[deleted]]| (4096,[3076],[1.0])|[-0.0026473123580…|(4196,[3076,4096,…|[1.63
908446370243…|[0.83741032174257…|       0.0|
|2015-05-01 00:00:06|      AskReddit|   cqug95f|     0|       JustMe80|
1|You're just tryin…|               1|            0|[you're, just,
tr…|[trying, get, fro…|(4096,[102,346,35…|[-0.0569235841433…|(4196,[102,
346,35…|[1.78134366700891…|[0.85586270227735…|       0.0|
|2015-05-01 00:00:09|      AskReddit|   cqug97r|     0|        zacmars|    1|
The Green Mile|               0|            0|  [the, green, mile]|
[green, mile]|(4096,[55,840],[1…|[-0.1286615307132…|(4196,[55,840,409…|[4.
87297271583519…|[0.99240749867798…|       0.0|
|2015-05-01 00:00:10|      AskReddit|   cqug98u|     0|    Gum_Disease|    1|I
don't know how …|               0|            0|[i, don't, know, …|[know,
read, comm…|(4096,[169,282,32…|[-0.0947679786869…|(4196,[169,282,32…|[5.
20433550227927…|[0.99453730582091…|       0.0|
|2015-05-01 00:00:14|leagueoflegends|   cqug9c0|     0|       Dragirby|    2|
Best Koopaling.|               0|            1|  [best, koopaling.]|  [best, koo
paling.]|(4096,[2187,2701]…|[-0.0381177179515…|(4196,[2187,2701,…|[0.41990
792194835…|[0.60346121618575…|       0.0|
|2015-05-01 00:00:14|leagueoflegends|   cqug9c6|     0|      Anxietyzx|
1|doesnt that make …|               0|            1|[doesnt, that,
ma…|[doesnt, make, sk…|(4096,[495,1717,2…|[-0.0530532707770…|(4196,[495,
1717,2…|[-4.1847123274737…|[0.01499821399332…|       1.0|
|2015-05-01 00:00:17|      AskReddit|   cqug9e2|     0|    ParadoxCity|
1|"""I don't believ…|            null|            0|["""i, don't,
bel…|["""i, believe, _…|(4096,[433,1021,1…|[-0.0069155909375…|(4196,[433
,1021,1…|[2.81113144078191…|[0.94327439048734…|       0.0|
|2015-05-01 00:00:20|      AskReddit|   cqug9gt|     0| Haruhi_Fujioka|    4|He
sounds like so…|               0|            0|[he, sounds, like…|[sounds,
like, je…|(4096,[209,523,34…|[-0.0652624456478…|(4196,[209,523,34…|[2.38
400510075870…|[0.91559945243792…|       0.0|
|2015-05-01 00:00:21|      AskReddit|   cqug9h3|     0|  punkballerina|    1|
```

```
"""Bud light|               0|            0|     ["""bud, light]|     ["""bud, l
ight]|(4096,[2712,3502]…|[-0.0915366038680…|(4196,[2712,3502,…|[3.52115885
198764…|[0.97128384389698…|       0.0|
|2015-05-01 00:00:24|      AskReddit|   cqug9jp|     0|       gwrjones|    1|But
they are both…|               0|            0|[but, they, are, …|[theories.,
scien…|(4096,[180,316,34…|[-0.0707536543313…|(4196,[180,316,34…|[2.16316
240873890…|[0.89689236415875…|       0.0|
|2015-05-01 00:00:26|      AskReddit|   cqug9ki|     0|Rogan_McFlubbin|
2|WHITE CIS MALES C…|               1|            0|[white, cis,
male…|[white, cis, male…|(4096,[1412,1843,…|[-0.2027047276496…|(4196,[14
12,1843,…|[2.95840464774700…|[0.95065921566876…|       0.0|
|2015-05-01 00:00:27|      AskReddit|   cqug9li|     0|  rainbowsanity|
514|Conceived 9 month…|               0|            0|[conceived, 9,
mo…|[conceived, 9, mo…|(4096,[1546,1816,…|[0.18289654627442…|(4196,[1546
,1816,…|[1.68256274421552…|[0.84324358255318…|       0.0|
|2015-05-01 00:00:27|leagueoflegends|   cqug9ld|     0|       Highfire|
3|&gt; People ought…|            null|            1|[&gt;, people,
ou…|[&gt;, people, ab…|(4096,[645,1239,3…|[-0.1706548016518…|(4196,[645,
1239,3…|[1.47445249316236…|[0.81373319993149…|       0.0|
|2015-05-01 00:00:29|      AskReddit|   cqug9mm|     0|  AutoModerator|
1|**PLEASE READ THI…|            null|            0|[**please, read,
…|[**please, read, …|(4096,[322,387,97…|[0.41265985406935…|(4196,[322,38
7,97…|[3.75078786453761…|[0.97704031053856…|       0.0|
+------------------+---------------+----------+------+--------------+-----+---
----------------+---------------+------------+--------------------+----------
----------+--------------------+--------------------+--------------------+------
-------------+--------------------+----------+
only showing top 20 rows
```

[20]:
```
# The predictions again seem to follow a fairly similar ratio to the actual␣
 ↪data!
prediction.groupBy('prediction').count().show()
```

```
+----------+------+
|prediction| count|
+----------+------+
|       0.0|210457|
|       1.0| 28066|
+----------+------+
```

[21]:
```
# reduce the predictions to only the necessary columns
predictionsAndLabels = prediction.drop('created_utc', 'subreddit',␣
 ↪'comment_id', 'gilded', 'controversiality', 'author', 'score', 'body',␣
 ↪'body_tokens', 'tokens_filtered', 'rawPrediction', 'probability', 'tf',␣
 ↪'w2v', 'features')
```

[22]:
```
# Format the predictions correctly
predictionsAndLabels = predictionsAndLabels.withColumnRenamed('subreddit_bin',␣
 ↪'label')
predictionsAndLabels = predictionsAndLabels.withColumn('label',␣
 ↪predictionsAndLabels.label.cast('float'))
predictionsAndLabels = predictionsAndLabels.withColumn('prediction',␣
 ↪predictionsAndLabels.prediction.cast('float'))
predictionsAndLabels.show()
```

```
+-----+----------+
|label|prediction|
+-----+----------+
|  0.0|       0.0|
|  1.0|       1.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       0.0|
|  1.0|       1.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       0.0|
|  0.0|       0.0|
+-----+----------+
only showing top 20 rows
```

[23]:
```
# Convert the predictions to RDD
rdd = predictionsAndLabels.rdd.map(tuple)
```

[24]:
```
# Run metrics on predictions
metrics = MulticlassMetrics(rdd)
```

[25]:
```
metrics.confusionMatrix().toArray()
```

[25]:
```
array([[185195.,  25262.],
       [  4546.,  23520.]])
```

[26]:
```
metrics.accuracy
```

[26]: 0.8750309194501159

[27]:
```
metrics.weightedPrecision
```

[27]: 0.9179263590174914

[28]:
```
metrics.weightedRecall
```

[28]: 0.875030919450116

[29]:
```
metrics.weightedFMeasure()
```

[29]: 0.8886404725156227

[30]:
```
metrics.weightedFMeasure(beta=0.5)
```

[30]: 0.904788745567817

[31]:
```
metrics.weightedFalsePositiveRate
```

[31]: 0.15704028448399157

Model biased toward predicting /r/Askreddit, could be a result of imbalance in classes in the training set. Will try downsampling.

[32]:
```
def downsample(df, target, positive_label, negative_label):

    # Split the data by class
    positive_df = df.filter(df[target] == positive_label)
    negative_df = df.filter(df[target] == negative_label)

    # Count the observations by class
    positive_count = positive_df.count()
    negative_count = negative_df.count()

    # Reduce the larger dataset by the appropriate ratio
    if positive_count > negative_count:
        positive_df = positive_df.sample(True, negative_count/positive_count)
    else:
        negative_df = negative_df.sample(True, positive_count/negative_count)

    # Recombine into a full dataset
    df_b = positive_df
    df_b = df_b.union(negative_df)
```

```
    return df_b
```

[33]:
```
# Downsample the data
df_train_ds = downsample(df_train, 'subreddit_bin', 1, 0)
```

[34]:
```
# Verify the new counts
df_train_ds.groupBy('subreddit_bin').count().show()
```

```
+-------------+------+
|subreddit_bin| count|
+-------------+------+
|            1|145672|
|            0|145393|
+-------------+------+
```

Nice! Much closer

[35]:
```
# The same pipeline can be reused... process follows...
model_ds = pipeline.fit(df_train_ds)
```

[36]:
```
prediction_ds = model_ds.transform(df_test)
```

[37]:
```
prediction_ds.groupBy('prediction').count().show()
```

```
+----------+------+
|prediction| count|
+----------+------+
|       0.0|166070|
|       1.0| 72453|
+----------+------+
```

[38]:
```
predictionsAndLabels_ds = prediction_ds.drop('created_utc', 'subreddit',␣
↪'comment_id', 'gilded', 'controversiality', 'author', 'score', 'body',␣
↪'body_tokens', 'tokens_filtered', 'rawPrediction', 'probability', 'tf',␣
↪'w2v', 'features')
```

[39]:
```
predictionsAndLabels_ds = predictionsAndLabels_ds.
↪withColumnRenamed('subreddit_bin', 'label')
predictionsAndLabels_ds = predictionsAndLabels_ds.withColumn('label',␣
↪predictionsAndLabels_ds.label.cast('float'))
predictionsAndLabels_ds = predictionsAndLabels_ds.withColumn('prediction',␣
↪predictionsAndLabels_ds.prediction.cast('float'))
```

[40]:
```
rdd_ds = predictionsAndLabels_ds.rdd.map(tuple)
```

[41]:
```
metrics_ds = MulticlassMetrics(rdd_ds)
```

[42]:
```
metrics_ds.confusionMatrix().toArray()
```

[42]:
```
array([[151830.,  14240.],
       [ 37911.,  34542.]])
```

[43]:
```
metrics_ds.accuracy
```

[43]:
```
0.7813586111192631
```

[44]:
```
metrics_ds.weightedPrecision
```

[44]:
```
0.7722179203921324
```

[45]:
```
metrics_ds.weightedRecall
```

[45]:
```
0.7813586111192631
```

[46]:
```
metrics_ds.weightedFMeasure()
```

[46]:
```
0.767286632177368
```

[47]:
```
metrics_ds.weightedFMeasure(beta=0.5)
```

[47]:
```
0.7674473102424092
```

[48]:
```
metrics_ds.weightedFalsePositiveRate
```

[48]:
```
0.3903551401706386
```

Less biased toward predicting /r/AskReddit but performed much worse overall. Maybe a more distinguishing feature? Anecdotally, gold and awards have always been given often on AskReddit for especially good or helpful answers, more than most other subreddits. Maybe would provide a good distinction.

[49]:
```
# Create a new pipeline step that includes the gilded column.
va2 = VectorAssembler(inputCols=["tf", "w2v", "gilded"], outputCol="features")
# And assemble a new pipeline
pipeline2 = Pipeline(stages=[tok, rem, htf, w2v, va2, lr])
```

[50]:
```
# Create a new model using the training data, repeat same process
model_more_feats = pipeline2.fit(df_train_ds)
prediction_more_feats = model_more_feats.transform(df_test)
```

[51]:
```
predictionsAndLabels_more_feats = prediction_more_feats.drop('created_utc',␣
↪'subreddit', 'comment_id', 'gilded', 'controversiality','author', 'score',␣
↪'body', 'body_tokens', 'tokens_filtered', 'rawPrediction', 'probability',␣
↪'tf', 'w2v', 'features')
```

[52]:
```
predictionsAndLabels_more_feats = predictionsAndLabels_more_feats.
 ↪withColumnRenamed('subreddit_bin', 'label')
predictionsAndLabels_more_feats = predictionsAndLabels_more_feats.
 ↪withColumn('label', predictionsAndLabels_more_feats.label.cast('float'))
predictionsAndLabels_more_feats = predictionsAndLabels_more_feats.
 ↪withColumn('prediction', predictionsAndLabels_more_feats.prediction.
 ↪cast('float'))
predictionsAndLabels_more_feats.show()
```

```
+-----+----------+
|label|prediction|
+-----+----------+
|  0.0|       0.0|
|  1.0|       1.0|
|  0.0|       1.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       0.0|
|  1.0|       1.0|
|  0.0|       0.0|
|  0.0|       1.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       0.0|
|  0.0|       0.0|
+-----+----------+
only showing top 20 rows
```

[53]:
```
rdd_more_feats = predictionsAndLabels_more_feats.rdd.map(tuple)
```

[54]:
```
metrics_more_feats = MulticlassMetrics(rdd_more_feats)
```

[55]:
```
metrics_more_feats.confusionMatrix().toArray()
```

[55]:
```
array([[151820.,  14235.],
       [ 37921.,  34547.]])
```

[56]:
```
metrics_more_feats.accuracy
```

[56]: 0.7813376487802015

[57]:
```
metrics_more_feats.weightedPrecision
```

[57]: 0.7722065775358125

[58]:
```
metrics_more_feats.weightedRecall
```

[58]: 0.7813376487802015

[59]:
```
metrics_more_feats.weightedFMeasure()
```

[59]: 0.7672583519488723

[60]:
```
metrics_more_feats.weightedFMeasure(beta=0.5)
```

[60]: 0.7674253169599519

[61]:
```
metrics_more_feats.weightedFalsePositiveRate
```

[61]: 0.3903414985887471

[62]:
```
# Save notebook as PDF document
!jupyter nbconvert --to pdf `pwd`/*.ipynb
```

```
[NbConvertApp] Converting notebook
/sfs/qumulo/qhome/bmf3bw/ds5110-project/ben.ipynb to pdf
[NbConvertApp] Writing 82591 bytes to notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', 'notebook.tex', '-quiet']
[NbConvertApp] Running bibtex 1 time: ['bibtex', 'notebook']
[NbConvertApp] WARNING | bibtex had problems, most likely because there were no
citations
[NbConvertApp] PDF successfully created
[NbConvertApp] Writing 68064 bytes to
/sfs/qumulo/qhome/bmf3bw/ds5110-project/ben.pdf
[NbConvertApp] Converting notebook
/sfs/qumulo/qhome/bmf3bw/ds5110-project/test_file.ipynb to pdf
[NbConvertApp] Writing 26544 bytes to notebook.tex
[NbConvertApp] Building PDF
[NbConvertApp] Running xelatex 3 times: ['xelatex', 'notebook.tex', '-quiet']
[NbConvertApp] CRITICAL | xelatex failed: ['xelatex', 'notebook.tex', '-quiet']
This is XeTeX, Version 3.14159265-2.6-0.99999 (TeX Live 2019/dev/Debian)
(preloaded format=xelatex)
 restricted \write18 enabled.
entering extended mode
(./notebook.tex
LaTeX2e <2018-12-01>
```

```
(/usr/share/texlive/texmf-dist/tex/latex/base/article.cls
Document Class: article 2018/09/03 v1.4i Standard LaTeX document class
(/usr/share/texlive/texmf-dist/tex/latex/base/size11.clo))
(/usr/share/texlive/texmf-dist/tex/latex/tcolorbox/tcolorbox.sty
(/usr/share/texlive/texmf-dist/tex/latex/pgf/basiclayer/pgf.sty
(/usr/share/texlive/texmf-dist/tex/latex/pgf/utilities/pgfrcs.sty
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfutil-common.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfutil-common-lists.t
ex)) (/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfutil-latex.def
(/usr/share/texlive/texmf-dist/tex/latex/ms/everyshi.sty))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfrcs.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/pgf.revision.tex)))
(/usr/share/texlive/texmf-dist/tex/latex/pgf/basiclayer/pgfcore.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics/graphicx.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics/keyval.sty)
(/usr/share/texlive/texmf-dist/tex/latex/graphics/graphics.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics/trig.sty)
(/usr/share/texlive/texmf-dist/tex/latex/graphics-cfg/graphics.cfg)
(/usr/share/texlive/texmf-dist/tex/latex/graphics-def/xetex.def)))
(/usr/share/texlive/texmf-dist/tex/latex/pgf/systemlayer/pgfsys.sty
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfkeys.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfkeysfiltered.code.t
ex)) (/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgf.cfg)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys-xetex.def
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys-dvipdfmx.def
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys-common-pdf.de
f))))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsyssoftpath.code.
tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsysprotocol.code.
tex)) (/usr/share/texlive/texmf-dist/tex/latex/xcolor/xcolor.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics-cfg/color.cfg))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcore.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmath.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathcalc.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathutil.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathparser.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.basic.code
.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.trigonomet
ric.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.random.cod
e.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.comparison
.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.base.code.
```

tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.round.code
.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.misc.code.
tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.integerari
thmetics.code.tex)))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfloat.code.tex))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorepoints.code.te
x)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorepathconstruct.
code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorepathusage.code
.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorescopes.code.te
x)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoregraphicstate.c
ode.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoretransformation
s.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorequick.code.tex
)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoreobjects.code.t
ex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorepathprocessing
.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorearrows.code.te
x)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoreshade.code.tex
)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoreimage.code.tex

(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoreexternal.code.
tex))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorelayers.code.te
x)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoretransparency.c
ode.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorepatterns.code.
tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorerdf.code.tex))
)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/modules/pgfmoduleshapes.code.tex
) (/usr/share/texlive/texmf-dist/tex/generic/pgf/modules/pgfmoduleplot.code.tex
)
(/usr/share/texlive/texmf-dist/tex/latex/pgf/compatibility/pgfcomp-version-0-65
.sty)
(/usr/share/texlive/texmf-dist/tex/latex/pgf/compatibility/pgfcomp-version-1-18

```
.sty)) (/usr/share/texlive/texmf-dist/tex/latex/tools/verbatim.sty)
(/usr/share/texlive/texmf-dist/tex/latex/environ/environ.sty
(/usr/share/texlive/texmf-dist/tex/latex/trimspaces/trimspaces.sty))
(/usr/share/texlive/texmf-dist/tex/latex/etoolbox/etoolbox.sty)
(/usr/share/texlive/texmf-dist/tex/latex/tcolorbox/tcbbreakable.code.tex
Library (tcolorbox): 'tcbbreakable.code.tex' version '4.15'
)) (/usr/share/texlive/texmf-dist/tex/latex/parskip/parskip.sty
(/usr/share/texlive/texmf-dist/tex/latex/oberdiek/kvoptions.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/ltxcmds.sty)
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/kvsetkeys.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/infwarerr.sty)
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/etexcmds.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/ifluatex.sty)))))
(/usr/share/texlive/texmf-dist/tex/generic/iftex/iftex.sty)
(/usr/share/texlive/texmf-dist/tex/latex/fontspec/fontspec.sty
(/usr/share/texlive/texmf-dist/tex/latex/l3packages/xparse/xparse.sty
(/usr/share/texlive/texmf-dist/tex/latex/l3kernel/expl3.sty
(/usr/share/texlive/texmf-dist/tex/latex/l3kernel/expl3-code.tex)
(/usr/share/texlive/texmf-dist/tex/latex/l3kernel/l3xdvipdfmx.def)))
(/usr/share/texlive/texmf-dist/tex/latex/fontspec/fontspec-xetex.sty
(/usr/share/texlive/texmf-dist/tex/latex/base/fontenc.sty
(/usr/share/texlive/texmf-dist/tex/latex/base/tuenc.def))
(/usr/share/texlive/texmf-dist/tex/latex/fontspec/fontspec.cfg)))
(/usr/share/texlive/texmf-dist/tex/latex/caption/caption.sty
(/usr/share/texlive/texmf-dist/tex/latex/caption/caption3.sty))
(/usr/share/texlive/texmf-dist/tex/latex/float/float.sty)
(/usr/share/texlive/texmf-dist/tex/latex/tools/enumerate.sty)
(/usr/share/texlive/texmf-dist/tex/latex/geometry/geometry.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/ifpdf.sty)
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/ifvtex.sty)
(/usr/share/texlive/texmf-dist/tex/generic/ifxetex/ifxetex.sty))
(/usr/share/texlive/texmf-dist/tex/latex/amsmath/amsmath.sty
For additional information on amsmath, use the `?' option.
(/usr/share/texlive/texmf-dist/tex/latex/amsmath/amstext.sty
(/usr/share/texlive/texmf-dist/tex/latex/amsmath/amsgen.sty))
(/usr/share/texlive/texmf-dist/tex/latex/amsmath/amsbsy.sty)
(/usr/share/texlive/texmf-dist/tex/latex/amsmath/amsopn.sty))
(/usr/share/texlive/texmf-dist/tex/latex/amsfonts/amssymb.sty
(/usr/share/texlive/texmf-dist/tex/latex/amsfonts/amsfonts.sty))
(/usr/share/texlive/texmf-dist/tex/latex/base/textcomp.sty
(/usr/share/texlive/texmf-dist/tex/latex/base/ts1enc.def))
(/usr/share/texlive/texmf-dist/tex/latex/upquote/upquote.sty)
(/usr/share/texlive/texmf-dist/tex/latex/eurosym/eurosym.sty)
(/usr/share/texlive/texmf-dist/tex/latex/ucs/ucs.sty
(/usr/share/texlive/texmf-dist/tex/latex/ucs/data/uni-global.def))
(/usr/share/texlive/texmf-dist/tex/latex/fancyvrb/fancyvrb.sty
Style option: `fancyvrb' v3.2a <2019/01/15> (tvz))
(/usr/share/texlive/texmf-dist/tex/latex/oberdiek/grffile.sty
```

```
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/pdftexcmds.sty))
(/usr/share/texlive/texmf-dist/tex/latex/adjustbox/adjustbox.sty
(/usr/share/texlive/texmf-dist/tex/latex/xkeyval/xkeyval.sty
(/usr/share/texlive/texmf-dist/tex/generic/xkeyval/xkeyval.tex
(/usr/share/texlive/texmf-dist/tex/generic/xkeyval/xkvutils.tex)))
(/usr/share/texlive/texmf-dist/tex/latex/adjustbox/adjcalc.sty)
(/usr/share/texlive/texmf-dist/tex/latex/adjustbox/trimclip.sty
(/usr/share/texlive/texmf-dist/tex/latex/collectbox/collectbox.sty)
(/usr/share/texlive/texmf-dist/tex/latex/adjustbox/tc-xetex.def))
(/usr/share/texlive/texmf-dist/tex/latex/ifoddpage/ifoddpage.sty)
(/usr/share/texlive/texmf-dist/tex/latex/varwidth/varwidth.sty))
(/usr/share/texlive/texmf-dist/tex/latex/hyperref/hyperref.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/hobsub-hyperref.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/hobsub-generic.sty))
(/usr/share/texlive/texmf-dist/tex/latex/oberdiek/auxhook.sty)
(/usr/share/texlive/texmf-dist/tex/latex/hyperref/pd1enc.def)
(/usr/share/texlive/texmf-dist/tex/latex/latexconfig/hyperref.cfg)
(/usr/share/texlive/texmf-dist/tex/latex/url/url.sty))
(/usr/share/texlive/texmf-dist/tex/latex/hyperref/hxetex.def
(/usr/share/texlive/texmf-dist/tex/latex/hyperref/puenc.def)
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/stringenc.sty)
(/usr/share/texlive/texmf-dist/tex/latex/oberdiek/rerunfilecheck.sty))
(/usr/share/texlive/texmf-dist/tex/latex/titling/titling.sty)
(/usr/share/texlive/texmf-dist/tex/latex/tools/longtable.sty)
(/usr/share/texlive/texmf-dist/tex/latex/booktabs/booktabs.sty)
(/usr/share/texlive/texmf-dist/tex/latex/enumitem/enumitem.sty)
(/usr/share/texlive/texmf-dist/tex/generic/ulem/ulem.sty)
(/usr/share/texlive/texmf-dist/tex/latex/jknapltx/mathrsfs.sty)
No file notebook.aux.
(/usr/share/texlive/texmf-dist/tex/latex/base/ts1cmr.fd)
ABD: EveryShipout initializing macros
(/usr/share/texlive/texmf-dist/tex/latex/caption/ltcaption.sty)
*geometry* driver: auto-detecting
*geometry* detected driver: xetex
*geometry* verbose mode - [ preamble ] result:
* driver: xetex
* paper: <default>
* layout: <same size as paper>
* layoutoffset:(h,v)=(0.0pt,0.0pt)
* modes:
* h-part:(L,W,R)=(72.26999pt, 469.75502pt, 72.26999pt)
* v-part:(T,H,B)=(72.26999pt, 650.43001pt, 72.26999pt)
* \paperwidth=614.295pt
* \paperheight=794.96999pt
* \textwidth=469.75502pt
* \textheight=650.43001pt
* \oddsidemargin=0.0pt
* \evensidemargin=0.0pt
```

```
* \topmargin=-37.0pt
* \headheight=12.0pt
* \headsep=25.0pt
* \topskip=11.0pt
* \footskip=30.0pt
* \marginparwidth=59.0pt
* \marginparsep=10.0pt
* \columnsep=10.0pt
* \skip\footins=10.0pt plus 4.0pt minus 2.0pt
* \hoffset=0.0pt
* \voffset=0.0pt
* \mag=1000
* \@twocolumnfalse
* \@twosidefalse
* \@mparswitchfalse
* \@reversemarginfalse
* (1in=72.27pt=25.4mm, 1cm=28.453pt)

(/usr/share/texlive/texmf-dist/tex/latex/ucs/ucsencs.def)
(/usr/share/texlive/texmf-dist/tex/latex/hyperref/nameref.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/gettitlestring.sty))

Package hyperref Warning: Rerun to get /PageLabels entry.

(/usr/share/texlive/texmf-dist/tex/latex/amsfonts/umsa.fd)
(/usr/share/texlive/texmf-dist/tex/latex/amsfonts/umsb.fd)
(/usr/share/texlive/texmf-dist/tex/latex/jknapltx/ursfs.fd)

LaTeX Warning: No \author given.

(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/se-ascii-print.def)

LaTeX Warning: File `attachment:27d67016-1a87-45ec-acc8-cf6a498607c6.png' not f
ound on input line 398.

! Unable to load picture or PDF file 'attachment:27d67016-1a87-45ec-acc8-cf6a49
8607c6.png'.
<to be read again>
                   }
l.398 …27d67016-1a87-45ec-acc8-cf6a498607c6.png}

?
! Emergency stop.
<to be read again>
                   }
l.398 …27d67016-1a87-45ec-acc8-cf6a498607c6.png}

No pages of output.
```

```
Transcript written on notebook.log.

Traceback (most recent call last):
  File "/opt/conda/bin/jupyter-nbconvert", line 11, in <module>
    sys.exit(main())
  File "/opt/conda/lib/python3.7/site-packages/jupyter_core/application.py",
line 254, in launch_instance
    return super(JupyterApp, cls).launch_instance(argv=argv, **kwargs)
  File "/opt/conda/lib/python3.7/site-packages/traitlets/config/application.py",
line 845, in launch_instance
    app.start()
  File "/opt/conda/lib/python3.7/site-packages/nbconvert/nbconvertapp.py", line
350, in start
    self.convert_notebooks()
  File "/opt/conda/lib/python3.7/site-packages/nbconvert/nbconvertapp.py", line
524, in convert_notebooks
    self.convert_single_notebook(notebook_filename)
  File "/opt/conda/lib/python3.7/site-packages/nbconvert/nbconvertapp.py", line
489, in convert_single_notebook
    output, resources = self.export_single_notebook(notebook_filename,
resources, input_buffer=input_buffer)
  File "/opt/conda/lib/python3.7/site-packages/nbconvert/nbconvertapp.py", line
418, in export_single_notebook
    output, resources = self.exporter.from_filename(notebook_filename,
resources=resources)
  File "/opt/conda/lib/python3.7/site-packages/nbconvert/exporters/exporter.py",
line 181, in from_filename
    return self.from_file(f, resources=resources, **kw)
  File "/opt/conda/lib/python3.7/site-packages/nbconvert/exporters/exporter.py",
line 199, in from_file
    return self.from_notebook_node(nbformat.read(file_stream, as_version=4),
resources=resources, **kw)
  File "/opt/conda/lib/python3.7/site-packages/nbconvert/exporters/pdf.py", line
183, in from_notebook_node
    self.run_latex(tex_file)
  File "/opt/conda/lib/python3.7/site-packages/nbconvert/exporters/pdf.py", line
154, in run_latex
    self.latex_count, log_error, raise_on_failure)
  File "/opt/conda/lib/python3.7/site-packages/nbconvert/exporters/pdf.py", line
143, in run_command
    command=command, output=out))
nbconvert.exporters.pdf.LatexFailed: PDF creating failed, captured latex output:
Failed to run "['xelatex', 'notebook.tex', '-quiet']" command:
This is XeTeX, Version 3.14159265-2.6-0.99999 (TeX Live 2019/dev/Debian)
(preloaded format=xelatex)
 restricted \write18 enabled.
entering extended mode
(./notebook.tex
```

```
LaTeX2e <2018-12-01>
(/usr/share/texlive/texmf-dist/tex/latex/base/article.cls
Document Class: article 2018/09/03 v1.4i Standard LaTeX document class
(/usr/share/texlive/texmf-dist/tex/latex/base/size11.clo))
(/usr/share/texlive/texmf-dist/tex/latex/tcolorbox/tcolorbox.sty
(/usr/share/texlive/texmf-dist/tex/latex/pgf/basiclayer/pgf.sty
(/usr/share/texlive/texmf-dist/tex/latex/pgf/utilities/pgfrcs.sty
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfutil-common.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfutil-common-lists.t
ex)) (/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfutil-latex.def
(/usr/share/texlive/texmf-dist/tex/latex/ms/everyshi.sty))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfrcs.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/pgf.revision.tex)))
(/usr/share/texlive/texmf-dist/tex/latex/pgf/basiclayer/pgfcore.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics/graphicx.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics/keyval.sty)
(/usr/share/texlive/texmf-dist/tex/latex/graphics/graphics.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics/trig.sty)
(/usr/share/texlive/texmf-dist/tex/latex/graphics-cfg/graphics.cfg)
(/usr/share/texlive/texmf-dist/tex/latex/graphics-def/xetex.def)))
(/usr/share/texlive/texmf-dist/tex/latex/pgf/systemlayer/pgfsys.sty
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfkeys.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/utilities/pgfkeysfiltered.code.t
ex)) (/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgf.cfg)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys-xetex.def
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys-dvipdfmx.def
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsys-common-pdf.de
f))))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsyssoftpath.code.
tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/systemlayer/pgfsysprotocol.code.
tex)) (/usr/share/texlive/texmf-dist/tex/latex/xcolor/xcolor.sty
(/usr/share/texlive/texmf-dist/tex/latex/graphics-cfg/color.cfg))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcore.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmath.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathcalc.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathutil.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathparser.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.code.tex
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.basic.code
.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.trigonomet
ric.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.random.cod
e.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.comparison
.code.tex)
```

(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.base.code.
tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.round.code
.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.misc.code.
tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfunctions.integerari
thmetics.code.tex)))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/math/pgfmathfloat.code.tex))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorepoints.code.te
x)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorepathconstruct.
code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorepathusage.code
.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorescopes.code.te
x)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoregraphicstate.c
ode.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoretransformation
s.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorequick.code.tex
)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoreobjects.code.t
ex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorepathprocessing
.code.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorearrows.code.te
x)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoreshade.code.tex
)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoreimage.code.tex

(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoreexternal.code.
tex))
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorelayers.code.te
x)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcoretransparency.c
ode.tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorepatterns.code.
tex)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/basiclayer/pgfcorerdf.code.tex))
)
(/usr/share/texlive/texmf-dist/tex/generic/pgf/modules/pgfmoduleshapes.code.tex
) (/usr/share/texlive/texmf-dist/tex/generic/pgf/modules/pgfmoduleplot.code.tex
)
(/usr/share/texlive/texmf-dist/tex/latex/pgf/compatibility/pgfcomp-version-0-65
.sty)

```
(/usr/share/texlive/texmf-dist/tex/latex/pgf/compatibility/pgfcomp-version-1-18
.sty)) (/usr/share/texlive/texmf-dist/tex/latex/tools/verbatim.sty)
(/usr/share/texlive/texmf-dist/tex/latex/environ/environ.sty
(/usr/share/texlive/texmf-dist/tex/latex/trimspaces/trimspaces.sty))
(/usr/share/texlive/texmf-dist/tex/latex/etoolbox/etoolbox.sty)
(/usr/share/texlive/texmf-dist/tex/latex/tcolorbox/tcbbreakable.code.tex
Library (tcolorbox): 'tcbbreakable.code.tex' version '4.15'
)) (/usr/share/texlive/texmf-dist/tex/latex/parskip/parskip.sty
(/usr/share/texlive/texmf-dist/tex/latex/oberdiek/kvoptions.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/ltxcmds.sty)
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/kvsetkeys.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/infwarerr.sty)
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/etexcmds.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/ifluatex.sty)))))
(/usr/share/texlive/texmf-dist/tex/generic/iftex/iftex.sty)
(/usr/share/texlive/texmf-dist/tex/latex/fontspec/fontspec.sty
(/usr/share/texlive/texmf-dist/tex/latex/l3packages/xparse/xparse.sty
(/usr/share/texlive/texmf-dist/tex/latex/l3kernel/expl3.sty
(/usr/share/texlive/texmf-dist/tex/latex/l3kernel/expl3-code.tex)
(/usr/share/texlive/texmf-dist/tex/latex/l3kernel/l3xdvipdfmx.def)))
(/usr/share/texlive/texmf-dist/tex/latex/fontspec/fontspec-xetex.sty
(/usr/share/texlive/texmf-dist/tex/latex/base/fontenc.sty
(/usr/share/texlive/texmf-dist/tex/latex/base/tuenc.def))
(/usr/share/texlive/texmf-dist/tex/latex/fontspec/fontspec.cfg)))
(/usr/share/texlive/texmf-dist/tex/latex/caption/caption.sty
(/usr/share/texlive/texmf-dist/tex/latex/caption/caption3.sty))
(/usr/share/texlive/texmf-dist/tex/latex/float/float.sty)
(/usr/share/texlive/texmf-dist/tex/latex/tools/enumerate.sty)
(/usr/share/texlive/texmf-dist/tex/latex/geometry/geometry.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/ifpdf.sty)
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/ifvtex.sty)
(/usr/share/texlive/texmf-dist/tex/generic/ifxetex/ifxetex.sty))
(/usr/share/texlive/texmf-dist/tex/latex/amsmath/amsmath.sty
For additional information on amsmath, use the `?' option.
(/usr/share/texlive/texmf-dist/tex/latex/amsmath/amstext.sty
(/usr/share/texlive/texmf-dist/tex/latex/amsmath/amsgen.sty))
(/usr/share/texlive/texmf-dist/tex/latex/amsmath/amsbsy.sty)
(/usr/share/texlive/texmf-dist/tex/latex/amsmath/amsopn.sty))
(/usr/share/texlive/texmf-dist/tex/latex/amsfonts/amssymb.sty
(/usr/share/texlive/texmf-dist/tex/latex/amsfonts/amsfonts.sty))
(/usr/share/texlive/texmf-dist/tex/latex/base/textcomp.sty
(/usr/share/texlive/texmf-dist/tex/latex/base/ts1enc.def))
(/usr/share/texlive/texmf-dist/tex/latex/upquote/upquote.sty)
(/usr/share/texlive/texmf-dist/tex/latex/eurosym/eurosym.sty)
(/usr/share/texlive/texmf-dist/tex/latex/ucs/ucs.sty
(/usr/share/texlive/texmf-dist/tex/latex/ucs/data/uni-global.def))
(/usr/share/texlive/texmf-dist/tex/latex/fancyvrb/fancyvrb.sty
Style option: `fancyvrb' v3.2a <2019/01/15> (tvz))
```

```
(/usr/share/texlive/texmf-dist/tex/latex/oberdiek/grffile.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/pdftexcmds.sty))
(/usr/share/texlive/texmf-dist/tex/latex/adjustbox/adjustbox.sty
(/usr/share/texlive/texmf-dist/tex/latex/xkeyval/xkeyval.sty
(/usr/share/texlive/texmf-dist/tex/generic/xkeyval/xkeyval.tex
(/usr/share/texlive/texmf-dist/tex/generic/xkeyval/xkvutils.tex)))
(/usr/share/texlive/texmf-dist/tex/latex/adjustbox/adjcalc.sty)
(/usr/share/texlive/texmf-dist/tex/latex/adjustbox/trimclip.sty
(/usr/share/texlive/texmf-dist/tex/latex/collectbox/collectbox.sty)
(/usr/share/texlive/texmf-dist/tex/latex/adjustbox/tc-xetex.def))
(/usr/share/texlive/texmf-dist/tex/latex/ifoddpage/ifoddpage.sty)
(/usr/share/texlive/texmf-dist/tex/latex/varwidth/varwidth.sty))
(/usr/share/texlive/texmf-dist/tex/latex/hyperref/hyperref.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/hobsub-hyperref.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/hobsub-generic.sty))
(/usr/share/texlive/texmf-dist/tex/latex/oberdiek/auxhook.sty)
(/usr/share/texlive/texmf-dist/tex/latex/hyperref/pd1enc.def)
(/usr/share/texlive/texmf-dist/tex/latex/latexconfig/hyperref.cfg)
(/usr/share/texlive/texmf-dist/tex/latex/url/url.sty))
(/usr/share/texlive/texmf-dist/tex/latex/hyperref/hxetex.def
(/usr/share/texlive/texmf-dist/tex/latex/hyperref/puenc.def)
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/stringenc.sty)
(/usr/share/texlive/texmf-dist/tex/latex/oberdiek/rerunfilecheck.sty))
(/usr/share/texlive/texmf-dist/tex/latex/titling/titling.sty)
(/usr/share/texlive/texmf-dist/tex/latex/tools/longtable.sty)
(/usr/share/texlive/texmf-dist/tex/latex/booktabs/booktabs.sty)
(/usr/share/texlive/texmf-dist/tex/latex/enumitem/enumitem.sty)
(/usr/share/texlive/texmf-dist/tex/generic/ulem/ulem.sty)
(/usr/share/texlive/texmf-dist/tex/latex/jknapltx/mathrsfs.sty)
No file notebook.aux.
(/usr/share/texlive/texmf-dist/tex/latex/base/ts1cmr.fd)
ABD: EveryShipout initializing macros
(/usr/share/texlive/texmf-dist/tex/latex/caption/ltcaption.sty)
*geometry* driver: auto-detecting
*geometry* detected driver: xetex
*geometry* verbose mode - [ preamble ] result:
* driver: xetex
* paper: <default>
* layout: <same size as paper>
* layoutoffset:(h,v)=(0.0pt,0.0pt)
* modes:
* h-part:(L,W,R)=(72.26999pt, 469.75502pt, 72.26999pt)
* v-part:(T,H,B)=(72.26999pt, 650.43001pt, 72.26999pt)
* \paperwidth=614.295pt
* \paperheight=794.96999pt
* \textwidth=469.75502pt
* \textheight=650.43001pt
* \oddsidemargin=0.0pt
```

```
* \evensidemargin=0.0pt
* \topmargin=-37.0pt
* \headheight=12.0pt
* \headsep=25.0pt
* \topskip=11.0pt
* \footskip=30.0pt
* \marginparwidth=59.0pt
* \marginparsep=10.0pt
* \columnsep=10.0pt
* \skip\footins=10.0pt plus 4.0pt minus 2.0pt
* \hoffset=0.0pt
* \voffset=0.0pt
* \mag=1000
* \@twocolumnfalse
* \@twosidefalse
* \@mparswitchfalse
* \@reversemarginfalse
* (1in=72.27pt=25.4mm, 1cm=28.453pt)

(/usr/share/texlive/texmf-dist/tex/latex/ucs/ucsencs.def)
(/usr/share/texlive/texmf-dist/tex/latex/hyperref/nameref.sty
(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/gettitlestring.sty))

Package hyperref Warning: Rerun to get /PageLabels entry.

(/usr/share/texlive/texmf-dist/tex/latex/amsfonts/umsa.fd)
(/usr/share/texlive/texmf-dist/tex/latex/amsfonts/umsb.fd)
(/usr/share/texlive/texmf-dist/tex/latex/jknapltx/ursfs.fd)

LaTeX Warning: No \author given.

(/usr/share/texlive/texmf-dist/tex/generic/oberdiek/se-ascii-print.def)

LaTeX Warning: File `attachment:27d67016-1a87-45ec-acc8-cf6a498607c6.png' not f
ound on input line 398.

! Unable to load picture or PDF file 'attachment:27d67016-1a87-45ec-acc8-cf6a49
8607c6.png'.
<to be read again>
                   }
l.398 …27d67016-1a87-45ec-acc8-cf6a498607c6.png}

?
! Emergency stop.
<to be read again>
                   }
l.398 …27d67016-1a87-45ec-acc8-cf6a498607c6.png}
```

```
No pages of output.
Transcript written on notebook.log.
```