Task-3

July 27, 2020

1 Dependencies

[1]:
```
import pandas as pd
import pyspark
import re
```

2 Setup

[2]:
```
pd.set_option('display.max_rows', 500)
pd.set_option('display.max_columns', 500)
pd.set_option('display.width', 1000)
pd.set_option('display.expand_frame_repr', False)
```

[3]:
```
spark.conf.get('spark.driver.memory')
```

[3]:
```
u'6g'
```

3 Task 3

Reading the saved parquet file for fitting 2 separate models

[4]:
```
shoesDfOriginal = spark.read.parquet("hdfs:///data/exercise/shoes.parquet")
```

3.0.1 Print schema

[5]:
```
shoesDfOriginal.printSchema()
```

```
root
 |-- product_id: integer (nullable = true)
 |-- name: string (nullable = true)
 |-- upc_id: string (nullable = true)
 |-- descr: string (nullable = true)
```

```
|-- vendor_catalog_url: string (nullable = true)
|-- buy_url: string (nullable = true)
|-- manufacturer_name: string (nullable = true)
|-- sale_price: decimal(38,18) (nullable = true)
|-- retail_price: decimal(38,18) (nullable = true)
|-- manufacturer_part_no: string (nullable = true)
|-- country: string (nullable = true)
|-- vendor_id: integer (nullable = true)
|-- category_name: string (nullable = true)
|-- category_code: string (nullable = true)
|-- category_id: integer (nullable = true)
```

3.1 Original Data

[6]:
```
pd.DataFrame(shoesDfOriginal.take(5), columns=shoesDfOriginal.columns)
```

[6]:
```
   product_id                                              name upc_id
descr                                  vendor_catalog_url
buy_url manufacturer_name              sale_price           retail_price
manufacturer_part_no country vendor_id category_name category_code  category_id
0      530987              i play Swim Shoes (Kids) - Aqua-4   None  i play Swim
Shoes Kids - Hot Pink Imported Imp…  http://www.shopstyle.com/p/i-play-swim-
shoes-k…  http://www.shopstyle.com/action/apiVisitRetail…             I Play
8.990000000000000000   9.990000000000000000                 None    None
None   Boys' Shoes    boys-shoes         1599
1      530519  K-Swiss 501 Classic Tennis Shoe (Little Kid)   None
K Swiss Classic Retro Shoes  http://www.shopstyle.com/p/k-swiss-501-classic…
http://www.shopstyle.com/action/apiVisitRetail…           K-Swiss
25.320000000000000000  48.000000000000000000                 None    None
None   Boys' Shoes    boys-shoes         1599
2      530581     K-Swiss 801 Classic Tennis Shoe (Big Kid)   None
K Swiss Classic Luxury Retro Shoes
http://www.shopstyle.com/p/k-swiss-801-classic…
http://www.shopstyle.com/action/apiVisitRetail…           K-Swiss
43.100000000000000000  50.000000000000000000                 None    None
None   Boys' Shoes    boys-shoes         1599
3      535774        Dinosoles Kids' Dinorama Veloci Raptor   None
Dinosoles Kids' Dinorama Veloci Raptor  http://www.shopstyle.com/p/dinosoles-
kids-dino…  http://www.shopstyle.com/action/apiVisitRetail…
Dinosoles  25.870000000000000000  36.950000000000000000                 None
None      None   Boys' Shoes    boys-shoes         1599
4      535775      Dinosoles Kids' Dino Ankylosaurus Tod/Pr   None
Dinosoles Kids' Dino Ankylosaurus Tod/Pr  http://www.shopstyle.com/p/dinosoles-
kids-dino…  http://www.shopstyle.com/action/apiVisitRetail…
Dinosoles  29.600000000000000000  37.000000000000000000                 None
```

```
None      None   Boys' Shoes    boys-shoes         1599
```

3.2 Keep only relevent data tha can help us classify a tweet to a product category

[7]:
```
shoesDf = shoesDfOriginal.select('name', 'descr', 'manufacturer_name',␣
↪'sale_price', 'country', 'category_name', 'category_code', 'category_id')
```

3.2.1 Number of Items per category

[8]:
```
shoesDf.groupBy("category_code", 'category_id').count().show()
```

```
+------------------+-----------+-----+
|      category_code|category_id|count|
+------------------+-----------+-----+
|         boys-shoes|       1599|15400|
|        girls-shoes|       1612|21632|
| mens-lace-up-shoes|       1564|12353|
|     shoes-athletic|       1976| 4899|
|mens-shoes-athletic|       1561| 7935|
|      evening-shoes|       1773|  901|
|       bridal-shoes|       1386|  848|
+------------------+-----------+-----+
```

3.3 Keep only relevent data tha can help us classify a tweet to a product category

[9]:
```
shoesDf = shoesDf.select('name', 'descr', 'manufacturer_name', 'sale_price',␣
↪'country', 'category_name', 'category_code', 'category_id')
```

3.3.1 Example of data

[10]:
```
pd.DataFrame(shoesDf.take(20), columns=shoesDf.columns)
```

[10]:
```
                                                name
descr manufacturer_name              sale_price country category_name
category_code  category_id
0                   i play Swim Shoes (Kids) - Aqua-4  i play Swim Shoes Kids -
Hot Pink Imported Imp…            I Play    8.990000000000000000    None
Boys' Shoes    boys-shoes         1599
1        K-Swiss 501 Classic Tennis Shoe (Little Kid)                       K
```

```
Swiss Classic Retro Shoes           K-Swiss   25.320000000000000000    None
Boys' Shoes    boys-shoes         1599
2           K-Swiss 801 Classic Tennis Shoe (Big Kid)                 K Swiss
Classic Luxury Retro Shoes           K-Swiss   43.100000000000000000    None
Boys' Shoes    boys-shoes         1599
3              Dinosoles Kids' Dinorama Veloci Raptor             Dinosoles
Kids' Dinorama Veloci Raptor         Dinosoles   25.870000000000000000    None
Boys' Shoes    boys-shoes         1599
4            Dinosoles Kids' Dino Ankylosaurus Tod/Pr           Dinosoles Kids'
Dino Ankylosaurus Tod/Pr         Dinosoles   29.600000000000000000    None
Boys' Shoes    boys-shoes         1599
5                Puma Drez S Jr - New Navy/Geranium-4        Puma Drez S Jr -
Dark Shadow/Limestone Gray              Puma   13.750000000000000000    None
Boys' Shoes    boys-shoes         1599
6            Puma Suede 2 Straps Kids - Black/White-4             Puma Suede 2
Straps Kids - Green Sheen              Puma   20.090000000000000000    None
Boys' Shoes    boys-shoes         1599
7                                   Slip-On Sneakers  <ul> <li>Textile/manmade
material</li> <li>Spo…           Crazy 8   11.190000000000000000    None
Boys' Shoes    boys-shoes         1599
8                                   Lace-Up Sneakers  <ul> <li>Textile/manmade
material</li> <li>Spo…           Crazy 8   11.190000000000000000    None
Boys' Shoes    boys-shoes         1599
9    OshKosh B'Gosh Octo Sneaker (Toddler/Little Kid)     OshKosh B'Gosh Octo
Sneaker Toddler/Little Kid          Osh Kosh   16.340000000000000000    None
Boys' Shoes    boys-shoes         1599
10                                  IL GUFO Ankle boots  IL GUFO Ankle boots.
sueded, solid color, lace…           Il Gufo   49.000000000000000000    None
Boys' Shoes    boys-shoes         1599
11                                       Munster Shoes  Night navy trendy London
style lace up shoes i…  Stella McCartney  175.000000000000000000    None
Boys' Shoes    boys-shoes         1599
12  EMU Australia Annie Boot (Toddler/Little Kid/B…  EMU Australia Annie Boot
Toddler/Little Kid/Bi…               Emu   42.810000000000000000    None
Boys' Shoes    boys-shoes         1599
13     Heelys Dual Up Skate Shoe (Little Kid/Big Kid)       Heelys Dual Up Skate
Shoe Little Kid/Big Kid            Heelys   42.170000000000000000    None
Boys' Shoes    boys-shoes         1599
14      Heelys Feisty Skate Shoe (Little Kid/Big Kid)
Non-marking outsole            Heelys   48.000000000000000000    None   Boys'
Shoes    boys-shoes         1599
15    KEEN Punky AK Youth Bootie (Little Kid/Big Kid)      KEEN Punky AK Youth
Bootie Little Kid/Big Kid              Keen   65.000000000000000000    None
Boys' Shoes    boys-shoes         1599
16            Vans Authentic (Inf/Tod) - Red-4 Infant  Vans Authentic - Red Vans
very first and most …              Vans   17.790000000000000000    None
Boys' Shoes    boys-shoes         1599
```

```
17   Vans Old Skool Crib Shoe - Navy/Navy-Navy/Navy-4  Vans Old Skool Crib Shoe
- Navy/Navy sizes ava…              Vans   17.600000000000000000    None
Boys' Shoes    boys-shoes         1599
18          Vans Classic Slip-On - T-Rex Blue/Black-4  Vans Classic Slip-On -
T-Rex Blue/Black sizes …              Vans   33.250000000000000000    None
Boys' Shoes    boys-shoes         1599
19                                    BIRKI'S Sandals  BIRKI&apos;S Sandals.
logo detail, solid color…           Birki's   44.000000000000000000    None
Boys' Shoes    boys-shoes         1599
```

3.4 Preprocessing

3.5 Custom Transformers

[11]:
```
from pyspark import keyword_only
from pyspark.ml import Transformer
from pyspark.ml.param.shared import HasInputCol, HasOutputCol
from pyspark.sql.functions import udf
from pyspark.sql.types import StringType
```

3.5.1 Define Custom Transformer to remove non alphanumeric charaters

[12]:
```
class CleanText(Transformer, HasInputCol, HasOutputCol):
    @keyword_only
    def __init__(self, inputCol=None, outputCol=None):
        super(CleanText, self).__init__()
        kwargs = self._input_kwargs
        self.setParams(**kwargs)

    def setParams(self, inputCol=None, outputCol=None):
        kwargs = self._input_kwargs
        return self._set(**kwargs)

    def _transform(self, dataset):
        def f(text):
            text = re.sub("\W", " ", text)
            text = re.sub(" +", " ", text)

            text = re.sub(r"’", "\'", text)
            text = re.sub(r'([^\x00-\x7f])',r'', text)
            return text

        t = StringType()
        out_col = self.getOutputCol()
```

```
        in_col = dataset[self.getInputCol()]
        return dataset.withColumn(out_col, udf(f, t)(in_col))
```

3.5.2 Define Custom Transformer to remove HTML

[13]:
```
from bs4 import BeautifulSoup

class BeutifulSoupTagRemover(Transformer, HasInputCol, HasOutputCol):

    @keyword_only
    def __init__(self, inputCol=None, outputCol=None):
        super(BeutifulSoupTagRemover, self).__init__()
        kwargs = self._input_kwargs
        self.setParams(**kwargs)

    @keyword_only
    def setParams(self, inputCol=None, outputCol=None):
        kwargs = self._input_kwargs
        return self._set(**kwargs)

    def _transform(self, dataset):

        def f(s):
            cleaned_post = BeautifulSoup(s).text
            return cleaned_post

        t = StringType()
        out_col = self.getOutputCol()
        in_col = dataset[self.getInputCol()]
        return dataset.withColumn(out_col, udf(f, t)(in_col))
```

3.5.3 Clone original dataframe to start preprocessing

[14]:
```
shoesDf = shoesDfOriginal
```

3.5.4 Clean Data

- make all text lowercase
- remove html tags from the name
- remove
- remove stopwords

[15]:
```
from pyspark.sql.functions import lower, col
shoesDf = shoesDf
shoesDf = shoesDf.withColumn("descr", lower(col("descr")))

cleanText  = CleanText(inputCol="name", outputCol="name1")
beutifulSoupTagRemover = BeutifulSoupTagRemover(inputCol="name1",␣
 ↪outputCol="name2")
tokenizer = pyspark.ml.feature.Tokenizer(inputCol="name2", outputCol="name3")
stopWordsRemover = pyspark.ml.feature.StopWordsRemover(inputCol="name3",␣
 ↪outputCol="name_preprocessed")

preprocessingPipeline = pyspark.ml.Pipeline(stages=[
    cleanText,
    beutifulSoupTagRemover,
    tokenizer,
    stopWordsRemover
])

shoesDf = preprocessingPipeline.fit(shoesDf).transform(shoesDf)
```

[16]:
```
shoesDf = shoesDf.persist(pyspark.StorageLevel.MEMORY_AND_DISK)
```

[17]:
```
pd.DataFrame(shoesDf.take(100), columns=shoesDf.columns)
```

[17]:
```
    product_id                                                  name upc_id
descr                                  vendor_catalog_url
buy_url manufacturer_name              sale_price            retail_price
manufacturer_part_no country vendor_id category_name category_code  category_id
name1                                                  name2
name3                                  name_preprocessed
0       530987                   i play Swim Shoes (Kids) - Aqua-4   None  i play
swim shoes kids - hot pink imported imp…  http://www.shopstyle.com/p/i-play-
swim-shoes-k…  http://www.shopstyle.com/action/apiVisitRetail…            I
Play    8.990000000000000000    9.990000000000000000                  None
None      None   Boys' Shoes    boys-shoes         1599                      i
play Swim Shoes Kids Aqua 4                      i play Swim Shoes Kids Aqua 4
[i, play, swim, shoes, kids, aqua, 4]                [play, swim, shoes, kids,
aqua, 4]
1       530519       K-Swiss 501 Classic Tennis Shoe (Little Kid)   None
k swiss classic retro shoes  http://www.shopstyle.com/p/k-swiss-501-classic…
http://www.shopstyle.com/action/apiVisitRetail…           K-Swiss
25.320000000000000000   48.000000000000000000                None    None
None   Boys' Shoes    boys-shoes         1599        K Swiss 501 Classic Tennis
Shoe Little Kid         K Swiss 501 Classic Tennis Shoe Little Kid   [k, swiss,
501, classic, tennis, shoe, little,…  [k, swiss, 501, classic, tennis, shoe,
little,…
2       530581          K-Swiss 801 Classic Tennis Shoe (Big Kid)   None
```

k swiss classic luxury retro shoes
http://www.shopstyle.com/p/k-swiss-801-classic…
http://www.shopstyle.com/action/apiVisitRetail… K-Swiss
43.100000000000000000 50.000000000000000000 None None
None Boys' Shoes boys-shoes 1599 K Swiss 801 Classic
Tennis Shoe Big Kid K Swiss 801 Classic Tennis Shoe Big Kid [k,
swiss, 801, classic, tennis, shoe, big, kid] [k, swiss, 801, classic, tennis,
shoe, big, kid]
3 535774 Dinosoles Kids' Dinorama Veloci Raptor None
dinosoles kids' dinorama veloci raptor http://www.shopstyle.com/p/dinosoles-
kids-dino… http://www.shopstyle.com/action/apiVisitRetail…
Dinosoles 25.870000000000000000 36.950000000000000000 None
None None Boys' Shoes boys-shoes 1599 Dinosoles
Kids Dinorama Veloci Raptor Dinosoles Kids Dinorama Veloci Raptor
[dinosoles, kids, dinorama, veloci, raptor] [dinosoles, kids, dinorama,
veloci, raptor]
4 535775 Dinosoles Kids' Dino Ankylosaurus Tod/Pr None
dinosoles kids' dino ankylosaurus tod/pr http://www.shopstyle.com/p/dinosoles-
kids-dino… http://www.shopstyle.com/action/apiVisitRetail…
Dinosoles 29.600000000000000000 37.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Dinosoles
Kids Dino Ankylosaurus Tod Pr Dinosoles Kids Dino Ankylosaurus Tod Pr
[dinosoles, kids, dino, ankylosaurus, tod, pr] [dinosoles, kids, dino,
ankylosaurus, tod, pr]
5 539893 Puma Drez S Jr - New Navy/Geranium-4 None
puma drez s jr - dark shadow/limestone gray http://www.shopstyle.com/p/puma-
drez-s-jr/4294… http://www.shopstyle.com/action/apiVisitRetail…
Puma 13.750000000000000000 55.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Puma
Drez S Jr New Navy Geranium 4 Puma Drez S Jr New Navy Geranium 4
[puma, drez, s, jr, new, navy, geranium, 4] [puma, drez, jr, new,
navy, geranium, 4]
6 539934 Puma Suede 2 Straps Kids - Black/White-4 None
puma suede 2 straps kids - green sheen http://www.shopstyle.com/p/puma-
suede-2-straps… http://www.shopstyle.com/action/apiVisitRetail…
Puma 20.090000000000000000 40.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Puma Suede 2
Straps Kids Black White 4 Puma Suede 2 Straps Kids Black White 4
[puma, suede, 2, straps, kids, black, white, 4] [puma, suede, 2, straps,
kids, black, white, 4]
7 531019 Slip-On Sneakers None
textile/manmade material spo… http://www.shopstyle.com/p/crazy-8
-slip-on-sne… http://www.shopstyle.com/action/apiVisitRetail…
Crazy 8 11.190000000000000000 19.880000000000000000 None
None None Boys' Shoes boys-shoes 1599
Slip On Sneakers Slip On Sneakers
[slip, on, sneakers] [slip, sneakers]

8 531020 Lace-Up Sneakers None <ul> textile/manmade material spo… http://www.shopstyle.com/p/crazy-8-lace-up-sne… http://www.shopstyle.com/action/apiVisitRetail… Crazy 8 11.190000000000000000 19.880000000000000000 None None None Boys' Shoes boys-shoes 1599 Lace Up Sneakers Lace Up Sneakers [lace, up, sneakers] [lace, sneakers]

9 531058 OshKosh B'Gosh Octo Sneaker (Toddler/Little Kid) None oshkosh b'gosh octo sneaker toddler/little kid http://www.shopstyle.com/p/oshkosh-octo-sneak… http://www.shopstyle.com/action/apiVisitRetail… Osh Kosh 16.340000000000000000 16.340000000000000000 None None None Boys' Shoes boys-shoes 1599 OshKosh B Gosh Octo Sneaker Toddler Little Kid OshKosh B Gosh Octo Sneaker Toddler Little Kid [oshkosh, b, gosh, octo, sneaker, toddler, lit… [oshkosh, b, gosh, octo, sneaker, toddler, lit…

10 531681 IL GUFO Ankle boots None il gufo ankle boots. sueded, solid color, lace… http://www.shopstyle.com/p/il-gufo-ankle-boots… http://www.shopstyle.com/action/apiVisitRetail… Il Gufo 49.000000000000000000 157.000000000000000000 None None None Boys' Shoes boys-shoes 1599 IL GUFO Ankle boots IL GUFO Ankle boots [il, gufo, ankle, boots] [il, gufo, ankle, boots]

11 532620 Munster Shoes None night navy trendy london style lace up shoes i… http://www.shopstyle.com/p/stella-mccartney-mu… http://www.shopstyle.com/action/apiVisitRetail… Stella McCartney 175.000000000000000000 175.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Munster Shoes Munster Shoes [munster, shoes] [munster, shoes]

12 533642 EMU Australia Annie Boot (Toddler/Little Kid/B… None emu australia annie boot toddler/little kid/bi… http://www.shopstyle.com/p/emu-annie-boot-todd… http://www.shopstyle.com/action/apiVisitRetail… Emu 42.810000000000000000 42.810000000000000000 None None None Boys' Shoes boys-shoes 1599 EMU Australia Annie Boot Toddler Little Kid Bi… EMU Australia Annie Boot Toddler Little Kid Bi… [emu, australia, annie, boot, toddler, little,… [emu, australia, annie, boot, toddler, little,…

13 533733 Heelys Dual Up Skate Shoe (Little Kid/Big Kid) None heelys dual up skate shoe little kid/big kid http://www.shopstyle.com/p/heelys-dual-up-skat… http://www.shopstyle.com/action/apiVisitRetail… Heelys 42.170000000000000000 42.170000000000000000 None None None Boys' Shoes boys-shoes 1599 Heelys Dual Up Skate Shoe Little Kid Big Kid Heelys Dual Up Skate Shoe Little Kid Big Kid [heelys, dual, up, skate, shoe, little, kid, b… [heelys, dual, skate, shoe, little, kid, big, …

14 533749 Heelys Feisty Skate Shoe (Little Kid/Big Kid) None non-marking outsole http://www.shopstyle.com/p/heelys-feisty-skate…

http://www.shopstyle.com/action/apiVisitRetail… Heelys
48.000000000000000000 48.000000000000000000 None None
None Boys' Shoes boys-shoes 1599 Heelys Feisty Skate Shoe
Little Kid Big Kid Heelys Feisty Skate Shoe Little Kid Big Kid [heelys,
feisty, skate, shoe, little, kid, big… [heelys, feisty, skate, shoe, little,
kid, big…
15 538249 KEEN Punky AK Youth Bootie (Little Kid/Big Kid) None
keen punky ak youth bootie little kid/big kid http://www.shopstyle.com/p/keen-
punky-ak-youth… http://www.shopstyle.com/action/apiVisitRetail…
Keen 65.000000000000000000 65.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 KEEN Punky AK Youth
Bootie Little Kid Big Kid KEEN Punky AK Youth Bootie Little Kid Big Kid
[keen, punky, ak, youth, bootie, little, kid, … [keen, punky, ak, youth,
bootie, little, kid, …
16 534425 Vans Authentic (Inf/Tod) - Red-4 Infant None vans
authentic - red vans very first and most … http://www.shopstyle.com/p/vans-
authentic/3931… http://www.shopstyle.com/action/apiVisitRetail…
Vans 17.790000000000000000 30.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Vans
Authentic Inf Tod Red 4 Infant Vans Authentic Inf Tod Red 4
Infant [vans, authentic, inf, tod, red, 4, infant] [vans,
authentic, inf, tod, red, 4, infant]
17 534426 Vans Old Skool Crib Shoe - Navy/Navy-Navy/Navy-4 None vans
old skool crib shoe - navy/navy sizes ava… http://www.shopstyle.com/p/vans-
old-skool-crib… http://www.shopstyle.com/action/apiVisitRetail…
Vans 17.600000000000000000 22.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Vans Old Skool Crib
Shoe Navy Navy Navy Navy 4 Vans Old Skool Crib Shoe Navy Navy Navy Navy 4
[vans, old, skool, crib, shoe, navy, navy, nav… [vans, old, skool, crib,
shoe, navy, navy, nav…
18 534427 Vans Classic Slip-On - T-Rex Blue/Black-4 None vans
classic slip-on - t-rex blue/black sizes … http://www.shopstyle.com/p/vans-
classic-slip-o… http://www.shopstyle.com/action/apiVisitRetail…
Vans 33.250000000000000000 35.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Vans Classic
Slip On T Rex Blue Black 4 Vans Classic Slip On T Rex Blue Black 4
[vans, classic, slip, on, t, rex, blue, black, 4] [vans, classic, slip,
rex, blue, black, 4]
19 535055 BIRKI'S Sandals None
birki's sandals. logo detail, solid color…
http://www.shopstyle.com/p/birki-s-sandals/461…
http://www.shopstyle.com/action/apiVisitRetail… Birki's
44.000000000000000000 44.000000000000000000 None None
None Boys' Shoes boys-shoes 1599
BIRKI S Sandals BIRKI S Sandals
[birki, s, sandals] [birki, sandals]
20 535056 BIRKI'S Sandals None

```
birki&amp;apos;s sandals. logo detail, solid color…
http://www.shopstyle.com/p/birki-s-sandals/461…
http://www.shopstyle.com/action/apiVisitRetail…           Birki's
44.000000000000000000   44.000000000000000000                None    None
None   Boys' Shoes    boys-shoes         1599
BIRKI S Sandals                                    BIRKI S Sandals
[birki, s, sandals]                                   [birki, sandals]
21      535261   Bailey Button Kids Infant Chestnut 5991t Che Ugg   None
bailey button kids infant chestnut 5991t che ugg  http://www.shopstyle.com/p
/ugg-bailey-button-k…  http://www.shopstyle.com/action/apiVisitRetail…
UGG  110.000000000000000000  110.000000000000000000                 None    None
None   Boys' Shoes    boys-shoes         1599   Bailey Button Kids Infant
Chestnut 5991t Che Ugg   Bailey Button Kids Infant Chestnut 5991t Che Ugg
[bailey, button, kids, infant, chestnut, 5991t…  [bailey, button, kids,
infant, chestnut, 5991t…
22      537473                             PUMA Low-tops & trainers   None  puma
low-tops & trainers. logo detail, solid c…  http://www.shopstyle.com/p/puma-
low-tops-train…  http://www.shopstyle.com/action/apiVisitRetail…
Puma   55.000000000000000000   55.000000000000000000                 None
None      None   Boys' Shoes    boys-shoes         1599
PUMA Low tops trainers                             PUMA Low tops trainers
[puma, low, tops, trainers]                        [puma, low, tops, trainers]
23      537474                             PUMA Low-tops & trainers   None  puma
low-tops & trainers. logo detail, solid c…  http://www.shopstyle.com/p/puma-
low-tops-train…  http://www.shopstyle.com/action/apiVisitRetail…
Puma   70.000000000000000000   70.000000000000000000                 None
None      None   Boys' Shoes    boys-shoes         1599
PUMA Low tops trainers                             PUMA Low tops trainers
[puma, low, tops, trainers]                        [puma, low, tops, trainers]
24      537561  PUMA Evospeed Lo Ferrari 1.3 JR Sneaker (Littl…   None
removable insole  http://www.shopstyle.com/p/puma-evospeed-lo-fe…
http://www.shopstyle.com/action/apiVisitRetail…              Puma
65.000000000000000000   65.000000000000000000                  None    None
None   Boys' Shoes    boys-shoes         1599  PUMA Evospeed Lo Ferrari 1 3 JR
Sneaker Little…  PUMA Evospeed Lo Ferrari 1 3 JR Sneaker Little…  [puma,
evospeed, lo, ferrari, 1, 3, jr, sneake…  [puma, evospeed, lo, ferrari, 1, 3,
jr, sneake…
25      537821  Reebok Classic Leather BTS Running Shoe (Littl…   None
reebok classic leather sneaker  http://www.shopstyle.com/p/reebok-classic-
leat…  http://www.shopstyle.com/action/apiVisitRetail…            Reebok
44.990000000000000000   45.000000000000000000                  None    None
None   Boys' Shoes    boys-shoes         1599  Reebok Classic Leather BTS
Running Shoe Little…  Reebok Classic Leather BTS Running Shoe Little…
[reebok, classic, leather, bts, running, shoe,…  [reebok, classic, leather,
bts, running, shoe,…
26      537822    Reebok Planes Zigkick Running Shoe (Little Kid)   None
reebok planes zigkick running shoe little kid  http://www.shopstyle.com/p
```

/reebok-planes-zigki… http://www.shopstyle.com/action/apiVisitRetail…
Reebok 59.990000000000000000 59.990000000000000000 None
None None Boys' Shoes boys-shoes 1599 Reebok Planes
Zigkick Running Shoe Little Kid Reebok Planes Zigkick Running Shoe Little
Kid [reebok, planes, zigkick, running, shoe, littl… [reebok, planes,
zigkick, running, shoe, littl…
27 537823 Reebok Cars Neon RN Running Shoe (Little Kid) None
reebok cars neon rn running shoe little kid http://www.shopstyle.com/p/reebok-
cars-neon-rn… http://www.shopstyle.com/action/apiVisitRetail…
Reebok 54.990000000000000000 54.990000000000000000 None
None None Boys' Shoes boys-shoes 1599 Reebok Cars Neon
RN Running Shoe Little Kid Reebok Cars Neon RN Running Shoe Little Kid
[reebok, cars, neon, rn, running, shoe, little… [reebok, cars, neon, rn,
running, shoe, little…
28 530661 Kid's Neon Leather Low-Top Sneaker None
neon yellow leather and perforated le… http://www.shopstyle.com/p/gucci-
neon-leather-… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 360.000000000000000000 360.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Kid s
Neon Leather Low Top Sneaker Kid s Neon Leather Low Top Sneaker
[kid, s, neon, leather, low, top, sneaker] [kid, neon, leather, low,
top, sneaker]
29 540706 OCA-LOCA Moccasins None oca-
loca moccasins. sueded, bow detailing, two… http://www.shopstyle.com/p/oca-
loca-moccasins/… http://www.shopstyle.com/action/apiVisitRetail…
Oca-Loca 73.000000000000000000 73.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
OCA LOCA Moccasins OCA LOCA Moccasins
[oca, loca, moccasins] [oca, loca, moccasins]
30 540707 OCA-LOCA Moccasins None oca-
loca moccasins. no appliqu s, solid color,… http://www.shopstyle.com/p/oca-
loca-moccasins/… http://www.shopstyle.com/action/apiVisitRetail…
Oca-Loca 58.000000000000000000 58.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
OCA LOCA Moccasins OCA LOCA Moccasins
[oca, loca, moccasins] [oca, loca, moccasins]
31 530769 GIESSWEIN Ankle boots None
giesswein ankle boots. flannel, logo detail, m… http://www.shopstyle.com/p
/giesswein-ankle-boo… http://www.shopstyle.com/action/apiVisitRetail…
Giesswein 36.000000000000000000 56.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
GIESSWEIN Ankle boots GIESSWEIN Ankle boots
[giesswein, ankle, boots] [giesswein, ankle, boots]
32 541546 Tsukihoshi Speed Sneaker - Steel/Lime-8 US Tod… None
tsukihoshi speed sneaker - white/white synthet… http://www.shopstyle.com/p
/tsukihoshi-speed-sn… http://www.shopstyle.com/action/apiVisitRetail…
Tsukihoshi 37.490000000000000000 58.000000000000000000 None

None None Boys' Shoes boys-shoes 1599 Tsukihoshi Speed Sneaker Steel Lime 8 US Toddler Tsukihoshi Speed Sneaker Steel Lime 8 US Toddler [tsukihoshi, speed, sneaker, steel, lime, 8, u… [tsukihoshi, speed, sneaker, steel, lime, 8, u…
33 541547 Tsukihoshi Lynx Sneaker - Graphite/Green-13 US… None tsukihoshi lynx sneaker - navy/silver syntheti… http://www.shopstyle.com/p/tsukihoshi-lynx-sne… http://www.shopstyle.com/action/apiVisitRetail… Tsukihoshi 31.500000000000000000 63.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Tsukihoshi Lynx Sneaker Graphite Green 13 US L… Tsukihoshi Lynx Sneaker Graphite Green 13 US L… [tsukihoshi, lynx, sneaker, graphite, green, 1… [tsukihoshi, lynx, sneaker, graphite, green, 1…
34 541548 Tsukihoshi Lynx Sneaker - Navy/Silver-2.5 US L… None tsukihoshi lynx sneaker - navy/silver syntheti… http://www.shopstyle.com/p/tsukihoshi-lynx-sne… http://www.shopstyle.com/action/apiVisitRetail… Tsukihoshi 31.500000000000000000 63.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Tsukihoshi Lynx Sneaker Navy Silver 2 5 US Lit… Tsukihoshi Lynx Sneaker Navy Silver 2 5 US Lit… [tsukihoshi, lynx, sneaker, navy, silver, 2, 5… [tsukihoshi, lynx, sneaker, navy, silver, 2, 5…
35 530770 BENSIMON Low-tops & trainers None bensimon low-tops & trainers. canvas, no appli… http://www.shopstyle.com/p/bensimon-low-tops-t… http://www.shopstyle.com/action/apiVisitRetail… Bensimon 28.000000000000000000 28.000000000000000000 None None None Boys' Shoes boys-shoes 1599 BENSIMON Low tops trainers BENSIMON Low tops trainers [bensimon, low, tops, trainers] [bensimon, low, tops, trainers]
36 545324 CAMPER Moccasins None camper moccasins. logo detail, solid color, ve… http://www.shopstyle.com/p/camper-moccasins/46… http://www.shopstyle.com/action/apiVisitRetail… Camper 111.000000000000000000 111.000000000000000000 None None None Boys' Shoes boys-shoes 1599 CAMPER Moccasins CAMPER Moccasins [camper, moccasins] [camper, moccasins]
37 530845 HYDROGEN Low-tops & trainers None hydrogen low-tops & trainers. sueded, leather … http://www.shopstyle.com/p/hydrogen-low-tops-t… http://www.shopstyle.com/action/apiVisitRetail… Hydrogen 68.000000000000000000 78.000000000000000000 None None None Boys' Shoes boys-shoes 1599 HYDROGEN Low tops trainers HYDROGEN Low tops trainers [hydrogen, low, tops, trainers] [hydrogen, low, tops, trainers]
38 530895 Marvel Captain America Slipper (Infant/Toddler) None marvel captain america slipper infant/toddler http://www.shopstyle.com/p/marvel-captain-amer… http://www.shopstyle.com/action/apiVisitRetail… Marvel 21.950000000000000000 21.950000000000000000 None

None None Boys' Shoes boys-shoes 1599 Marvel Captain America Slipper Infant Toddler Marvel Captain America Slipper Infant Toddler [marvel, captain, america, slipper, infant, to… [marvel, captain, america, slipper, infant, to…
39 545443 CAMPER Sandals None camper sandals. sueded, logo detail, solid col… http://www.shopstyle.com/p/camper-sandals/4547… http://www.shopstyle.com/action/apiVisitRetail…
Camper 32.000000000000000000 57.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
CAMPER Sandals CAMPER Sandals
[camper, sandals] [camper, sandals]
40 545508 crocs 10190 Baya Kids Clog (Toddler/Little Kid) None
crocs 10190 baya kids clog toddler/little kid
http://www.shopstyle.com/p/crocs-10190-baya-cl…
http://www.shopstyle.com/action/apiVisitRetail… Crocs
21.370000000000000000 21.370000000000000000 None None
None Boys' Shoes boys-shoes 1599 crocs 10190 Baya Kids Clog Toddler Little Kid crocs 10190 Baya Kids Clog Toddler Little Kid [crocs, 10190, baya, kids, clog, toddler, litt… [crocs, 10190, baya, kids, clog, toddler, litt…
41 530933 START RITE Ankle boots None start rite ankle boots. sueded, no appliqu s, … http://www.shopstyle.com/p/start-rite-ankle-bo… http://www.shopstyle.com/action/apiVisitRetail… Start Rite 49.000000000000000000 49.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
START RITE Ankle boots START RITE Ankle boots
[start, rite, ankle, boots] [start, rite, ankle, boots]
42 530934 POM D'API Ankle boots None pom d'api ankle boots. sueded, no appliqu… http://www.shopstyle.com/p/pom-d-api-ankle-boo… http://www.shopstyle.com/action/apiVisitRetail… Pom D'Api 172.000000000000000000 172.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
POM D API Ankle boots POM D API Ankle boots
[pom, d, api, ankle, boots] [pom, d, api, ankle, boots]
43 530935 POM D'API Ankle boots None pom d'api ankle boots. lacing, camouflage… http://www.shopstyle.com/p/pom-d-api-ankle-boo… http://www.shopstyle.com/action/apiVisitRetail… Pom D'Api 91.000000000000000000 166.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
POM D API Ankle boots POM D API Ankle boots
[pom, d, api, ankle, boots] [pom, d, api, ankle, boots]
44 545555 crocs 12213 Duet Plus Clg K Clog (Toddler/Litt… None adjustable heel strap http://www.shopstyle.com/p/crocs-12213-duet-pl…
http://www.shopstyle.com/action/apiVisitRetail… Crocs
16.250000000000000000 16.250000000000000000 None None
None Boys' Shoes boys-shoes 1599 crocs 12213 Duet Plus Clg K Clog Toddler Littl… crocs 12213 Duet Plus Clg K Clog Toddler Littl… [crocs,

12213, duet, plus, clg, k, clog, toddl… [crocs, 12213, duet, plus, clg, k, clog, toddl…
45 545620 Patent Lace-Up Shoes None derby style lace up closure remov… http://www.shopstyle.com/p/dolce-gabbana-paten… http://www.shopstyle.com/action/apiVisitRetail… Dolce & Gabbana 350.000000000000000000 350.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Patent Lace Up Shoes Patent Lace Up Shoes [patent, lace, up, shoes] [patent, lace, shoes]
46 531287 FALCOTTO BY NATURINO Ankle boots None falcotto by naturino ankle boots. logo detail,… http://www.shopstyle.com/p/naturino-falcotto-b… http://www.shopstyle.com/action/apiVisitRetail… Naturino 36.000000000000000000 75.000000000000000000 None None None Boys' Shoes boys-shoes 1599 FALCOTTO BY NATURINO Ankle boots FALCOTTO BY NATURINO Ankle boots [falcotto, by, naturino, ankle, boots] [falcotto, naturino, ankle, boots]
47 531288 NATURINO Moccasins None naturino moccasins. sueded, textured leather, … http://www.shopstyle.com/p/naturino-moccasins/… http://www.shopstyle.com/action/apiVisitRetail… Naturino 43.000000000000000000 79.000000000000000000 None None None Boys' Shoes boys-shoes 1599 NATURINO Moccasins NATURINO Moccasins [naturino, moccasins] [naturino, moccasins]
48 531396 MERRELL Sandals None merrell sandals. sueded, logo detail, two-tone… http://www.shopstyle.com/p/merrell-sandals/450… http://www.shopstyle.com/action/apiVisitRetail… Merrell 33.000000000000000000 53.000000000000000000 None None None Boys' Shoes boys-shoes 1599 MERRELL Sandals MERRELL Sandals [merrell, sandals] [merrell, sandals]
49 530627 Toddler Leather High-Top Sneaker With Web Detail None <ul> brown microguccissma leather <li… http://www.shopstyle.com/p/gucci-leather-high-… http://www.shopstyle.com/action/apiVisitRetail… Gucci 275.000000000000000000 275.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Toddler Leather High Top Sneaker With Web Detail Toddler Leather High Top Sneaker With Web Detail [toddler, leather, high, top, sneaker, with, w… [toddler, leather, high, top, sneaker, web, de…
50 530486 Toddler Shearling High Top Sneaker None <ul> red shearling red patent lea… http://www.shopstyle.com/p/gucci-shearling-hig… http://www.shopstyle.com/action/apiVisitRetail… Gucci 275.000000000000000000 275.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Toddler Shearling High Top Sneaker Toddler Shearling High Top Sneaker [toddler, shearling, high, top, sneaker] [toddler, shearling, high, top, sneaker]

51 530487 Toddler Gg Print Lace-Up Sneaker None
beige/ebony original gg fabric with g… http://www.shopstyle.com/p/gucci-
gg-print-lace… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 195.000000000000000000 195.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
Toddler Gg Print Lace Up Sneaker Toddler Gg Print Lace Up
Sneaker [toddler, gg, print, lace, up, sneaker]
[toddler, gg, print, lace, sneaker]
52 530488 toddler brown GG rubber rain boot None
brown gg transparent rubber … http://www.shopstyle.com/p/gucci-
brown-gg-rubb… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 145.000000000000000000 145.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 toddler
brown GG rubber rain boot toddler brown GG rubber rain boot
[toddler, brown, gg, rubber, rain, boot] [toddler, brown, gg, rubber,
rain, boot]
53 530489 Jumping Jacks Cheerleader III Saddle Shoe Oxfo… None
perfect for special occasions and durable enou… http://www.shopstyle.com/p
/jumping-jacks-cheer… http://www.shopstyle.com/action/apiVisitRetail…
Jumping Jacks 38.420000000000000000 43.160000000000000000
None None None Boys' Shoes boys-shoes 1599 Jumping Jacks
Cheerleader III Saddle Shoe Oxfo… Jumping Jacks Cheerleader III Saddle Shoe
Oxfo… [jumping, jacks, cheerleader, iii, saddle, sho… [jumping, jacks,
cheerleader, iii, saddle, sho…
54 530490 Kid's Gg Print Slip-On Sneaker None
beige/ebony original gg fabric with b… http://www.shopstyle.com/p/gucci-
gg-print-slip… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 195.000000000000000000 195.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Kid
s Gg Print Slip On Sneaker Kid s Gg Print Slip On Sneaker
[kid, s, gg, print, slip, on, sneaker] [kid, gg, print, slip,
sneaker]
55 530491 Kid's Blue Suede Driver With Web Detail None
navy blue suede with blue/red/blue si… http://www.shopstyle.com/p/gucci-
kid-s-blue-su… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 195.000000000000000000 195.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Kid s Blue
Suede Driver With Web Detail Kid s Blue Suede Driver With Web Detail
[kid, s, blue, suede, driver, with, web, detail] [kid, blue, suede,
driver, web, detail]
56 530492 Toddler Gg Star Supreme High-Top Sneaker None
beige, ebony and yellow micro gg supr… http://www.shopstyle.com/p/gucci-
gg-star-supre… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 295.000000000000000000 295.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Toddler Gg
Star Supreme High Top Sneaker Toddler Gg Star Supreme High Top Sneaker
[toddler, gg, star, supreme, high, top, sneaker] [toddler, gg, star, supreme,

```
high, top, sneaker]
57      530493                                    GF FERRE' Bootees   None  gf
ferre&amp;apos; bootees. jersey, stretch, strip…  http://www.shopstyle.com/p
/gianfranco-ferre-bo…  http://www.shopstyle.com/action/apiVisitRetail…
Gianfranco Ferre   30.000000000000000000  104.000000000000000000
None    None      None   Boys' Shoes    boys-shoes         1599
GF FERRE Bootees                                    GF FERRE Bootees
[gf, ferre, bootees]                                [gf, ferre, bootees]
58      530494                                    GF FERRE' Bootees   None  gf
ferre&amp;apos; bootees. jersey, solid color, r…  http://www.shopstyle.com/p
/gianfranco-ferre-bo…  http://www.shopstyle.com/action/apiVisitRetail…
Gianfranco Ferre   22.000000000000000000   29.000000000000000000
None    None      None   Boys' Shoes    boys-shoes         1599
GF FERRE Bootees                                    GF FERRE Bootees
[gf, ferre, bootees]                                [gf, ferre, bootees]
59      530495                              HAVAIANAS Thong sandals   None
havaianas thong sandals. logo detail, solid co…  http://www.shopstyle.com/p
/havaianas-thong-san…  http://www.shopstyle.com/action/apiVisitRetail…
Havaianas   12.000000000000000000   24.000000000000000000                 None
None      None   Boys' Shoes    boys-shoes         1599
HAVAIANAS Thong sandals                             HAVAIANAS Thong sandals
[havaianas, thong, sandals]                         [havaianas, thong, sandals]
60      530496                              HAVAIANAS Thong sandals   None
havaianas thong sandals. logo detail, solid co…  http://www.shopstyle.com/p
/havaianas-thong-san…  http://www.shopstyle.com/action/apiVisitRetail…
Havaianas   12.000000000000000000   31.000000000000000000                 None
None      None   Boys' Shoes    boys-shoes         1599
HAVAIANAS Thong sandals                             HAVAIANAS Thong sandals
[havaianas, thong, sandals]                         [havaianas, thong, sandals]
61      530497                              HAVAIANAS Thong sandals   None
havaianas thong sandals. solid color, elastici…  http://www.shopstyle.com/p
/havaianas-flip-flop…  http://www.shopstyle.com/action/apiVisitRetail…
Havaianas   12.000000000000000000   29.000000000000000000                 None
None      None   Boys' Shoes    boys-shoes         1599
HAVAIANAS Thong sandals                             HAVAIANAS Thong sandals
[havaianas, thong, sandals]                         [havaianas, thong, sandals]
62      530498                              HAVAIANAS Thong sandals   None
havaianas thong sandals. logo detail, solid co…  http://www.shopstyle.com/p
/havaianas-thong-san…  http://www.shopstyle.com/action/apiVisitRetail…
Havaianas   12.000000000000000000   31.000000000000000000                 None
None      None   Boys' Shoes    boys-shoes         1599
HAVAIANAS Thong sandals                             HAVAIANAS Thong sandals
[havaianas, thong, sandals]                         [havaianas, thong, sandals]
63      530499                              HAVAIANAS Thong sandals   None
havaianas thong sandals. no appliqu s, solid c…  http://www.shopstyle.com/p
/havaianas-thong-san…  http://www.shopstyle.com/action/apiVisitRetail…
Havaianas   24.000000000000000000   31.000000000000000000                 None
```

None None Boys' Shoes boys-shoes 1599
HAVAIANAS Thong sandals HAVAIANAS Thong sandals
[havaianas, thong, sandals] [havaianas, thong, sandals]
64 530500 HOGAN JUNIOR Sandals None hogan
junior sandals. sueded, techno fabric, l… http://www.shopstyle.com/p/hogan-
junior-sandal… http://www.shopstyle.com/action/apiVisitRetail…
Hogan 100.000000000000000000 100.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
HOGAN JUNIOR Sandals HOGAN JUNIOR Sandals
[hogan, junior, sandals] [hogan, junior, sandals]
65 530501 GUCCI Lace-up shoes None gucci
lace-up shoes. sueded, logo detail, soli… http://www.shopstyle.com/p/gucci-
lace-up-shoes… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 104.000000000000000000 143.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
GUCCI Lace up shoes GUCCI Lace up shoes
[gucci, lace, up, shoes] [gucci, lace, shoes]
66 530502 LA PERLA Newborn shoes None la
perla newborn shoes. plain weave, no appliq… http://www.shopstyle.com/p/la-
perla-newborn-sh… http://www.shopstyle.com/action/apiVisitRetail…
La Perla 24.000000000000000000 24.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
LA PERLA Newborn shoes LA PERLA Newborn shoes
[la, perla, newborn, shoes] [la, perla, newborn, shoes]
67 530504 HOGAN REBEL High-tops & trainers None hogan
rebel high-tops & trainers. canvas, sued… http://www.shopstyle.com/p/hogan-
rebel-high-to… http://www.shopstyle.com/action/apiVisitRetail…
Hogan 139.000000000000000000 139.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
HOGAN REBEL High tops trainers HOGAN REBEL High tops
trainers [hogan, rebel, high, tops, trainers]
[hogan, rebel, high, tops, trainers]
68 530505 GF FERRE' Moccasins None gf
ferre' moccasins. sueded, logo detail,… http://www.shopstyle.com/p
/gianfranco-ferre-mo… http://www.shopstyle.com/action/apiVisitRetail…
Gianfranco Ferre 79.000000000000000000 132.000000000000000000
None None None Boys' Shoes boys-shoes 1599
GF FERRE Moccasins GF FERRE Moccasins
[gf, ferre, moccasins] [gf, ferre, moccasins]
69 530506 HAVAIANAS Thong sandals None
havaianas thong sandals. logo detail, solid co… http://www.shopstyle.com/p
/havaianas-thong-san… http://www.shopstyle.com/action/apiVisitRetail…
Havaianas 12.000000000000000000 31.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
HAVAIANAS Thong sandals HAVAIANAS Thong sandals
[havaianas, thong, sandals] [havaianas, thong, sandals]
70 530507 HOGAN JUNIOR Sandals None hogan

junior sandals. techno fabric, sueded, l… http://www.shopstyle.com/p/hogan-junior-sandal… http://www.shopstyle.com/action/apiVisitRetail…
Hogan 109.000000000000000000 109.000000000000000000 None None None Boys' Shoes boys-shoes 1599
HOGAN JUNIOR Sandals HOGAN JUNIOR Sandals
[hogan, junior, sandals] [hogan, junior, sandals]
71 530508 Toddler Red Leather High-Top Sneaker None <ul> red leather made in italy</l… http://www.shopstyle.com/p/gucci-red-leather-h… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 275.000000000000000000 275.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Toddler Red Leather High Top Sneaker Toddler Red Leather High Top Sneaker
[toddler, red, leather, high, top, sneaker] [toddler, red, leather, high, top, sneaker]
72 530509 Toddler Gg Star Print Rubber Rain Boot None <ul> yellow and black gg star print rubber… http://www.shopstyle.com/p/gucci-gg-star-print… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 145.000000000000000000 145.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Toddler Gg Star Print Rubber Rain Boot Toddler Gg Star Print Rubber Rain Boot
[toddler, gg, star, print, rubber, rain, boot] [toddler, gg, star, print, rubber, rain, boot]
73 530510 Toddler Neon Leather Low-Top Sneaker None <ul> neon yellow leather and perforated le… http://www.shopstyle.com/p/gucci-neon-leather-… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 310.000000000000000000 310.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Toddler Neon Leather Low Top Sneaker Toddler Neon Leather Low Top Sneaker
[toddler, neon, leather, low, top, sneaker] [toddler, neon, leather, low, top, sneaker]
74 530511 havaianas Boys' Monsters Inc. Printed Flip Flo… None this monsters inc. printed pair will ignite yo… http://www.shopstyle.com/p/havaianas-boys-mons… http://www.shopstyle.com/action/apiVisitRetail…
Havaianas 9.970000000000000000 19.000000000000000000 None None None Boys' Shoes boys-shoes 1599 havaianas Boys Monsters Inc Printed Flip Flops… havaianas Boys Monsters Inc Printed Flip Flops…
[havaianas, boys, monsters, inc, printed, flip… [havaianas, boys, monsters, inc, printed, flip…
75 530512 Jumping Jacks Glitter Zip Boot (Toddler/Little… None boot featuring glittered shaft, dual buckle st… http://www.shopstyle.com/p/jumping-jacks-glitt… http://www.shopstyle.com/action/apiVisitRetail…
Jumping Jacks 44.950000000000000000 44.950000000000000000 None None None Boys' Shoes boys-shoes 1599 Jumping Jacks Glitter Zip Boot Toddler Little … Jumping Jacks Glitter Zip Boot Toddler Little …
[jumping, jacks, glitter, zip, boot, toddler, … [jumping, jacks, glitter, zip, boot, toddler, …
76 530513 Hello Kitty Kelli Sneaker (Little Kid) None both

no fuss, and cute enough for her! have yo… http://www.shopstyle.com/p/hello-kitty-kelli-s… http://www.shopstyle.com/action/apiVisitRetail… Hello Kitty 29.990000000000000000 29.990000000000000000 None None None Boys' Shoes boys-shoes 1599 Hello Kitty Kelli Sneaker Little Kid Hello Kitty Kelli Sneaker Little Kid [hello, kitty, kelli, sneaker, little, kid] [hello, kitty, kelli, sneaker, little, kid]
77 530514 K-Swiss 51514 Lozan Strap DX Children Sneaker … None breathable leather upper flexible sole non mar… http://www.shopstyle.com/p/k-swiss-51514-lozan… http://www.shopstyle.com/action/apiVisitRetail… K-Swiss 48.000000000000000000 48.000000000000000000 None None None Boys' Shoes boys-shoes 1599 K Swiss 51514 Lozan Strap DX Children Sneaker … K Swiss 51514 Lozan Strap DX Children Sneaker … [k, swiss, 51514, lozan, strap, dx, children, … [k, swiss, 51514, lozan, strap, dx, children, …
78 530515 Jumping Jacks Titan Sneaker (Toddler/Little Kid) None for 60 years, jumping jacks has been committed… http://www.shopstyle.com/p/jumping-jacks-titan… http://www.shopstyle.com/action/apiVisitRetail… Jumping Jacks 38.950000000000000000 38.950000000000000000 None None None Boys' Shoes boys-shoes 1599 Jumping Jacks Titan Sneaker Toddler Little Kid Jumping Jacks Titan Sneaker Toddler Little Kid [jumping, jacks, titan, sneaker, toddler, litt… [jumping, jacks, titan, sneaker, toddler, litt…
79 530516 Hello Kitty Lil Lindsey Mary Jane Sneaker (Tod… None this hello kitty-themed shoe is flat-out adora… http://www.shopstyle.com/p/hello-kitty-lil-lin… http://www.shopstyle.com/action/apiVisitRetail… Hello Kitty 24.990000000000000000 24.990000000000000000 None None None Boys' Shoes boys-shoes 1599 Hello Kitty Lil Lindsey Mary Jane Sneaker Todd… Hello Kitty Lil Lindsey Mary Jane Sneaker Todd… [hello, kitty, lil, lindsey, mary, jane, sneak… [hello, kitty, lil, lindsey, mary, jane, sneak…
80 530517 Havaianas Monsters Inc Flip Flop (Toddler/Litt… None havaianas - kids monsters inc sandal- black http://www.shopstyle.com/p/havaianas-monsters-… http://www.shopstyle.com/action/apiVisitRetail… Havaianas 7.980000000000000000 7.980000000000000000 None None None Boys' Shoes boys-shoes 1599 Havaianas Monsters Inc Flip Flop Toddler Littl… Havaianas Monsters Inc Flip Flop Toddler Littl… [havaianas, monsters, inc, flip, flop, toddler… [havaianas, monsters, inc, flip, flop, toddler…
81 530518 Jumping Jacks Finish Line Sneaker (Toddler) None charming baby & toddler athletic inspired boy'… http://www.shopstyle.com/p/jumping-jacks-finis… http://www.shopstyle.com/action/apiVisitRetail… Jumping Jacks 20.610000000000000000 20.610000000000000000 None None None Boys' Shoes boys-shoes 1599 Jumping Jacks Finish Line Sneaker Toddler Jumping Jacks Finish Line Sneaker Toddler [jumping, jacks, finish, line, sneaker, toddler] [jumping, jacks,

finish, line, sneaker, toddler]
82 530520 Havaianas Power Flip Flop (Toddler/Little Kid) None
havaianas stands for style, comfort, affordabl… http://www.shopstyle.com/p
/havaianas-power-fli… http://www.shopstyle.com/action/apiVisitRetail…
Havaianas 12.070000000000000000 23.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Havaianas Power
Flip Flop Toddler Little Kid Havaianas Power Flip Flop Toddler Little Kid
[havaianas, power, flip, flop, toddler, little… [havaianas, power, flip,
flop, toddler, little…
83 530521 baby original GG canvas sneaker with signature… None
beige/ebony original gg canvas and br… http://www.shopstyle.com/p/gucci-
baby-original… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 195.000000000000000000 195.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 baby original GG canvas
sneaker with signature… baby original GG canvas sneaker with signature…
[baby, original, gg, canvas, sneaker, with, si… [baby, original, gg, canvas,
sneaker, signatur…
84 530522 Toddler Leather High-Top Sneaker With Web Detail None
red microguccissma leather r… http://www.shopstyle.com/p/gucci-
leather-high-… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 275.000000000000000000 275.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 Toddler Leather High
Top Sneaker With Web Detail Toddler Leather High Top Sneaker With Web Detail
[toddler, leather, high, top, sneaker, with, w… [toddler, leather, high, top,
sneaker, web, de…
85 530523 kid's blue GG rubber rain boot None
blue gg transparent rubber b… http://www.shopstyle.com/p/gucci-
blue-gg-rubbe… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 145.000000000000000000 145.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 kid
s blue GG rubber rain boot kid s blue GG rubber rain boot
[kid, s, blue, gg, rubber, rain, boot] [kid, blue, gg, rubber,
rain, boot]
86 530524 Toddler Gg Print Lace-Up Sneaker None
beige/blue original gg fabric with bl… http://www.shopstyle.com/p/gucci-
lace-up-sneak… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 195.000000000000000000 195.000000000000000000 None
None None Boys' Shoes boys-shoes 1599
Toddler Gg Print Lace Up Sneaker Toddler Gg Print Lace Up
Sneaker [toddler, gg, print, lace, up, sneaker]
[toddler, gg, print, lace, sneaker]
87 530525 kid's beige/blue original GG lace-up sneaker None
beige/blue original gg fabric with bl… http://www.shopstyle.com/p/gucci-
lace-up-sneak… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 200.000000000000000000 200.000000000000000000 None
None None Boys' Shoes boys-shoes 1599 kid s beige blue
original GG lace up sneaker kid s beige blue original GG lace up sneaker

[kid, s, beige, blue, original, gg, lace, up, … [kid, beige, blue, original, gg, lace, sneaker]
88 530526 Toddler High-Top Sneaker With Web Detail None <ul> white leather with green/red/green we… http://www.shopstyle.com/p/gucci-high-top-lace… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 245.000000000000000000 245.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Toddler High Top Sneaker With Web Detail Toddler High Top Sneaker With Web Detail [toddler, high, top, sneaker, with, web, detail] [toddler, high, top, sneaker, web, detail]
89 530527 Baby Suede Driver None <ul> red suede green/red/green si… http://www.shopstyle.com/p/gucci-baby-suede-dr… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 185.000000000000000000 185.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Baby Suede Driver Baby Suede Driver [baby, suede, driver] [baby, suede, driver]
90 530528 Baby Lace-Up Sneaker None <ul> beige/ebony original gg fabric with g… http://www.shopstyle.com/p/gucci-baby-lace-up-… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 185.000000000000000000 185.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Baby Lace Up Sneaker Baby Lace Up Sneaker [baby, lace, up, sneaker] [baby, lace, sneaker]
91 530529 Toddler Leather Sandal With Signature Web Straps None <ul> black leather with green/red/green si… http://www.shopstyle.com/p/gucci-sandal-with-s… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 210.000000000000000000 210.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Toddler Leather Sandal With Signature Web Straps Toddler Leather Sandal With Signature Web Straps [toddler, leather, sandal, with, signature, we… [toddler, leather, sandal, signature, web, str…
92 530530 Toddler Leather High-Top Sneaker With Web Detail None <ul> black microguccissma leather <li… http://www.shopstyle.com/p/gucci-leather-high-… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 275.000000000000000000 275.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Toddler Leather High Top Sneaker With Web Detail Toddler Leather High Top Sneaker With Web Detail [toddler, leather, high, top, sneaker, with, w… [toddler, leather, high, top, sneaker, web, de…
93 530531 Baby Suede Driver None <ul> navy suede blue/red/blue sig… http://www.shopstyle.com/p/gucci-baby-suede-dr… http://www.shopstyle.com/action/apiVisitRetail…
Gucci 185.000000000000000000 185.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Baby Suede Driver Baby Suede Driver [baby, suede, driver] [baby, suede, driver]

94 530532 Kid's Rubber Rain Boot None <ul> red orange rubber made in it… http://www.shopstyle.com/p/gucci-rubber-rain-b… http://www.shopstyle.com/action/apiVisitRetail… Gucci 165.000000000000000000 165.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Kid s Rubber Rain Boot Kid s Rubber Rain Boot [kid, s, rubber, rain, boot] [kid, rubber, rain, boot]
95 530533 toddler blue GG rubber rain boot None <ul> blue gg transparent rubber b… http://www.shopstyle.com/p/gucci-blue-gg-rubbe… http://www.shopstyle.com/action/apiVisitRetail… Gucci 145.000000000000000000 145.000000000000000000 None None None Boys' Shoes boys-shoes 1599 toddler blue GG rubber rain boot toddler blue GG rubber rain boot [toddler, blue, gg, rubber, rain, boot] [toddler, blue, gg, rubber, rain, boot]
96 530534 kid's GG supreme canvas high-top sneaker None <ul> beige/ebony gg supreme canvas, made u… http://www.shopstyle.com/p/gucci-gg-supreme-ca… http://www.shopstyle.com/action/apiVisitRetail… Gucci 285.000000000000000000 285.000000000000000000 None None None Boys' Shoes boys-shoes 1599 kid s GG supreme canvas high top sneaker kid s GG supreme canvas high top sneaker [kid, s, gg, supreme, canvas, high, top, sneaker] [kid, gg, supreme, canvas, high, top, sneaker]
97 530535 kid's brown GG rubber rain boot None <ul> brown gg transparent rubber … http://www.shopstyle.com/p/gucci-brown-gg-rubb… http://www.shopstyle.com/action/apiVisitRetail… Gucci 145.000000000000000000 145.000000000000000000 None None None Boys' Shoes boys-shoes 1599 kid s brown GG rubber rain boot kid s brown GG rubber rain boot [kid, s, brown, gg, rubber, rain, boot] [kid, brown, gg, rubber, rain, boot]
98 530536 Toddler Black Leather Lace-Up High-Tops None <ul> black leather made in italy<… http://www.shopstyle.com/p/gucci-black-leather… http://www.shopstyle.com/action/apiVisitRetail… Gucci 275.000000000000000000 275.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Toddler Black Leather Lace Up High Tops Toddler Black Leather Lace Up High Tops [toddler, black, leather, lace, up, high, tops] [toddler, black, leather, lace, high, tops]
99 530537 Kid's Gg Print Slip-On Sneaker None <ul> beige/blue original gg fabric with bl… http://www.shopstyle.com/p/gucci-slip-on-sneak… http://www.shopstyle.com/action/apiVisitRetail… Gucci 195.000000000000000000 195.000000000000000000 None None None Boys' Shoes boys-shoes 1599 Kid s Gg Print Slip On Sneaker Kid s Gg Print Slip On Sneaker [kid, s, gg, print, slip, on, sneaker] [kid, gg, print, slip, sneaker]

3.5.5 Split dataset

- 0.8 % used for training
- 0.2 % used for testing

[18]:
```
# .sample(True, 0.1)
(training, test) = shoesDf.sample(True, 0.1).randomSplit([0.8, 0.2], seed=1987)
```

3.6 Verbose Cross Validation

[19]:
```
import numpy as np

from pyspark.ml.tuning import CrossValidator, CrossValidatorModel
from pyspark.sql.functions import rand


class CrossValidatorVerbose(CrossValidator):

    def _fit(self, dataset):
        est = self.getOrDefault(self.estimator)
        epm = self.getOrDefault(self.estimatorParamMaps)
        numModels = len(epm)

        eva = self.getOrDefault(self.evaluator)
        metricName = eva.getMetricName()

        nFolds = self.getOrDefault(self.numFolds)
        seed = self.getOrDefault(self.seed)
        h = 1.0 / nFolds

        randCol = self.uid + "_rand"
        df = dataset.select("*", rand(seed).alias(randCol))
        metrics = [0.0] * numModels

        for i in range(nFolds):
            foldNum = i + 1
            print("Comparing models on fold %d" % foldNum)

            validateLB = i * h
            validateUB = (i + 1) * h
            condition = (df[randCol] >= validateLB) & (df[randCol] < validateUB)
            validation = df.filter(condition)
            train = df.filter(~condition)

            for j in range(numModels):
                paramMap = epm[j]
```

```
                model = est.fit(train, paramMap)
                metric = eva.evaluate(model.transform(validation, paramMap))
                metrics[j] += metric

                avgSoFar = metrics[j] / foldNum
                print("params: %s\t%s: %f\tavg: %f" % (
                    {param.name: val for (param, val) in paramMap.items()},
                    metricName, metric, avgSoFar))

        if eva.isLargerBetter():
            bestIndex = np.argmax(metrics)
        else:
            bestIndex = np.argmin(metrics)

        bestParams = epm[bestIndex]
        bestModel = est.fit(dataset, bestParams)
        avgMetrics = [m / nFolds for m in metrics]
        bestAvg = avgMetrics[bestIndex]
        print("Best model:\nparams: %s\t%s: %f" % (
            {param.name: val for (param, val) in bestParams.items()},
            metricName, bestAvg))

        return self._copyValues(CrossValidatorModel(bestModel, avgMetrics))
```

3.7 Logistic Regression

3.7.1 Model Training

[20]:
```
%%time

from pyspark.ml.feature import *

label_indexer_lr = StringIndexer(inputCol="category_id",␣
 ↪outputCol="category_id_indexed")
hashingTF_lr = pyspark.ml.feature.HashingTF(inputCol="name_preprocessed",␣
 ↪outputCol="features", numFeatures=10)
lr = pyspark.ml.classification.LogisticRegression(maxIter=10, regParam=0.01,␣
 ↪featuresCol='features', labelCol="category_id_indexed")

pipeline_lr = pyspark.ml.Pipeline(stages=[
    label_indexer_lr,
    hashingTF_lr,
    lr
])
```

```
lrModel = pipeline_lr.fit(training)
```

```
CPU times: user 41.2 ms, sys: 7.39 ms, total: 48.6 ms
Wall time: 3.67 s
```

[21]:
```
pipeline_lr.getStages()
```

[21]:
```
[StringIndexer_cc2b72f174f4,
 HashingTF_878c485240c2,
 LogisticRegression_d81bdbd9ccb9]
```

3.7.2 Model Parameters

[22]:
```
# Print the coefficients and intercept for multinomial logistic regression
print("Coefficients: \n" + str(lrModel.stages[-1].coefficientMatrix))
print("Intercept: " + str(lrModel.stages[-1].interceptVector))
```

```
Coefficients:
DenseMatrix([[-0.26715459,  0.36778312,  0.24932515, -0.34870887,  0.3189703 ,
               0.16421328,  0.05211823,  0.0688311 , -0.10108216,  0.41872814],
             [-0.21438976,  0.28528888,  0.27902749, -0.36274238, -0.11659786,
               0.10042673,  0.30232345,  0.01820837,  0.08177082,  0.39796326],
             [ 0.25143419, -0.25856971, -0.21349432,  0.21844803,  0.01020006,
              -0.12719524, -0.56723017, -0.31709821, -0.51459766,  0.08954914],
             [ 0.07504628, -0.22496379, -0.15153154,  0.82886636, -0.12154805,
              -0.08233997, -0.54114432, -0.32882842,  0.32210137, -0.40721314],
             [-0.07873442, -0.37277331, -0.23107384,  0.11781403, -0.17950036,
              -0.18106532,  0.67221275, -0.22428094,  0.29831999, -0.22275605],
             [ 0.10605562,  0.06610271,  0.07960076, -0.21948921,  0.04909961,
               0.10150021,  0.04376101,  0.26992509, -0.01424266, -0.11627448],
             [ 0.12774269,  0.13713209, -0.01185371, -0.23418796,  0.03937631,
               0.02446033,  0.03795904,  0.51324301, -0.0722697 , -0.15999688]])
Intercept: [0.876749838765395,0.5061226442946448,1.6516994384885242,0.4140584009
796802,-0.04881115356349259,-1.5581527479131865,-1.8416664210515648]
```

3.7.3 Check training summary

[23]:
```
trainingSummary = lrModel.stages[-1].summary

objectiveHistory = trainingSummary.objectiveHistory
print("objectiveHistory:")
for objective in objectiveHistory:
    print(objective)

print("False positive rate by label:")
```

```
for i, rate in enumerate(trainingSummary.falsePositiveRateByLabel):
    print("label %d: %s" % (i, rate))

print("True positive rate by label:")
for i, rate in enumerate(trainingSummary.truePositiveRateByLabel):
    print("label %d: %s" % (i, rate))

print("Precision by label:")
for i, prec in enumerate(trainingSummary.precisionByLabel):
    print("label %d: %s" % (i, prec))

print("Recall by label:")
for i, rec in enumerate(trainingSummary.recallByLabel):
    print("label %d: %s" % (i, rec))

print("F-measure by label:")
for i, f in enumerate(trainingSummary.fMeasureByLabel()):
    print("label %d: %s" % (i, f))

accuracy = trainingSummary.accuracy
falsePositiveRate = trainingSummary.weightedFalsePositiveRate
truePositiveRate = trainingSummary.weightedTruePositiveRate
fMeasure = trainingSummary.weightedFMeasure()
precision = trainingSummary.weightedPrecision
recall = trainingSummary.weightedRecall
print("Accuracy: %s\nFPR: %s\nTPR: %s\nF-measure: %s\nPrecision: %s\nRecall: %s"
      % (accuracy, falsePositiveRate, truePositiveRate, fMeasure, precision,␣
 ↪recall))
```

```
objectiveHistory:
1.598841274
1.52788208073
1.42776507336
1.4195461772
1.40702601954
1.40422951834
1.39570606091
1.38558738319
1.38369218579
1.37294935707
1.3704019748
False positive rate by label:
label 0: 0.47983014862
label 1: 0.0975544922913
label 2: 0.11709253287
label 3: 0.0662251655629
label 4: 0.0130208333333
```

```
label 5: 0.0
label 6: 0.0
True positive rate by label:
label 0: 0.73562537048
label 1: 0.218494271686
label 2: 0.457502623295
label 3: 0.380165289256
label 4: 0.103723404255
label 5: 0.0
label 6: 0.0
Precision by label:
label 0: 0.439603258944
label 1: 0.421135646688
label 2: 0.480176211454
label 3: 0.442307692308
label 4: 0.393939393939
label 5: 0.0
label 6: 0.0
Recall by label:
label 0: 0.73562537048
label 1: 0.218494271686
label 2: 0.457502623295
label 3: 0.380165289256
label 4: 0.103723404255
label 5: 0.0
label 6: 0.0
F-measure by label:
label 0: 0.550332594235
label 1: 0.287715517241
label 2: 0.46856528748
label 3: 0.408888888889
label 4: 0.164210526316
label 5: 0.0
label 6: 0.0
Accuracy: 0.444020866774
FPR: 0.217744039451
TPR: 0.444020866774
F-measure: 0.408439627445
Precision: 0.427280046865
Recall: 0.444020866774
```

3.7.4 Model evaluation

[24]:
```
prediction = lrModel.transform(test)
```

```
evaluator_lr = pyspark.ml.evaluation.
 ↪MulticlassClassificationEvaluator(labelCol="category_id_indexed",␣
 ↪predictionCol="prediction", metricName="accuracy")

accuracy = evaluator_lr.evaluate(prediction)
print("Accuracy = %g " % (accuracy))
print("Test Error = %g " % (1.0 - accuracy))
```

```
Accuracy = 0.451349
Test Error = 0.548651
```

3.8 Logistic Regression Cross Validation

[25]:
```
from pyspark.ml.tuning import CrossValidator, ParamGridBuilder
from pyspark.ml.tuning import  TrainValidationSplit
```

[26]:
```
hashingTF_lr_cross = pyspark.ml.feature.HashingTF(inputCol="name_preprocessed",␣
 ↪outputCol="features")
lr_cross = pyspark.ml.classification.LogisticRegression(maxIter=10,␣
 ↪featuresCol='features', labelCol="category_id_indexed")

pipeline_lr_cross = pyspark.ml.Pipeline(stages=[
    label_indexer_lr,
    hashingTF_lr_cross,
    lr_cross
])

paramGrid_lr = ParamGridBuilder() \
    .addGrid(hashingTF_lr_cross.numFeatures, [100, 1000, 1500, 2000, 3000]) \
    .addGrid(lr_cross.maxIter, [10, 20, 30]) \
    .addGrid(lr_cross.regParam, [0.1, 0.01, 0.001]) \
    .build()

evaluator_lr_cross = pyspark.ml.evaluation.MulticlassClassificationEvaluator(
    labelCol="category_id_indexed",
    predictionCol="prediction",
    metricName="accuracy"
)

crossval_lr = CrossValidatorVerbose(
    estimator=pipeline_lr_cross,
    estimatorParamMaps=paramGrid_lr,
    evaluator=evaluator_lr_cross,
    numFolds=3
)
```

3.8.1 Run CrossValidator

[27]:
```
# Run cross-validation, and choose the best set of parameters.
crossval_lr_model = crossval_lr.fit(training)
```

```
Comparing models on fold 1
params: {'regParam': 0.1, 'numFeatures': 100, 'maxIter': 10}    accuracy:
0.610259      avg: 0.610259
params: {'regParam': 0.1, 'numFeatures': 100, 'maxIter': 20}    accuracy:
0.617925      avg: 0.617925
params: {'regParam': 0.1, 'numFeatures': 100, 'maxIter': 30}    accuracy:
0.617335      avg: 0.617335
params: {'regParam': 0.01, 'numFeatures': 100, 'maxIter': 10}   accuracy:
0.617925      avg: 0.617925
params: {'regParam': 0.01, 'numFeatures': 100, 'maxIter': 20}   accuracy:
0.621462      avg: 0.621462
params: {'regParam': 0.01, 'numFeatures': 100, 'maxIter': 30}   accuracy:
0.620283      avg: 0.620283
params: {'regParam': 0.001, 'numFeatures': 100, 'maxIter': 10}  accuracy:
0.613797      avg: 0.613797
params: {'regParam': 0.001, 'numFeatures': 100, 'maxIter': 20}  accuracy:
0.611439      avg: 0.611439
params: {'regParam': 0.001, 'numFeatures': 100, 'maxIter': 30}  accuracy:
0.612028      avg: 0.612028
params: {'regParam': 0.1, 'numFeatures': 1000, 'maxIter': 10}   accuracy:
0.709906      avg: 0.709906
params: {'regParam': 0.1, 'numFeatures': 1000, 'maxIter': 20}   accuracy:
0.715802      avg: 0.715802
params: {'regParam': 0.1, 'numFeatures': 1000, 'maxIter': 30}   accuracy:
0.715212      avg: 0.715212
params: {'regParam': 0.01, 'numFeatures': 1000, 'maxIter': 10}  accuracy:
0.708137      avg: 0.708137
params: {'regParam': 0.01, 'numFeatures': 1000, 'maxIter': 20}  accuracy:
0.706368      avg: 0.706368
params: {'regParam': 0.01, 'numFeatures': 1000, 'maxIter': 30}  accuracy:
0.713443      avg: 0.713443
params: {'regParam': 0.001, 'numFeatures': 1000, 'maxIter': 10} accuracy:
0.709906      avg: 0.709906
params: {'regParam': 0.001, 'numFeatures': 1000, 'maxIter': 20} accuracy:
0.702241      avg: 0.702241
params: {'regParam': 0.001, 'numFeatures': 1000, 'maxIter': 30} accuracy:
0.696344      avg: 0.696344
params: {'regParam': 0.1, 'numFeatures': 1500, 'maxIter': 10}   accuracy:
0.705778      avg: 0.705778
params: {'regParam': 0.1, 'numFeatures': 1500, 'maxIter': 20}   accuracy:
0.706368      avg: 0.706368
params: {'regParam': 0.1, 'numFeatures': 1500, 'maxIter': 30}   accuracy:
```

```
0.705189      avg: 0.705189
params: {'regParam': 0.01, 'numFeatures': 1500, 'maxIter': 10}  accuracy:
0.698113      avg: 0.698113
params: {'regParam': 0.01, 'numFeatures': 1500, 'maxIter': 20}  accuracy:
0.702241      avg: 0.702241
params: {'regParam': 0.01, 'numFeatures': 1500, 'maxIter': 30}  accuracy:
0.708137      avg: 0.708137
params: {'regParam': 0.001, 'numFeatures': 1500, 'maxIter': 10} accuracy:
0.692217      avg: 0.692217
params: {'regParam': 0.001, 'numFeatures': 1500, 'maxIter': 20} accuracy:
0.698113      avg: 0.698113
params: {'regParam': 0.001, 'numFeatures': 1500, 'maxIter': 30} accuracy:
0.705778      avg: 0.705778
params: {'regParam': 0.1, 'numFeatures': 2000, 'maxIter': 10}   accuracy:
0.709316      avg: 0.709316
params: {'regParam': 0.1, 'numFeatures': 2000, 'maxIter': 20}   accuracy:
0.709906      avg: 0.709906
params: {'regParam': 0.1, 'numFeatures': 2000, 'maxIter': 30}   accuracy:
0.709906      avg: 0.709906
params: {'regParam': 0.01, 'numFeatures': 2000, 'maxIter': 10}  accuracy:
0.702241      avg: 0.702241
params: {'regParam': 0.01, 'numFeatures': 2000, 'maxIter': 20}  accuracy:
0.698703      avg: 0.698703
params: {'regParam': 0.01, 'numFeatures': 2000, 'maxIter': 30}  accuracy:
0.701061      avg: 0.701061
params: {'regParam': 0.001, 'numFeatures': 2000, 'maxIter': 10} accuracy:
0.699292      avg: 0.699292
params: {'regParam': 0.001, 'numFeatures': 2000, 'maxIter': 20} accuracy:
0.689269      avg: 0.689269
params: {'regParam': 0.001, 'numFeatures': 2000, 'maxIter': 30} accuracy:
0.695165      avg: 0.695165
params: {'regParam': 0.1, 'numFeatures': 3000, 'maxIter': 10}   accuracy:
0.705778      avg: 0.705778
params: {'regParam': 0.1, 'numFeatures': 3000, 'maxIter': 20}   accuracy:
0.706368      avg: 0.706368
params: {'regParam': 0.1, 'numFeatures': 3000, 'maxIter': 30}   accuracy:
0.707547      avg: 0.707547
params: {'regParam': 0.01, 'numFeatures': 3000, 'maxIter': 10}  accuracy:
0.707547      avg: 0.707547
params: {'regParam': 0.01, 'numFeatures': 3000, 'maxIter': 20}  accuracy:
0.706368      avg: 0.706368
params: {'regParam': 0.01, 'numFeatures': 3000, 'maxIter': 30}  accuracy:
0.706368      avg: 0.706368
params: {'regParam': 0.001, 'numFeatures': 3000, 'maxIter': 10} accuracy:
0.701061      avg: 0.701061
params: {'regParam': 0.001, 'numFeatures': 3000, 'maxIter': 20} accuracy:
0.697524      avg: 0.697524
params: {'regParam': 0.001, 'numFeatures': 3000, 'maxIter': 30} accuracy:
```

```
0.695165      avg: 0.695165
Comparing models on fold 2
params: {'regParam': 0.1, 'numFeatures': 100, 'maxIter': 10}    accuracy:
0.612194      avg: 0.611227
params: {'regParam': 0.1, 'numFeatures': 100, 'maxIter': 20}    accuracy:
0.612194      avg: 0.615059
params: {'regParam': 0.1, 'numFeatures': 100, 'maxIter': 30}    accuracy:
0.613451      avg: 0.615393
params: {'regParam': 0.01, 'numFeatures': 100, 'maxIter': 10}   accuracy:
0.615965      avg: 0.616945
params: {'regParam': 0.01, 'numFeatures': 100, 'maxIter': 20}   accuracy:
0.623507      avg: 0.622485
params: {'regParam': 0.01, 'numFeatures': 100, 'maxIter': 30}   accuracy:
0.624136      avg: 0.622209
params: {'regParam': 0.001, 'numFeatures': 100, 'maxIter': 10}  accuracy:
0.620993      avg: 0.617395
params: {'regParam': 0.001, 'numFeatures': 100, 'maxIter': 20}  accuracy:
0.629793      avg: 0.620616
params: {'regParam': 0.001, 'numFeatures': 100, 'maxIter': 30}  accuracy:
0.630421      avg: 0.621225
params: {'regParam': 0.1, 'numFeatures': 1000, 'maxIter': 10}   accuracy:
0.729101      avg: 0.719503
params: {'regParam': 0.1, 'numFeatures': 1000, 'maxIter': 20}   accuracy:
0.730987      avg: 0.723394
params: {'regParam': 0.1, 'numFeatures': 1000, 'maxIter': 30}   accuracy:
0.731615      avg: 0.723414
params: {'regParam': 0.01, 'numFeatures': 1000, 'maxIter': 10}  accuracy:
0.721559      avg: 0.714848
params: {'regParam': 0.01, 'numFeatures': 1000, 'maxIter': 20}  accuracy:
0.714016      avg: 0.710192
params: {'regParam': 0.01, 'numFeatures': 1000, 'maxIter': 30}  accuracy:
0.716530      avg: 0.714987
params: {'regParam': 0.001, 'numFeatures': 1000, 'maxIter': 10} accuracy:
0.712759      avg: 0.711332
params: {'regParam': 0.001, 'numFeatures': 1000, 'maxIter': 20} accuracy:
0.697046      avg: 0.699643
params: {'regParam': 0.001, 'numFeatures': 1000, 'maxIter': 30} accuracy:
0.702703      avg: 0.699524
params: {'regParam': 0.1, 'numFeatures': 1500, 'maxIter': 10}   accuracy:
0.719673      avg: 0.712726
params: {'regParam': 0.1, 'numFeatures': 1500, 'maxIter': 20}   accuracy:
0.716530      avg: 0.711449
params: {'regParam': 0.1, 'numFeatures': 1500, 'maxIter': 30}   accuracy:
0.718416      avg: 0.711802
params: {'regParam': 0.01, 'numFeatures': 1500, 'maxIter': 10}  accuracy:
0.707731      avg: 0.702922
params: {'regParam': 0.01, 'numFeatures': 1500, 'maxIter': 20}  accuracy:
0.716530      avg: 0.709386
```

```
params: {'regParam': 0.01, 'numFeatures': 1500, 'maxIter': 30}  accuracy:
0.717788      avg: 0.712962
params: {'regParam': 0.001, 'numFeatures': 1500, 'maxIter': 10} accuracy:
0.697046      avg: 0.694631
params: {'regParam': 0.001, 'numFeatures': 1500, 'maxIter': 20} accuracy:
0.699560      avg: 0.698837
params: {'regParam': 0.001, 'numFeatures': 1500, 'maxIter': 30} accuracy:
0.711502      avg: 0.708640
params: {'regParam': 0.1, 'numFeatures': 2000, 'maxIter': 10}   accuracy:
0.714016      avg: 0.711666
params: {'regParam': 0.1, 'numFeatures': 2000, 'maxIter': 20}   accuracy:
0.717788      avg: 0.713847
params: {'regParam': 0.1, 'numFeatures': 2000, 'maxIter': 30}   accuracy:
0.717159      avg: 0.713532
params: {'regParam': 0.01, 'numFeatures': 2000, 'maxIter': 10}  accuracy:
0.714016      avg: 0.708128
params: {'regParam': 0.01, 'numFeatures': 2000, 'maxIter': 20}  accuracy:
0.711502      avg: 0.705103
params: {'regParam': 0.01, 'numFeatures': 2000, 'maxIter': 30}  accuracy:
0.714016      avg: 0.707539
params: {'regParam': 0.001, 'numFeatures': 2000, 'maxIter': 10} accuracy:
0.707731      avg: 0.703512
params: {'regParam': 0.001, 'numFeatures': 2000, 'maxIter': 20} accuracy:
0.707102      avg: 0.698186
params: {'regParam': 0.001, 'numFeatures': 2000, 'maxIter': 30} accuracy:
0.702074      avg: 0.698620
params: {'regParam': 0.1, 'numFeatures': 3000, 'maxIter': 10}   accuracy:
0.726587      avg: 0.716183
params: {'regParam': 0.1, 'numFeatures': 3000, 'maxIter': 20}   accuracy:
0.727844      avg: 0.717106
params: {'regParam': 0.1, 'numFeatures': 3000, 'maxIter': 30}   accuracy:
0.727844      avg: 0.717696
params: {'regParam': 0.01, 'numFeatures': 3000, 'maxIter': 10}  accuracy:
0.728473      avg: 0.718010
params: {'regParam': 0.01, 'numFeatures': 3000, 'maxIter': 20}  accuracy:
0.724073      avg: 0.715220
params: {'regParam': 0.01, 'numFeatures': 3000, 'maxIter': 30}  accuracy:
0.722187      avg: 0.714278
params: {'regParam': 0.001, 'numFeatures': 3000, 'maxIter': 10} accuracy:
0.709617      avg: 0.705339
params: {'regParam': 0.001, 'numFeatures': 3000, 'maxIter': 20} accuracy:
0.712759      avg: 0.705141
params: {'regParam': 0.001, 'numFeatures': 3000, 'maxIter': 30} accuracy:
0.722187      avg: 0.708676
Comparing models on fold 3
params: {'regParam': 0.1, 'numFeatures': 100, 'maxIter': 10}    accuracy:
0.631703      avg: 0.618052
params: {'regParam': 0.1, 'numFeatures': 100, 'maxIter': 20}    accuracy:
```

```
0.635239      avg: 0.621786
params: {'regParam': 0.1, 'numFeatures': 100, 'maxIter': 30}    accuracy:
0.635239      avg: 0.622008
params: {'regParam': 0.01, 'numFeatures': 100, 'maxIter': 10}   accuracy:
0.620507      avg: 0.618132
params: {'regParam': 0.01, 'numFeatures': 100, 'maxIter': 20}   accuracy:
0.631703      avg: 0.625557
params: {'regParam': 0.01, 'numFeatures': 100, 'maxIter': 30}   accuracy:
0.634060      avg: 0.626160
params: {'regParam': 0.001, 'numFeatures': 100, 'maxIter': 10}  accuracy:
0.610489      avg: 0.615093
params: {'regParam': 0.001, 'numFeatures': 100, 'maxIter': 20}  accuracy:
0.621685      avg: 0.620972
params: {'regParam': 0.001, 'numFeatures': 100, 'maxIter': 30}  accuracy:
0.626989      avg: 0.623146
params: {'regParam': 0.1, 'numFeatures': 1000, 'maxIter': 10}   accuracy:
0.714791      avg: 0.717933
params: {'regParam': 0.1, 'numFeatures': 1000, 'maxIter': 20}   accuracy:
0.716559      avg: 0.721116
params: {'regParam': 0.1, 'numFeatures': 1000, 'maxIter': 30}   accuracy:
0.716559      avg: 0.721129
params: {'regParam': 0.01, 'numFeatures': 1000, 'maxIter': 10}  accuracy:
0.697113      avg: 0.708936
params: {'regParam': 0.01, 'numFeatures': 1000, 'maxIter': 20}  accuracy:
0.703595      avg: 0.707993
params: {'regParam': 0.01, 'numFeatures': 1000, 'maxIter': 30}  accuracy:
0.705952      avg: 0.711975
params: {'regParam': 0.001, 'numFeatures': 1000, 'maxIter': 10} accuracy:
0.690631      avg: 0.704432
params: {'regParam': 0.001, 'numFeatures': 1000, 'maxIter': 20} accuracy:
0.688273      avg: 0.695853
params: {'regParam': 0.001, 'numFeatures': 1000, 'maxIter': 30} accuracy:
0.690631      avg: 0.696559
params: {'regParam': 0.1, 'numFeatures': 1500, 'maxIter': 10}   accuracy:
0.693577      avg: 0.706343
params: {'regParam': 0.1, 'numFeatures': 1500, 'maxIter': 20}   accuracy:
0.697113      avg: 0.706670
params: {'regParam': 0.1, 'numFeatures': 1500, 'maxIter': 30}   accuracy:
0.697113      avg: 0.706906
params: {'regParam': 0.01, 'numFeatures': 1500, 'maxIter': 10}  accuracy:
0.690041      avg: 0.698628
params: {'regParam': 0.01, 'numFeatures': 1500, 'maxIter': 20}  accuracy:
0.691809      avg: 0.703527
params: {'regParam': 0.01, 'numFeatures': 1500, 'maxIter': 30}  accuracy:
0.692398      avg: 0.706108
params: {'regParam': 0.001, 'numFeatures': 1500, 'maxIter': 10} accuracy:
0.682970      avg: 0.690744
params: {'regParam': 0.001, 'numFeatures': 1500, 'maxIter': 20} accuracy:
```

```
0.682381      avg: 0.693351
params: {'regParam': 0.001, 'numFeatures': 1500, 'maxIter': 30} accuracy:
0.684738      avg: 0.700673
params: {'regParam': 0.1, 'numFeatures': 2000, 'maxIter': 10}   accuracy:
0.705362      avg: 0.709565
params: {'regParam': 0.1, 'numFeatures': 2000, 'maxIter': 20}   accuracy:
0.705952      avg: 0.711215
params: {'regParam': 0.1, 'numFeatures': 2000, 'maxIter': 30}   accuracy:
0.706541      avg: 0.711202
params: {'regParam': 0.01, 'numFeatures': 2000, 'maxIter': 10}  accuracy:
0.696523      avg: 0.704260
params: {'regParam': 0.01, 'numFeatures': 2000, 'maxIter': 20}  accuracy:
0.697702      avg: 0.702636
params: {'regParam': 0.01, 'numFeatures': 2000, 'maxIter': 30}  accuracy:
0.701237      avg: 0.705438
params: {'regParam': 0.001, 'numFeatures': 2000, 'maxIter': 10} accuracy:
0.684738      avg: 0.697254
params: {'regParam': 0.001, 'numFeatures': 2000, 'maxIter': 20} accuracy:
0.695345      avg: 0.697239
params: {'regParam': 0.001, 'numFeatures': 2000, 'maxIter': 30} accuracy:
0.690041      avg: 0.695760
params: {'regParam': 0.1, 'numFeatures': 3000, 'maxIter': 10}   accuracy:
0.701827      avg: 0.711397
params: {'regParam': 0.1, 'numFeatures': 3000, 'maxIter': 20}   accuracy:
0.701827      avg: 0.712013
params: {'regParam': 0.1, 'numFeatures': 3000, 'maxIter': 30}   accuracy:
0.702416      avg: 0.712602
params: {'regParam': 0.01, 'numFeatures': 3000, 'maxIter': 10}  accuracy:
0.704773      avg: 0.713598
params: {'regParam': 0.01, 'numFeatures': 3000, 'maxIter': 20}  accuracy:
0.695934      avg: 0.708792
params: {'regParam': 0.01, 'numFeatures': 3000, 'maxIter': 30}  accuracy:
0.695934      avg: 0.708163
params: {'regParam': 0.001, 'numFeatures': 3000, 'maxIter': 10} accuracy:
0.698880      avg: 0.703186
params: {'regParam': 0.001, 'numFeatures': 3000, 'maxIter': 20} accuracy:
0.703595      avg: 0.704626
params: {'regParam': 0.001, 'numFeatures': 3000, 'maxIter': 30} accuracy:
0.698291      avg: 0.705214
Best model:
params: {'regParam': 0.1, 'numFeatures': 1000, 'maxIter': 30}   accuracy:
0.721129
```

3.8.2 Get Predictions from Best Model

[28]:
```
# Make predictions on test documents. cvModel uses the best model found␣
 ↪(lrModel).
crossval_lr_prediction = crossval_lr_model.transform(test)

crossval_lr_accuracy = evaluator_lr_cross.evaluate(crossval_lr_prediction)
print("Test Accuracy = %g " % (crossval_lr_accuracy))
print("Test Error = %g " % (1.0 - crossval_lr_accuracy))

#get the best model
best_lr = crossval_lr_model.bestModel
```

```
Test Accuracy = 0.736713
Test Error = 0.263287
```

3.8.3 Best Model Summary

[29]:
```
#show the sets of params to be evaluated
print crossval_lr_model.explainParams()

#show the stages of the model
print best_lr.stages

#show num of features of best model
print best_lr.stages[1].getNumFeatures()

print crossval_lr_model.bestModel.stages[0].explainParams()
```

```
estimator: estimator to be cross-validated (current: Pipeline_e714827feb8c)
estimatorParamMaps: estimator param maps (current:
[{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 100, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 100, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 100, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
```

```
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 100, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 100, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 100, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 100, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 100, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 100, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
```

```
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1500, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1500, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1500, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1500, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
```

features.'): 1500, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1500, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1500, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1500, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1500, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',

```
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 3000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 3000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 3000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.1},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 3000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 3000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 3000, Param(parent=u'LogisticRegression_71f931934eef',
```

```
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.01},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 3000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 10,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 3000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 20,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 3000, Param(parent=u'LogisticRegression_71f931934eef',
name='maxIter', doc='max number of iterations (>= 0).'): 30,
Param(parent=u'LogisticRegression_71f931934eef', name='regParam',
doc='regularization parameter (>= 0).'): 0.001}])
evaluator: evaluator used to select hyper-parameters that maximize the validator
metric (current: MulticlassClassificationEvaluator_61ba1cab3aa4)
seed: random seed. (default: -8706542896013686583)
[StringIndexer_cc2b72f174f4, HashingTF_e2ec5e790eca, LogisticRegressionModel:
uid = LogisticRegression_71f931934eef, numClasses = 7, numFeatures = 1000]
1000
handleInvalid: how to handle invalid data (unseen or NULL values) in features
and label column of string type. Options are 'skip' (filter out rows with
invalid data), error (throw an error), or 'keep' (put invalid data in a special
additional bucket, at index numLabels). (default: error)
inputCol: input column name. (current: category_id)
outputCol: output column name. (default: StringIndexer_cc2b72f174f4__output,
current: category_id_indexed)
stringOrderType: How to order labels of string column. The first label after
ordering is assigned an index of 0. Supported options: frequencyDesc,
frequencyAsc, alphabetDesc, alphabetAsc. (default: frequencyDesc)
```

3.8.4 Checking training summary on best model

[30]:
```
crossval_lr_model_trainingSummary = crossval_lr_model.bestModel.stages[-1].
 ↪summary

# Obtain the objective per iteration
objectiveHistory = crossval_lr_model_trainingSummary.objectiveHistory
print("objectiveHistory:")
for objective in objectiveHistory:
    print(objective)
```

```
# for multiclass, we can inspect metrics on a per-label basis
print("False positive rate by label:")
for i, rate in enumerate(crossval_lr_model_trainingSummary.
 ↪falsePositiveRateByLabel):
    print("label %d: %s" % (i, rate))

print("True positive rate by label:")
for i, rate in enumerate(crossval_lr_model_trainingSummary.
 ↪truePositiveRateByLabel):
    print("label %d: %s" % (i, rate))

print("Precision by label:")
for i, prec in enumerate(crossval_lr_model_trainingSummary.precisionByLabel):
    print("label %d: %s" % (i, prec))

print("Recall by label:")
for i, rec in enumerate(crossval_lr_model_trainingSummary.recallByLabel):
    print("label %d: %s" % (i, rec))

print("F-measure by label:")
for i, f in enumerate(crossval_lr_model_trainingSummary.fMeasureByLabel()):
    print("label %d: %s" % (i, f))

accuracy = crossval_lr_model_trainingSummary.accuracy
falsePositiveRate = crossval_lr_model_trainingSummary.weightedFalsePositiveRate
truePositiveRate = crossval_lr_model_trainingSummary.weightedTruePositiveRate
fMeasure = crossval_lr_model_trainingSummary.weightedFMeasure()
precision = crossval_lr_model_trainingSummary.weightedPrecision
recall = crossval_lr_model_trainingSummary.weightedRecall
print("Accuracy: %s\nFPR: %s\nTPR: %s\nF-measure: %s\nPrecision: %s\nRecall: %s"
      % (accuracy, falsePositiveRate, truePositiveRate, fMeasure, precision,␣
 ↪recall))
```

```
objectiveHistory:
1.598841274
0.928998326892
0.772294172088
0.762976547499
0.726487377031
0.721514151343
0.71650480228
0.712727945294
0.711422197322
0.711146692195
0.71099175979
0.710809495681
0.71062603517
```

```
0.710301185005
0.710157040228
0.709935082915
0.709854840973
0.70983839907
0.709826099087
0.709821732247
0.709817600177
0.709815083754
0.709809843448
0.709803904523
0.7097958271
0.70978334007
0.709777868864
0.709775761776
0.709774833304
0.709774580831
0.709774358937
False positive rate by label:
label 0: 0.0958447073097
label 1: 0.0555555555556
label 2: 0.0131481022079
label 3: 0.00867778031514
label 4: 0.00303819444444
label 5: 0.000815660685155
label 6: 0.000203128173878
True positive rate by label:
label 0: 0.895672791938
label 1: 0.788870703764
label 2: 0.970619097587
label 3: 0.902479338843
label 4: 0.837765957447
label 5: 0.675
label 6: 0.55737704918
Precision by label:
label 0: 0.827038861522
label 1: 0.821824381927
label 2: 0.945807770961
label 3: 0.934931506849
label 4: 0.957446808511
label 5: 0.931034482759
label 6: 0.971428571429
Recall by label:
label 0: 0.895672791938
label 1: 0.788870703764
label 2: 0.970619097587
label 3: 0.902479338843
label 4: 0.837765957447
```

```
label 5: 0.675
label 6: 0.55737704918
F-measure by label:
label 0: 0.859988616961
label 1: 0.805010438413
label 2: 0.958052822372
label 3: 0.918418839361
label 4: 0.893617021277
label 5: 0.782608695652
label 6: 0.708333333333
Accuracy: 0.872592295345
FPR: 0.0498754240368
TPR: 0.872592295345
F-measure: 0.871791396283
Precision: 0.874841938794
Recall: 0.872592295345
```

[77]:
```
# Accessing _java_obj shouldn't be necessary in Spark 2.3+
{x._java_obj.getOutputCol(): x.labels for x in best_lr.stages if isinstance(x,␣
 ↪StringIndexerModel)}
```

[77]:
```
{u'category_id_indexed': [u'1612',
  u'1599',
  u'1564',
  u'1561',
  u'1976',
  u'1773',
  u'1386']}
```

3.8.5 Save Logistinc Regression Best Model

[31]:
```
best_lr.write().overwrite().save("hdfs:///data/exercise/LogRegCrossvalModel")
```

3.9 Random Forest

[32]:
```
from pyspark.ml.classification import RandomForestClassifier
```

3.9.1 Define Simple Model Pipeline

[33]:
```
%%time

from pyspark.ml.feature import *
```

```
label_indexer_rf = StringIndexer(inputCol="category_id",␣
 ↪outputCol="category_id_index")
hashingTF_rf = pyspark.ml.feature.HashingTF(inputCol="name_preprocessed",␣
 ↪outputCol="features", numFeatures=1000)
rf = pyspark.ml.classification.RandomForestClassifier(numTrees=1000,␣
 ↪featuresCol='features', labelCol="category_id_index")

pipeline_rf = pyspark.ml.Pipeline(stages=[
    label_indexer_rf,
    hashingTF_rf,
    rf
])

rf_model = pipeline_rf.fit(training)
```

```
CPU times: user 29.6 ms, sys: 21.8 ms, total: 51.4 ms
Wall time: 21.9 s
```

[34]:
```
rf_predictions = rf_model.transform(test)

# Select example rows to display.
# rf_predictions.select("predictedLabel", "label", "features").show(5)

# Select (prediction, true label) and compute test error
rf_evaluator = pyspark.ml.evaluation.
 ↪MulticlassClassificationEvaluator(labelCol="category_id_index",␣
 ↪predictionCol="prediction", metricName="accuracy")
rf_accuracy = rf_evaluator.evaluate(rf_predictions)
print("Test Accuracy = %g" % (rf_accuracy))
print("Test Error = %g" % (1.0 - rf_accuracy))
```

```
Test Accuracy = 0.522486
Test Error = 0.477514
```

3.10 CrossValidation for Random Forest

[37]:
```
label_indexer_rf = StringIndexer(inputCol="category_id",␣
 ↪outputCol="category_id_index")
hashingTF_rf_cross = pyspark.ml.feature.HashingTF(inputCol="name_preprocessed",␣
 ↪outputCol="features")
rf_cross = pyspark.ml.classification.
 ↪RandomForestClassifier(featuresCol='features', labelCol="category_id_index")

pipeline_rf_cross = pyspark.ml.Pipeline(stages=[
    label_indexer_rf,
    hashingTF_rf_cross,
```

```
    rf_cross
])

paramGrid_rf = ParamGridBuilder() \
    .addGrid(hashingTF_lr_cross.numFeatures, [1000, 1500, 2000]) \
    .addGrid(rf_cross.numTrees, [1000, 1500]) \
    .build()

evaluator_rf_cross = pyspark.ml.evaluation.MulticlassClassificationEvaluator(
    labelCol="category_id_index",
    predictionCol="prediction",
    metricName="accuracy"
)

crossval_rf = CrossValidatorVerbose(
    estimator=pipeline_rf_cross,
    estimatorParamMaps=paramGrid_rf,
    evaluator=evaluator_rf_cross,
    numFolds=3
)
```

3.11 Run Random Forest CrossValidation

[38]:
```
# Run cross-validation, and choose the best set of parameters.
crossval_rf_model = crossval_rf.fit(training)
```

```
Comparing models on fold 1
params: {'numTrees': 1000, 'numFeatures': 1000} accuracy: 0.341981      avg:
0.341981
params: {'numTrees': 1500, 'numFeatures': 1000} accuracy: 0.341981      avg:
0.341981
params: {'numTrees': 1000, 'numFeatures': 1500} accuracy: 0.341981      avg:
0.341981
params: {'numTrees': 1500, 'numFeatures': 1500} accuracy: 0.341981      avg:
0.341981
params: {'numTrees': 1000, 'numFeatures': 2000} accuracy: 0.341981      avg:
0.341981
params: {'numTrees': 1500, 'numFeatures': 2000} accuracy: 0.341981      avg:
0.341981
Comparing models on fold 2
params: {'numTrees': 1000, 'numFeatures': 1000} accuracy: 0.339409      avg:
0.340695
params: {'numTrees': 1500, 'numFeatures': 1000} accuracy: 0.339409      avg:
0.340695
params: {'numTrees': 1000, 'numFeatures': 1500} accuracy: 0.339409      avg:
0.340695
```

```
params: {'numTrees': 1500, 'numFeatures': 1500} accuracy: 0.339409      avg:
0.340695
params: {'numTrees': 1000, 'numFeatures': 2000} accuracy: 0.339409      avg:
0.340695
params: {'numTrees': 1500, 'numFeatures': 2000} accuracy: 0.339409      avg:
0.340695
Comparing models on fold 3
params: {'numTrees': 1000, 'numFeatures': 1000} accuracy: 0.334119      avg:
0.338503
params: {'numTrees': 1500, 'numFeatures': 1000} accuracy: 0.334119      avg:
0.338503
params: {'numTrees': 1000, 'numFeatures': 1500} accuracy: 0.334119      avg:
0.338503
params: {'numTrees': 1500, 'numFeatures': 1500} accuracy: 0.334119      avg:
0.338503
params: {'numTrees': 1000, 'numFeatures': 2000} accuracy: 0.334119      avg:
0.338503
params: {'numTrees': 1500, 'numFeatures': 2000} accuracy: 0.334119      avg:
0.338503
Best model:
params: {'numTrees': 1000, 'numFeatures': 1000} accuracy: 0.338503
```

3.12 Get Predictions from Best Model

[39]:
```
# Make predictions on test documents. cvModel uses the best model found␣
↪(lrModel).
crossval_rf_prediction = crossval_rf_model.transform(test)

crossval_rf_accuracy = evaluator_rf_cross.evaluate(crossval_rf_prediction)
print("Test Accuracy = %g " % (crossval_rf_accuracy))
print("Test Error = %g " % (1.0 - crossval_rf_accuracy))

#get the best model
best_rf = crossval_rf_model.bestModel
```

```
Test Accuracy = 0.345871
Test Error = 0.654129
```

3.13 Best Model Summary

[48]:
```
#show the sets of params to be evaluated
print crossval_rf_model.explainParams()

#show the stages of the model
print best_rf.stages
```

```
#show num of features of best model
# print best_rf.stages[1].getNumFeatures()

print crossval_rf_model.bestModel.stages[0].explainParams()
```

```
estimator: estimator to be cross-validated (current: Pipeline_4e153972edf6)
estimatorParamMaps: estimator param maps (current:
[{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'RandomForestClassifier_e42647b054a7',
name='numTrees', doc='Number of trees to train (>= 1).'): 1000},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1000, Param(parent=u'RandomForestClassifier_e42647b054a7',
name='numTrees', doc='Number of trees to train (>= 1).'): 1500},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1500, Param(parent=u'RandomForestClassifier_e42647b054a7',
name='numTrees', doc='Number of trees to train (>= 1).'): 1000},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 1500, Param(parent=u'RandomForestClassifier_e42647b054a7',
name='numTrees', doc='Number of trees to train (>= 1).'): 1500},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'RandomForestClassifier_e42647b054a7',
name='numTrees', doc='Number of trees to train (>= 1).'): 1000},
{Param(parent=u'HashingTF_e2ec5e790eca', name='numFeatures', doc='number of
features.'): 2000, Param(parent=u'RandomForestClassifier_e42647b054a7',
name='numTrees', doc='Number of trees to train (>= 1).'): 1500}])
evaluator: evaluator used to select hyper-parameters that maximize the validator
metric (current: MulticlassClassificationEvaluator_de8673419391)
seed: random seed. (default: -8706542896013686583)
[StringIndexer_824de986c9b2, HashingTF_44ff780b8d5f,
RandomForestClassificationModel (uid=RandomForestClassifier_e42647b054a7) with
1000 trees]
handleInvalid: how to handle invalid data (unseen or NULL values) in features
and label column of string type. Options are 'skip' (filter out rows with
invalid data), error (throw an error), or 'keep' (put invalid data in a special
additional bucket, at index numLabels). (default: error)
inputCol: input column name. (current: category_id)
outputCol: output column name. (default: StringIndexer_824de986c9b2__output,
current: category_id_index)
stringOrderType: How to order labels of string column. The first label after
ordering is assigned an index of 0. Supported options: frequencyDesc,
frequencyAsc, alphabetDesc, alphabetAsc. (default: frequencyDesc)
```

[50]:
```
# crossval_rf_model_trainingSummary = crossval_rf_model.bestModel.stages[-1].
 ↪summary

# # Obtain the objective per iteration
# objectiveHistory = crossval_rf_model_trainingSummary.objectiveHistory
```

```
# print("objectiveHistory:")
# for objective in objectiveHistory:
#     print(objective)

# # for multiclass, we can inspect metrics on a per-label basis
# print("False positive rate by label:")
# for i, rate in enumerate(crossval_rf_model_trainingSummary.
 ↪falsePositiveRateByLabel):
#     print("label %d: %s" % (i, rate))

# print("True positive rate by label:")
# for i, rate in enumerate(crossval_rf_model_trainingSummary.
 ↪truePositiveRateByLabel):
#     print("label %d: %s" % (i, rate))

# print("Precision by label:")
# for i, prec in enumerate(crossval_rf_model_trainingSummary.precisionByLabel):
#     print("label %d: %s" % (i, prec))

# print("Recall by label:")
# for i, rec in enumerate(crossval_rf_model_trainingSummary.recallByLabel):
#     print("label %d: %s" % (i, rec))

# print("F-measure by label:")
# for i, f in enumerate(crossval_rf_model_trainingSummary.fMeasureByLabel()):
#     print("label %d: %s" % (i, f))

# accuracy = crossval_rf_model_trainingSummary.accuracy
# falsePositiveRate = crossval_rf_model_trainingSummary.
 ↪weightedFalsePositiveRate
# truePositiveRate = crossval_rf_model_trainingSummary.weightedTruePositiveRate
# fMeasure = crossval_rf_model_trainingSummary.weightedFMeasure()
# precision = crossval_rf_model_trainingSummary.weightedPrecision
# recall = crossval_rf_model_trainingSummary.weightedRecall
# print("Accuracy: %s\nFPR: %s\nTPR: %s\nF-measure: %s\nPrecision: %s\nRecall:␣
 ↪%s"
#       % (accuracy, falsePositiveRate, truePositiveRate, fMeasure, precision,␣
 ↪recall))
```

3.13.1 Save Best Random Forest Model

[51]:
```
best_rf.write().overwrite().save("hdfs:///data/exercise/
 ↪RandomForestCrossvalModel")
```

3.14 OneVsRest Logistic Regression

3.14.1 The best logistic regression has these parameters

params: {'regParam': 0.01, 'numFeatures': 1000, 'maxIter': 30} accuracy: 0.716705

3.14.2 Define Simple Model Pipeline

[53]:
```
from pyspark.ml.feature import *

label_indexer_ovr_lr = StringIndexer(inputCol="category_id",␣
 ↪outputCol="category_id_index")
hashingTF_ovr_lr = pyspark.ml.feature.HashingTF(inputCol="name_preprocessed",␣
 ↪outputCol="features", numFeatures=1000)
lr = pyspark.ml.classification.LinearSVC(maxIter=30, regParam=0.01)
ovr_lr = pyspark.ml.classification.OneVsRest(classifier=lr,␣
 ↪featuresCol='features', labelCol="category_id_index")

pipeline_ovr_lr = pyspark.ml.Pipeline(stages=[
    label_indexer_ovr_lr,
    hashingTF_ovr_lr,
    ovr_lr
])

ovr_lr_model = pipeline_ovr_lr.fit(training)
```

3.14.3 Get Pipeline Stage

[54]:
```
pipeline_ovr_lr.getStages()
```

[54]:
```
[StringIndexer_fb1643014cb8, HashingTF_cd320ad4981d, OneVsRest_4c455b24aefc]
```

3.14.4 Model Parameters

3.14.5 Model evaluation

[63]:
```
prediction = ovr_lr_model.transform(test)

evaluator_lr = pyspark.ml.evaluation.
 ↪MulticlassClassificationEvaluator(labelCol="category_id_index",␣
 ↪predictionCol="prediction", metricName="accuracy")

accuracy = evaluator_lr.evaluate(prediction)
print("Accuracy = %g " % (accuracy))
```

```
print("Test Error = %g " % (1.0 - accuracy))
```

```
Accuracy = 0.736713
Test Error = 0.263287
```

3.14.6 Save OneVsRest with Best LogisticRegression Model

[62]:
```
pipeline_ovr_lr.write().overwrite().save("hdfs:///data/exercise/
↪OneVsRestLogRegModel")
```