ALS_Modeling

May 8, 2023

[1]:
```
# from google.colab import drive
# drive.mount('/content/drive')
```

```
Mounted at /content/drive
```

[1]:
```
!pip install kaggle
!mkdir ~/.kaggle
!cp kaggle.json ~/.kaggle/
!chmod 600 ~/.kaggle/kaggle.json
!kaggle datasets download -d yelp-dataset/yelp-dataset
!unzip yelp-dataset.zip
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Requirement already satisfied: kaggle in /usr/local/lib/python3.10/dist-packages
(1.5.13)
Requirement already satisfied: urllib3 in /usr/local/lib/python3.10/dist-
packages (from kaggle) (1.26.15)
Requirement already satisfied: python-slugify in /usr/local/lib/python3.10/dist-
packages (from kaggle) (8.0.1)
Requirement already satisfied: requests in /usr/local/lib/python3.10/dist-
packages (from kaggle) (2.27.1)
Requirement already satisfied: python-dateutil in
/usr/local/lib/python3.10/dist-packages (from kaggle) (2.8.2)
Requirement already satisfied: tqdm in /usr/local/lib/python3.10/dist-packages
(from kaggle) (4.65.0)
Requirement already satisfied: certifi in /usr/local/lib/python3.10/dist-
packages (from kaggle) (2022.12.7)
Requirement already satisfied: six>=1.10 in /usr/local/lib/python3.10/dist-
packages (from kaggle) (1.16.0)
Requirement already satisfied: text-unidecode>=1.3 in
/usr/local/lib/python3.10/dist-packages (from python-slugify->kaggle) (1.3)
Requirement already satisfied: charset-normalizer~=2.0.0 in
/usr/local/lib/python3.10/dist-packages (from requests->kaggle) (2.0.12)
Requirement already satisfied: idna<4,>=2.5 in /usr/local/lib/python3.10/dist-
packages (from requests->kaggle) (3.4)
Downloading yelp-dataset.zip to /content
100% 4.07G/4.07G [00:40<00:00, 101MB/s]
100% 4.07G/4.07G [00:40<00:00, 107MB/s]
```

```
Archive:  yelp-dataset.zip
  inflating: Dataset_User_Agreement.pdf
  inflating: yelp_academic_dataset_business.json
  inflating: yelp_academic_dataset_checkin.json
  inflating: yelp_academic_dataset_review.json
  inflating: yelp_academic_dataset_tip.json
  inflating: yelp_academic_dataset_user.json
```

[2]:
```
!pip install pyspark
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-
wheels/public/simple/
Collecting pyspark
  Downloading pyspark-3.4.0.tar.gz (310.8 MB)
                             310.8/310.8
MB 3.8 MB/s eta 0:00:00
  Preparing metadata (setup.py) … done
Requirement already satisfied: py4j==0.10.9.7 in /usr/local/lib/python3.10/dist-
packages (from pyspark) (0.10.9.7)
Building wheels for collected packages: pyspark
  Building wheel for pyspark (setup.py) … done
  Created wheel for pyspark: filename=pyspark-3.4.0-py2.py3-none-any.whl
size=311317145
sha256=3591afdce7b574640bcc1b20fe5550b0776486411105ec6d0d9a65b7fc39fdea
  Stored in directory: /root/.cache/pip/wheels/7b/1b/4b/3363a1d04368e7ff0d408e57
ff57966fcdf00583774e761327
Successfully built pyspark
Installing collected packages: pyspark
Successfully installed pyspark-3.4.0
```

1 Importing Libraries

[3]:
```
%matplotlib inline
from pyspark import SparkConf
from pyspark.sql import SparkSession
from pyspark.sql import SQLContext
from pyspark.sql.types import StructType, StructField, StringType, MapType
import pyspark.sql.functions as F
import matplotlib.pyplot as plt
import seaborn as sns
import pandas as pd
from pyspark.ml.feature import VectorAssembler
from pyspark.ml import Pipeline
# from pyspark.ml.feature import StringIndexer, OneHotEncoderEstimator
from pyspark.sql import Window
from pyspark.ml.evaluation import RegressionEvaluator
sns.set_theme(style="whitegrid", palette="pastel")
```

[4]:
```
from sklearn.metrics import confusion_matrix, classification_report,␣
↪accuracy_score
```

[5]:
```
# conf = SparkConf().set("spark.kryoserializer.buffer.max", "4g")
spark = SparkSession.builder.getOrCreate()
spark_context = spark.sparkContext
sqlContext = SQLContext(spark_context)
```

```
/usr/local/lib/python3.10/dist-packages/pyspark/sql/context.py:112:
FutureWarning: Deprecated in 3.0.0. Use SparkSession.builder.getOrCreate()
instead.
  warnings.warn(
```

2 Data Loading

2.1 Loading Businesses Dataset

[6]:
```
businesses = spark.read.json("yelp_academic_dataset_business.json")
```

2.2 Loading Reviews Dataset

[7]:
```
reviews =  spark.read.json("yelp_academic_dataset_review.json")
```

2.3 Loading Users Dataset

[8]:
```
users =  spark.read.json("yelp_academic_dataset_user.json")
```

3 Data Cleaning

3.1 Cleaning Businesses Dataset

Renaming and Dropping Columns

[9]:
```
#change name for starts to avoid duplicates
businesses=businesses.withColumnRenamed("stars", "Restaurant_stars")
businesses=businesses.withColumnRenamed("name", "Restaurant_name")
businesses=businesses.filter(F.col('categories').rlike('Restaurants'))
businesses = businesses.select('*','attributes.*','hours.*')
columns_to_drop = ['address','postal_code','review_count','attributes','hours']
businesses = businesses.drop(*columns_to_drop)
```

Renaming and Filtering Columns

[10]:
```
businesses = businesses.withColumn('categories', F.regexp_replace(F.
 ↪col("categories"), "(,?\ ?Restaurants,?)", ""))
businesses = businesses.withColumn('categories', F.regexp_replace(F.
 ↪col("categories"), "( ?)", ""))
businesses = businesses.filter(F.col("is_open").contains("1"))
```

[11]:
```
businesses = businesses.filter(F.col('state')=='PA')
```

3.2 Cleaning Reviews Dataset

Dropping Unrequired Columns

[12]:
```
columns_to_drop = ['cool','funny','average_stars']
reviews = reviews.drop(*columns_to_drop)
```

3.3 Cleaning Users Dataset

Dropping and renaming columns

[13]:
```
columns_to_drop =␣
 ↪['elite','useful','yelping_since','review_count','average_stars']
users = users.drop(*columns_to_drop)
users=users.withColumnRenamed("name", "user_name")
```

4 Data Transformation and Merging

Converting Ids(uuid/hex) to int

[14]:
```
w = Window().orderBy('business_id')
businesses= businesses.withColumn("business_id_int", F.row_number().over(w))
w = Window().orderBy('user_id')
users= users.withColumn("user_id_int", F.row_number().over(w))
```

[15]:
```
#joining three tables into one table in case need for future
df = reviews.join(businesses,on ='business_id', how = 'inner')
df = df.join(users,on ='user_id', how = 'inner')
```

[16]:
```
# df.show(5) # Takes too much time to execute, so commenting it for now
```

```
+--------------------+--------------------+--------------------+----------------
---+-----+--------------------+------+--------------------+-----------+-------
+------------+-------------+--------------------+---------------+-----+------
---------+----------+----------+--------------------+----+----------+-------
------------+----------+--------------------+-------------------------+----
--------------+----------------+------+---------+-------+------------------
+----------+---------+-------------+----------+--------------------+---------
--------+---------+-----+--------------------+----------+----------+----------
---+----------------+-------------------------+------------------+-----------
-----------+---------------------+---------------------+--------------------
---+------------------+-------+--------------------+-------+----------+-------+-
---------+---------+----------+----------+----------+---------------+----------
----+---------------+---------------+-------------+--------------+----------
----+---------------+----------------+----------------+------------------+-----
-----------+----+----+-------+-----+---------+-----------+
|             user_id|         business_id|               date|
review_id|stars|                text|useful|          categories|
city|is_open|     latitude|     longitude|
Restaurant_name|Restaurant_stars|state|AcceptsInsurance|AgesAllowed|    Alcohol|
Ambience|BYOB|BYOBCorkage|
BestNights|BikeParking|BusinessAcceptsBitcoin|BusinessAcceptsCreditCards|     Bu
sinessParking|ByAppointmentOnly|Caters|CoatCheck|Corkage|DietaryRestrictions|Dog
sAllowed|DriveThru|GoodForDancing|GoodForKids|
GoodForMeal|HairSpecializesIn|HappyHour|HasTV|               Music|NoiseLevel|Op
en24Hours|OutdoorSeating|RestaurantsAttire|RestaurantsCounterService|Restaurants
Delivery|RestaurantsGoodForGroups|RestaurantsPriceRange2|RestaurantsReservations
|RestaurantsTableService|RestaurantsTakeOut|Smoking|WheelchairAccessible|
WiFi|    Friday| Monday|  Saturday|   Sunday|  Thursday|   Tuesday| Wednesday|bu
siness_id_int|compliment_cool|compliment_cute|compliment_funny|compliment_hot|co
mpliment_list|compliment_more|compliment_note|compliment_photos|compliment_plain
|compliment_profile|compliment_writer|cool|fans|friends|funny|user_name|user_id_
int|
+--------------------+--------------------+--------------------+----------------
---+-----+--------------------+------+--------------------+-----------+-------
+------------+-------------+--------------------+---------------+-----+------
---------+----------+----------+--------------------+----+----------+-------
------------+----------+--------------------+-------------------------+----
--------------+----------------+------+---------+-------+------------------
+----------+---------+-------------+----------+--------------------+---------
--------+---------+-----+--------------------+----------+----------+----------
---+----------------+-------------------------+------------------+-----------
-----------+---------------------+---------------------+--------------------
---+------------------+-------+--------------------+-------+----------+-------+-
---------+---------+----------+----------+----------+---------------+----------
----+---------------+---------------+-------------+--------------+----------
----+---------------+----------------+----------------+------------------+-----
-----------+----+----+-------+-----+---------+-----------+
|--UhENQdbuWEh0mU5…|K_s-9Wd6vXSfnxYFz…|2017-08-06
```

```
02:42:02|dghJt1TSuyFkmLddu…|  5.0|When im first arr…|     0|
SushiBars|Philadelphia|      1|    39.944413|   -75.1707392|Kei Sushi
Restaurant|             4.5|   PA|            null|       null|
u'none'|{'touristy': Fals…|True|       null|                null|      False|
False|                      True|{'garage': False,…|             null|  True|
null|   null|               null|      False|     null|          null|
False|{'dessert': None,…|             null|     null|False|
null| 'average'|       null|         False|        u'casual'|
null|              False|                     True|                     2|
True|                   True|              True|   null|                True|
u'no'|11:30-22:0|0:0-0:0|11:30-22:0|16:0-21:0|11:30-21:0|11:30-21:0|11:30-21:0|
2730|              0|              0|               0|             0|
0|              0|              0|                0|              0|
0|                0|   1|   0|   None|    1|    Sonny|        217|
|--cjT1ICjm_ajiwSK…|G9LZoNlCfRH941q87…|2017-08-05
13:41:10|2nDLXAISthMftjVgB…|  5.0|This place was so…|     0|Tex-
MexFood,Night…|Philadelphia|      1|39.9502217062|-75.1665529981|    Mission
Taqueria|             4.0|   PA|            null|
null|u'full_bar'|{u'divey': False,…|null|       null|                null|
True|                  null|                     True|{'garage': False,…|
null| False|     null|   null|               null|      False|     None|
null|       True|{'dessert': None,…|             null|     True| True|
null|   u'loud'|       null|          True|        u'casual'|
null|               True|                     True|                     2|
False|                  True|              True|   null|
False|u'free'| 11:0-23:0|0:0-0:0| 11:0-23:0|     null| 11:0-22:0| 11:0-22:0|
11:0-22:0|           2185|              0|              0|               0|
0|              0|              0|              0|               0|
0|                 0|                0|   1|   0|   None|    2|   Shayla|
302|
|--cjT1ICjm_ajiwSK…|9PZxjhTIU7OgPIzuG…|2017-08-05
13:34:35|FN8qUbNlb9ulfoGtn…|  4.0|The decor in this…|
0|Lounges,Bars,Nigh…|Philadelphia|      1|39.9497020026|-75.1617702842|
El Vez|             4.0|   PA|            null|       null|
'full_bar'|{'touristy': Fals…|null|       'no'|               null|
True|                  null|                     True|{'garage': False,…|
null| False|     True|  False|               null|      False|     None|
null|       True|{'dessert': False…|             null|     True|False|
null|   u'loud'|       null|          True|         'casual'|
null|               True|                     True|                     2|
True|                   True|              True|   null|                null|
u'no'| 16:0-0:30|0:0-0:0| 16:0-22:0|12:0-22:0| 12:0-22:0| 12:0-22:0| 12:0-22:0|
1318|              0|              0|               0|             0|
0|              0|              0|                0|              0|
0|                0|   1|   0|   None|    2|   Shayla|        302|
|-2G_aOeur5RTmI-vc…|Cj4SH7N9HKtPhG_wp…|2018-12-22
23:22:38|L9KxbORVgQkTSfqDz…|  5.0|Best restaurant i…|
1|Bars,NightlifeAme…|  Warrington|      1|   40.2638126|   -75.1305032|
```

```
TJ Smith's|             3.0|   PA|            null|
null|u'full_bar'|{'touristy': Fals…|null|       null|{'monday': True, …|
False|                 False|                    True|{'garage': False,…|
null| False|     null|   null|               null|       null|     null|
False|       True|{u'breakfast': Fa…|            null|     null| True|{'dj':
False, 'ba…|u'average'|       null|          True|        u'casual'|
null|               True|                    True|                     2|
False|                   null|              True|   null|
null|u'free'| 11:0-23:0|0:0-0:0| 11:0-23:0|12:0-22:0| 11:0-22:0| 11:0-22:0|
11:0-22:0|           1769|              0|              0|               0|
0|              0|              0|              0|               0|
0|                 0|                0|   0|   0|   None|    1|    Kelly|
1632|
|-2IwE8_2pTsIKlEjE…|T6fI_rbI6iq_k9uIl…|2018-07-29
22:46:36|ee2cqOC8Am1QbLNPV…|  5.0|Great place, flav…|     0|
TacosMexican|  Royersford|      1|    40.224723|    -75.513763|Frida's Mexican
Food|             4.5|   PA|            null|       null|    u'none'|{u'divey':
False,…|True|       null|               null|      False|
False|                     True|{'garage': False,…|            null|  True|
null|  False|               null|      False|     null|          null|
True|{'dessert': None,…|            null|     True|False|
null|   'quiet'|       null|         False|        u'casual'|
null|               True|                    True|                     2|
False|                   True|              True|   null|
True|u'free'| 10:0-21:0|0:0-0:0| 10:0-21:0|10:0-20:0| 11:0-21:0| 11:0-21:0|
11:0-21:0|           3765|              0|              0|               0|
0|              0|              0|              0|               0|
0|                 0|                0|   0|   0|   None|    0|     Juan|
1648|
+--------------------+--------------------+------------------+----------------
---+-----+--------------------+------+--------------------+-----------+-------
+-------------+-------------+------------------+---------------+-----+------
----------+-----------+-----------+--------------------+----+-----------+-------
------------+-----------+----------------------+--------------------------+----
---------------+-----------------+------+---------+-------+------------------
+-----------+---------+-------------+-----------+--------------------+---------
--------+---------+-----+--------------------+----------+-----------+-----------
---+-----------------+--------------------------+------------------+------------
-----------+----------------------+----------------------+--------------------
---+------------------+-------+--------------------+-------+----------+-------+-
---------+---------+----------+----------+----------+--------------+-----------
----+---------------+---------------+--------------+--------------+-----------
----+---------------+-----------------+----------------+------------------+-----
-----------+----+----+-------+-----+---------+-----------+
only showing top 5 rows
```

5 ASL Model

[17]:
```
# Importing ALS Library
from pyspark.ml.recommendation import ALS
```

Selecting Required Columns

[18]:
```
ratings = df.
 ↪select('user_id_int','business_id_int','Restaurant_name','user_name','stars')
```

[19]:
```
train_df, test_df = ratings.randomSplit([.8,.2],seed=1)
als = ALS(maxIter=10, regParam=0.3, userCol="user_id_int",␣
 ↪itemCol="business_id_int", ratingCol="stars",
          coldStartStrategy="drop", rank=10, nonnegative = True)
```

[51]:
```
model = als.fit(df)
predictions = model.transform(df)
```

[21]:
```
predictions.show(5)
```

```
+-----------+--------------+--------------------+---------+-----+----------+
|user_id_int|business_id_int|     Restaurant_name|user_name|stars|prediction|
+-----------+--------------+--------------------+---------+-----+----------+
|         44|           1070|Dragon & Phoenix …|     Bert|  4.0|  3.567086|
|         44|           7572|     JJ Thai Cuisine|     Bert|  5.0| 4.3269286|
|         44|           7639|        Vetri Cucina|     Bert|  4.0|  4.421284|
|         44|           7680|            El Limon|     Bert|  5.0|  4.299594|
|        127|           1378|The Farm and Fish…|  Michael|  1.0|0.97495496|
+-----------+--------------+--------------------+---------+-----+----------+
only showing top 5 rows
```

[22]:
```
evaluator = RegressionEvaluator(metricName='rmse', labelCol='stars')
rmse = evaluator.evaluate(predictions)
print("Root-mean-square error = " + str(rmse))
```

```
Root-mean-square error = 1.3014098322040726
```

[53]:
```
userRecs = model.recommendForAllUsers(3)
```

[24]:
```
userRecs.show(5)
```

```
+-----------+--------------------+
|user_id_int|     recommendations|
+-----------+--------------------+
|          5|[{643, 3.7245593}…|
|         19|[{554, 2.2800214}…|
|         31|[{554, 4.4678254}…|
|         41|[{3671, 1.139868}…|
|         44|[{554, 5.274683},…|
```

```
+-----------+--------------------+
only showing top 5 rows
```

[55]:
```
userRecs_DF = (userRecs
  .select("user_id_int", F.explode("recommendations")
  .alias("recommendation"))
  .select("user_id_int", "recommendation.*")
)
userRecs_DF2 = userRecs_DF.join(users.select('user_id_int','user_name'),␣
 ↪on='user_id_int', how ='inner').join(businesses.
 ↪select('business_id_int','Restaurant_name'), on='business_id_int', how␣
 ↪='inner')
userRecs_DF2_pd = userRecs_DF2.toPandas()
```

[26]:
```
userRecs_DF2_pd.sort_values(['user_name','rating'],ascending=[True,False]).
 ↪head(20)
```

[26]:

	business_id_int	user_id_int	rating	user_name	Restaurant_name
17	554	48	5.577898	Amber	Frog Commissary
16	1912	48	5.396627	Amber	Sunny Chang's Pizza & More
15	3671	48	5.381920	Amber	The Chilly Banana
5	554	19	2.280021	Andrew	Frog Commissary
4	4859	19	2.212214	Andrew	Chef Jeff’s Hot Meals To-Go
3	6083	19	2.191591	Andrew	Cherish Philly
8	554	31	4.467825	Anthony	
7	610	31	4.245423	Anthony	
6	3068	31	4.216107	Anthony	
14	554	44	5.274683	Bert	
13	610	44	4.991056	Bert	
12	3888	44	4.984607	Bert	
2	643	5	3.724559	Brian	
1	1912	5	3.657573	Brian	
0	3800	5	3.537057	Brian	
23	554	101	2.267021	Mattthew	
22	4859	101	2.217390	Mattthew	
21	8018	101	2.197032	Mattthew	
26	554	111	5.495425	Mike	
25	6755	111	5.408771	Mike	

```
8                Frog Commissary
7                Academic Bistro
6          Umi Sushi And Seafood
14               Frog Commissary
13               Academic Bistro
12                      Maat Zip
2                   Akiko Sushi
1     Sunny Chang's Pizza & More
0               Hong Kong Garden
23               Frog Commissary
22  Chef Jeff’s Hot Meals To-Go
21                    Dew's Deli
26               Frog Commissary
25                      Yogorino
```

[42]:
```
def get_business_id(user_id):
    result = df.filter(df.user_id == user_id).select("user_id_int").collect()
    if result:
        return result[0].user_id_int
    else:
        return None
```

6 Recommendations for Desirae

[57]:
```
userId = get_business_id("6O4seIFz_buDGYXCOIT03A")
userRecs_DF2_pd[userRecs_DF2_pd["user_id_int"] ==␣
 ↪userId][["user_name","Restaurant_name","rating"]]
```

[57]:

	user_name	Restaurant_name	rating
46803	Desirae	Steel Penny Cafe	5.619937
46804	Desirae	Otolith Sustainable Seafood	5.552305
46805	Desirae	The Chilly Banana	5.521576

7 Recommendations for Aaron

[58]:
```
userId = get_business_id("A3DrdXmkNb1I6x-lSbj96g")
userRecs_DF2_pd[userRecs_DF2_pd["user_id_int"] ==␣
 ↪userId][["user_name","Restaurant_name","rating"]]
```

[58]:

	user_name	Restaurant_name	rating
70068	Aaron	Sunny Chang's Pizza & More	1.044408
70069	Aaron	Steel Penny Cafe	1.038287
70070	Aaron	Academic Bistro	1.024865

8 Recommendations for Brett

[59]:
```
userId = get_business_id("pou3BbKsIozfH50rxmnMew")
userRecs_DF2_pd[userRecs_DF2_pd["user_id_int"] ==␣
 ↪userId][["user_name","Restaurant_name","rating"]]
```

[59]:

	user_name	Restaurant_name	rating
743433	Brett	Academic Bistro	4.982213
743434	Brett	Otolith Sustainable Seafood	4.952850
743435	Brett	El Primo Produce	4.901747

[ ]:

[ ]:

[ ]:

[57]:

[ ]:

[ ]:

[ ]:

[ ]: