Pyspark_Basic_and_Linear_Regression

September 2, 2021

[1]:
```
!pip install pyspark
```

```
Requirement already satisfied: pyspark in /usr/local/lib/python3.7/dist-packages
(3.1.2)
Requirement already satisfied: py4j==0.10.9 in /usr/local/lib/python3.7/dist-
packages (from pyspark) (0.10.9)
```

[2]:
```
from pyspark.sql import SparkSession
spark= SparkSession.builder.appName('Customers').getOrCreate()
```

[3]:
```
spark
```

[3]:
```
<pyspark.sql.session.SparkSession at 0x7f14dac69910>
```

[5]:
```
from pyspark.ml.regression import LinearRegression
```

[6]:
```
dataset=spark.read.csv("/content/Ecommerce_Customers.
 ↪csv",inferSchema=True,header=True)
```

[8]:
```
type(dataset)
```

[8]:
```
pyspark.sql.dataframe.DataFrame
```

[9]:
```
dataset.dtypes
```

[9]:
```
[('Email', 'string'),
 ('Address', 'string'),
 ('Avg Session Length', 'double'),
 ('Time on App', 'double'),
 ('Time on Website', 'double'),
 ('Length of Membership', 'double'),
 ('Yearly Amount Spent', 'double')]
```

[10]:
```
dataset.select(['Email','Address']).show()
```

```
+--------------------+--------------------+
|               Email|             Address|
+--------------------+--------------------+
|mstephenson@ferna...|835 Frank TunnelW...|
|   hduke@hotmail.com|4547 Archer Commo...|
|    pallen@yahoo.com|24645 Valerie Uni...|
|riverarebecca@gma...|1414 David Throug...|
|mstephens@davidso...|14023 Rodriguez P...|
|alvareznancy@luca...|645 Martha Park A...|
|katherine20@yahoo...|68388 Reyes Light...|
|  awatkins@yahoo.com|Unit 6538 Box 898...|
|vchurch@walter-ma...|860 Lee KeyWest D...|
|    bonnie69@lin.biz|PSC 2734, Box 525...|
|andrew06@peterson...|26104 Alexander G...|
|ryanwerner@freema...|Unit 2413 Box 034...|
|   knelson@gmail.com|6705 Miller Orcha...|
|wrightpeter@yahoo...|05302 Dunlap Ferr...|
|taylormason@gmail...|7773 Powell Sprin...|
| jstark@anderson.com|49558 Ramirez Roa...|
| wjennings@gmail.com|6362 Wilson Mount...|
|rebecca45@hale-ba...|8982 Burton RowWi...|
|alejandro75@hotma...|64475 Andre Club ...|
|samuel46@love-wes...|544 Alexander Hei...|
+--------------------+--------------------+
only showing top 20 rows
```

[11]:
```
dataset
```

[11]:
```
DataFrame[Email: string, Address: string, Avg Session Length: double, Time on
App: double, Time on Website: double, Length of Membership: double, Yearly
Amount Spent: double]
```

[12]:
```
dataset.show()
```

```
+--------------------+--------------------+------------------+-----------+------
---------+--------------------+-------------------+
|               Email|             Address|Avg Session Length|Time on App|Time
on Website|Length of Membership|Yearly Amount Spent|
+--------------------+--------------------+------------------+-----------+------
---------+--------------------+-------------------+
|mstephenson@ferna...|835 Frank TunnelW...|       34.49726773|12.65565115|
39.57766802|         4.082620633|         587.951054|
|   hduke@hotmail.com|4547 Archer Commo...|       31.92627203|11.10946073|
37.26895887|         2.664034182|        392.2049334|
|    pallen@yahoo.com|24645 Valerie Uni...|       33.00091476|11.33027806|
37.11059744|         4.104543202|        487.5475049|
|riverarebecca@gma...|1414 David Throug...|       34.30555663|13.71751367|
36.72128268|         3.120178783|         581.852344|
|mstephens@davidso...|14023 Rodriguez P...|       33.33067252|12.79518855|
37.5366533|         4.446308318|         599.406092|
|alvareznancy@luca...|645 Martha Park A...|       33.87103788|12.02692534|
34.47687763|         5.493507201|        637.1024479|
|katherine20@yahoo...|68388 Reyes Light...|        32.0215955|11.36634831|
36.68377615|         4.685017247|        521.5721748|
|  awatkins@yahoo.com|Unit 6538 Box 898...|       32.73914294|12.35195897|
37.37335886|         4.434273435|        549.9041461|
|vchurch@walter-ma...|860 Lee KeyWest D...|        33.9877729|13.38623528|
37.53449734|         3.273433578|         570.200409|
|    bonnie69@lin.biz|PSC 2734, Box 525...|       31.93654862|11.81412829|
37.14516822|         3.202806072|        427.1993849|
|andrew06@peterson...|26104 Alexander G...|       33.99257277|13.33897545|
37.22580613|         2.482607771|        492.6060127|
|ryanwerner@freema...|Unit 2413 Box 034...|       33.87936082|  11.584783|
37.08792607|         3.713209203|        522.3374046|
|   knelson@gmail.com|6705 Miller Orcha...|       29.53242897| 10.9612984|
37.42021558|         4.046423164|        408.6403511|
|wrightpeter@yahoo...|05302 Dunlap Ferr...|       33.19033404|12.95922609|
36.1446667|         3.918541839|        573.4158673|
|taylormason@gmail...|7773 Powell Sprin...|       32.38797585|13.14872569|
36.61995708|         2.494543647|        470.4527333|
| jstark@anderson.com|49558 Ramirez Roa...|       30.73772037|12.63660605|
36.21376309|         3.357846842|        461.7807422|
| wjennings@gmail.com|6362 Wilson Mount...|        32.1253869|11.73386169|
34.89409275|         3.136132716|        457.8476959|
|rebecca45@hale-ba...|8982 Burton RowWi...|       32.33889932|12.01319469|
38.38513659|         2.420806161|        407.7045475|
|alejandro75@hotma...|64475 Andre Club ...|       32.18781205|14.71538754|
38.24411459|         1.516575581|        452.3156755|
|samuel46@love-wes...|544 Alexander Hei...|       32.61785606|13.98959256|
37.1905038|          4.06454855|        605.0610388|
+--------------------+--------------------+------------------+-----------+------
---------+--------------------+-------------------+
only showing top 20 rows
```

[13]:
```
dataset.head(10)
```

[13]:
```
[Row(Email='mstephenson@fernandez.com', Address='835 Frank TunnelWrightmouth, MI
82180-9605', Avg Session Length=34.49726773, Time on App=12.65565115, Time on
Website=39.57766802, Length of Membership=4.082620633, Yearly Amount
Spent=587.951054),
 Row(Email='hduke@hotmail.com', Address='4547 Archer CommonDiazchester, CA
06566-8576', Avg Session Length=31.92627203, Time on App=11.10946073, Time on
Website=37.26895887, Length of Membership=2.664034182, Yearly Amount
Spent=392.2049334),
 Row(Email='pallen@yahoo.com', Address='24645 Valerie Unions Suite
582Cobbborough, DC 99414-7564', Avg Session Length=33.00091476, Time on
App=11.33027806, Time on Website=37.11059744, Length of Membership=4.104543202,
Yearly Amount Spent=487.5475049),
 Row(Email='riverarebecca@gmail.com', Address='1414 David ThroughwayPort Jason,
OH 22070-1220', Avg Session Length=34.30555663, Time on App=13.71751367, Time on
Website=36.72128268, Length of Membership=3.120178783, Yearly Amount
Spent=581.852344),
 Row(Email='mstephens@davidson-herman.com', Address='14023 Rodriguez PassagePort
Jacobville, PR 37242-1057', Avg Session Length=33.33067252, Time on
App=12.79518855, Time on Website=37.5366533, Length of Membership=4.446308318,
Yearly Amount Spent=599.406092),
 Row(Email='alvareznancy@lucas.biz', Address='645 Martha Park Apt.
611Jeffreychester, MN 67218-7250', Avg Session Length=33.87103788, Time on
App=12.02692534, Time on Website=34.47687763, Length of Membership=5.493507201,
Yearly Amount Spent=637.1024479),
 Row(Email='katherine20@yahoo.com', Address='68388 Reyes Lights Suite
692Josephbury, WV 92213-0247', Avg Session Length=32.0215955, Time on
App=11.36634831, Time on Website=36.68377615, Length of Membership=4.685017247,
Yearly Amount Spent=521.5721748),
 Row(Email='awatkins@yahoo.com', Address='Unit 6538 Box 8980DPO AP 09026-4941',
Avg Session Length=32.73914294, Time on App=12.35195897, Time on
Website=37.37335886, Length of Membership=4.434273435, Yearly Amount
Spent=549.9041461),
 Row(Email='vchurch@walter-martinez.com', Address='860 Lee KeyWest Debra, SD
97450-0495', Avg Session Length=33.9877729, Time on App=13.38623528, Time on
Website=37.53449734, Length of Membership=3.273433578, Yearly Amount
Spent=570.200409),
 Row(Email='bonnie69@lin.biz', Address='PSC 2734, Box 5255APO AA 98456-7482',
Avg Session Length=31.93654862, Time on App=11.81412829, Time on
Website=37.14516822, Length of Membership=3.202806072, Yearly Amount
Spent=427.1993849)]
```

[14]:

```
dataset.printSchema()
```

```
root
 |-- Email: string (nullable = true)
 |-- Address: string (nullable = true)
 |-- Avg Session Length: double (nullable = true)
 |-- Time on App: double (nullable = true)
 |-- Time on Website: double (nullable = true)
 |-- Length of Membership: double (nullable = true)
 |-- Yearly Amount Spent: double (nullable = true)
```

[15]:

```
from pyspark.ml.linalg import Vectors
from pyspark.ml.feature import VectorAssembler
```

[16]:

```
featureassembler=VectorAssembler(inputCols=["Avg Session Length","Time on␣
 ↪App","Time on Website","Length of Membership"],outputCol="Independent␣
 ↪Features")
```

[17]:

```
output=featureassembler.transform(dataset)
```

[18]:

```
output.show()
```

```
+--------------------+--------------------+------------------+-----------+------
---------+--------------------+-------------------+--------------------+
|               Email|             Address|Avg Session Length|Time on App|Time
on Website|Length of Membership|Yearly Amount Spent|Independent Features|
+--------------------+--------------------+------------------+-----------+------
---------+--------------------+-------------------+--------------------+
|mstephenson@ferna...|835 Frank TunnelW...|       34.49726773|12.65565115|
39.57766802|         4.082620633|         587.951054|[34.49726773,12.6...|
|   hduke@hotmail.com|4547 Archer Commo...|       31.92627203|11.10946073|
37.26895887|         2.664034182|        392.2049334|[31.92627203,11.1...|
|    pallen@yahoo.com|24645 Valerie Uni...|       33.00091476|11.33027806|
37.11059744|         4.104543202|        487.5475049|[33.00091476,11.3...|
|riverarebecca@gma...|1414 David Throug...|       34.30555663|13.71751367|
36.72128268|         3.120178783|         581.852344|[34.30555663,13.7...|
|mstephens@davidso...|14023 Rodriguez P...|       33.33067252|12.79518855|
37.5366533|         4.446308318|         599.406092|[33.33067252,12.7...|
|alvareznancy@luca...|645 Martha Park A...|       33.87103788|12.02692534|
34.47687763|         5.493507201|        637.1024479|[33.87103788,12.0...|
|katherine20@yahoo...|68388 Reyes Light...|        32.0215955|11.36634831|
36.68377615|         4.685017247|        521.5721748|[32.0215955,11.36...|
|  awatkins@yahoo.com|Unit 6538 Box 898...|       32.73914294|12.35195897|
37.37335886|         4.434273435|        549.9041461|[32.73914294,12.3...|
|vchurch@walter-ma...|860 Lee KeyWest D...|        33.9877729|13.38623528|
37.53449734|         3.273433578|         570.200409|[33.9877729,13.38...|
|    bonnie69@lin.biz|PSC 2734, Box 525...|       31.93654862|11.81412829|
37.14516822|         3.202806072|        427.1993849|[31.93654862,11.8...|
|andrew06@peterson...|26104 Alexander G...|       33.99257277|13.33897545|
37.22580613|         2.482607771|        492.6060127|[33.99257277,13.3...|
|ryanwerner@freema...|Unit 2413 Box 034...|       33.87936082|  11.584783|
37.08792607|         3.713209203|        522.3374046|[33.87936082,11.5...|
|   knelson@gmail.com|6705 Miller Orcha...|       29.53242897| 10.9612984|
37.42021558|         4.046423164|        408.6403511|[29.53242897,10.9...|
|wrightpeter@yahoo...|05302 Dunlap Ferr...|       33.19033404|12.95922609|
36.1446667|         3.918541839|        573.4158673|[33.19033404,12.9...|
|taylormason@gmail...|7773 Powell Sprin...|       32.38797585|13.14872569|
36.61995708|         2.494543647|        470.4527333|[32.38797585,13.1...|
| jstark@anderson.com|49558 Ramirez Roa...|       30.73772037|12.63660605|
36.21376309|         3.357846842|        461.7807422|[30.73772037,12.6...|
| wjennings@gmail.com|6362 Wilson Mount...|        32.1253869|11.73386169|
34.89409275|         3.136132716|        457.8476959|[32.1253869,11.73...|
|rebecca45@hale-ba...|8982 Burton RowWi...|       32.33889932|12.01319469|
38.38513659|         2.420806161|        407.7045475|[32.33889932,12.0...|
|alejandro75@hotma...|64475 Andre Club ...|       32.18781205|14.71538754|
38.24411459|         1.516575581|        452.3156755|[32.18781205,14.7...|
|samuel46@love-wes...|544 Alexander Hei...|       32.61785606|13.98959256|
37.1905038|          4.06454855|        605.0610388|[32.61785606,13.9...|
+--------------------+--------------------+------------------+-----------+------
---------+--------------------+-------------------+--------------------+
only showing top 20 rows
```

[19]:
```
output.select("Independent Features").show()
```

```
+--------------------+
|Independent Features|
+--------------------+
|[34.49726773,12.6...|
|[31.92627203,11.1...|
|[33.00091476,11.3...|
|[34.30555663,13.7...|
|[33.33067252,12.7...|
|[33.87103788,12.0...|
|[32.0215955,11.36...|
|[32.73914294,12.3...|
|[33.9877729,13.38...|
|[31.93654862,11.8...|
|[33.99257277,13.3...|
|[33.87936082,11.5...|
|[29.53242897,10.9...|
|[33.19033404,12.9...|
|[32.38797585,13.1...|
|[30.73772037,12.6...|
|[32.1253869,11.73...|
|[32.33889932,12.0...|
|[32.18781205,14.7...|
|[32.61785606,13.9...|
+--------------------+
only showing top 20 rows
```

[20]:
```
output.columns
```

[20]:
```
['Email',
 'Address',
 'Avg Session Length',
 'Time on App',
 'Time on Website',
 'Length of Membership',
 'Yearly Amount Spent',
 'Independent Features']
```

[21]:
```
finalized_data=output.select("Independent Features","Yearly Amount Spent")
```

[22]:
```
finalized_data.show()
```

```
+--------------------+-------------------+
|Independent Features|Yearly Amount Spent|
+--------------------+-------------------+
|[34.49726773,12.6...|         587.951054|
|[31.92627203,11.1...|        392.2049334|
|[33.00091476,11.3...|        487.5475049|
|[34.30555663,13.7...|         581.852344|
|[33.33067252,12.7...|         599.406092|
|[33.87103788,12.0...|        637.1024479|
|[32.0215955,11.36...|        521.5721748|
|[32.73914294,12.3...|        549.9041461|
|[33.9877729,13.38...|         570.200409|
|[31.93654862,11.8...|        427.1993849|
|[33.99257277,13.3...|        492.6060127|
|[33.87936082,11.5...|        522.3374046|
|[29.53242897,10.9...|        408.6403511|
|[33.19033404,12.9...|        573.4158673|
|[32.38797585,13.1...|        470.4527333|
|[30.73772037,12.6...|        461.7807422|
|[32.1253869,11.73...|        457.8476959|
|[32.33889932,12.0...|        407.7045475|
|[32.18781205,14.7...|        452.3156755|
|[32.61785606,13.9...|        605.0610388|
+--------------------+-------------------+
only showing top 20 rows
```

[23]:
```
train_data,test_data=finalized_data.randomSplit([0.75,0.25])
```

[24]:
```
train_data.count()
```

[24]:
```
371
```

[25]:
```
test_data.count()
```

[25]:
```
129
```

[26]:
```
regressor=LinearRegression(featuresCol='Independent Features', labelCol='Yearly Amount Spent')
regressor=regressor.fit(train_data)
```

[27]:
```
regressor.coefficients
```

[27]:
```
DenseVector([25.7461, 38.7417, 0.6197, 61.8607])
```

[28]:
```
regressor.intercept
```

[28]:
```
-1060.3177002417435
```

[29]:
```
pred_results=regressor.evaluate(test_data)
```

[30]:
```
pred_results.predictions.show(40)
```

```
+--------------------+-------------------+------------------+
|Independent Features|Yearly Amount Spent|        prediction|
+--------------------+-------------------+------------------+
|[30.73772037,12.6...|        461.7807422| 450.7820313354307|
|[30.97167564,11.7...|        494.6386098| 487.7478527969274|
|[31.06132516,12.3...|        487.5554581|493.52645424555953|
|[31.06621816,11.7...|        448.9332932|461.73355783107104|
|[31.12397435,12.3...|        486.9470538| 508.2472622160801|
|[31.12809005,13.2...|        557.2526867| 564.8989607807507|
|[31.1695068,13.97...|        427.3565308| 416.5622690271264|
|[31.28344748,12.7...|        591.7810894| 569.7292310946609|
|[31.3123496,11.68...|         463.591418| 444.8841309095378|
|[31.38958548,10.9...|        410.0696111|  409.482350316935|
|[31.57020083,13.3...|        545.9454921| 563.5906634180401|
|[31.6005122,12.22...|        479.1728515| 460.8528601619648|
|[31.60983957,12.7...|        444.5455497|426.89591906209057|
|[31.62536013,13.1...|        376.3369008| 380.2428684489839|
|[31.6739155,12.32...|        475.7250679| 502.1463897936337|
|[31.72420252,13.1...|        503.3878873|509.29468290470277|
|[31.73663569,10.7...|        496.9334463| 494.4691440932404|
|[31.81861657,11.2...|        446.4186734| 448.2676900870197|
|[31.86483255,13.4...|        439.8912805| 449.8090620297878|
|[31.93654862,11.8...|        427.1993849|440.76923509934136|
|[31.95630056,12.8...|        547.1259317| 565.1221877993498|
|[31.97648006,10.7...|         330.594446| 324.6782285180991|
|[32.01807401,10.0...|        357.7831107| 340.0715885085715|
|[32.03054972,12.6...|        594.2744834| 589.4118194928258|
|[32.06377462,10.7...|        378.3309069| 389.7227212508187|
|[32.07894758,12.7...|        357.8637186| 351.7244799748105|
|[32.09610899,10.8...|        375.3984554| 374.7877162231441|
|[32.11511907,11.9...|        350.0582002|341.93967506234753|
|[32.17550124,13.3...|        588.7126055| 577.5140167273655|
|[32.19249883,13.3...|         616.660286|  620.019300582682|
|[32.20465465,12.4...|         478.584286| 478.5699557810224|
|[32.22729914,13.7...|        613.5993234| 621.9200532049815|
|[32.24635,11.3055...|        327.3779526| 336.5272685471018|
|[32.25997327,14.1...|        571.2160048| 573.4501095539897|
|[32.27184828,13.4...|          511.97986| 507.1811966944365|
|[32.2917561,12.19...|        494.5518611| 499.6979308744485|
|[32.34279623,11.4...|        486.0834255|476.13430175671715|
|[32.38103459,12.4...|        532.7248055|  546.469223234534|
|[32.40173183,12.0...|        506.5473071|505.60822907734405|
|[32.4071483,13.80...|        662.9610878| 643.9841202307359|
+--------------------+-------------------+------------------+
only showing top 40 rows
```

[31]:

```
pred_results.meanAbsoluteError,pred_results.meanSquaredError
```

[31]:

```
(8.678809773032123, 114.7037925117276)
```