databricksLinear Regression with pyspark

```
file_location= "/FileStore/tables/tips-1.csv"
file_type="csv"

df =spark.read.csv(file_location,header=True,inferSchema=True)
df.show()
```

```
+----------+----+------+------+---+------+----+
|total_bill| tip|   sex|smoker|day|  time|size|
+----------+----+------+------+---+------+----+
|     16.99|1.01|Female|    No|Sun|Dinner|   2|
|     10.34|1.66|  Male|    No|Sun|Dinner|   3|
|     21.01| 3.5|  Male|    No|Sun|Dinner|   3|
|     23.68|3.31|  Male|    No|Sun|Dinner|   2|
|     24.59|3.61|Female|    No|Sun|Dinner|   4|
|     25.29|4.71|  Male|    No|Sun|Dinner|   4|
|      8.77| 2.0|  Male|    No|Sun|Dinner|   2|
|     26.88|3.12|  Male|    No|Sun|Dinner|   4|
|     15.04|1.96|  Male|    No|Sun|Dinner|   2|
|     14.78|3.23|  Male|    No|Sun|Dinner|   2|
|     10.27|1.71|  Male|    No|Sun|Dinner|   2|
|     35.26| 5.0|Female|    No|Sun|Dinner|   4|
|     15.42|1.57|  Male|    No|Sun|Dinner|   2|
|     18.43| 3.0|  Male|    No|Sun|Dinner|   4|
|     14.83|3.02|Female|    No|Sun|Dinner|   2|
|     21.58|3.92|  Male|    No|Sun|Dinner|   2|
|     10.33|1.67|Female|    No|Sun|Dinner|   3|
|     16.29|3.71|  Male|    No|Sun|Dinner|   3|
```

```
df.printSchema()
```

```
root
 |-- total_bill: double (nullable = true)
 |-- tip: double (nullable = true)
 |-- sex: string (nullable = true)
 |-- smoker: string (nullable = true)
 |-- day: string (nullable = true)
 |-- time: string (nullable = true)
 |-- size: integer (nullable = true)
```

```
df.columns
```

```
Out[4]: ['total_bill', 'tip', 'sex', 'smoker', 'day', 'time', 'size']
```

```
from pyspark.ml.feature import StringIndexer
```

```
indexer =StringIndexer(inputCol='sex',outputCol='sex_indexed')
df_r=indexer.fit(df).transform(df)
```

```
df_r.show()
```

```
+----------+----+------+------+---+------+----+-----------+
|total_bill| tip|   sex|smoker|day|  time|size|sex_indexed|
+----------+----+------+------+---+------+----+-----------+
|     16.99|1.01|Female|    No|Sun|Dinner|   2|        1.0|
|     10.34|1.66|  Male|    No|Sun|Dinner|   3|        0.0|
|     21.01| 3.5|  Male|    No|Sun|Dinner|   3|        0.0|
|     23.68|3.31|  Male|    No|Sun|Dinner|   2|        0.0|
|     24.59|3.61|Female|    No|Sun|Dinner|   4|        1.0|
|     25.29|4.71|  Male|    No|Sun|Dinner|   4|        0.0|
|      8.77| 2.0|  Male|    No|Sun|Dinner|   2|        0.0|
|     26.88|3.12|  Male|    No|Sun|Dinner|   4|        0.0|
|     15.04|1.96|  Male|    No|Sun|Dinner|   2|        0.0|
|     14.78|3.23|  Male|    No|Sun|Dinner|   2|        0.0|
|     10.27|1.71|  Male|    No|Sun|Dinner|   2|        0.0|
|     35.26| 5.0|Female|    No|Sun|Dinner|   4|        1.0|
|     15.42|1.57|  Male|    No|Sun|Dinner|   2|        0.0|
|     18.43| 3.0|  Male|    No|Sun|Dinner|   4|        0.0|
|     14.83|3.02|Female|    No|Sun|Dinner|   2|        1.0|
|     21.58|3.92|  Male|    No|Sun|Dinner|   2|        0.0|
|     10.33|1.67|Female|    No|Sun|Dinner|   3|        1.0|
|     16.29|3.71|  Male|    No|Sun|Dinner|   3|        0.0|
```

```
indexer=StringIndexer(inputCols=["sex","smoker","day","time"],outputCols=
["sex_indexed","smoker_indexed","day_indexed","time_indexed"])
df_r=indexer.fit(df).transform(df)
df_r.show()
```

```
+----------+----+------+------+---+------+----+-----------+--------------+--
---------+------------+
|total_bill| tip|   sex|smoker|day|  time|size|sex_indexed|smoker_indexed|da
y_indexed|time_indexed|
+----------+----+------+------+---+------+----+-----------+--------------+--
---------+------------+
|     16.99|1.01|Female|    No|Sun|Dinner|   2|        1.0|           0.0|
1.0|         0.0|
|     10.34|1.66|  Male|    No|Sun|Dinner|   3|        0.0|           0.0|
1.0|         0.0|
|     21.01| 3.5|  Male|    No|Sun|Dinner|   3|        0.0|           0.0|
1.0|         0.0|
|     23.68|3.31|  Male|    No|Sun|Dinner|   2|        0.0|           0.0|
1.0|         0.0|
|     24.59|3.61|Female|    No|Sun|Dinner|   4|        1.0|           0.0|
1.0|         0.0|
|     25.29|4.71|  Male|    No|Sun|Dinner|   4|        0.0|           0.0|
1.0|         0.0|
|      8.77| 2.0|  Male|    No|Sun|Dinner|   2|        0.0|           0.0|
1.0|         0.0|
```

```
##VectorAssembler
from pyspark.ml.feature import VectorAssembler
featureassembler=VectorAssembler(inputCols=
['tip','size','sex_indexed','smoker_indexed','time_indexed','day_indexed'],o
utputCol="Independent Features")
output=featureassembler.transform(df_r)
```

```
output.show()
```

```
+----------+----+------+------+---+------+----+-----------+--------------+--
---------+------------+--------------------+
|total_bill| tip|   sex|smoker|day|  time|size|sex_indexed|smoker_indexed|da
y_indexed|time_indexed|Independent Features|
+----------+----+------+------+---+------+----+-----------+--------------+--
---------+------------+--------------------+
|     16.99|1.01|Female|    No|Sun|Dinner|   2|        1.0|           0.0|
1.0|         0.0|[1.01,2.0,1.0,0.0...|
|     10.34|1.66|  Male|    No|Sun|Dinner|   3|        0.0|           0.0|
1.0|         0.0|[1.66,3.0,0.0,0.0...|
|     21.01| 3.5|  Male|    No|Sun|Dinner|   3|        0.0|           0.0|
1.0|         0.0|[3.5,3.0,0.0,0.0,...|
|     23.68|3.31|  Male|    No|Sun|Dinner|   2|        0.0|           0.0|
1.0|         0.0|[3.31,2.0,0.0,0.0...|
|     24.59|3.61|Female|    No|Sun|Dinner|   4|        1.0|           0.0|
1.0|         0.0|[3.61,4.0,1.0,0.0...|
|     25.29|4.71|  Male|    No|Sun|Dinner|   4|        0.0|           0.0|
1.0|         0.0|[4.71,4.0,0.0,0.0...|
|      8.77| 2.0|  Male|    No|Sun|Dinner|   2|        0.0|           0.0|
1.0|         0.0|[2.0,2.0,0.0,0.0,...|
|     26.88|3.12|  Male|    No|Sun|Dinner|   4|        0.0|           0.0|
```

```
output.select("Independent Features").show()
```

```
+--------------------+
|Independent Features|
+--------------------+
|[1.01,2.0,1.0,0.0...|
|[1.66,3.0,0.0,0.0...|
|[3.5,3.0,0.0,0.0,...|
|[3.31,2.0,0.0,0.0...|
|[3.61,4.0,1.0,0.0...|
|[4.71,4.0,0.0,0.0...|
|[2.0,2.0,0.0,0.0,...|
|[3.12,4.0,0.0,0.0...|
|[1.96,2.0,0.0,0.0...|
|[3.23,2.0,0.0,0.0...|
|[1.71,2.0,0.0,0.0...|
|[5.0,4.0,1.0,0.0,...|
|[1.57,2.0,0.0,0.0...|
|[3.0,4.0,0.0,0.0,...|
|[3.02,2.0,1.0,0.0...|
|[3.92,2.0,0.0,0.0...|
|[1.67,3.0,1.0,0.0...|
|[3.71,3.0,0.0,0.0...|
```

```
finalized_data.show()
```

```
+--------------------+----------+
|Independent Features|total_bill|
+--------------------+----------+
|[1.01,2.0,1.0,0.0...|     16.99|
|[1.66,3.0,0.0,0.0...|     10.34|
|[3.5,3.0,0.0,0.0,...|     21.01|
|[3.31,2.0,0.0,0.0...|     23.68|
|[3.61,4.0,1.0,0.0...|     24.59|
|[4.71,4.0,0.0,0.0...|     25.29|
|[2.0,2.0,0.0,0.0,...|      8.77|
|[3.12,4.0,0.0,0.0...|     26.88|
|[1.96,2.0,0.0,0.0...|     15.04|
|[3.23,2.0,0.0,0.0...|     14.78|
|[1.71,2.0,0.0,0.0...|     10.27|
|[5.0,4.0,1.0,0.0,...|     35.26|
|[1.57,2.0,0.0,0.0...|     15.42|
|[3.0,4.0,0.0,0.0,...|     18.43|
|[3.02,2.0,1.0,0.0...|     14.83|
|[3.92,2.0,0.0,0.0...|     21.58|
|[1.67,3.0,1.0,0.0...|     10.33|
|[3.71,3.0,0.0,0.0...|     16.29|
```

```
finalized_data= output.select("Independent Features","total_bill")
```

```
from pyspark.ml.regression import LinearRegression
##train test split
train_data,test_data=finalized_data.randomSplit([0.75,0.25])
regressor=LinearRegression(featuresCol='Independent
Features',labelCol='total_bill')
regressor=regressor.fit(train_data)
```

```
regressor.coefficients
```

```
Out[34]: DenseVector([2.5813, 3.839, -0.6162, 2.3309, -2.546, 0.3996])
```

```
regressor.intercept
```

```
Out[36]: 1.8855151295689612
```

```
pred_results = regressor.evaluate(test_data)
```

```
pred_results.predictions.show()
```

```
+--------------------+----------+------------------+
|Independent Features|total_bill|        prediction|
+--------------------+----------+------------------+
|(6,[0,1],[1.25,2.0])|     10.07| 12.79014430860621|
|(6,[0,1],[1.97,2.0])|     12.02| 14.64865912014842|
| (6,[0,1],[2.0,2.0])|     13.37| 14.72609723729601|
| (6,[0,1],[3.6,3.0])|     24.06|22.695150634111375|
|(6,[0,1],[6.73,4.0])|     48.27|34.613548005453936|
|[1.0,1.0,1.0,0.0,...|      7.25| 7.689592627284103|
|[1.0,2.0,1.0,1.0,...|      5.75|15.058279045180068|
|[1.25,2.0,1.0,0.0...|      8.51|10.427057023440192|
|[1.48,2.0,0.0,0.0...|      8.52|11.636962811053445|
|[1.5,2.0,0.0,0.0,...|     12.46|14.634205296158603|
|[1.5,2.0,1.0,0.0,...|     10.65| 11.07237466633679|
|[1.61,2.0,1.0,1.0...|     10.59|15.434110749191978|
|[1.63,2.0,1.0,0.0...|     11.87|11.407939840643023|
|[1.67,3.0,1.0,0.0...|     10.33|  17.4966659896865|
|[1.8,2.0,1.0,0.0,...|     12.43|11.846755837812712|
|[2.0,2.0,0.0,0.0,...|     13.81|15.125678352181275|
|[2.0,2.0,0.0,1.0,...|     17.89| 17.45660094314432|
|[2.0,2.0,1.0,1.0,...|     27.18|16.440806272110674|
```

```
Out[39]: (0.658005200052283, 4.1423052881176465, 29.32498865908518)
```