In [1]:
```
from pyspark.sql import SparkSession
spark=SparkSession.builder.appName('logisticregression').getOrCreate()
```

In [19]:
```
customer_data = spark.read.csv('gs://bigdatabucket30/customer_churn.csv',inferSch

customer_data.printSchema()
```

```
root
 |-- Names: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- Total_Purchase: double (nullable = true)
 |-- Account_Manager: integer (nullable = true)
 |-- Years: double (nullable = true)
 |-- Num_Sites: double (nullable = true)
 |-- Onboard_date: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Company: string (nullable = true)
 |-- Churn: integer (nullable = true)
```

In [20]:
```
customer_data.describe().show()
```

```
+-------+-------------+-----------------+-----------------+------------------+-
----------------+-----------------+------------------+--------------------+--
------------------+-------------------+
|summary|        Names|              Age|   Total_Purchase|   Account_Manager|
Years|        Num_Sites|       Onboard_date|            Location|
Company|              Churn|
+-------+-------------+-----------------+-----------------+------------------+-
----------------+-----------------+------------------+--------------------+--
------------------+-------------------+
|  count|          900|              900|              900|               900|
900|              900|               900|                 900|
900|               900|
|   mean|         null|41.81666666666667|10062.82403333334|0.4811111111111111|
5.27315555555555| 8.587777777777777|              null|                null|
null|0.16666666666666666|
| stddev|         null|6.127560416916251|2408.644531858096|0.4999208935073339|
1.274449013194616|1.7648355920350969|              null|                null|
null| 0.3728852122772358|
|    min|   Aaron King|             22.0|            100.0|                 0|
1.0|              3.0|2006-01-02 04:16:13|00103 Jeffrey Cre...|     Abbott-Tho
mpson|                  0|
|    max|Zachary Walsh|             65.0|         18026.01|                 1|
9.15|             14.0|2016-12-28 04:07:38|Unit 9800 Box 287...|Zuniga, Clark
and...|                  1|
+-------+-------------+-----------------+-----------------+------------------+-
----------------+-----------------+------------------+--------------------+--
------------------+-------------------+
```

In [21]:
```
customer_data.columns
```

Out[21]:
```
['Names',
 'Age',
 'Total_Purchase',
 'Account_Manager',
 'Years',
 'Num_Sites',
 'Onboard_date',
 'Location',
 'Company',
 'Churn']
```

In [22]:
```
from pyspark.ml.feature import VectorAssembler
vector_assembler = VectorAssembler(inputCols=['Age','Total_Purchase','Account_Mar
output = vector_assembler.transform(customer_data)

output
```

Out[22]:
```
DataFrame[Names: string, Age: double, Total_Purchase: double, Account_Manager:
int, Years: double, Num_Sites: double, Onboard_date: string, Location: string,
Company: string, Churn: int, features: vector]
```

In [23]:
```
originaldata = output.select('features','churn')
training_data,test_data = originaldata.randomSplit([0.8,0.2])
```

In [35]:
```
from pyspark.ml.classification import LogisticRegression
logistic_regression = LogisticRegression(labelCol='churn')
model = logistic_regression.fit(training_data)

training_summary = model.summary
training_summary.predictions.describe().show()
```

```
+-------+-------------------+-------------------+
|summary|              churn|         prediction|
+-------+-------------------+-------------------+
|  count|                708|                708|
|   mean| 0.1483050847457627|0.10310734463276836|
| stddev|0.35565340410241986| 0.3043140170821189|
|    min|                0.0|                0.0|
|    max|                1.0|                1.0|
+-------+-------------------+-------------------+
```

In [26]:
```
from pyspark.ml.evaluation import BinaryClassificationEvaluator
prediction = model.evaluate(test_data)

prediction.predictions.show()
```

```
+--------------------+-----+--------------------+--------------------+---------
-+
|            features|churn|       rawPrediction|         probability|predictio
n|
+--------------------+-----+--------------------+--------------------+---------
-+
|[25.0,9672.03,0.0...|    0|[4.49480858114781...|[0.98895650277304...|       0.
0|
|[28.0,9090.43,1.0...|    0|[1.69619749840099...|[0.84503745357781...|       0.
0|
|[28.0,11245.38,0....|    0|[3.49941128083266...|[0.97067101376560...|       0.
0|
|[29.0,8688.17,1.0...|    1|[2.79425225888153...|[0.94236443557162...|       0.
0|
|[29.0,9617.59,0.0...|    0|[4.27445197497661...|[0.98627142237888...|       0.
0|
|[30.0,10744.14,1....|    1|[1.74693264366352...|[0.85156550087731...|       0.
0|
|[31.0,8688.21,0.0...|    0|[6.49262917259758...|[0.99848772870366...|       0.
0|
|[31.0,10058.87,1....|    0|[4.32613202563539...|[0.98695387338568...|       0.
0|
|[32.0,8617.98,1.0...|    1|[1.21111437751044...|[0.77049606591240...|       0.
0|
|[32.0,10716.75,0....|    0|[4.28614191735786...|[0.98642880895794...|       0.
0|
|[32.0,11715.72,0....|    0|[3.37063219108148...|[0.96677400441422...|       0.
0|
|[33.0,7720.61,1.0...|    0|[1.95346301944155...|[0.87582375813060...|       0.
0|
|[33.0,8556.73,0.0...|    0|[3.77672922495252...|[0.97761509595302...|       0.
0|
|[33.0,10309.71,1....|    0|[6.40937793293958...|[0.99835665688589...|       0.
0|
|[33.0,10709.39,1....|    0|[6.11005907880787...|[0.99778449971115...|       0.
0|
|[33.0,12249.96,0....|    0|[5.49273610720445...|[0.99590031124971...|       0.
0|
|[34.0,6461.86,1.0...|    0|[4.39812733244304...|[0.98784910744526...|       0.
0|
|[34.0,7324.32,0.0...|    0|[1.13700037743060...|[0.75712847908829...|       0.
0|
|[34.0,7818.13,0.0...|    0|[3.69447835457799...|[0.97574262977386...|       0.
0|
|[34.0,9845.35,0.0...|    0|[5.50969890348254...|[0.99596898897955...|       0.
0|
+--------------------+-----+--------------------+--------------------+---------
-+
only showing top 20 rows
```

In [27]:
```
binaryCE = BinaryClassificationEvaluator(rawPredictionCol='prediction',labelCol='
evaluation = binaryCE.evaluate(prediction.predictions)
evaluation
```

Out[27]:
```
0.7675736961451247
```

In [32]:
```
new_logistic_regression = logistic_regression.fit(data)
new_customer_data = spark.read.csv('gs://bigdatabucket30/new_customers.csv',infer

new_customer_data.printSchema()
```

```
root
 |-- Names: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- Total_Purchase: double (nullable = true)
 |-- Account_Manager: integer (nullable = true)
 |-- Years: double (nullable = true)
 |-- Num_Sites: double (nullable = true)
 |-- Onboard_date: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Company: string (nullable = true)
```

In [33]:
```
new_test_data = vector_assembler.transform(new_customer_data)

new_test_data.printSchema()
```

```
root
 |-- Names: string (nullable = true)
 |-- Age: double (nullable = true)
 |-- Total_Purchase: double (nullable = true)
 |-- Account_Manager: integer (nullable = true)
 |-- Years: double (nullable = true)
 |-- Num_Sites: double (nullable = true)
 |-- Onboard_date: string (nullable = true)
 |-- Location: string (nullable = true)
 |-- Company: string (nullable = true)
 |-- features: vector (nullable = true)
```

In [36]:
```
final_output = new_logistic_regression.transform(new_test_data)

final_output.select('Company','prediction').show()
```

```
+----------------+----------+
|         Company|prediction|
+----------------+----------+
|        King Ltd|       0.0|
|   Cannon-Benson|       1.0|
|Barron-Robertson|       1.0|
|   Sexton-Golden|       1.0|
|        Wood LLC|       0.0|
|   Parks-Robbins|       1.0|
+----------------+----------+
```

In [ ]:
```

```