Initiate a Spark Session & Import Data

In [1]:
```
import findspark
findspark.init('/home/ubuntu/spark-2.4.5-bin-hadoop2.7')
import pyspark
from pyspark.sql import SparkSession
spark = SparkSession.builder.appName('lr_example').getOrCreate()
```

In [2]:
```
df = spark.read.csv('customer_churn.csv', inferSchema=True, header=True)
```

In [3]:
```
df.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: timestamp (nullable = true)
 |-- Location: string (nullable = true)
 |-- Company: string (nullable = true)
 |-- Churn: integer (nullable = true)
```

In [4]:
```
df.show()
```

```
+------------------+----+-------------+--------------+-----+---------+--
----------------+--------------------+--------------------+-----+
|             Names| Age|Total_Purchase|Account_Manager|Years|Num_Sites|
Onboard_date|            Location|             Company|Churn|
+------------------+----+-------------+--------------+-----+---------+--
----------------+--------------------+--------------------+-----+
|   Cameron Williams|42.0|       11066.8|             0| 7.22|      8.0|20
13-08-30 07:00:40|10265 Elizabeth M...|         Harvey LLC|    1|
|      Kevin Mueller|41.0|      11916.22|             0|  6.5|     11.0|20
13-08-13 00:38:46|6157 Frank Garden...|         Wilson PLC|    1|
|        Eric Lozano|38.0|      12884.75|             0| 6.67|     12.0|20
16-06-29 06:20:07|1331 Keith Court ...|Miller, Johnson a...|    1|
|      Phillip White|42.0|       8010.76|             0| 6.71|     10.0|20
14-04-22 12:43:12|13120 Daniel Moun...|           Smith Inc|    1|
|     Cynthia Norton|37.0|       9191.58|             0| 5.56|      9.0|20
16-01-19 15:31:15|765 Tricia Row Ka...|         Love-Jones|    1|
|   Jessica Williams|48.0|      10356.02|             0| 5.12|      8.0|20
09-03-03 23:13:37|6187 Olson Mounta...|        Kelly-Warren|    1|
|        Eric Butler|44.0|      11331.58|             1| 5.23|     11.0|20
16-12-05 03:35:43|4846 Savannah Roa...|   Reynolds-Sheppard|    1|
|      Zachary Walsh|32.0|       9885.12|             1| 6.92|      9.0|20
06-03-09 14:50:20|25271 Roy Express...|         Singh-Cole|    1|
|        Ashlee Carr|43.0|       14062.6|             1| 5.46|     11.0|20
11-09-29 05:47:23|3725 Caroline Str...|          Lopez PLC|    1|
|     Jennifer Lynch|40.0|       8066.94|             1| 7.11|     11.0|20
06-03-28 15:42:45|363 Sandra Lodge ...|       Reed-Martinez|    1|
|       Paula Harris|30.0|      11575.37|             1| 5.22|      8.0|20
16-11-13 13:13:01|Unit 8120 Box 916...|Briggs, Lamb and ...|    1|
|     Bruce Phillips|45.0|       8771.02|             1| 6.64|     11.0|20
15-05-28 12:14:03|Unit 1895 Box 094...|    Figueroa-Maynard|    1|
|       Craig Garner|45.0|       8988.67|             1| 4.84|     11.0|20
11-02-16 08:10:47|897 Kelley Overpa...|     Abbott-Thompson|    1|
|       Nicole Olson|40.0|       8283.32|             1|  5.1|     13.0|20
12-11-22 05:35:03|11488 Weaver Cape...|Smith, Kim and Ma...|    1|
|     Harold Griffin|41.0|       6569.87|             1|  4.3|     11.0|20
15-03-28 02:13:44|1774 Peter Row Ap...|Snyder, Lee and M...|    1|
|       James Wright|38.0|      10494.82|             1| 6.81|     12.0|20
15-07-22 08:38:40|45408 David Path ...|      Sanders-Pierce|    1|
|      Doris Wilkins|45.0|       8213.41|             1| 7.35|     11.0|20
06-09-03 06:13:55|28216 Wright Moun...|Andrews, Adams an...|    1|
|Katherine Carpenter|43.0|      11226.88|             0| 8.08|     12.0|20
06-10-22 04:42:38|Unit 4948 Box 481...|Morgan, Phillips ...|    1|
|     Lindsay Martin|53.0|       5515.09|             0| 6.85|      8.0|20
15-10-07 00:27:10|69203 Crosby Divi...|      Villanueva LLC|    1|
|        Kathy Curry|46.0|        8046.4|             1| 5.69|      8.0|20
14-11-06 23:47:14|9569 Caldwell Cre...|Berry, Orr and Ca...|    1|
+------------------+----+-------------+--------------+-----+---------+--
----------------+--------------------+--------------------+-----+
only showing top 20 rows
```

In [5]:
```
df.columns
```

Out[5]:
```
['Names',
 'Age',
 'Total_Purchase',
 'Account_Manager',
 'Years',
 'Num_Sites',
 'Onboard_date',
 'Location',
 'Company',
 'Churn']
```

In [6]:
```
df.describe('Age', 'Total_Purchase', 'Years', 'Num_Sites').show()
```

```
+-------+-----------------+-----------------+-----------------+-----------------+
|summary|              Age|   Total_Purchase|            Years|        Num_Sites|
+-------+-----------------+-----------------+-----------------+-----------------+
|  count|              900|              900|              900|              900|
|   mean|41.81666666666667|10062.82403333334| 5.27315555555555| 8.587777777777777|
| stddev|6.127560416916251|2408.644531858096|1.274449013194616|1.7648355920350969|
|    min|             22.0|            100.0|              1.0|              3.0|
|    max|             65.0|         18026.01|             9.15|             14.0|
+-------+-----------------+-----------------+-----------------+-----------------+
```

Feature Engineering

In [7]:
```
from pyspark.sql.functions import month, year
```

In [8]:
```
df = df.withColumn(colName='Onboard_year', col=year('Onboard_date'))
```

In [9]:
```
df.head(2)
```

Out[9]:
```
[Row(Names='Cameron Williams', Age=42.0, Total_Purchase=11066.8, Account_Ma
nager=0, Years=7.22, Num_Sites=8.0, Onboard_date=datetime.datetime(2013, 8,
30, 7, 0, 40), Location='10265 Elizabeth Mission Barkerburgh, AK 89518', Co
mpany='Harvey LLC', Churn=1, Onboard_year=2013),
 Row(Names='Kevin Mueller', Age=41.0, Total_Purchase=11916.22, Account_Mana
ger=0, Years=6.5, Num_Sites=11.0, Onboard_date=datetime.datetime(2013, 8, 1
3, 0, 38, 46), Location='6157 Frank Gardens Suite 019 Carloshaven, RI 1775
6', Company='Wilson PLC', Churn=1, Onboard_year=2013)]
```

In [10]:
```
df.select('Onboard_year').distinct().count()
```

Out[10]:
```
11
```

In [11]:
```
from pyspark.ml.feature import StringIndexer, OneHotEncoderEstimator, VectorA
```

In [12]:
```
stage1 = StringIndexer(inputCol='Onboard_year', outputCol='onboard_index')
```

In [13]:
```
stage2 = OneHotEncoderEstimator(inputCols=[stage1.getOutputCol()], outputCols
```

In [14]:
```
stage3 = VectorAssembler(inputCols=['Age', 'Total_Purchase', 'Account_Manager
                         outputCol='features')
```

In [15]:
```
### Building ML Pipeline
```

In [16]:
```
from pyspark.ml import Pipeline
from pyspark.ml.classification import LogisticRegression
```

In [17]:
```
regression = LogisticRegression(featuresCol='features', labelCol='Churn')
```

In [18]:
```
pipeline = Pipeline(stages=[stage1, stage2, stage3, regression])
```

In [19]:
```
test_data, train_data = df.randomSplit([0.7, 0.3])
```

In [20]:
```
model = pipeline.fit(train_data)
```

In [21]:
```
result = model.transform(test_data)
```

In [22]:
```
result.select('prediction', 'Churn').show()
```

```
+----------+-----+
|prediction|Churn|
+----------+-----+
|       0.0|    0|
|       0.0|    0|
|       0.0|    0|
|       1.0|    1|
|       0.0|    0|
|       0.0|    0|
|       0.0|    1|
|       1.0|    1|
|       0.0|    0|
|       0.0|    0|
|       0.0|    0|
|       0.0|    1|
|       1.0|    0|
|       0.0|    0|
|       0.0|    0|
|       0.0|    0|
|       0.0|    0|
|       0.0|    0|
|       0.0|    0|
|       0.0|    0|
+----------+-----+
only showing top 20 rows
```

In [23]:
```
from pyspark.ml.evaluation import BinaryClassificationEvaluator
```

In [24]:
```
evaluator = BinaryClassificationEvaluator(rawPredictionCol='prediction', labe
```

In [25]:
```
AUC = evaluator.evaluate(result)
```

In [26]:
```
AUC
```

Out[26]: 0.7799909204403586

In [ ]:
```

```