Predicting Flights Delays

Spark Version

In [2]:

```
spark.version
```

```
'2.4.7-amzn-0'
```

Load PySpark Libraries

In [3]:

```
from pyspark.sql.types import *
from pyspark.sql.functions import monotonically_increasing_id, col, expr,
when, concat, lit, isnan
from pyspark.ml.linalg import Vectors
from pyspark.ml.regression import GeneralizedLinearRegression
from pyspark.ml.classification import RandomForestClassifier,
LogisticRegression
from pyspark.ml.feature import VectorIndexer, VectorAssembler, StringIndexer,
OneHotEncoder
from pyspark.ml.evaluation import MulticlassClassificationEvaluator,
RegressionEvaluator, BinaryClassificationEvaluator
from pyspark.ml import Pipeline
```

Load Data

In [4]:

```
train_df  = spark.read.load("s3a://sparklab123/train_df.csv", "csv",
delimiter=",", inferSchema=True, header=True)
train_df.createOrReplaceTempView("train_df")
```

Looking at the datase

In [5]:

```
sqlDF = spark.sql("select * from train_df").show()
```

```
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+
|YEAR|MONTH|DAY_OF_MONTH|DAY_OF_WEEK|CARRIER|FL_NUM|ORIGIN|DEST|DEP_TIME|DEP_
DELAY|ARR_TIME|ARR_DELAY|CANCELLED|CANCELLATION_CODE|AIR_TIME|DISTANCE|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+
|2015|    1|           1|          4|     AA|     1|   JFK| LAX|     855|
-5|    1237|        7|        0|             null|     378|    2475|
|2015|    1|           2|          5|     AA|     1|   JFK| LAX|     850|
-10|    1211|      -19|        0|              null|     357|    2475|
|2015|    1|           3|          6|     AA|     1|   JFK| LAX|     853|
-7|    1151|      -39|        0|             null|     330|    2475|
|2015|    1|           4|          7|     AA|     1|   JFK| LAX|     853|
-7|    1218|      -12|        0|             null|     352|    2475|
```

```
|2015|    1|           5|          1|     AA|     1|   JFK| LAX|     853|
-7|    1222|       -8|        0|            null|     338|    2475|
|2015|    1|           6|          2|     AA|     1|   JFK| LAX|     856|
-4|    1300|       25|        0|            null|     335|    2475|
|2015|    1|           7|          3|     AA|     1|   JFK| LAX|     859|
-1|    1221|      -14|        0|            null|     341|    2475|
|2015|    1|           8|          4|     AA|     1|   JFK| LAX|     856|
-4|    1158|      -37|        0|            null|     333|    2475|
|2015|    1|           9|          5|     AA|     1|   JFK| LAX|     901|
1|    1241|        6|        0|              null|     353|    2475|
|2015|    1|          10|          6|     AA|     1|   JFK| LAX|     903|
3|    1235|        0|        0|              null|     345|    2475|
|2015|    1|          11|          7|     AA|     1|   JFK| LAX|     854|
-6|    1159|      -36|        0|            null|     337|    2475|
|2015|    1|          12|          1|     AA|     1|   JFK| LAX|     853|
-7|    1230|       -5|        0|            null|     354|    2475|
|2015|    1|          13|          2|     AA|     1|   JFK| LAX|     854|
-6|    1155|      -40|        0|            null|     336|    2475|
|2015|    1|          14|          3|     AA|     1|   JFK| LAX|     856|
-4|    1126|      -69|        0|            null|     302|    2475|
|2015|    1|          15|          4|     AA|     1|   JFK| LAX|     852|
-8|    1231|       -4|        0|            null|     320|    2475|
|2015|    1|          16|          5|     AA|     1|   JFK| LAX|     900|
0|    1206|      -29|        0|              null|     321|    2475|
|2015|    1|          17|          6|     AA|     1|   JFK| LAX|     852|
-8|    1157|      -38|        0|            null|     343|    2475|
|2015|    1|          18|          7|     AA|     1|   JFK| LAX|     916|
16|    1236|        1|        0|            null|     340|    2475|
|2015|    1|          19|          1|     AA|     1|   JFK| LAX|     854|
-6|    1218|      -17|        0|            null|     359|    2475|
|2015|    1|          20|          2|     AA|     1|   JFK| LAX|     853|
-7|    1203|      -32|        0|            null|     349|    2475|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+
only showing top 20 rows
```

Create a "WEEKEND" flag

In [6]:

```
train_df = train_df.withColumn("WEEKEND", when((train_df.DAY_OF_WEEK == 5) | (train_df.DAY_OF_WEEK == 6) | (train_df.DAY_OF_WEEK == 7), 1).otherwise(0))
train_df.show()
```

```
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+
|YEAR|MONTH|DAY_OF_MONTH|DAY_OF_WEEK|CARRIER|FL_NUM|ORIGIN|DEST|DEP_TIME|DEP_
DELAY|ARR_TIME|ARR_DELAY|CANCELLED|CANCELLATION_CODE|AIR_TIME|DISTANCE|WEEKEN
D|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+
|2015|    1|           1|          4|     AA|     1|   JFK| LAX|     855|
-5|    1237|        7|        0|            null|     378|    2475|      0|
|2015|    1|           2|          5|     AA|     1|   JFK| LAX|     850|
-10|    1211|      -19|        0|             null|     357|    2475|      1|
```

```
|2015|    1|           3|          6|     AA|     1|   JFK| LAX|     853|
-7|    1151|      -39|        0|            null|     330|    2475|      1|
|2015|    1|           4|          7|     AA|     1|   JFK| LAX|     853|
-7|    1218|      -12|        0|            null|     352|    2475|      1|
|2015|    1|           5|          1|     AA|     1|   JFK| LAX|     853|
-7|    1222|       -8|        0|            null|     338|    2475|      0|
|2015|    1|           6|          2|     AA|     1|   JFK| LAX|     856|
-4|    1300|       25|        0|            null|     335|    2475|      0|
|2015|    1|           7|          3|     AA|     1|   JFK| LAX|     859|
-1|    1221|      -14|        0|            null|     341|    2475|      0|
|2015|    1|           8|          4|     AA|     1|   JFK| LAX|     856|
-4|    1158|      -37|        0|            null|     333|    2475|      0|
|2015|    1|           9|          5|     AA|     1|   JFK| LAX|     901|
1|    1241|        6|        0|              null|     353|    2475|      1|
|2015|    1|          10|          6|     AA|     1|   JFK| LAX|     903|
3|    1235|        0|        0|              null|     345|    2475|      1|
|2015|    1|          11|          7|     AA|     1|   JFK| LAX|     854|
-6|    1159|      -36|        0|            null|     337|    2475|      1|
|2015|    1|          12|          1|     AA|     1|   JFK| LAX|     853|
-7|    1230|       -5|        0|            null|     354|    2475|      0|
|2015|    1|          13|          2|     AA|     1|   JFK| LAX|     854|
-6|    1155|      -40|        0|            null|     336|    2475|      0|
|2015|    1|          14|          3|     AA|     1|   JFK| LAX|     856|
-4|    1126|      -69|        0|            null|     302|    2475|      0|
|2015|    1|          15|          4|     AA|     1|   JFK| LAX|     852|
-8|    1231|       -4|        0|            null|     320|    2475|      0|
|2015|    1|          16|          5|     AA|     1|   JFK| LAX|     900|
0|    1206|      -29|        0|              null|     321|    2475|      1|
|2015|    1|          17|          6|     AA|     1|   JFK| LAX|     852|
-8|    1157|      -38|        0|            null|     343|    2475|      1|
|2015|    1|          18|          7|     AA|     1|   JFK| LAX|     916|
16|    1236|        1|        0|            null|     340|    2475|      1|
|2015|    1|          19|          1|     AA|     1|   JFK| LAX|     854|
-6|    1218|      -17|        0|            null|     359|    2475|      0|
|2015|    1|          20|          2|     AA|     1|   JFK| LAX|     853|
-7|    1203|      -32|        0|            null|     349|    2475|      0|
+----+-----+-----------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+
only showing top 20 rows
```

Extract Departure Hour

In [7]:

```
train_df = train_df.withColumn("DEP_HOUR",
(train_df.DEP_TIME/float(100)).cast('int') )
train_df.show(20)
```

```
+----+-----+-----------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+---------+
|YEAR|MONTH|DAY_OF_MONTH|DAY_OF_WEEK|CARRIER|FL_NUM|ORIGIN|DEST|DEP_TIME|DEP_
DELAY|ARR_TIME|ARR_DELAY|CANCELLED|CANCELLATION_CODE|AIR_TIME|DISTANCE|WEEKEN
D|DEP_HOUR|
```

```
+----+-----+-----------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+
|2015|    1|          1|          4|     AA|     1|   JFK| LAX|     855|
-5|    1237|        7|        0|             null|     378|    2475|      0|
8|
|2015|    1|          2|          5|     AA|     1|   JFK| LAX|     850|
-10|    1211|      -19|        0|              null|     357|    2475|      1|
8|
|2015|    1|          3|          6|     AA|     1|   JFK| LAX|     853|
-7|    1151|      -39|        0|             null|     330|    2475|      1|
8|
|2015|    1|          4|          7|     AA|     1|   JFK| LAX|     853|
-7|    1218|      -12|        0|             null|     352|    2475|      1|
8|
|2015|    1|          5|          1|     AA|     1|   JFK| LAX|     853|
-7|    1222|       -8|        0|             null|     338|    2475|      0|
8|
|2015|    1|          6|          2|     AA|     1|   JFK| LAX|     856|
-4|    1300|       25|        0|             null|     335|    2475|      0|
8|
|2015|    1|          7|          3|     AA|     1|   JFK| LAX|     859|
-1|    1221|      -14|        0|             null|     341|    2475|      0|
8|
|2015|    1|          8|          4|     AA|     1|   JFK| LAX|     856|
-4|    1158|      -37|        0|             null|     333|    2475|      0|
8|
|2015|    1|          9|          5|     AA|     1|   JFK| LAX|     901|
1|    1241|        6|        0|               null|     353|    2475|      1|
9|
|2015|    1|         10|          6|     AA|     1|   JFK| LAX|     903|
3|    1235|        0|        0|               null|     345|    2475|      1|
9|
|2015|    1|         11|          7|     AA|     1|   JFK| LAX|     854|
-6|    1159|      -36|        0|             null|     337|    2475|      1|
8|
|2015|    1|         12|          1|     AA|     1|   JFK| LAX|     853|
-7|    1230|       -5|        0|             null|     354|    2475|      0|
8|
|2015|    1|         13|          2|     AA|     1|   JFK| LAX|     854|
-6|    1155|      -40|        0|             null|     336|    2475|      0|
8|
|2015|    1|         14|          3|     AA|     1|   JFK| LAX|     856|
-4|    1126|      -69|        0|             null|     302|    2475|      0|
8|
|2015|    1|         15|          4|     AA|     1|   JFK| LAX|     852|
-8|    1231|       -4|        0|             null|     320|    2475|      0|
8|
|2015|    1|         16|          5|     AA|     1|   JFK| LAX|     900|
0|    1206|      -29|        0|               null|     321|    2475|      1|
9|
|2015|    1|         17|          6|     AA|     1|   JFK| LAX|     852|
-8|    1157|      -38|        0|             null|     343|    2475|      1|
8|
|2015|    1|         18|          7|     AA|     1|   JFK| LAX|     916|
16|    1236|        1|        0|             null|     340|    2475|      1|
9|
```

```
|2015|    1|          19|          1|     AA|     1|   JFK| LAX|     854|
-6|    1218|      -17|        0|            null|     359|    2475|      0|
8|
|2015|    1|          20|          2|     AA|     1|   JFK| LAX|     853|
-7|    1203|      -32|        0|            null|     349|    2475|      0|
8|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+---------+
only showing top 20 rows
```

Create DELAY_LABELED

In [8]:

```
# DELAY_LABELED: Has a value of 1 if the arrival delay (ARR_DELAY) is greater
than 15 minutes and 0 if ARR_DELAY is less than or equal to 15 minutes.

train_df = train_df.withColumn("DELAY_LABELED", when( (train_df.ARR_DELAY >
15), 1).otherwise(0))
train_df.show(10)
```

```
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+---------+-------------+
|YEAR|MONTH|DAY_OF_MONTH|DAY_OF_WEEK|CARRIER|FL_NUM|ORIGIN|DEST|DEP_TIME|DEP_
DELAY|ARR_TIME|ARR_DELAY|CANCELLED|CANCELLATION_CODE|AIR_TIME|DISTANCE|WEEKEN
D|DEP_HOUR|DELAY_LABELED|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+---------+-------------+
|2015|    1|           1|          4|     AA|     1|   JFK| LAX|     855|
-5|    1237|        7|        0|            null|     378|    2475|      0|
8|           0|
|2015|    1|           2|          5|     AA|     1|   JFK| LAX|     850|
-10|    1211|      -19|        0|            null|     357|    2475|      1|
8|           0|
|2015|    1|           3|          6|     AA|     1|   JFK| LAX|     853|
-7|    1151|      -39|        0|            null|     330|    2475|      1|
8|           0|
|2015|    1|           4|          7|     AA|     1|   JFK| LAX|     853|
-7|    1218|      -12|        0|            null|     352|    2475|      1|
8|           0|
|2015|    1|           5|          1|     AA|     1|   JFK| LAX|     853|
-7|    1222|       -8|        0|            null|     338|    2475|      0|
8|           0|
|2015|    1|           6|          2|     AA|     1|   JFK| LAX|     856|
-4|    1300|       25|        0|            null|     335|    2475|      0|
8|           1|
|2015|    1|           7|          3|     AA|     1|   JFK| LAX|     859|
-1|    1221|      -14|        0|            null|     341|    2475|      0|
8|           0|
|2015|    1|           8|          4|     AA|     1|   JFK| LAX|     856|
-4|    1158|      -37|        0|            null|     333|    2475|      0|
8|           0|
```

```
|2015|    1|           9|          5|     AA|     1|   JFK| LAX|     901|
1|    1241|        6|        0|             null|     353|    2475|      1|
9|            0|
|2015|    1|          10|          6|     AA|     1|   JFK| LAX|     903|
3|    1235|        0|        0|             null|     345|    2475|      1|
9|            0|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+-------------+
only showing top 10 rows
```

Filter out Cancelled Flights

In [9]:

```
#We will keep only those flight records where it did not get cancelled.
#Keeping those records where the value of CANCELLED is equal to 0. (a value
of 1 == CANCELLED)

train_df = train_df.filter(train_df.CANCELLED == 0)
train_df.show(10)
```

```
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+-------------+
|YEAR|MONTH|DAY_OF_MONTH|DAY_OF_WEEK|CARRIER|FL_NUM|ORIGIN|DEST|DEP_TIME|DEP_
DELAY|ARR_TIME|ARR_DELAY|CANCELLED|CANCELLATION_CODE|AIR_TIME|DISTANCE|WEEKEN
D|DEP_HOUR|DELAY_LABELED|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+-------------+
|2015|    1|           1|          4|     AA|     1|   JFK| LAX|     855|
-5|    1237|        7|        0|             null|     378|    2475|      0|
8|            0|
|2015|    1|           2|          5|     AA|     1|   JFK| LAX|     850|
-10|    1211|      -19|        0|             null|     357|    2475|      1|
8|            0|
|2015|    1|           3|          6|     AA|     1|   JFK| LAX|     853|
-7|    1151|      -39|        0|             null|     330|    2475|      1|
8|            0|
|2015|    1|           4|          7|     AA|     1|   JFK| LAX|     853|
-7|    1218|      -12|        0|             null|     352|    2475|      1|
8|            0|
|2015|    1|           5|          1|     AA|     1|   JFK| LAX|     853|
-7|    1222|       -8|        0|             null|     338|    2475|      0|
8|            0|
|2015|    1|           6|          2|     AA|     1|   JFK| LAX|     856|
-4|    1300|       25|        0|             null|     335|    2475|      0|
8|            1|
|2015|    1|           7|          3|     AA|     1|   JFK| LAX|     859|
-1|    1221|      -14|        0|             null|     341|    2475|      0|
8|            0|
|2015|    1|           8|          4|     AA|     1|   JFK| LAX|     856|
-4|    1158|      -37|        0|             null|     333|    2475|      0|
8|            0|
```

```
|2015|    1|           9|          5|     AA|     1|   JFK| LAX|     901|
1|    1241|        6|        0|             null|     353|    2475|      1|
9|            0|
|2015|    1|          10|          6|     AA|     1|   JFK| LAX|     903|
3|    1235|        0|        0|             null|     345|    2475|      1|
9|            0|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+-------------+
only showing top 10 rows
```

Remove ARR_DELAY NAs

In [10]:

```
#There are lot of NA values in ARR_DELAY column. We should keep only those
where we have valid readings of ARR_DELAY (remove NAs)

print("Initial Train_DF Count: " + str(train_df.count()))
train_df = train_df.filter( train_df.ARR_DELAY != "NA" )
print("New Train_DF Count:     " + str(train_df.count()))
train_df.show(10)
```

```
Initial Train_DF Count: 1033485
New Train_DF Count:     1030606
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+-------------+
|YEAR|MONTH|DAY_OF_MONTH|DAY_OF_WEEK|CARRIER|FL_NUM|ORIGIN|DEST|DEP_TIME|DEP_
DELAY|ARR_TIME|ARR_DELAY|CANCELLED|CANCELLATION_CODE|AIR_TIME|DISTANCE|WEEKEN
D|DEP_HOUR|DELAY_LABELED|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+-------------+
|2015|    1|           1|          4|     AA|     1|   JFK| LAX|     855|
-5|    1237|        7|        0|             null|     378|    2475|      0|
8|            0|
|2015|    1|           2|          5|     AA|     1|   JFK| LAX|     850|
-10|    1211|      -19|        0|             null|     357|    2475|      1|
8|            0|
|2015|    1|           3|          6|     AA|     1|   JFK| LAX|     853|
-7|    1151|      -39|        0|             null|     330|    2475|      1|
8|            0|
|2015|    1|           4|          7|     AA|     1|   JFK| LAX|     853|
-7|    1218|      -12|        0|             null|     352|    2475|      1|
8|            0|
|2015|    1|           5|          1|     AA|     1|   JFK| LAX|     853|
-7|    1222|       -8|        0|             null|     338|    2475|      0|
8|            0|
|2015|    1|           6|          2|     AA|     1|   JFK| LAX|     856|
-4|    1300|       25|        0|             null|     335|    2475|      0|
8|            1|
|2015|    1|           7|          3|     AA|     1|   JFK| LAX|     859|
-1|    1221|      -14|        0|             null|     341|    2475|      0|
8|            0|
```

```
|2015|    1|           8|          4|     AA|     1|   JFK| LAX|     856|
-4|    1158|      -37|        0|            null|     333|    2475|      0|
8|            0|
|2015|    1|           9|          5|     AA|     1|   JFK| LAX|     901|
1|    1241|        6|        0|              null|     353|    2475|      1|
9|            0|
|2015|    1|          10|          6|     AA|     1|   JFK| LAX|     903|
3|    1235|        0|        0|              null|     345|    2475|      1|
9|            0|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+-------------+
only showing top 10 rows
```

What is the Schema? How many records?

In [11]:

```
print('\r\nTotal Records: ' + str(train_df.count()) + '\r\n\r\n')
for i in train_df.dtypes: print(i)
```

```
Total Records: 1030606


('YEAR', 'int')
('MONTH', 'int')
('DAY_OF_MONTH', 'int')
('DAY_OF_WEEK', 'int')
('CARRIER', 'string')
('FL_NUM', 'int')
('ORIGIN', 'string')
('DEST', 'string')
('DEP_TIME', 'string')
('DEP_DELAY', 'string')
('ARR_TIME', 'string')
('ARR_DELAY', 'string')
('CANCELLED', 'int')
('CANCELLATION_CODE', 'string')
('AIR_TIME', 'string')
('DISTANCE', 'int')
('WEEKEND', 'int')
('DEP_HOUR', 'int')
('DELAY_LABELED', 'int')
```

Convert from String to Int

In [12]:

```
#Convert the type of ARR_DELAY and DEP_DELAY from string to integer so that
we can perform mathematical operations on that.

train_df = train_df.withColumn("ARR_DELAY",     col("ARR_DELAY").cast("int"))
\
                   .withColumn("DEP_DELAY",     col("DEP_DELAY").cast("int"))
\
```

```
                    .withColumn("AIR_TIME",      col("AIR_TIME").cast("int"))
\
                    .withColumn("DAY_OF_WEEK",
col("DAY_OF_WEEK").cast("string"))

train_df.dtypes
```

```
[('YEAR', 'int'), ('MONTH', 'int'), ('DAY_OF_MONTH', 'int'), ('DAY_OF_WEEK',
'string'), ('CARRIER', 'string'), ('FL_NUM', 'int'), ('ORIGIN', 'string'), ('
DEST', 'string'), ('DEP_TIME', 'string'), ('DEP_DELAY', 'int'), ('ARR_TIME',
'string'), ('ARR_DELAY', 'int'), ('CANCELLED', 'int'), ('CANCELLATION_CODE',
'string'), ('AIR_TIME', 'int'), ('DISTANCE', 'int'), ('WEEKEND', 'int'), ('DE
P_HOUR', 'int'), ('DELAY_LABELED', 'int')]
```

Create Temporary Table/View to Query it with SQL

In [13]:

```
train_df.createOrReplaceTempView("train_df")
```

Exploratory Analysis

Create Delay DataFrame (% of on-time vs delayed flights)

In [14]:

```
#We create a new dataframe called delay which will have two columns,
DELAY_LABELED and the count of it.
#Basically it will have a count of delayed flights and ontime flights. We
will be using aggregate where we group the
#dataframe by DELAY_LABELED and calculating the count using n().

#Create the delay dataframe first
delay = train_df.groupBy( train_df.DELAY_LABELED ).count() \
                .withColumn("STATUS", when( col("DELAY_LABELED") == 0,
"ontime").otherwise("delayed") ) \
                .drop("DELAY_LABELED")

delay.show()


#To calculate the total, we use the collect () function, which returns all
the elements of the dataset as an array
total = delay.select("count").collect()[0][0] +
delay.select("count").collect()[1][0]

print("Total Count: " + str(total))

delay_r = delay.withColumn("PERCENTAGE", (col("count")/float(total)*100) )

delay_r.show()
```

```
+------+-------+
```

```
| count| STATUS|
+------+-------+
|187216|delayed|
|843390| ontime|
+------+-------+

Total Count: 1030606
+------+-------+-----------------+
| count| STATUS|       PERCENTAGE|
+------+-------+-----------------+
|187216|delayed|18.16562294417071|
|843390| ontime| 81.8343770558293|
+------+-------+-----------------+
```

Compare Delays vs. Non-Delays by Day of Week (1=Monday, 7=Sunday)

In [15]:

```
delay_flights_count = train_df.filter( train_df.DELAY_LABELED == 1 ) \
                              .groupby( train_df.DAY_OF_WEEK ).count() \
                              .withColumnRenamed( "count", "DELAY_COUNT")

non_delay_flights_count = train_df.filter( train_df.DELAY_LABELED == 0 ) \
                              .groupby( train_df.DAY_OF_WEEK ).count() \
                              .withColumnRenamed( "count", "NON_DELAY_COUNT")

dayofweek_count = delay_flights_count.join(non_delay_flights_count,
delay_flights_count["DAY_OF_WEEK"] == non_delay_flights_count["DAY_OF_WEEK"]
) \
                                     .drop(
non_delay_flights_count.DAY_OF_WEEK ) \
                                     .sort( col("DAY_OF_WEEK") )

dayofweek_count.createOrReplaceTempView("dayofweek_count")
dayofweek_count.show()
```

```
+-----------+-----------+---------------+
|DELAY_COUNT|DAY_OF_WEEK|NON_DELAY_COUNT|
+-----------+-----------+---------------+
|      31396|          1|         120871|
|      26323|          2|         120897|
|      24874|          3|         124656|
|      30381|          4|         127328|
|      29230|          5|         127520|
|      19393|          6|         103732|
|      25619|          7|         118386|
+-----------+-----------+---------------+
```

Display Descriptive Stats

In [16]:

```
train_df.describe(['day_of_month','day_of_week','dep_delay','arr_delay','canc
elled','air_time','distance','delay_labeled']).show()
```

```
+-------+-----------------+------------------+------------------+------------
------+---------+------------------+------------------+------------------+
|summary|     day_of_month|       day_of_week|         dep_delay|         arr
_delay|cancelled|          air_time|          distance|      delay_labeled|
+-------+-----------------+------------------+------------------+------------
------+---------+------------------+------------------+------------------+
|  count|          1030606|           1030606|           1030606|           1
030606|  1030606|           1030606|           1030606|            1030606|
|   mean|15.39944071740316|3.9361967619051317|10.939452128165371|4.4964680974
106495|      0.0|137.35455644543114|1027.7749382402199|0.18165622944170712|
| stddev|8.730470469828326|1.9823334094478096|  41.6874499687493| 44.15147169
489554|      0.0|  79.0332443431303| 667.7328895728119| 0.3855611339196749|
|    min|                1|                 1|                -68|
-87|        0|                17|                74|                 0|
|    max|               31|                 7|              1988|
1971|        0|               683|              4983|                 1|
+-------+-----------------+------------------+------------------+------------
------+---------+------------------+------------------+------------------+
```

Print the most current train_df table

In [17]:

```
train_df.show(10)
```

```
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+-------------+
|YEAR|MONTH|DAY_OF_MONTH|DAY_OF_WEEK|CARRIER|FL_NUM|ORIGIN|DEST|DEP_TIME|DEP_
DELAY|ARR_TIME|ARR_DELAY|CANCELLED|CANCELLATION_CODE|AIR_TIME|DISTANCE|WEEKEN
D|DEP_HOUR|DELAY_LABELED|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+-------------+
|2015|    1|           1|          4|     AA|     1|   JFK| LAX|     855|
-5|    1237|        7|        0|             null|     378|    2475|      0|
8|           0|
|2015|    1|           2|          5|     AA|     1|   JFK| LAX|     850|
-10|    1211|      -19|        0|             null|     357|    2475|      1|
8|           0|
|2015|    1|           3|          6|     AA|     1|   JFK| LAX|     853|
-7|    1151|      -39|        0|             null|     330|    2475|      1|
8|           0|
|2015|    1|           4|          7|     AA|     1|   JFK| LAX|     853|
-7|    1218|      -12|        0|             null|     352|    2475|      1|
8|           0|
|2015|    1|           5|          1|     AA|     1|   JFK| LAX|     853|
-7|    1222|       -8|        0|             null|     338|    2475|      0|
8|           0|
|2015|    1|           6|          2|     AA|     1|   JFK| LAX|     856|
-4|    1300|       25|        0|             null|     335|    2475|      0|
8|           1|
|2015|    1|           7|          3|     AA|     1|   JFK| LAX|     859|
-1|    1221|      -14|        0|             null|     341|    2475|      0|
8|           0|
```

```
|2015|    1|           8|          4|     AA|     1|   JFK| LAX|     856|
-4|    1158|      -37|        0|            null|     333|    2475|      0|
8|            0|
|2015|    1|           9|          5|     AA|     1|   JFK| LAX|     901|
1|    1241|        6|        0|              null|     353|    2475|      1|
9|            0|
|2015|    1|          10|          6|     AA|     1|   JFK| LAX|     903|
3|    1235|        0|        0|              null|     345|    2475|      1|
9|            0|
+----+-----+------------+-----------+-------+------+------+----+--------+----
-----+--------+---------+---------+-----------------+--------+--------+------
-+--------+-------------+
only showing top 10 rows
```

Modeling

Processing Train and Test Dataframes

In [18]:

```
# Process Training DataFrame
model_train_df = train_df
# Filter for certain Airports (based on airport code) for both "Origin" and
"Dest" variables.
model_train_df =
model_train_df[model_train_df.ORIGIN.isin("LAX","JFK","LGA","ORD","ATL","SFO"
,"DTW","SLC","CLT","DEN","LAS","SEA","MCO","MIA","PHX","IAH")]
model_train_df =
model_train_df[model_train_df.DEST.isin("LAX","JFK","LGA","ORD","ATL","SFO","
DTW","SLC","CLT","DEN","LAS","SEA","MCO","MIA","PHX","IAH")]
# Stratified Sampling using sampleBy(x, col, fractions, seed) function
# x = A SparkDataFrame, col = column that defines strata, fractions = A named
list giving sampling fraction for each stratum, seed = random seed
model_train_df = model_train_df.sampleBy("delay_labeled", fractions={0: 0.4,
1: 0.99}, seed=1111)
print("After sampling:")
model_train_df.groupBy( train_df.DELAY_LABELED ).count() \
                .withColumn("STATUS", when( col("DELAY_LABELED") == 0,
"ontime").otherwise("delayed") ) \
                .drop("DELAY_LABELED").show()
```

```
After sampling:
+-----+-------+
|count| STATUS|
+-----+-------+
|54250|delayed|
|86837| ontime|
+-----+-------+
```

In [20]:

```
# Process Test DataFrame
model_test_df = spark.read.load("s3a://sparklab123/test_df.csv", "csv",
delimiter=",", inferSchema=True, header=True)
```

```
model_test_df =
model_test_df[model_test_df.ORIGIN.isin("LAX","JFK","LGA","ORD","ATL","SFO","
DTW","SLC","CLT","DEN","LAS","SEA","MCO","MIA","PHX","IAH")]
model_test_df =
model_test_df[model_test_df.DEST.isin("LAX","JFK","LGA","ORD","ATL","SFO","DT
W","SLC","CLT","DEN","LAS","SEA","MCO","MIA","PHX","IAH")]
model_test_df = model_test_df.withColumn("WEEKEND",
when((model_test_df.DAY_OF_WEEK == 5) | (model_test_df.DAY_OF_WEEK == 6) |
(model_test_df.DAY_OF_WEEK == 7), 1).otherwise(0))
model_test_df = model_test_df.withColumn("DEP_HOUR",
(model_test_df.DEP_TIME/float(100)).cast('int') )
model_test_df = model_test_df.withColumn("DELAY_LABELED", when(
(model_test_df.ARR_DELAY > 15), 1).otherwise(0))
model_test_df = model_test_df.filter(model_test_df.CANCELLED == 0)
model_test_df = model_test_df.filter( model_test_df.ARR_DELAY != "NA" )
model_test_df = model_test_df.withColumn("ARR_DELAY",
col("ARR_DELAY").cast("int")) \
                              .withColumn("DEP_DELAY",
col("DEP_DELAY").cast("int")) \
                              .withColumn("AIR_TIME",
col("AIR_TIME").cast("int"))  \
                              .withColumn("DAY_OF_WEEK",
col("DAY_OF_WEEK").cast("string"))


print("Training Record Count:  " + str(model_train_df.count()))
print("Testing Record Count:   " + str(model_test_df.count()))
```

```
Training Record Count:  141505
Testing Record Count:   192422
```

Exploring target variable

In [21]:

```
model_train_df.groupBy("delay_labeled").count().show()
```

```
+-------------+-----+
|delay_labeled|count|
+-------------+-----+
|            1|54278|
|            0|87227|
+-------------+-----+
```

Model Preparation and Transformations

In [23]:

```
#Documentation can be found here:
https://docs.databricks.com/applications/machine-learning/mllib/binary-
classification-mllib-pipelines.html

# Since we are going to try algorithms like Logistic Regression, we will have
to convert the
```

```
# categorical variables in the dataset into numeric variables. There are 2
ways we can do this:
# Category Indexing: This is basically assigning a numeric value to each
category from {0, 1, 2, ...numCategories-1}.
# This introduces an implicit ordering among your categories, and is more
suitable for ordinal variables (eg: Poor: 0, Average: 1, Good: 2)
# One-Hot Encoding: This converts categories into binary vectors with at most
one nonzero value
# (eg: (Blue: [1, 0]), (Green: [0, 1]), (Red: [0, 0]))
# Here, we will use a combination of StringIndexer and OneHotEncoderEstimator
to convert the categorical variables.
# The OneHotEncoderEstimator will return a SparseVector.

# Category Indexing with StringIndexer: create Index string variables (using
string indexer)
si1  = StringIndexer(inputCol="DAY_OF_WEEK", outputCol="DAY_OF_WEEK_index")
si2  = StringIndexer(inputCol="CARRIER",     outputCol="CARRIER_index")
si3  = StringIndexer(inputCol="ORIGIN",      outputCol="ORIGIN_index")
si4  = StringIndexer(inputCol="DEST",        outputCol="DEST_index")
si5  = StringIndexer(inputCol="DEP_HOUR",    outputCol="DEP_HOUR_index")
# Use OneHotEncoder to convert categorical variables into binary
SparseVectors
ohe1 = OneHotEncoder(inputCol="DAY_OF_WEEK_index",
outputCol="DAY_OF_WEEK_ohe")
ohe2 = OneHotEncoder(inputCol="CARRIER_index",      outputCol="CARRIER_ohe")
ohe3 = OneHotEncoder(inputCol="ORIGIN_index",       outputCol="ORIGIN_ohe")
ohe4 = OneHotEncoder(inputCol="DEST_index",         outputCol="DEST_ohe")
ohe5 = OneHotEncoder(inputCol="DEP_HOUR_index",     outputCol="DEP_HOUR_ohe")

col_target   =  'DELAY_LABELED'
col_features =  [

'DAY_OF_WEEK_index','CARRIER_index','ORIGIN_index','DEST_index','DEP_HOUR_ind
ex',
                'AIR_TIME','DISTANCE','WEEKEND'
                ]

# Use a VectorAssembler to combine all the feature columns into a single
vector column.
va  = VectorAssembler(inputCols=col_features, outputCol="features")

#Since we will have more than 1 stage of feature transformations, we use a
Pipeline to tie the stages together.
rfc = RandomForestClassifier(featuresCol="features", labelCol=col_target,
predictionCol="prediction", probabilityCol="probability", numTrees=25,
maxDepth=5, maxBins=32, seed=12345)
pipeline =
Pipeline(stages=[si1,si2,si3,si4,si5,ohe1,ohe2,ohe3,ohe4,ohe5,va,rfc])
```

Fit and Evaluate the Models

In [24]:

```
model = pipeline.fit(model_train_df)

predictions = model.transform(model_test_df)
predictions.createOrReplaceTempView("predictions")

# Multiclass Evaluator
mc_evaluator = MulticlassClassificationEvaluator(labelCol=col_target,
predictionCol="prediction",
metricName="accuracy")#f1|weightedPrecision|weightedRecall|accuracy
accuracy     = mc_evaluator.evaluate(predictions)
print("Accuracy:       " + str(accuracy))

# Binary Evaluator to evaluate our model
bi_evaluator = BinaryClassificationEvaluator(labelCol=col_target,
metricName='areaUnderROC')  # areaUnderROC | areaUnderPR
areaunderroc = bi_evaluator.evaluate(predictions)
print("Area Under ROC: " + str(areaunderroc))

# Print True Positive vs. False Positives
predictions.groupBy('delay_labeled','prediction').count().show()

# Print Feature Importance
feature_importance_vars = sorted([(col_features[i],feature) for i,feature in
enumerate(model.stages[-1].featureImportances)], key=lambda x: x[1],
reverse=True)
print('Feature Importances (descending):')
for f in feature_importance_vars:
    print(f)
```

```
Accuracy:       0.8103179470122959
Area Under ROC: 0.6409473426806589
+-------------+----------+------+
|delay_labeled|prediction| count|
+-------------+----------+------+
|            1|       1.0|  3293|
|            0|       0.0|152630|
|            0|       1.0|  5990|
|            1|       0.0| 30509|
+-------------+----------+------+

Feature Importances (descending):
('DEP_HOUR_index', 0.6256826068339603)
('CARRIER_index', 0.13180390143864693)
('ORIGIN_index', 0.08539310070416628)
('DEST_index', 0.05614584765920664)
('AIR_TIME', 0.04180464235570858)
('DAY_OF_WEEK_index', 0.029710568705850066)
('DISTANCE', 0.029459332302461477)
('WEEKEND', 0.0)
```