Binary Classification using Spark - Ning Han

- Motivation

The dataset used in the exercise is from the Data Science Institute's Capstone project with the Department of Motor Vehicle. The original dataset contains 579,229 crashes from 2015 to 2017. The dataset used in this exercise is the cleaned and encoded training data we used in the capstone project. The motivation of this exercise is to explore the capabilities of Spark and see if we can further employ it in our capstone project.

Exploratory Steps

- Initialize Context
- Read Spark DataFrame From CSV
- Write Parquet to S3
- Demo DataFrame
- VectorAssembler
- Train and Fit Logistic Model and Random Forest
- Model Evaluation

In [4]:
```
#Read into spark dataframe from csv in s3
role = get_execution_role()
bucket='odl-spark19spds6003-001'
data_key = 'nh4mq/crash_test_afterdummy.csv'
data_location = 's3://{}/{}'.format(bucket, data_key)

```

In [5]:
```
pd.read_csv(data_location).head()
```

Out[5]:

	Unnamed: 0	Latitude	Longitude	VehicleYear	VehicleSpeedBeforeCrash	VehicleSpeedLimit	VehicleMaximumSafeSpeed	AgeAtCrash	CountOffense	Week	...
0	0	36.84859	-76.20993	2004	10.0	30.0	30.0	21.0	0	4	...
1	1	38.67771	-77.23463	2012	0.0	55.0	0.0	19.0	0	11	...
2	2	36.78789	-76.42436	2006	0.0	60.0	0.0	65.0	0	51	...
3	3	38.12925	-78.89123	2013	10.0	35.0	0.0	39.0	0	3	...
4	4	37.55637	-77.44213	2006	25.0	35.0	35.0	49.0	0	10	...

In [13]:
```
df.printSchema()
```

```
root
 |-- Latitude: double (nullable = true)
 |-- Longitude: double (nullable = true)
 |-- VehicleYear: long (nullable = true)
 |-- VehicleSpeedBeforeCrash: double (nullable = true)
 |-- VehicleSpeedLimit: double (nullable = true)
 |-- VehicleMaximumSafeSpeed: double (nullable = true)
 |-- AgeAtCrash: double (nullable = true)
 |-- CountOffense: long (nullable = true)
 |-- Week: long (nullable = true)
 |-- Month: long (nullable = true)
 |-- Year: long (nullable = true)
 |-- TestHour: long (nullable = true)
 |-- TestMinutes: long (nullable = true)
 |-- CrashDayOfWeekId_2: long (nullable = true)
 |-- CrashDayOfWeekId_3: long (nullable = true)
 |-- CrashDayOfWeekId_4: long (nullable = true)
 |-- CrashDayOfWeekId_5: long (nullable = true)
 |-- CrashDayOfWeekId_6: long (nullable = true)
 |-- CrashDayOfWeekId_7: long (nullable = true)
 |-- RoadwaySurfaceTypeID_Unknown: long (nullable = true)
 |-- RoadwaySurfaceTypeID_Bad: long (nullable = true)
 |-- RoadwayDefectID_Defects: long (nullable = true)
 |-- LightConditionID_Darkness: long (nullable = true)
 |-- LightConditionID_Day: long (nullable = true)
 |-- LightConditionID_DarknessUnknownLighting: long (nullable = true)
 |-- LightConditionID_Unknown: long (nullable = true)
 |-- RoadwayDescriptionID_OnewayUndivided: long (nullable = true)
 |-- RoadwayDescriptionID_TwowayDivided: long (nullable = true)
 |-- RoadwayDescriptionID_TwowayUndivided: long (nullable = true)
 |-- WeatherConditionID_AdverseCondtions: long (nullable = true)
 |-- WeatherConditionID_NoAdverseCondtion: long (nullable = true)
 |-- RoadwayAlignmentID_Other: long (nullable = true)
 |-- RoadwayAlignmentID_Hillcrest: long (nullable = true)
 |-- RoadwayAlignmentID_Dip: long (nullable = true)
 |-- RoadwayAlignmentID_Curve: long (nullable = true)
 |-- RoadwayAlignmentID_Grade: long (nullable = true)
 |-- SchoolZoneID_No: long (nullable = true)
 |-- IntersectionTypeID_Intersection: long (nullable = true)
```

Copy

Visualization

- The results are:

Test: Area Under ROC for Logistic Model :
0.7977774039278601
Test: Area Under ROC for Random Forest:
0.8053983341482011

- The graph shows the ROC curve of the logistics model.