CONGESTION PREDICTION – Dublin Bus dataset

**Data Description:**

Dublin Bus GPS sample data from Dublin City Council. Bus GPS Data Dublin Bus GPS data across Dublin City, from Dublin City Council traffic control, in csv format. Each data point (row in the CSV file) has the following entries:

1. Timestamp micro since 1970 01 01 00:00:00 GMT
2. Line ID
3. Direction
4. Journey Pattern ID
5. Time Frame (The start date of the production time table - in Dublin the production time table starts at 6am and ends at 3am)
6. Vehicle Journey ID (A given run on the journey pattern)
7. Operator (Bus operator, not the driver)
8. Congestion [0=no,1=yes]
9. Lon WGS84, Lat WGS84, Delay (seconds, negative if bus is ahead of schedule)
10. Block ID (a section ID of the journey pattern)
11. Vehicle ID
12. Stop ID
13. At Stop [0=no,1=yes]

**Data set Link:**

https://data.gov.ie/dataset/dublin-bus-gps-sample-data-from-dublin-city-council-insight-project

**Project Implementation:**

In this project, I tried to predict if there is congestion in route based on feature attributes such as operator, congestion, at stop, direction with congestion being class attribute. I choose logistic regression approach to achieve this.

Following are the steps involved:

1. Creating spark session and reading the data

2. Preprocessing

- Filtering the data by converting the string into list of numbers

- Dropping unnecessary columns

3. Creating the categorical variables

- Unlike python, attributes cannot be directly passed to train\_test\_split function in pyspark. We need to explicitly create data frames, define which of those attributes are categorical, index them using sting indexer, encode them and convert them into RDD again.

4. Splitting the training and testing data

- Randomly splitting the data into 80-20%

5. Training the model

6. Evaluating the model on testing data

7. Calculating the f-measure

**Results and Observations:**

After performing above steps, we evaluate the results by considering confusion matrix and calculating F – measure. In our data set I have observed the F- measure to be 0.9944495455289587 with TP, FP, TN and FN printed along.

**Scope and Extension:**

There is a problem of class imbalance, meaning unequal weightage of binary class attributes (Ex: 70% of '0' and 30% of '1') when you choose such large real world datasets. This can be dealt with using several sampling techniques. Due to complex implementations in spark, I would like to perform strategies like under-sampling or over-sampling later on as an extension for this project.