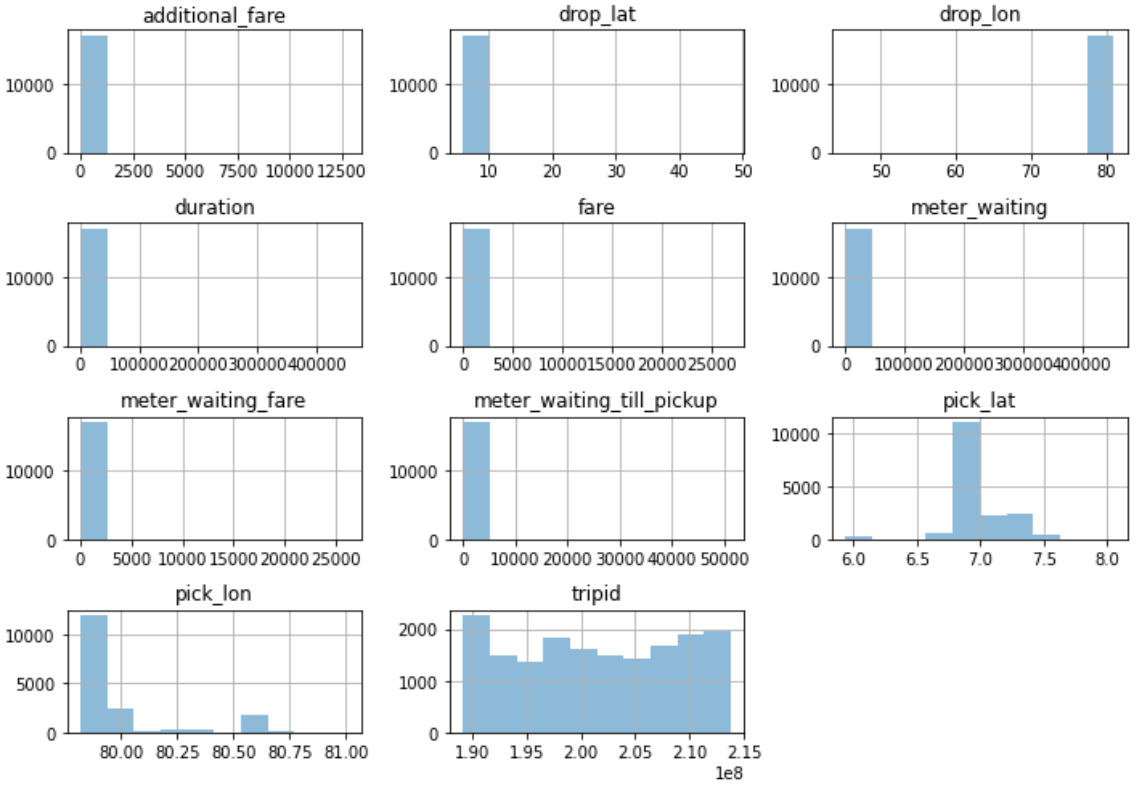
Name Nimesha Asintha Muthunayake
Index No 209359G
Git Repository Link https://github.com/nimeshacs/Fare-Classification-DS-MSc2020
Data Set Ride Fare Classification

Fare Classification

First Look of Data (analysis of data)

If we look at the dimensions of the data there are 8,576 entries and 14 columns in the data set and I noticed several missing data, which is a great reminder that data collected in the real-world will never be perfect. hist() function is used to draw one histogram of the DataFrame's columns. A histogram is a representation of the distribution of data.



Data Preparation

Since we have empty (null) data we have to remove these NaN values and less important data columns form the learning model used df.dropna() to remove the NaN values and removed label,pickup_time,drop_time training data set

Random Forest for Classification

Random forests is a supervised learning algorithm. It can be used both for classification and regression and Random Forest Scikit-Learn API is used for this classification problem

Creating a Random Forest Model

- **Encoding Data**

The first step for us is known as encoding of the data. This process takes categorical variables, mainly the label is converted into a numerical representation without an arbitrary ordering

df['label'].map({'correct': 1, 'incorrect': 0 })

- **Training and Testing Sets**

Generally, when training a model, we randomly split the data into training and testing sets to get a representation of all data points ,used 0.33% of data for the test set and used 0 as random state.

After all the work of data preparation, creating and training the model is pretty simple using Scikit-learn and Instantiate model with 1000 decision trees with 100 Max depth

```
38 rfc = RandomForestClassifier(n_estimators=1000, max_depth=100, max_features='sqrt')
39 rfc.fit(X_train,y_train)
40 rfc_predict = rfc.predict(X_test)
41
```

Evaluating Performance

In here in the data set we need to predict whether the fare is correct (1) or incorrect(0).

The confusion matrix is useful for giving the false positives and false negatives. The classification report tells the accuracy of the model.

The Accuracy of the model was 0.94

```
Accuracy  : 0.9405357142857143


              === Confusion Matrix ===
              [[ 233  276]
               [  55 5036]]


              === Classification Report ===
                           precision    recall  f1-score   support

                        0       0.81      0.46      0.58       509
                        1       0.95      0.99      0.97      5091

                 accuracy                           0.94      5600
                macro avg       0.88      0.72      0.78      5600
             weighted avg       0.94      0.94      0.93      5600
```