Capstone Project 1 In Depth Analysis (Machine Learning)

1. The Model:

The target (dependent) variable will be “Pri_Problem” column. This column will be classified by using text columns. There are 17 different reason for air conflict issues. The classification of the variable and value counts as below:

```
Human Factors                                67311
Ambiguous                                     5259
Procedure                                     4122
None                                          3598
Weather                                       3017
Company Policy                                2582
Aircraft                                      1428
ATC Equipment / Nav Facility / Buildings       910
Chart Or Publication                           894
Airspace Structure                             881
Airport                                        799
Environment - Non Weather Related              553
Staffing                                       109
Equipment / Tooling                            104
Manuals                                         80
MEL                                             12
Incorrect / Not Installed / Unavailable Part     8
Logbook Entry                                    2
Name: Pri_Problem, dtype: int64
```

For the model development; there are three questions to be answered. I will show the questions below and will give the answers by explaining the progressive parts of the project.

Question 1: There are two text columns for the air conflict narrative. One is “Synopsis” and the other one is “Narrative”. Narrative column is the exact text from each report filled by the crew. In general, this column is longer than the Synopsis column. Synopsis is a quality summary of Narrative text. This column is focusing the most critical information of the incident and represents the version of Narrative column in which story part was chopped off. Now the first question is which column should be used for a better NLP model for the classification purposes.

Question 2: There are 17 categories in Pri_Problem column. The classification model can be run either by each category or a multiclass model. As it is seen above the amount of each class is either very low or very high. When we compare “Human Factors” with “Logbook Entry” won’t

be reasonable. I will try binomial models for each individual category and try a top 3 and top 6 multiclass model.

Question 3: Which classifier should be use for best performance? The classifiers which I will use for multiclass will be MNB, TFIDF, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Model. The best performer model will run with hyper-tuning with parameters and bag of words vectorizer.

2. Findings:

Firstly, I create a copy of the dataframe and run a Multinomial Naïve Bayes model for only "Human Factors" values in the target column and tried to predict it in a model. In first run I used Synopsis column and in the second run I tried Narrative column. The accuracy scores for both training set and test set was approximately 80% for both columns. However, Synopsis column resulted better almost 1.5%. Most likely for Narrative column there were more noise than the Synopsis column. The results are as below for Synopsis and Narrative column respectively:

```
Accuracy Scores for the Training set : 0.8058524874693831  and Test Set : 0.79218075845886
True
```

```
Accuracy Scores for the Training set : 0.790215575272916  and Test Set : 0.7801074861857543
True
```

As a result, I decided to use Synopsis column for the remaining parts of the model.

As the next step I run MNB models for each category. The results as accuracy score was very high, however the amount of true positive was very low for the categories other than top three. As an example, I will depict the results for "Environment – Non Weather Related" category below:

```
Accuracy Scores for the Training set : 0.9936414215964574  and Test Set : 0.9929982590265688
True
[[26237    22]
 [  163     0]]
```

So, the answer for the second question will be Multiclass classification. I will run that for both top 3 and top 6 categories.

Up to this step, we decided to use "Synopsis" column with a multiclass classifier for the air conflict issues dataframe. In the progressive steps, for top 3 categories I will run MNB, TFIDF, Logistic Regression, Decision Tree, Random Forest, Gradient Boosting Model, will find the best performer and I will compare the model with top 6 category model and choose one. And lastly I will hypertune the model with bag of words and other parameters.

In order to restructure the dataframe the steps will be executed as it is exhibited in the below picture.

TOP 3 Classification

In [13]:
```
top3=['Human Factors', 'Ambiguous','Procedure']
df_top3=df[df.Pri_Problem.isin(top3)]
```

In [14]:
```
df_top3.shape
```

Out[14]: (76692, 3)

In [15]:
```
df_top3.head()
```

Out[15]:

idx	Pri_Problem	Synopsis	Narrative
1	Human Factors	SMA PENETRATED TCA ON CLIMB OUT.	THIS WAS MY FIRST DEP FROM BFI ON 31L. MY TURN...
2	Human Factors	SMT PLT DESCENDED TO ARPT UNDERLYING TCA; ACCU...	A VFR FLT; BEING CONDUCTED UNDER FAR PART 91; ...
6	Human Factors	LESS THAN STANDARD SEPARATION BETWEEN TWO ACR ...	ACR Y CLIMBING TO FL210 WAS STOPPED AT 160 FOR...
7	Human Factors	ACR LTT LANDED AT THE WRONG ARPT; DESTINATION ...	SOME TIME HAD PASSED AFTER WE HAD PASSED THE R...
8	Human Factors	LESS THAN STANDARD SEPARATION BETWEEN FLT OF 2...	ACFT X ON FINAL FOR RWY 30L (FLT OF 2) MISSED ...

In [16]:
```
df_top3.Pri_Problem.unique()
```

Out[16]: array(['Human Factors', 'Ambiguous', 'Procedure'], dtype=object)

Results for MNB;

```
Accuracy Scores for the Training set : 0.8250689218389091  and Test Set : 0.8140212100139083
True
[[  467   862   253]
 [ 1427 17276  1502]
 [  217    18   986]]

Classification Report
              precision    recall  f1-score   support

    Ambiguous       0.22      0.30      0.25      1582
Human Factors       0.95      0.86      0.90     20205
    Procedure       0.36      0.81      0.50      1221

    micro avg       0.81      0.81      0.81     23008
    macro avg       0.51      0.65      0.55     23008
 weighted avg       0.87      0.81      0.83     23008
```

Results for TFIDF;

```
Accuracy Scores for the Training set : 0.8832054243349974  and Test Set : 0.8786509040333796
True
[[    2  1577     6]
 [    6 20143    34]
 [    2  1167    71]]

Classification Report
              precision    recall  f1-score   support

    Ambiguous       0.20      0.00      0.00      1585
Human Factors       0.88      1.00      0.94     20183
    Procedure       0.64      0.06      0.11      1240

    micro avg       0.88      0.88      0.88     23008
    macro avg       0.57      0.35      0.35     23008
 weighted avg       0.82      0.88      0.83     23008
```

Results for Logistic Regression;

```
Accuracy Scores for the Training set : 0.9219506743163699  and Test Set : 0.885778859527121
True
[[  207  1287    91]
 [  250 19642   291]
 [  112   597   531]]

Classification Report
               precision    recall  f1-score   support

    Ambiguous       0.36      0.13      0.19      1585
Human Factors       0.91      0.97      0.94     20183
    Procedure       0.58      0.43      0.49      1240

    micro avg       0.89      0.89      0.89     23008
    macro avg       0.62      0.51      0.54     23008
 weighted avg       0.86      0.89      0.87     23008
```

Results for Decision Tree;

```
Accuracy Scores for the Training set : 0.8861113180836003  and Test Set : 0.882519123783032
True
[[   24  1535    26]
 [   13 19978   192]
 [   26   911   303]]

Classification Report
               precision    recall  f1-score   support

    Ambiguous       0.38      0.02      0.03      1585
Human Factors       0.89      0.99      0.94     20183
    Procedure       0.58      0.24      0.34      1240

    micro avg       0.88      0.88      0.88     23008
    macro avg       0.62      0.42      0.44     23008
 weighted avg       0.84      0.88      0.84     23008
```

Results for Random forest and GBM;

```
Accuracy Scores for the Training set : 0.9988637210342002  and Test Set : 0.8858657858136301
True
[[   33  1549     3]
 [   16 20105    62]
 [   17   979   244]]

Classification Report
               precision    recall  f1-score   support

    Ambiguous       0.50      0.02      0.04      1585
Human Factors       0.89      1.00      0.94     20183
    Procedure       0.79      0.20      0.32      1240

    micro avg       0.89      0.89      0.89     23008
    macro avg       0.73      0.40      0.43     23008
 weighted avg       0.86      0.89      0.84     23008


Accuracy Scores for the Training set : 0.8949221369495567  and Test Set : 0.8867350486787204
True
[[   80  1483    22]
 [   55 19939   189]
 [   45   812   383]]

Classification Report
               precision    recall  f1-score   support

    Ambiguous       0.44      0.05      0.09      1585
Human Factors       0.90      0.99      0.94     20183
    Procedure       0.64      0.31      0.42      1240

    micro avg       0.89      0.89      0.89     23008
    macro avg       0.66      0.45      0.48     23008
 weighted avg       0.85      0.89      0.85     23008
```

Random Forest and GBM are chosen as the best performer models. The accuracy score of the test set for the Top 6 category model decreased 6%.

3. Conclusion:

Eventually, our model will be formed by the following criteria:

- NLP text will be gathered from “Synopsis” column.
- A multiclass classifier model will be used.
- Random Forest classifier and Gradient Boosting classifier will be used.
- Top three categories will be used for the classification problem.