Model Metrics File

Model features

Since we used dummy variables the actual number of features is too large to list but provided here are the variables that were used to create the dummy variables.

'CRASH DATE', 'CRASH TIME', 'BOROUGH', 'ZIP CODE', 'LATITUDE',
'LONGITUDE', 'ON STREET NAME', 'CROSS STREET NAME', 'OFF STREET NAME',
'NUMBER OF PERSONS INJURED', 'NUMBER OF PEDESTRIANS INJURED',
'NUMBER OF CYCLIST INJURED', 'NUMBER OF MOTORIST INJURED',
'CONTRIBUTING FACTOR VEHICLE 1', 'COLLISION_ID', 'MM-DD', 'ON HOLIDAY',
'MM', 'Weekday', 'Time', 'DEATH OCCURED', 'DD', 'begin_of_month', 'midddle_of_month',
'end_of_month', 'morning', 'day', 'evening',
'night', 'winter', 'spring', 'summer', 'autumn']

Hyperparameters

For each model we first ran a test with general hyperparameters before using Random search to find the best hyperparameters.

Random Forest:

First test parameters:

(bootstrap=True,n_estimators=50,criterion='entropy', random_state =1)

Random Search parameters:

(min_samples_split=16, n_estimators=50, n_jobs=-1, random_state=1)

XGBoost:

First test parameters:

(n_estimators=2, objective= 'binary:logistic', eval_metric= 'error', random_state=1)

Random Search parameters:

early_stopping_rounds=8, eval_metric='error', learning_rate=0.5, n_estimators=25, random_state=1)

Logistic Regression:

First test parameters:

(penalty = 'l2', C = .1 ,random_state = 40)

Random Search parameters

penalty = 'l2',random_state = 40, C= 1)

Performance metrics

We collected precision, recall, f1-score, and support. We also cross-validated all the precision scores to give us standard deviations of our model to show us how accurate it was when run multiple times. Confusion matrices were used to show the true-false rate of both death and non-death classification. ROC- curves were also implemented to create a visual representation of how well the model performed.