Predicting if aircraft crash has fatalities or not

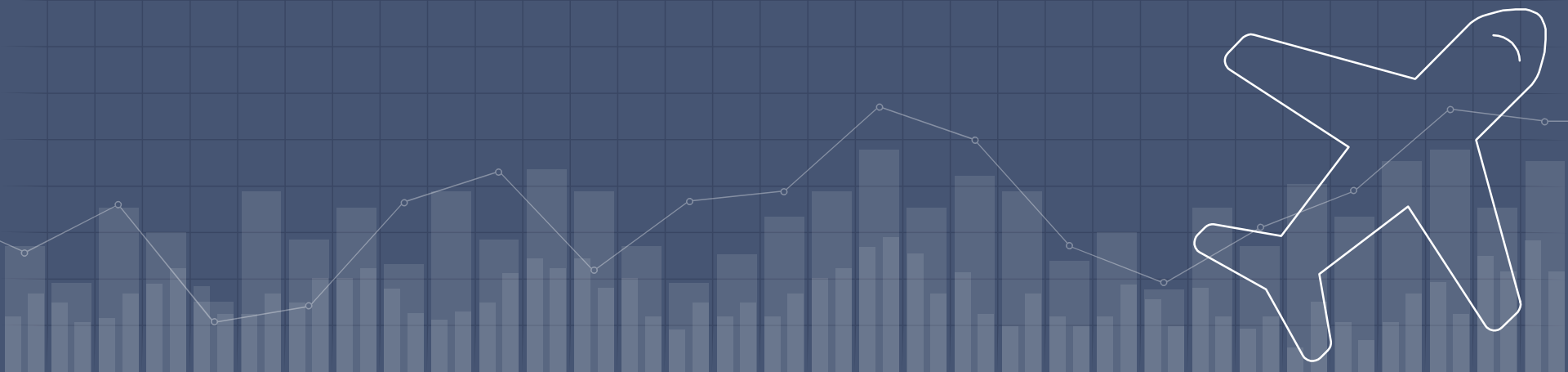
Yunjin Bak

Background

- Aviation travel has increased dramatically over the last couple of decades
- It is important to find out significant factors that could predict if fatality occurs in an aircraft crash
- The results may help to prevent or minimize accidents, which can save lives

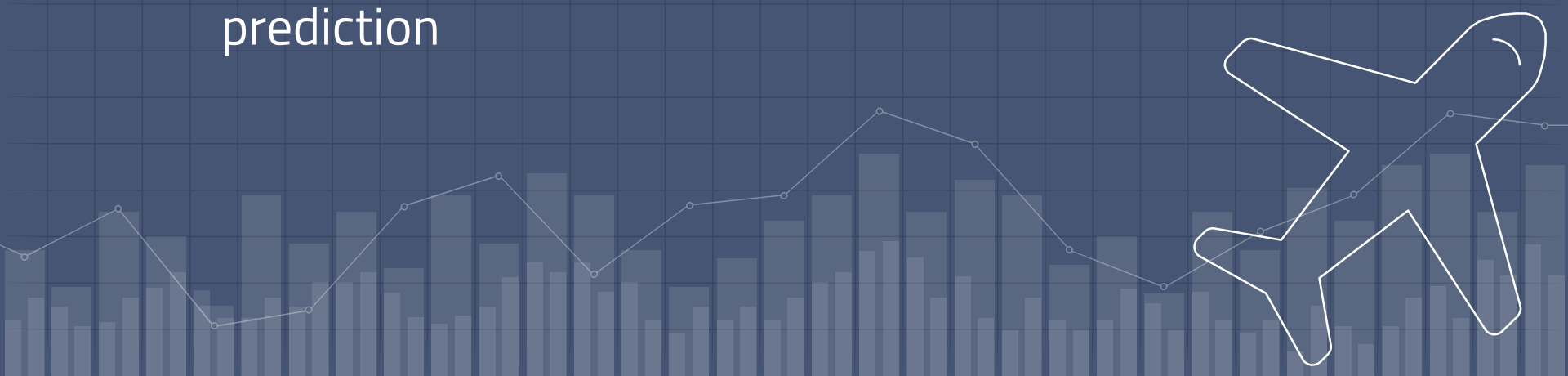
DATA

- Acquired from National Transportation Safety Board
- Information from 1962 to Feb, 2019 about civil aviation accidents and incidents
- Data have 32 variables and 162,800 data points



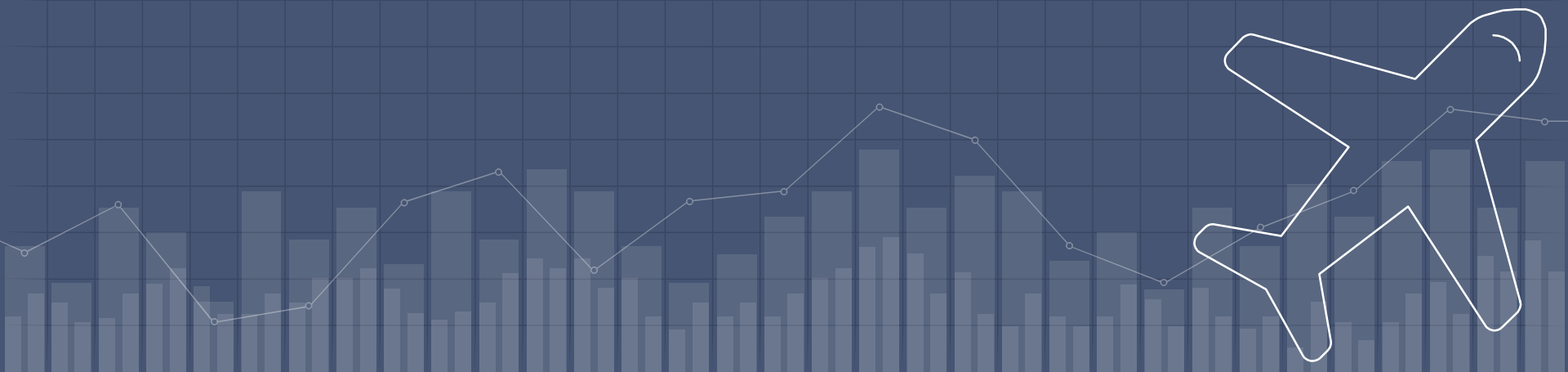
Goal

- Predict successfully whether the crash is fatal or not
- Compare between the models, and find out the best model
- Examine which variables are important for the prediction



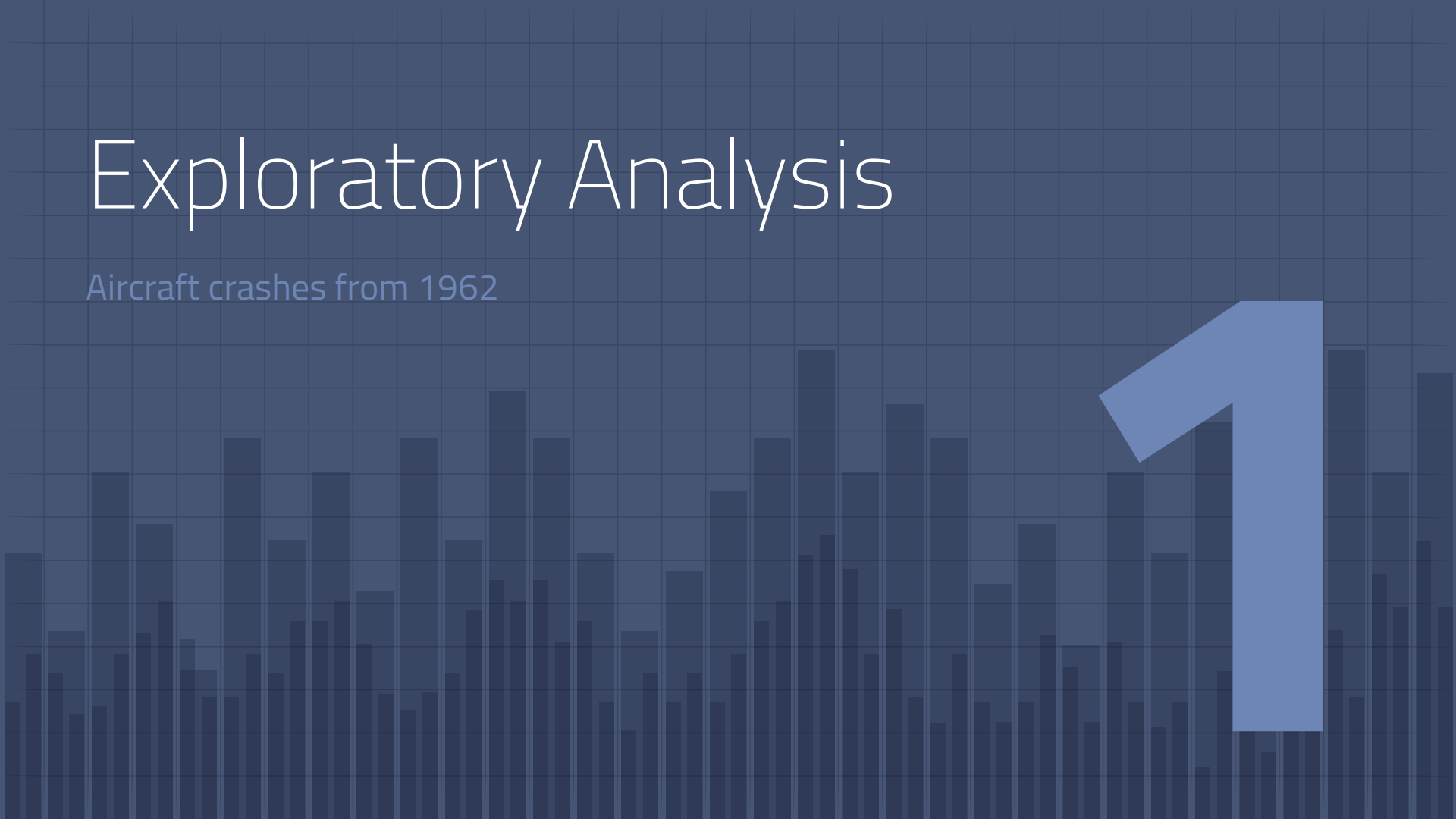
Preprocessing

- 9 out of 32 variables which have over 50 % of missing values were excluded in features
- Then, missing values were dropped

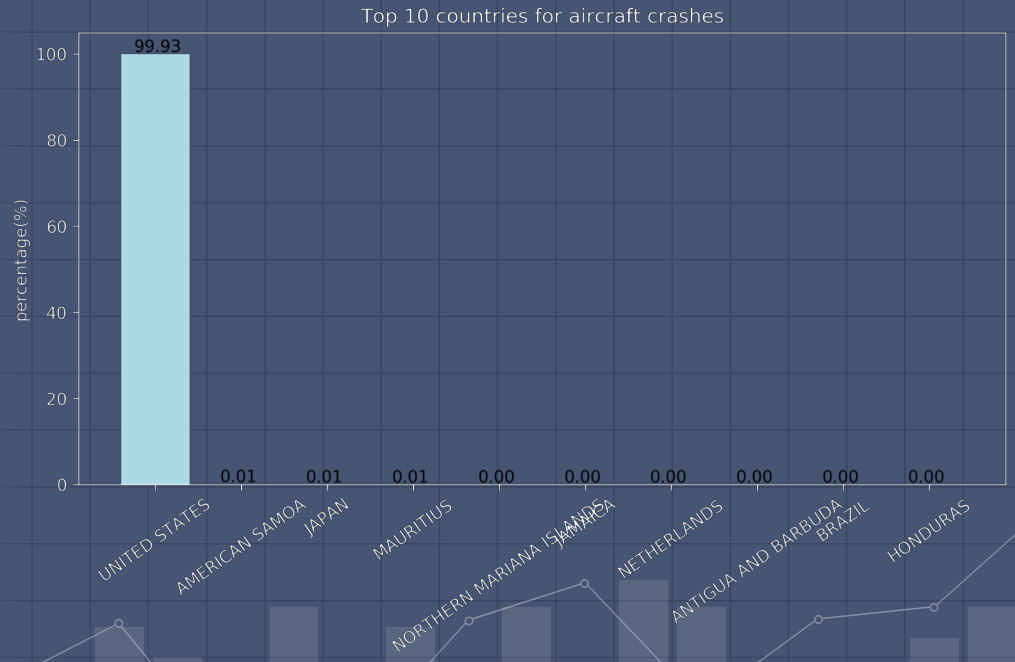


Exploratory Analysis

Aircraft crashes from 1962



Country

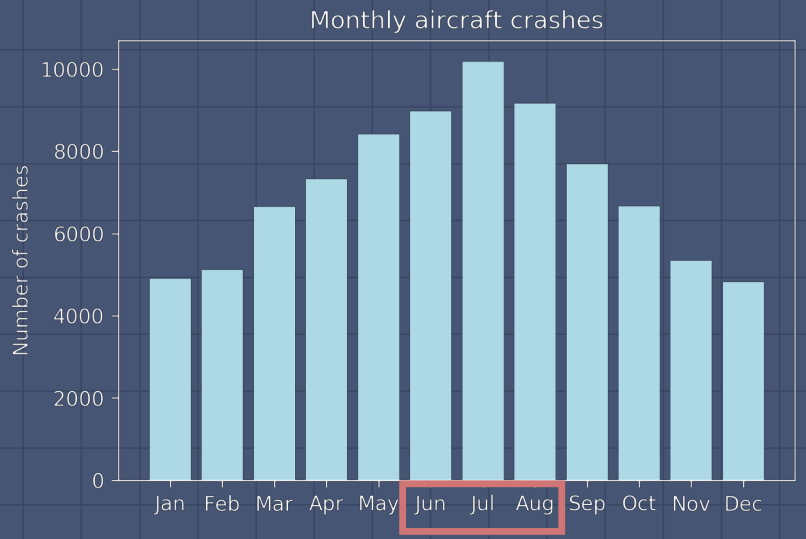


- Over 99% of the data are the crash that happened in the US.
- The data from other countries were excluded for further analysis.

Event Date (Year)

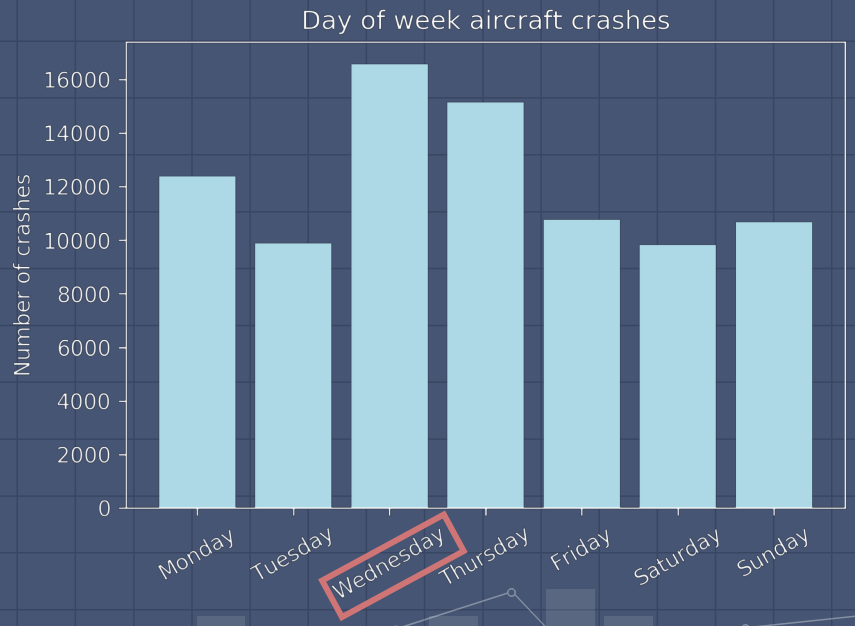
- The trend for yearly aircraft crashes looks random.

Event Date (Month)



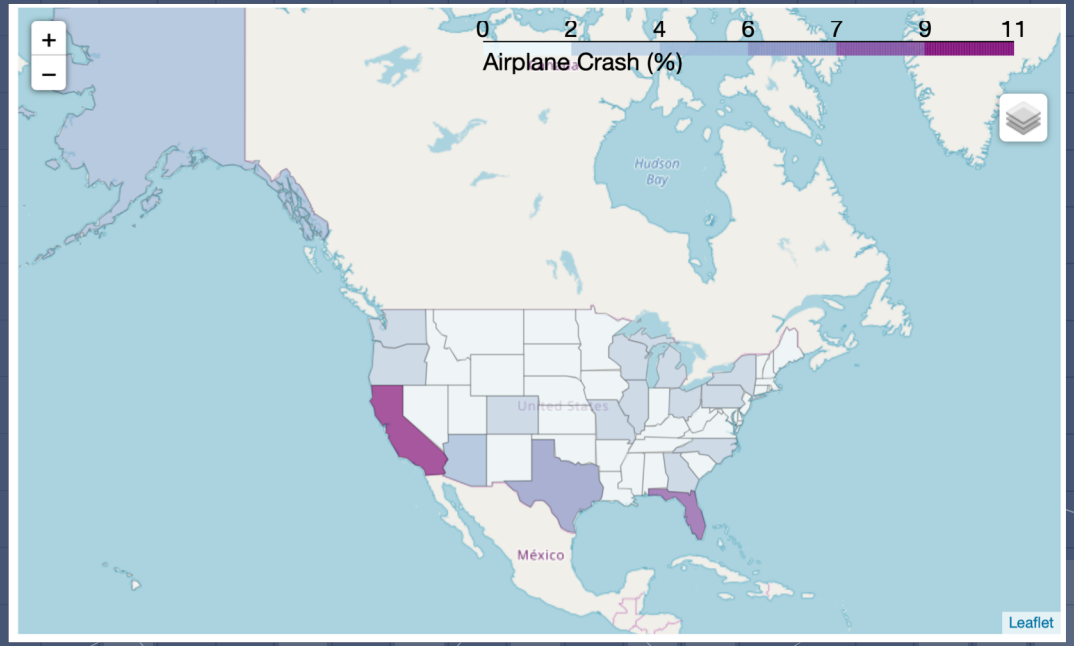
- Air crashes tend to occur mostly during the summer.
- This might be due to high demand on air travel during summer months.

Event Date (Day of Week)



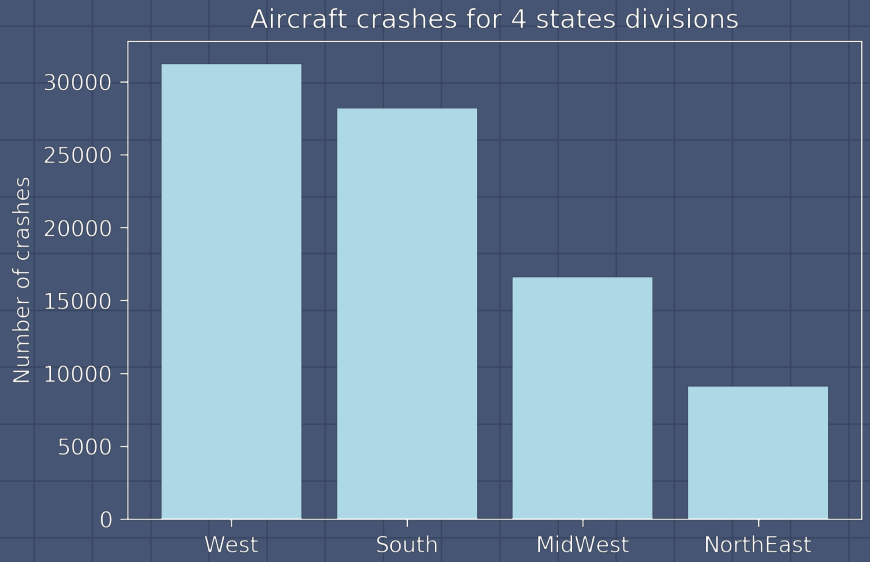
- Air crashes occurred on Wednesday the most.
- This also may be due to more flights on middle of week.
- Monthly and daily crashes were only included, since year information is not suitable for predicting future aviation crashes.

Location (States)



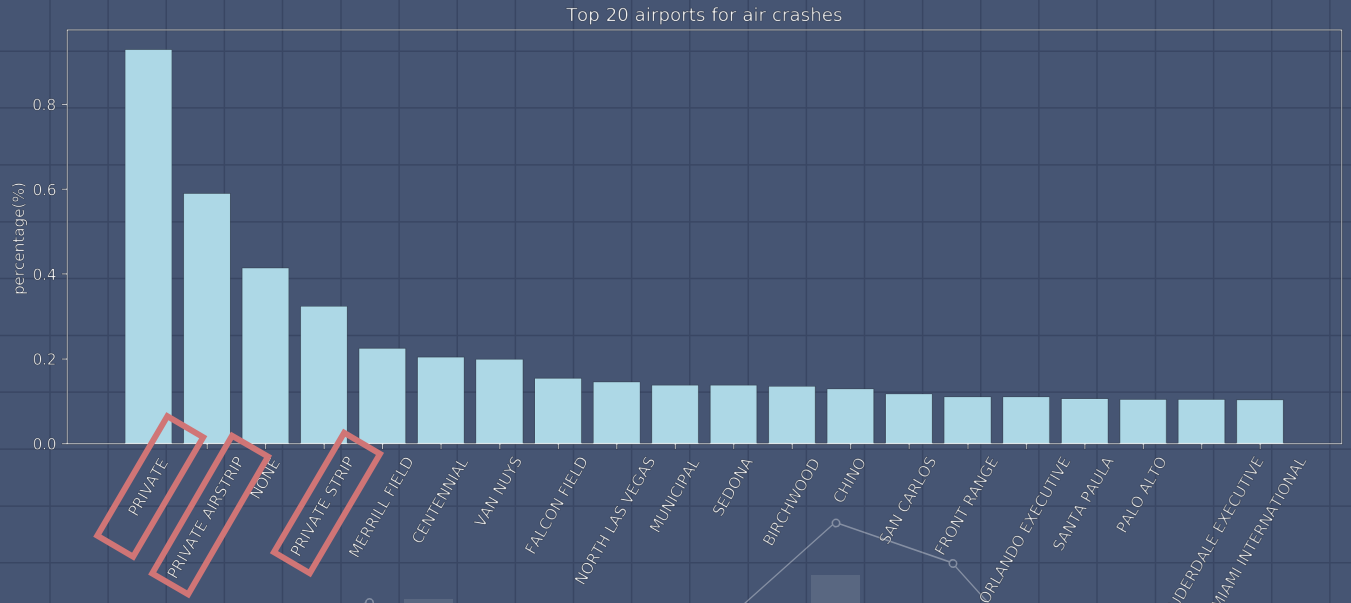
- California has the most airplane crashes, and Florida is the next.
- In order to reduce dimension in states variable, states were divided into 4 categories.

Location (States)



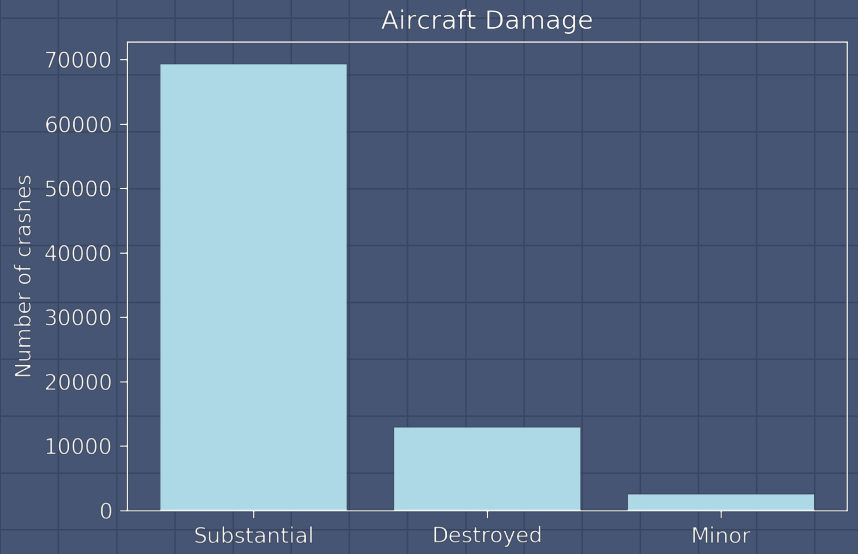
- West has the most frequent number of crashes

Airport



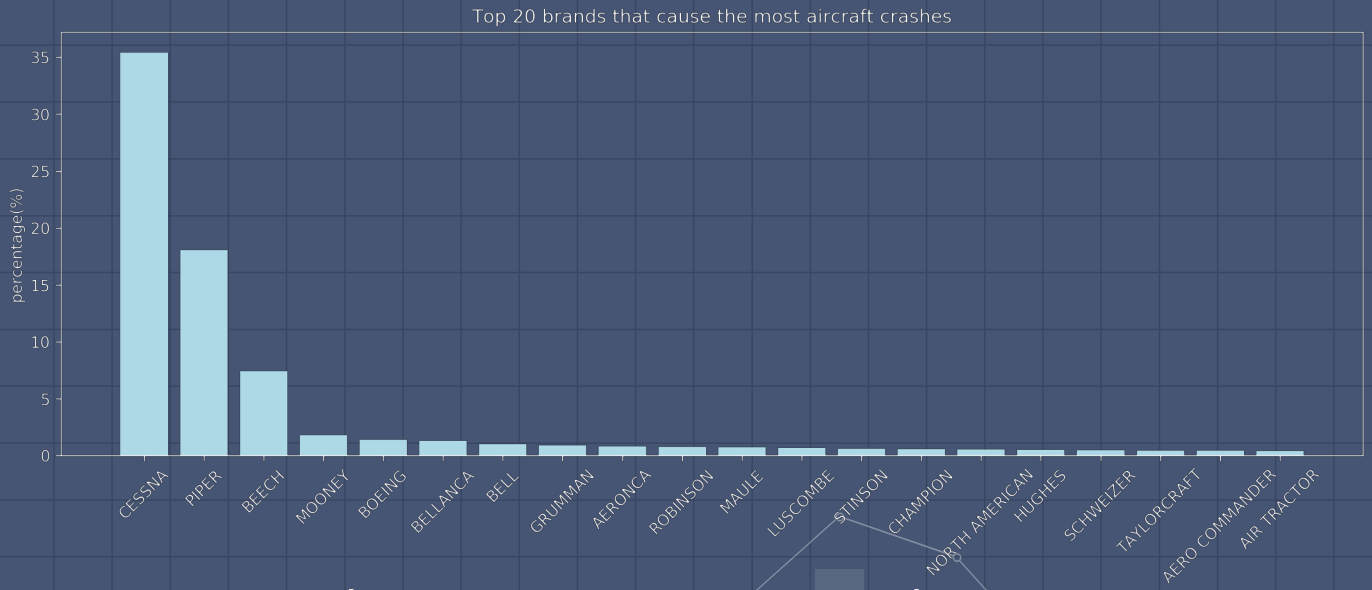
- Aircraft crashes occurred in private airport the most
- The feature was simplified as ‘private or public’

Aircraft Damage



- Most of crashes resulted in substantial damage of aircraft
- This variable was excluded because of target leakage

Brands



- Top 3 brands (Cessna, Piper, Beech) take over 50%
- To reduce the levels, brands were categorized into 4 different values (Cessna, Piper, Beech, Others).

Amateur Built & Number of Engines

Amateur Built

Whether aircraft is a homebuilt (Y/N)

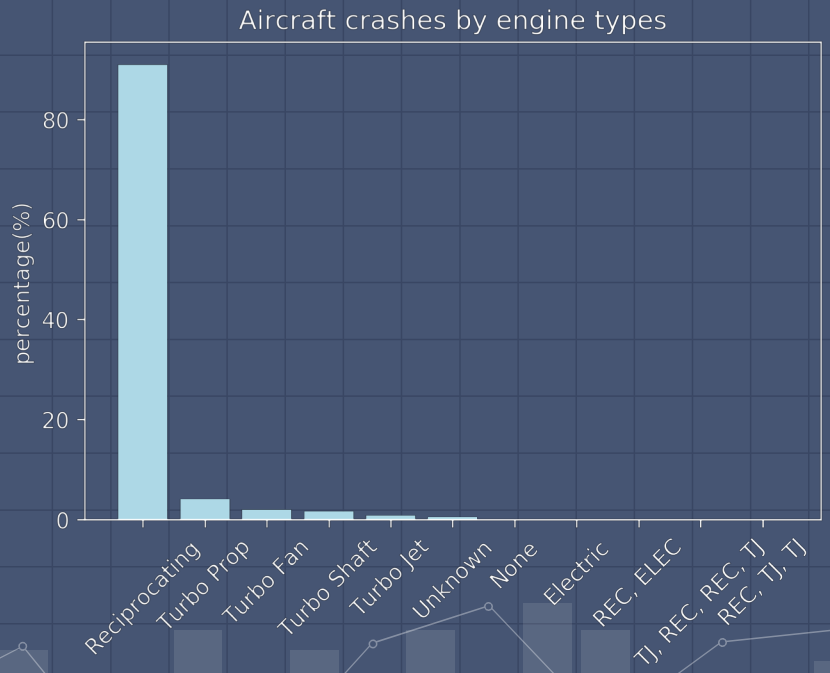
No 75376

Yes 9348

The Number of Engines

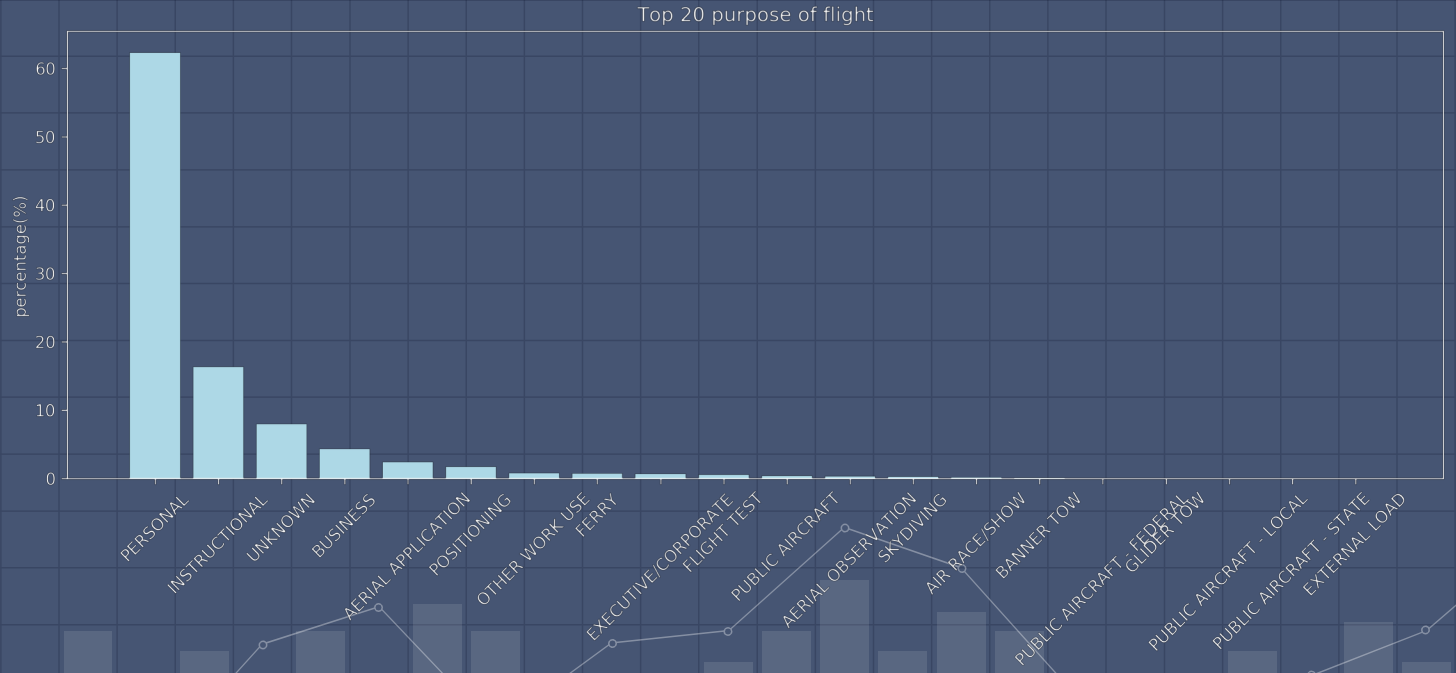
The number of Engines	
1	72102
2	11499
3	451
0	363
4	307
18	2

Engine Type



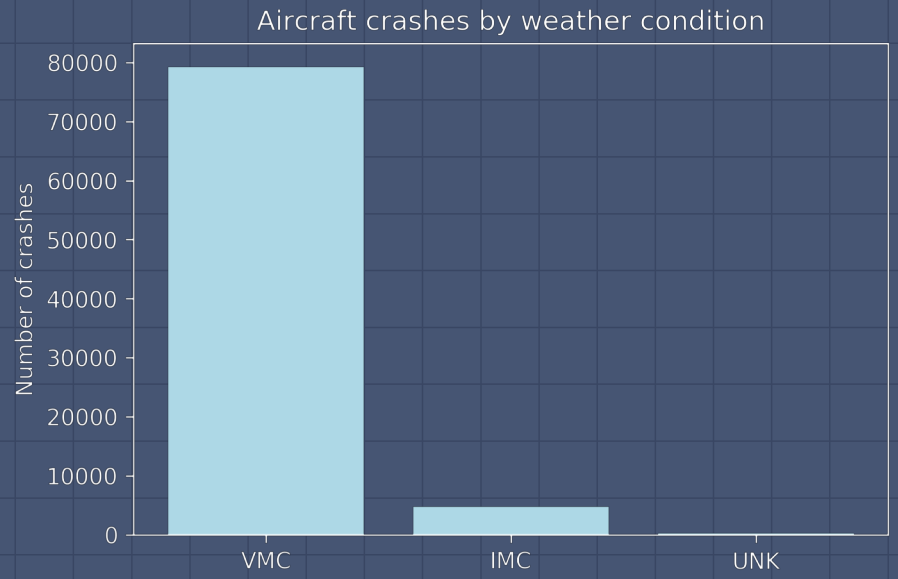
- Over 90% of crashes were caused by planes with reciprocating engine
- The reason for this might be that this engine is the most common type

Purpose of Flight



- Purpose of flight was re-categorized into 3 values (Personal, Instructional, Other)

Weather Condition

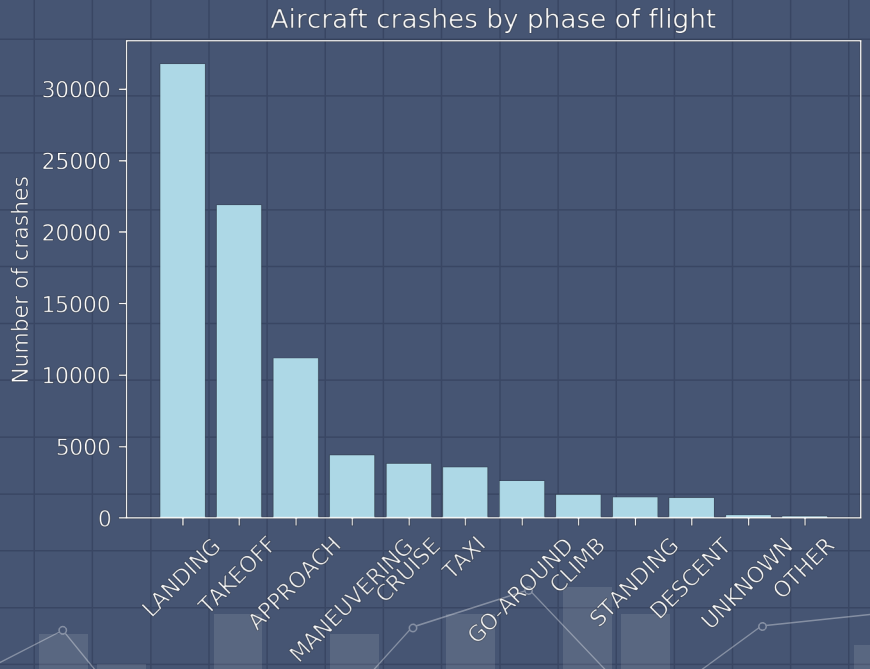


- Most of crashes occurred when there is sufficient visibility
- It might be due to less number of flights in IMC condition

VMC : Visual Meteorological Condition

IMC : Instrument Meteorological Condition

Broad Phase of Flight



- Most of crashes occurred during landing phase

Outcome Variable : Injury Severity

- Fatal / non-fatal
-

Fatal	11558
Non-Fatal	72258

- In order to solve class imbalance problem, Down Sampling was performed

Classification

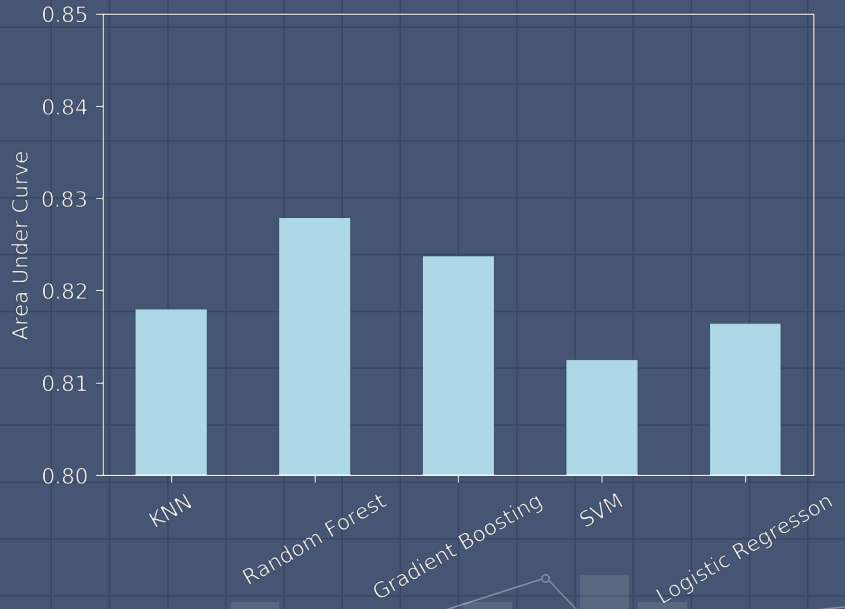
Let’s play with different models

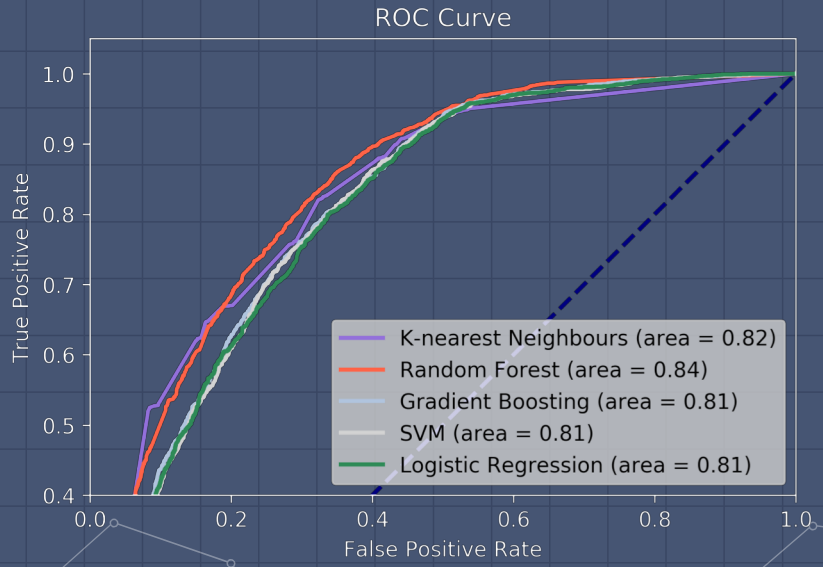
2

Model Comparison

- 10 features (1 numerical, 9 categorical)
- KNN, Random Forest, Gradient Boosting, SVM, and Logistic Regression were used and compared.

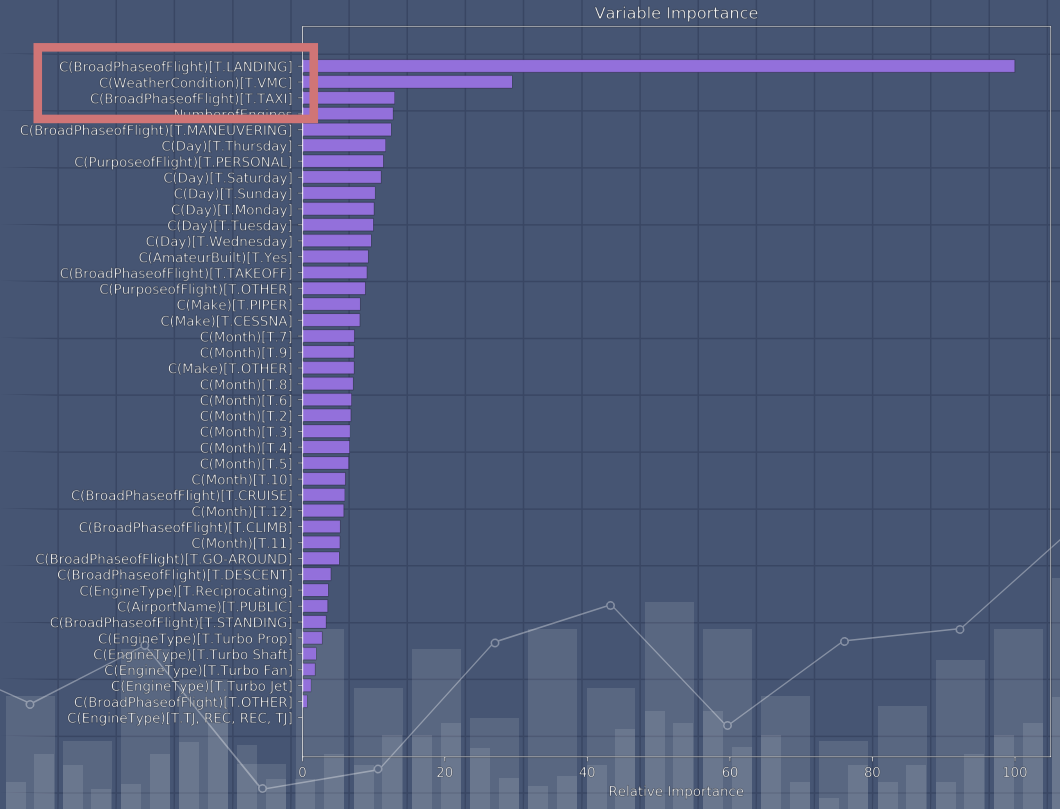
Model Comparison





- Obtained AUC value using cross validation
- Random Forest performed the best

Feature Importance



- Obtained from Random Forest model
- Top 3 : Landing, VMC (Good weather), Taxi

Feature Importance

```
corrmat['Injury Severity'].sort_values()[:5]

C(BroadPhaseofFlight)[T.LANDING]    -0.277294
C(WeatherCondition)[T.VMC]          -0.258879
C(Make)[T.CESSNA]                   -0.084449
C(BroadPhaseofFlight)[T.TAXI]       -0.077133
C(EngineType)[T.Reciprocating]      -0.030680
Name: Injury Severity, dtype: float64
```

BroadPhaseofFlight	Outcome	
APPROACH	1	3071
	0	1276
CLIMB	1	508
	0	163
CRUISE	1	789
	0	476
DESCENT	1	344
	0	170
GO-AROUND	1	593
	0	330
LANDING	0	5128
	1	493

- Landing has negative correlation with outcome variable
- VMC feature could be interpreted intuitively

Conclusion

- All 5 models showed over 80 % AUC score
- Random Forest performed the best with about 85 % accuracy
- Landing phase and good weather condition were the most important features

THANKS!

Any questions?

You can find me at

- github.com/datasciyj
- yunjin.bak1@gmail.com