Introduction

- In US, flights delays can result in making passengers late for personal scheduled events.
- Could a delay flight can be predicted?

Objective

- Identify factors that influence the flight delay
- Predict Fights that will be delayed

Data set

- Title: US flights
- Source: https://mlcourse.ai/assignments
- Description: The Bureau of transportation Statistics tracks the on time performance of flights operated by large air carriers.
- Number of rows: 7 million

For simplicity, in this project a sample of 8.5 % is used for modelling.

- Number of variables: 10

Variable Description

	Variables	Description	Type
1.	Year	A number between 2004 and 2007	int
2.	Month	A number between 1 and 12	int
3.	DayofMonth	A number between 1 and 31	int
4.	DayOfWeek	A number between 1 and 7	int
5.	DepTime	A number between 0 and 2400	int
6.	UniqueCarrier	Two-character airline code	cat
7.	Origin	Three-letter departure airport code	cat
8.	Dest	Three-letter destination airport code	cat
9.	Distance	Flight distance in miles	int
10.	DelayTime	Y/N indicating a delay of > 20 min.	bin

Approaches

- Data preparation: R programming
- Data Analysis: R, Python
- Model Building:

1) Decision Trees using R
2) Random Forest using Python
3) Gradient Boosting using R

- Model evaluation: Accuracy, Sensitivity, Specificity

No Delay or Delay Prediction

Based on the variable selection, used Month, Origin, Destination, Unique Carrier, Departure Hour to train the single decision tree.

Decision tree

	Predicted NoDelay	Predicted Delay
Actual No-Delay	tn =688537	fp =64883
Actual Delay	fn = 369438	tp =148169

Sensitivity(recall)	**(probability to identify a true positive):**	**69.55%**
Specificity	(probability to identify a true negative):	65.08%
Accuracy:	(probability of a correct identification):	65.83%

Random Forest

	Pred No-Delay	Pred Delay
Actual No-Delay	tn =221555	fp =12275
Actual Delay	fn = 32582	tp =11061

Sensitivity(recall)	**84%**
Specificity	36%
Precision:	81%

Gradient Boosting

	Predict No-Delay	Predict Delay
Actual No-Delay	tn = 663824	fp = 57826
Actual Delay	fn = 394151	tp = 155226

Sensitivity(recall)	**72.86%**
Specificity	62.74%
Accuracy:	67.80%