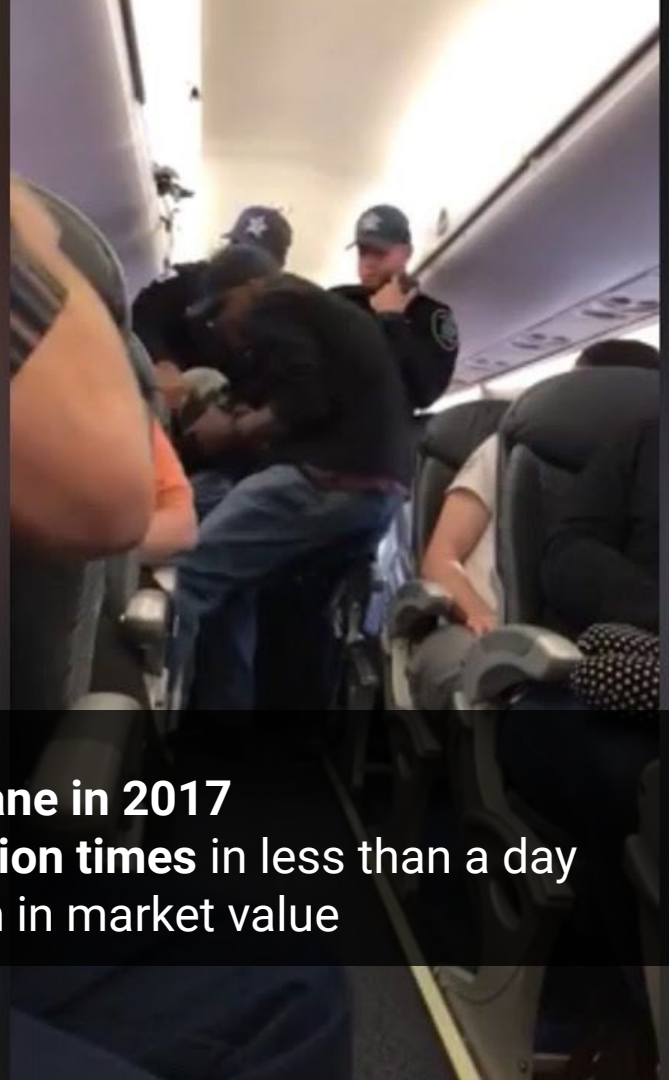
Kaggle competition

Exploratory data analysis and prediction of passenger satisfaction for a US airline company

François LE GAC

2020

The Kaggle competition : https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction



A passenger dragged off a plane in 2017

The video was viewed **6.8 million times** in less than a day

The airline company lost **$1Bn** in market value

Problem statement

The previous controversy shows how ensuring good customer experience is critical for a company

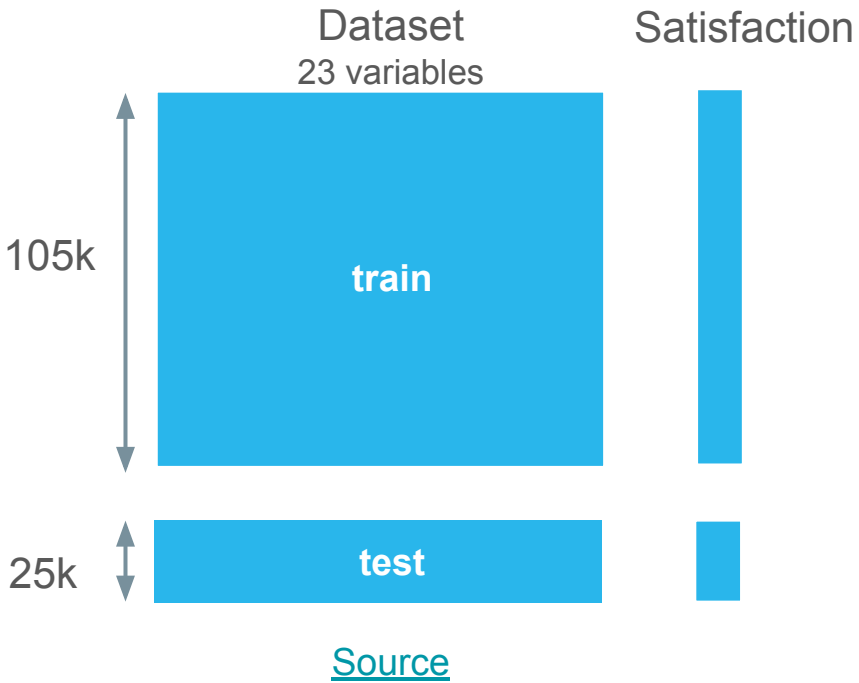
Objectives: Understand what drives passenger overall satisfaction and predict if a passenger will be satisfied or not.

Three high stakes:

1. higher retention rates
2. build a solid reputation
3. acquiring new customers

Dataset description

Kaggle Dataset



Source

Type of variables

- **Client variables**: Gender, Age, Customer type (loyal vs disloyal)
- **Flight info**: Class, Flight distance, Type of travel (Business vs Personal) etc...
- **Satisfaction variables (0-5):** inflight wifi service, cleanliness, seat comfort etc...

Hypothesis, the most important variables are: flight delay, baggage handling and on-board service

Dealing with missing values

Only one variable contains missing value: Arrival delay, with **393** missing values

How to deal with it?

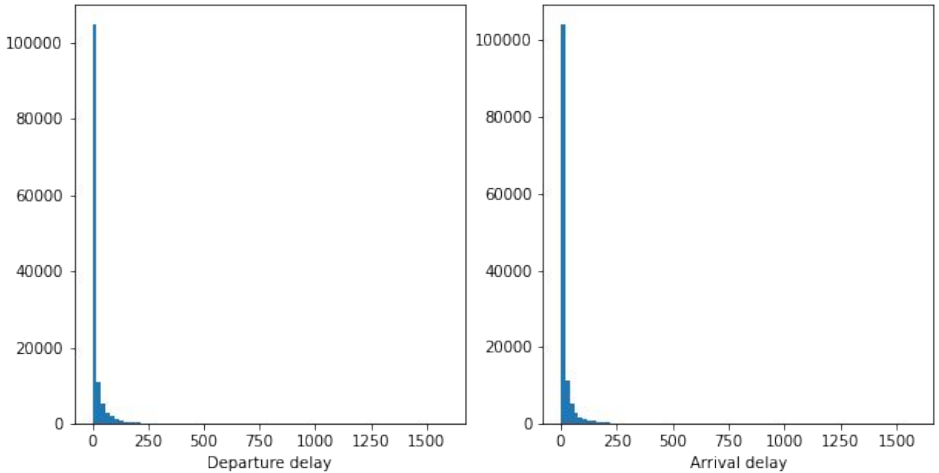


Fig.1 - Full histograms of departure and arrival delays

Two histograms almost layered ➡ we fill the variable accordingly

Univariate data analysis

The target: satisfaction

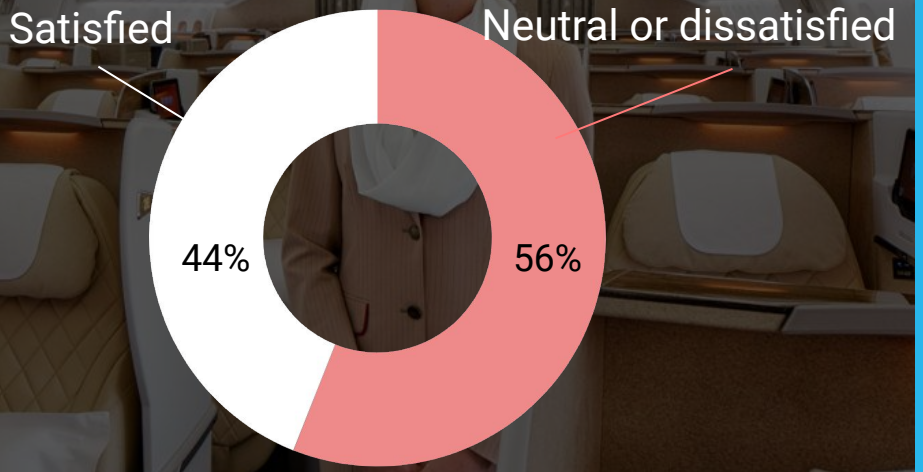


Fig.2 - Pie chart of the satisfaction variable

Univariate data analysis

- Gender - Males vs Females 50:50
- Class - Business, Eco vs Eco+ 45:45:10
- Loyal vs disloyal customers 80:20
- Business flights vs personal flights 70:30

Univariate data analysis

Flight distances

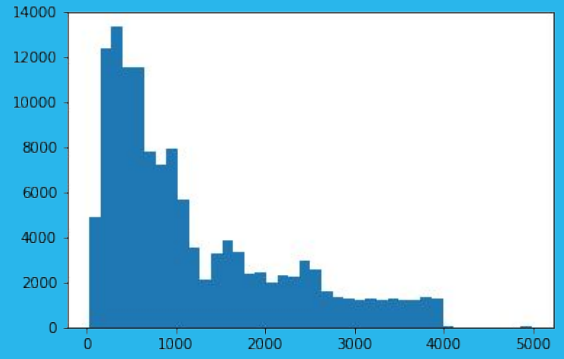


Fig.3 - Histogram of flight distances

Max: ~5,000km
Min: ~ 30km
Creation of a new variable:

- short-haul [0-1,000 km] 60%
- medium-haul [1,000-3,000 km] 35%
- long-haul [> 3,000 km] 5%

Age

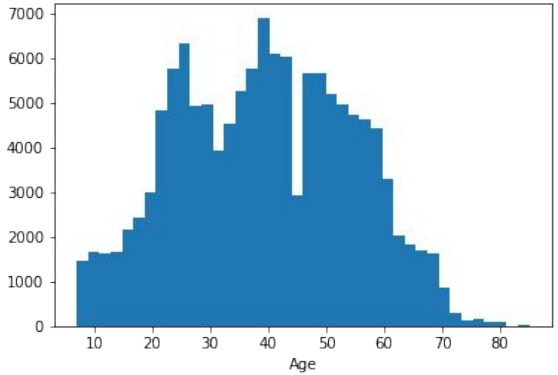


Fig.4 - Histogram of age

Mean: 40 years old

Satisfaction variables

People were asked to grade from 0 to 5 each service

Customers liked these experiences most

Variable name	Mean sat.
Inflight service	3.64
Baggage handling	3.63
Seat comfort	3.44

...these not so much

Variable name	Mean sat.
Gate location	2.97
Ease of online booking	2.75
Inflight wifi service	2.72

Before giving recommendations, are these services important?

A quick example

	Neutral or dissatisfied			Satisfied	
Type of Travel	**Size**	**Percent**		**Size**	**Percent**
Business travel	37337	51	Business travel	52356	93
Personal Travel	36115	49	Personal Travel	4072	7

Fig.14 -Type of travel vs satisfaction

Satisfied customer

- Usually a man
- Fly in business class
- Older than dissatisfied customers ~ 30-50 years old.

- Fly longer distances with a high percentage of long and medium-hauls.
- Almost all traveling for business purposes.
- Usually experiencing less delays

Loyal customer

- Usually a women
- Equally flying for business purposes as for personal purposes

- Fly Long distances
- Business class while most of the disloyal customers are in eco class.

Logistic regression

- Model of Yi knowing Xi:

$\mathbb{P}(y_i = 1 \mid x_i) = \sigma(x_i^\top w + b)$ with $\sigma(z) = \frac{1}{1+e^{-z}}$ the sigmoid function

- We find the best parameters using the likelihood $L(w,b) = \frac{1}{n}\sum_{i=1}^{n} l(y_i, x_i^\top w + b)$

Where $\ell(y, y') = \log(1 + e^{-yy'})$ is the cost function

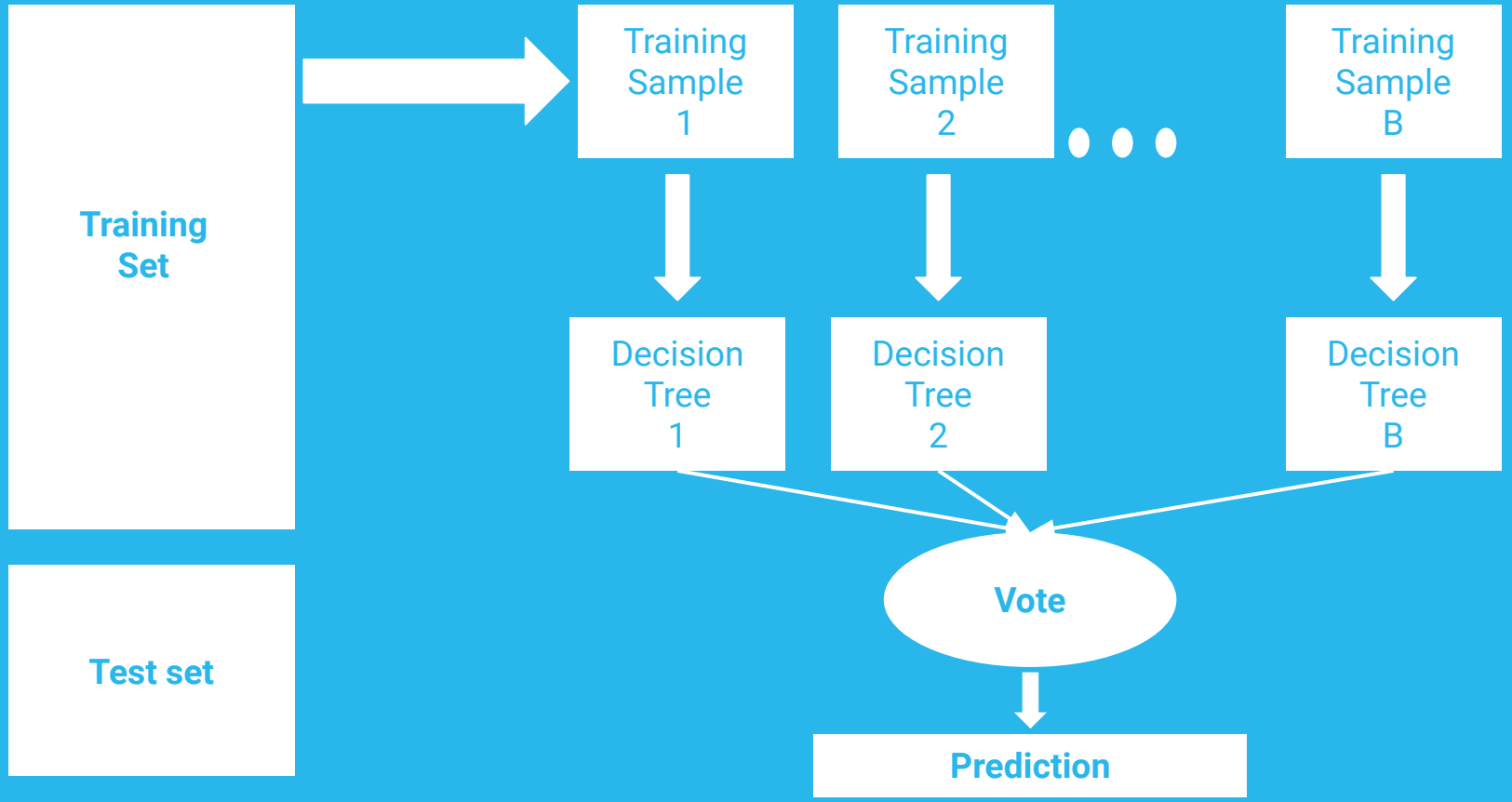
We minimize $argmin(-\log(L(w,b)) + C * \|w\|)$

log-likelihood

Parameter to adjust

The penalty l1 or l2

Random Forest



Performances

Two types of errors:

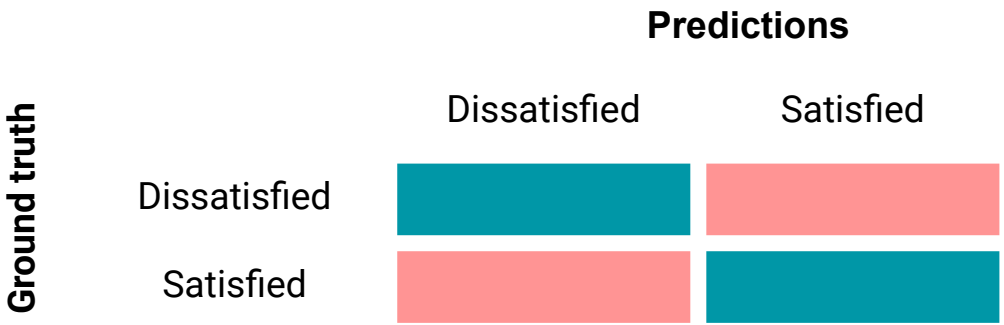


Fig.5 - 2 types of errors

Logistic regression: **87% accuracy***
We improve with Random forest: **93% accuracy**

**accuracy = % of good responses*

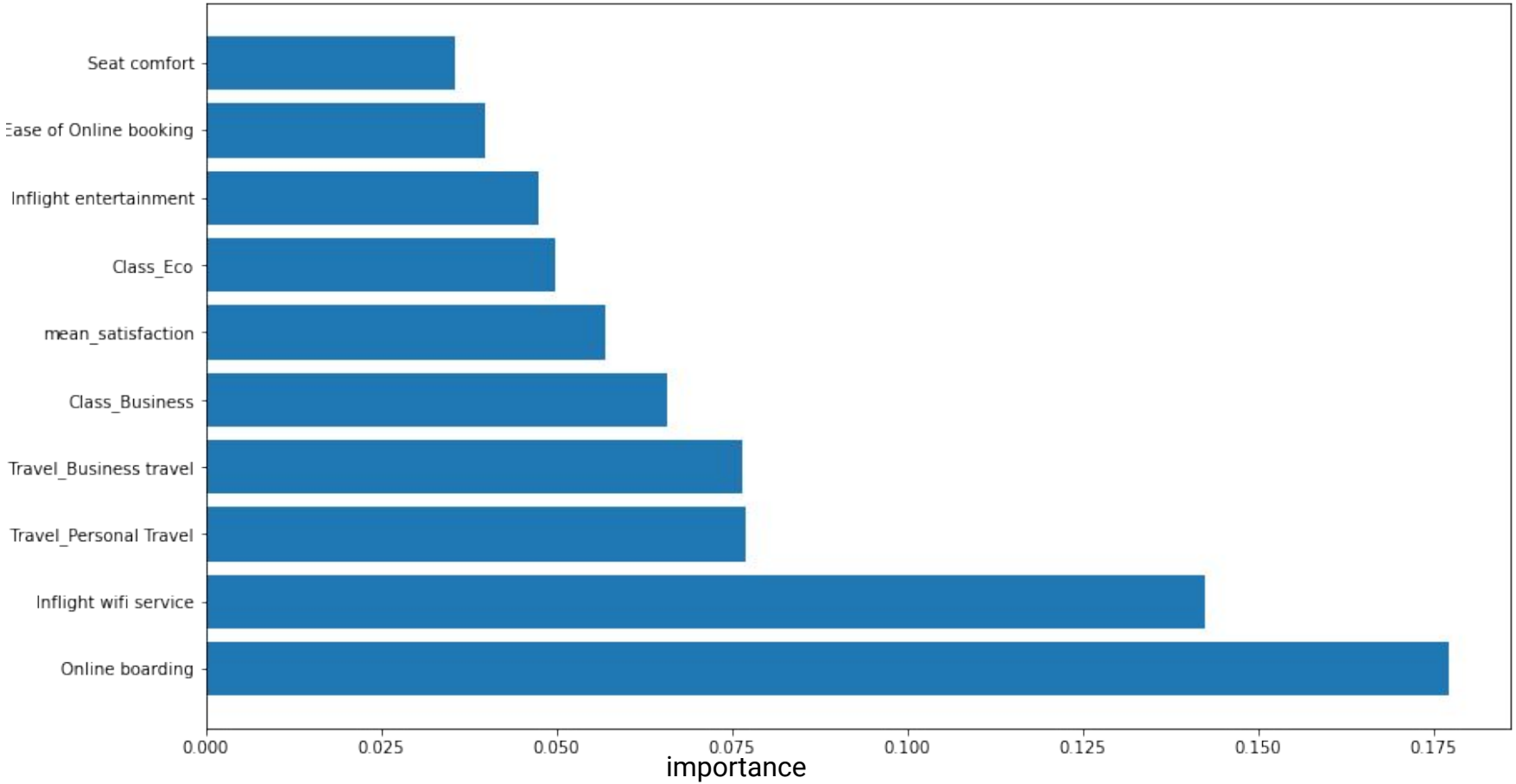


Fig.6 - 10 most important features (from RF)

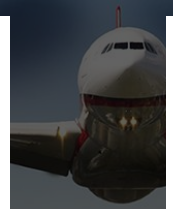
Recommendations

Online boarding experience

Mean satisfaction score: **3.25**

Inflight wifi service

Mean satisfaction score: **2.72** (worst service)



Maintain Inflight entertainment quality

Mean satisfaction score: **3.35**

ANNEXE

Type	Variable	Description
Customer info	Gender	Gender of the passengers (Female, Male)
	Customer Type	The customer type (Loyal customer, disloyal customer)
	Age	The actual age of the passengers
Flight info	Type of Travel	Purpose of the flight of the passengers (Personal Travel, Business Travel)
	Class	Travel class in the plane of the passengers (Business, Eco, Eco Plus)
	Flight Distance	The flight distance of this journey
	Departure Delay in Minutes	Minutes delayed when departure
	Arrival Delay in Minutes	Minutes delayed when Arrival
Satisfaction	Inflight wifi service	Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
	Departure/Arrival time convenient	Satisfaction level of Departure/Arrival time convenient
	Ease of Online booking	Satisfaction level of online booking
	Gate location	Satisfaction level of Gate location
	Food and drink	Satisfaction level of Food and drink
	Online boarding	Satisfaction level of online boarding
	Seat comfort	Satisfaction level of Seat comfort
	Inflight entertainment	Satisfaction level of inflight entertainment
	On-board service	Satisfaction level of On-board service
	Leg room service	Satisfaction level of Leg room service
	Baggage handling	Satisfaction level of baggage handling
	Checkin service	Satisfaction level of Check-in service
	Inflight service	Satisfaction level of inflight service

Dummy variables

First example:

Gender
Male
Female
Female
Male

Gender_male
1
0
0
1

Second example:

Class
Business
Eco
Eco +
Business

Class_Business	Class_Eco
1	0
0	1
0	0
1	0