



# Flight price prediction in the USA

Group 18 Google transmission

111550114 趙堉安 111550060 劉千慈 111550074 陳映竹



01

# Introduction

# Introduction

- **Overview**

Developing an model that can predict airline ticket prices between 16 airports in the USA

- **Motivation**

1. Evaluating the going rate of economy classes in all airline.
2. Planning a surprise visit to our friend.

# Introduction

Using airport ATL to be an example:



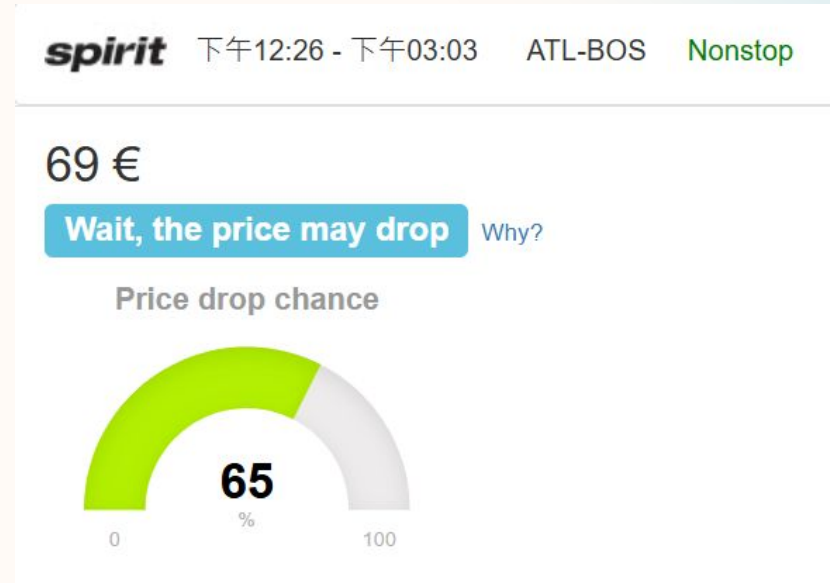


02

# Related work

# Airfare prediction websites

- Most of them predict the price trend in the coming weeks or whether to buy the tickets
- The difference is that we focus on the current going rate



# Papers about flight price prediction

- Most papers include all classes of tickets and use a numerous dataset to predict flight price.
- We want to train our model with the following three characters:
  1. Only using the data of economy class.
  2. Using a smaller dataset with limit time interval.



03

# Dataset/Platform

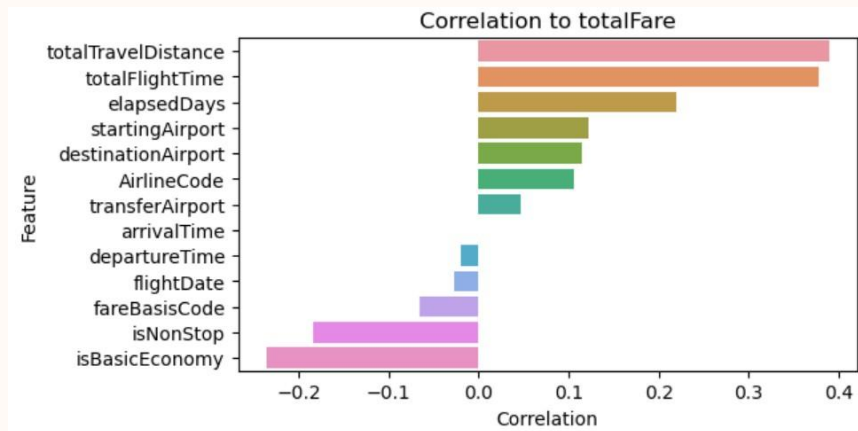


# Dataset/Platform

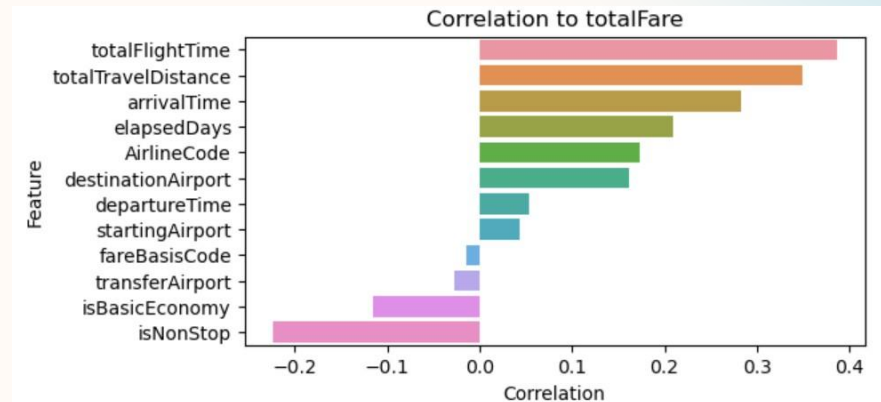
kaggle

- Fetching the dataset from Kaggle
- Preprocessing the data
  1. Selecting the flights on 2022-04-17 and economy class.
    - size: 60000000 -> 80000 -> 7000 rows
  2. Keeping the effective features.
    - Ex: total flight time, travel distance...
- Separating the data into train and test. (4:1)

# Initial data analysis



80000-row dataset



7000-row dataset



04

**Baseline**

# Baseline

- **Decision tree algorithm**

- Using a single tree to sort all data points into several sections.

- STEP:



## **Initialization and split into groups**

- Calculate the loss and find the best splitting point with MSE.
- Split the dataset into at most `min_samples_split` groups.



## **Iterative the process above.**

- It will continue until the depth is equal to `max_depth`.



## **Obtaining a tree with prediction.**

- Data will be separated into several leaves and each leaf will represent the prediction under that specific situation.

# Baseline

- **Decision tree algorithm**
- Reason:
  1. Complex non-linear relationships between features in data.
  2. It is suitable in classification and regression problems.
- Limitation:
  1. Risk of overfitting
  2. It can't properly handle high dimension data or data with lots of missing.



05

# Main Approach

# Input & Output

## Input

**1. Transfer the non-number data to number with pandas**

**2. features in dataset:**

total distance, total flight time, elapsed day, destination, airline, transfer airport, arrival time, departure time, flight date, fare basis code, nonstop and basic economy

**3. Separate two sub dataset for training and testing**

## Output

- The predicted price of flight

# Method

## 1. Random Forest

- It is constructed by `n_estimator` random trees.
- STEP:

- **Selecting a dataset randomly from the original dataset**

- size will be the same as the original one
- and the data can be repeated

- **Applying decision tree method & construct a forest.**

- In our case, the forest contains 50 trees.

- **obtaining the final result of prediction**

- we take the average of each prediction result of the trees



# Method

## 2. Gradient boosting

- Combining simple models to build a strong model.
- STEP:
  - **Getting initial loss value.**
    - loss value between current result and the real value will be calculated.
  - **Iterative training.**
    - take part of loss value to update and optimize the prediction.
    - Training model with `n_estimator` times
  - **Updating the prediction every round.**



06

# Evaluation Metric

# Evaluation Metric

## 1. Quantitative

- RMSE
- $R^2$  score
- Accuracy

## 2. Qualitative

- It is not suitable to use qualitative evaluation metric, since our implementation is totally constructed by numerical value.

# Quantitative

## 1. RMSE

Measure the average difference between the model's predicted values and the actual values.

$$RMSE = \sqrt{\sum_{i=1}^n \frac{(\hat{y}_i - y_i)^2}{n}}$$

## 2. R<sup>2</sup> score

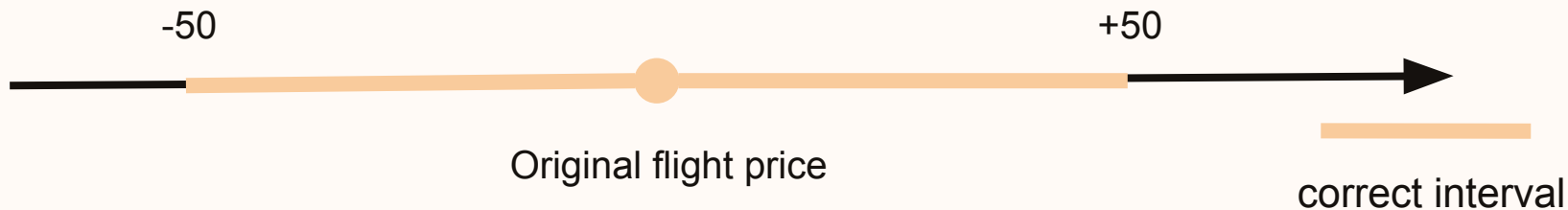
Measure the extent to which the model explains the variance in the data.

$$R^2 = 1 - \frac{SS_{RES}}{SS_{TOT}} = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

# Quantitative

## 3. Accuracy

it can be used to measure the model's performance in predicting the correct interval.



we compare the prediction result to the original total fare.

If the error is like above, we assume it is accurate and calculate the whole accuracy.



07

# Results & Analysis

# Result & analysis

- Decision tree:

```
RMSE: 154.64495591166593  
R2 score: 0.49395996907805195  
Accuracy: 53.806781829814454 %
```

- Random forest:

```
RMSE: 127.68198368653677  
R2 score: 0.6550369348689498  
Accuracy: 61.228406909788866 %
```

- Gradient boosting:

```
RMSE: 119.30057359567645  
R2 score: 0.6988392160832606  
Accuracy: 69.80166346769033 %
```

- Conclusion:

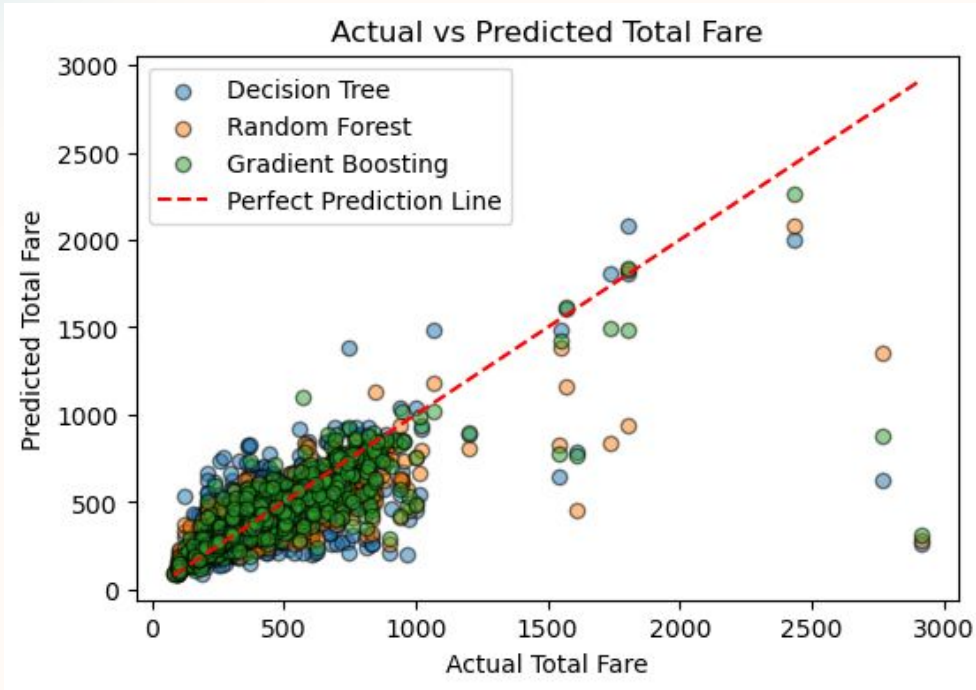
- Performance rank:

1. Gradient boosting
2. Random forest
3. Decision tree

- R<sup>2</sup> score v.s. Accuracy

1. The outcome of last two is similar in R<sup>2</sup> score, but different in accuracy.
2. It may be caused by the number of tree.

# Result & analysis



- Conclusion:

- Distribution:

1. Gradient boosting is more scattered than others
2. Random forest and Gradient boosting are slightly different, but better than the decision tree.



# Result & analysis

Selected row:

	startingAirport	destinationAirport	isBasicEconomy	isNonStop	\
3342	EWR	SFO	False	False	
	AirlineCode	transferAirport	departureTime	arrivalTime	totalFlightTime \
3342	UA DL DL	DCA MSP	1650222000	1650263340	29760
	elapsedDays	totalTravelDistance	fareBasisCode	baseFare	totalFare
3342	0	2725	HAA0AFEN	536.74	605.1

Predicted totalFare:

Decision Tree: 514.09

Random Forest: 648.3927333333336

Gradient Boosting 550.4990379737137

- Conclusion:

By observing, it is unusual that the random forest performs better.

The reason may be setting of hyper parameters, learning rate and subsamples.



08

**Github link**

**“Our Github link.”**

– Click it, thanks!



09

# Reference

# Reference

## Related paper

- RWA: A Regression-based Scheme for Flight Price Prediction

## Dataset

- Flight Prices (kaggle.com)

## Related knowledge

- RMSE
- R2\_SCORE



10

**Contribution of each member**

# Contribution of each member

<b>111550114 趙堉安</b>	Preprocessing data, Programming, topic conceiving. (33.33%)
<b>111550060 劉千慈</b>	Preprocessing data, making slides, topic conceiving. (33.33%)
<b>111550074 陳映竹</b>	Preprocessing data, Presentation, topic conceiving. (33.33%)

# Thanks for listening!

CREDITS: This presentation template was created by **Slidesgo**, and includes icons by **Flaticon** and infographics & images by **Freepik**