This is the dataset from the paper To buy or not to buy: mining airfare data to minimize ticket purchase price (2003) by Oren Etzioni , Rattapoom Tuchinda , Craig A. Knoblock , Alexander Yates

BibTex

@INPROCEEDINGS{Etzioni03tobuy,  
    author = {Oren Etzioni and Rattapoom Tuchinda and Craig A. Knoblock and Alexander Yates},  
    title = {To buy or not to buy: mining airfare data to minimize ticket purchase price},  
    booktitle = {In Proceedings of KDD’03},  
    year = {2003},  
    pages = {119--128},  
    publisher = {ACM Press}  
}

Throughout the years people have asked for this dataset, we decide to release it so interested researcher can play around with it. The data is provided as is and it is not possible to provide the code that runs it because the IP and Patents have been licensed to Microsoft (see the PricePredictor on bing.com/travel).

**About the data:**

1. The data set contains only direct flight information in two routes (LAX-BOS and SEA IAD). The reason for using two routes is so that the model created is not tilted toward the behavior of a particular single route.

2. There are about 12000 data points. These are divided into two set. Training set and testing set. Figure 1 will help you understand more

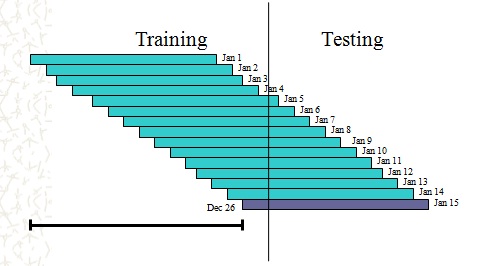


Fig 1

There are many facets of data. For example, we are collecting flights that depart on Jan1, Jan2,….. Jan15. The flight number might be the same but the departure date is different. We can use this to help mine the characteristic of a particular flight or a particular airline.

Anyway, to build a prediction model, we cannot look ahead into the future, so we divide the training and testing set along those lines. As a result, there are some portion of flights that belong in both training and testing set.

You can divide up the training/testing as you’d like. Just be aware that to make a prediction model, you shouldn’t look ahead into the future.

3. Now let’s explain the data itself

id: AA192-null-aa223-null means it’s a direct flight AA192 (depart) - AA223 (return). We had null because it’s a format we use to extract flights with 2 legs. It’s not really used in this dataset.

date\_time: the time the data is recorded.

best\_price: the best price you can possibly get if you have 100% knowledge of the future. This is not used in the creation of model. It’s used to measure the performance of the algorithm after the fact.

Observation\_count: how many ticks until the flight take off. One tick is 15 minutes.

ripper\_predict: we use multiple algorithm to decide whether to wait or to buy. We also use ensemble learning. ripper\_predict is one of the output of the learning algorithm. It’s by no means a complete, but we leave it there as an example.