Flights Delay Prediction

Bing, Rita, Eric, Zac

Outline

- Business Scenario
- Business Problem
- Machine Problem
- Data Set Information

Business Scenario

- 航空公司想要提供一項服務
- 旅客在購買機票時便告知航班是否會因天氣而導致誤點，以便提前做好準備，提高顧客體驗(減少損失)。
- 給予2013-2018的航班data，希望藉由ML的方式預測航班是否會因天氣而誤點。

Scenario Analysis

- 適合用ML來解決這個問題的原因
    - 難以用人力去閱讀數據以及預測
    - 大量資料需要處裡以及自動化
- 商業上的影響
    - 誤點太久旅客可要求全額退費，對航空公司造成損失

航班誤點 5 小時可全額退費 交通部修正調處辦法

法源編輯室/ 2017-09-18 [評論數 0 篇]

交通部日前修正「民用航空乘客與航空器運送人運送糾紛調處辦法」，航班誤點超過五小時，若旅客不接受航空公司安排可要求全額退費，且不得收取手續費；即便是颱風天等不可控因素也適用。

民航局指出，以往並無明文規定延誤多久可以退票，若旅客不願接受航空公司安排也無法退費，因此新修民用航空乘客與航空器運送人運送糾紛調處辦法第 3 條第 2 項，明訂延誤五小時為退費時限。除台灣，歐盟也有類似做法。另，同法條第 1 項規定，在確定航班無法按時啓程後，境內航線延誤十五分鐘以上、國際航線延誤三十分鐘以上，或變更航線、起降點，航空公司都應向乘客說明。

民航局進一步說明，又考量航空器因故轉降其他機場後，如無法短期內飛往原目的地機場，為使乘客儘早抵達目的地或因應其特殊狀況需求（如健康因素），增訂民用航空乘客與航空器運送人運送糾紛調處辦法第 3 條之 1 規定，於機場條件允許下，可由業者協調機場相關單位，安排乘客由轉降機場下機或入境，以符實需。

Business Problem

- 飛機因為天氣而誤點太久造成對航空公司的損失
- 透過預測是否會誤點，在旅客買票時提前告知可能會誤點，讓旅客有心理準備
- 若在告知有可能誤點的情況下旅客仍選擇購買，則航空公司可選擇不予退費(or 部分退費)
- 簡化問題所以忽略部分退費的情況，考慮若delay超過15分鐘則退費

Business Goal

- 減少因天氣誤點而對航空公司造成的損失(i.e 顧客退費)
- Success metrics: 因天氣而退費給顧客金額比例降低

ML Problem

- 給予航班的data，預測是否班機會delay超過15分鐘
- 適合使用二元分類演算法(是否會delay超過15分鐘)
- 使用recall作為評估指標 (因為航空公司希望能夠盡量減少誤點造成的損失)
- False Positive 較沒影響(因為沒誤點，皆大歡喜)

DataSet Information

- 日期
- 起飛/降落機場
- 預計到達時間/實際到達時間
- Delay Time 相關attribute

CRSDepTime	CRS Departure Time (local time: hhmm)
DepTime	Actual Departure Time (local time: hhmm)
DepDelay	Difference in minutes between scheduled and actual departure time. Early departures show negative numbers.
DepDelayMinutes	Difference in minutes between scheduled and actual departure time. Early departures set to 0.
DepDel15	Departure Delay Indicator, 15 Minutes or More (1=Yes)
DepartureDelayGroups	Departure Delay intervals, every (15 minutes from <-15 to >180)
DepTimeBlk	CRS Departure Time Block, Hourly Intervals
TaxiOut	Taxi Out Time, in Minutes
WheelsOff	Wheels Off Time (local time: hhmm)
WheelsOn	Wheels On Time (local time: hhmm)
TaxiIn	Taxi In Time, in Minutes
CRSArrTime	CRS Arrival Time (local time: hhmm)
ArrTime	Actual Arrival Time (local time: hhmm)
ArrDelay	Difference in minutes between scheduled and actual arrival time. Early arrivals show negative numbers.
ArrDelayMinutes	Difference in minutes between scheduled and actual arrival time. Early arrivals set to 0.
ArrDel15	Arrival Delay Indicator, 15 Minutes or More (1=Yes)
ArrivalDelayGroups	Arrival Delay intervals, every (15-minutes from <-15 to >180)
ArrTimeBlk	CRS Arrival Time Block, Hourly Intervals

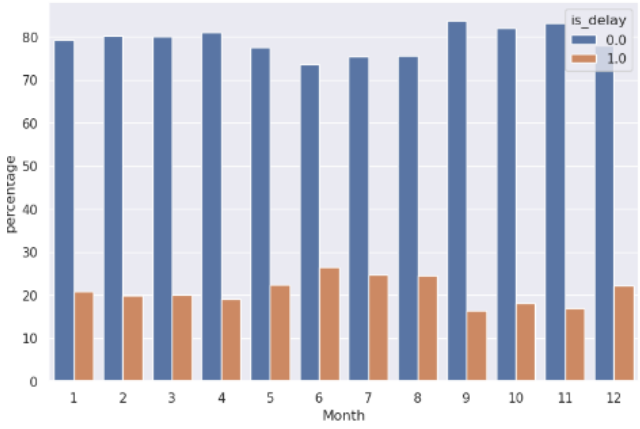
```
print("The #rows and #columns are ", data.shape[0] , " and ", data.shape[1])
print("The years in this dataset are: ", list(data.Year.unique()))
print("The months covered in this dataset are: ", sorted(list(data.Month.unique())))
print("The date range for data is :" , min(data.FlightDate), " to ", max(data.FlightDate))
print("The airlines covered in this dataset are: ", list(data.Reporting_Airline.unique()))
print("The Origin airports covered are: ", list(data.Origin.unique()))
print("The Destination airports covered are: ", list(data.Dest.unique()))
```

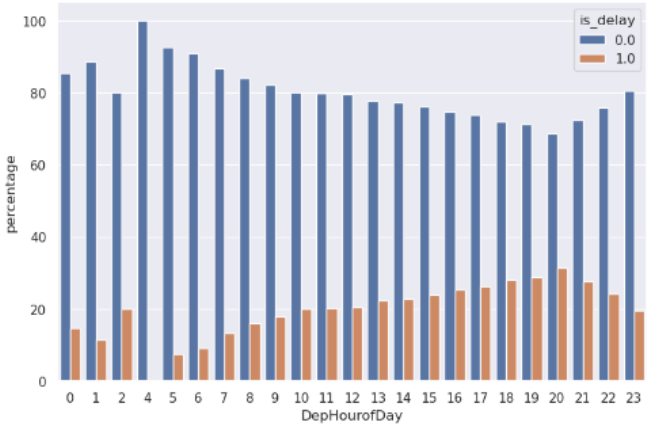
```
The #rows and #columns are  1658130  and  20
The years in this dataset are:  [2015, 2016, 2014, 2018, 2017]
The months covered in this dataset are:  [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12]
The date range for data is :  2014-01-01  to  2018-12-31
The airlines covered in this dataset are:  ['AA', 'DL', 'UA', 'WN', 'OO']
The Origin airports covered are:  ['CLT', 'DFW', 'ORD', 'LAX', 'SFO', 'PHX', 'IAH', 'DEN', 'ATL']
The Destination airports covered are:  ['DFW', 'ORD', 'ATL', 'PHX', 'SFO', 'LAX', 'IAH', 'DEN', 'CLT']
```

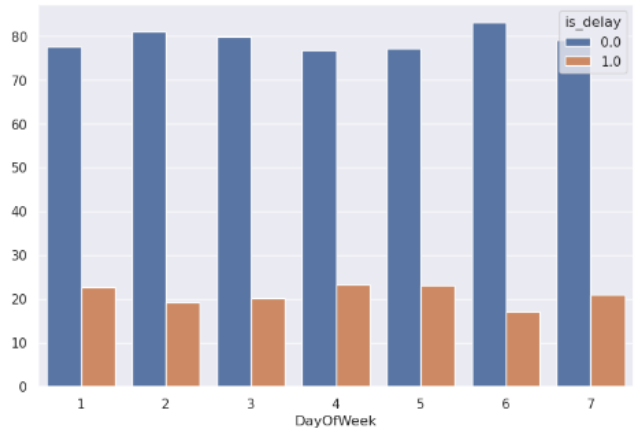
Thoughts

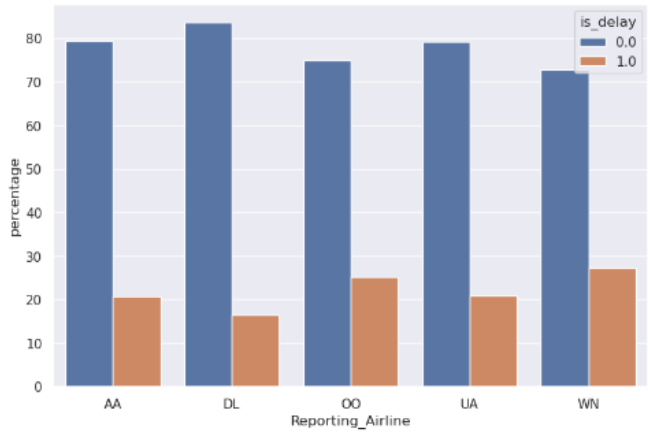
- 天氣通常與季節循環有關(日期,時間)

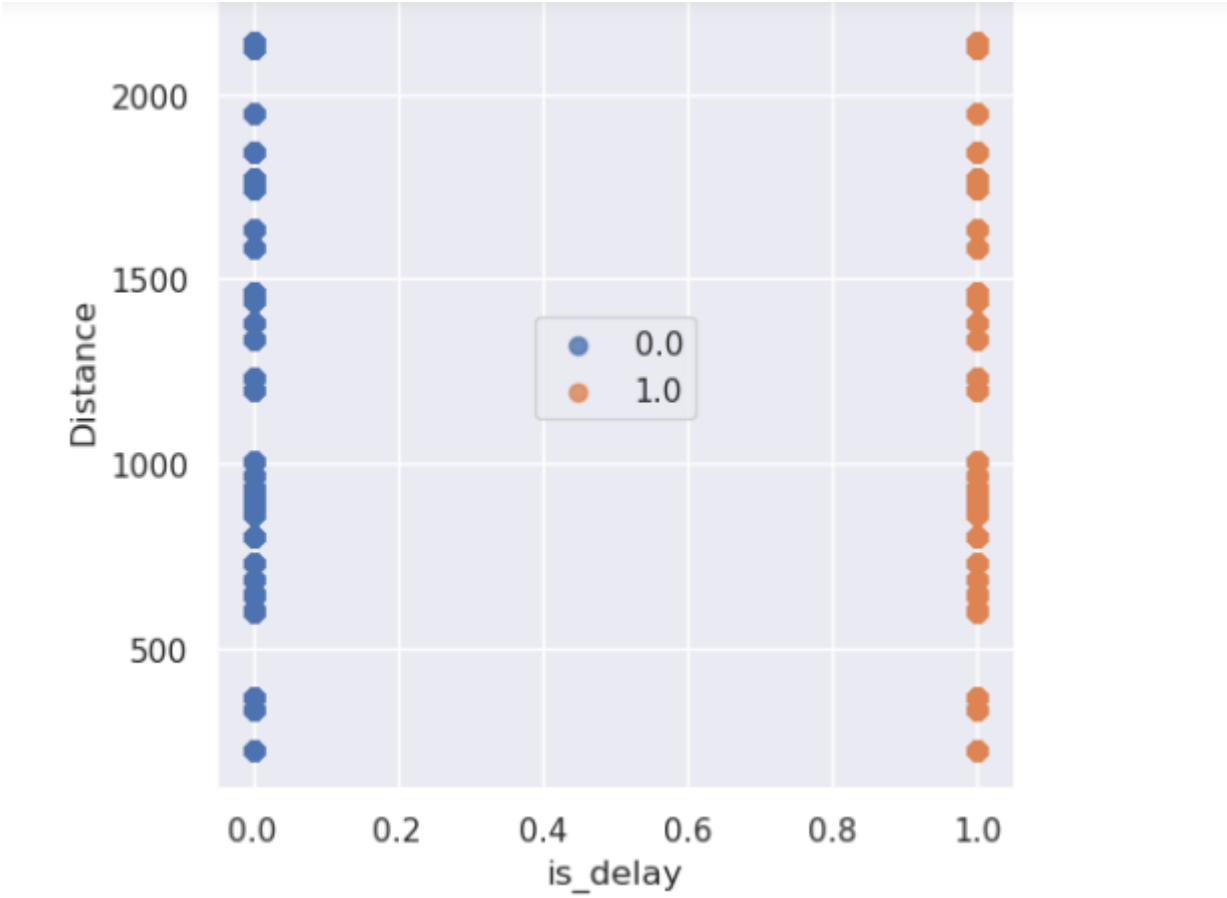
DataSet Analysis











Model Evaluation