



Cancellation Prediction Model

Group 7

Ploy, Philip, Sheila, Thuy

Data: 2013 - 2017

Source: [Bureau of Transportation Statistics](#)

Agenda

1. Executive Summary
2. Descriptive Analysis - Key Findings
3. Prediction Model
4. Recommendations



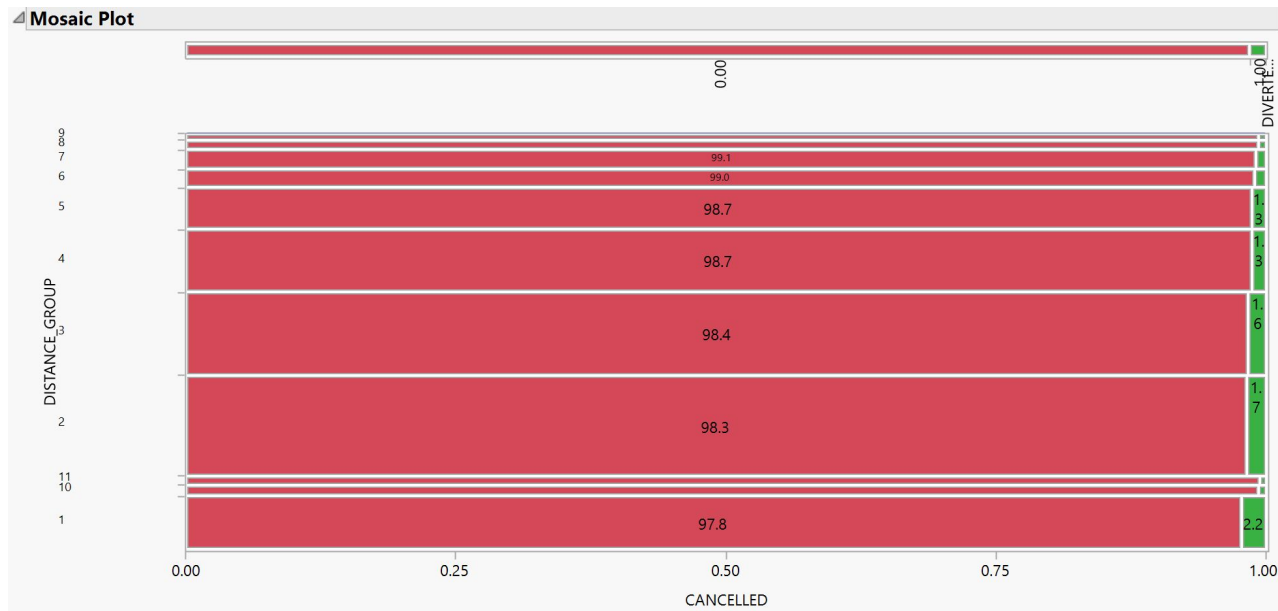
Executive Summary

- Analyze as Data Scientists & Product Managers at Skyscanner
- **Business Initiative:** *Develop and monetize a new feature in Flight Search Engine that predicts flight cancellation.*
- **Historical data:** 2013 - 2017, 47 million rows, 4GB .csv file

Descriptive Analysis



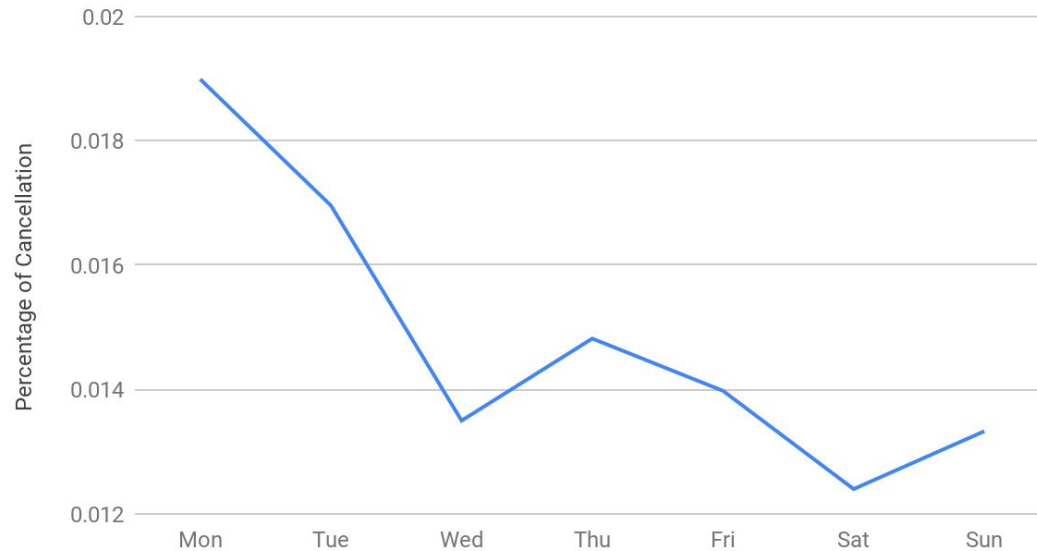
Cancellation and Distance



Distance group 1 (less than 250 miles) had **highest** percent of cancellation

Percentage of Cancellation by Day of Week

Percentage of Cancellation



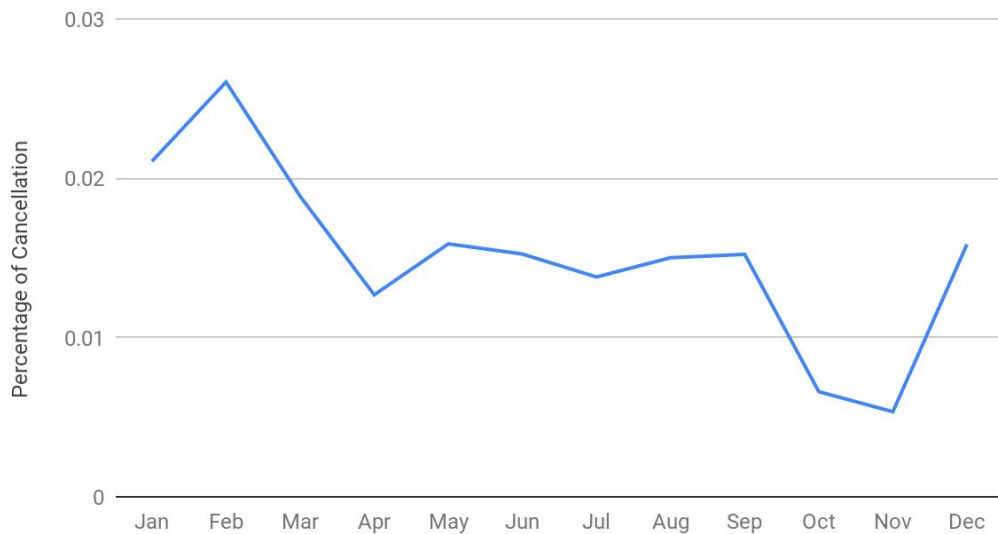
Monday = highest cancellation

Saturday = lowest cancellation

Weekend had the **better** cancellation performance **than weekdays**

Percentage of Cancellation by month

Percentage of Cancellation by Month

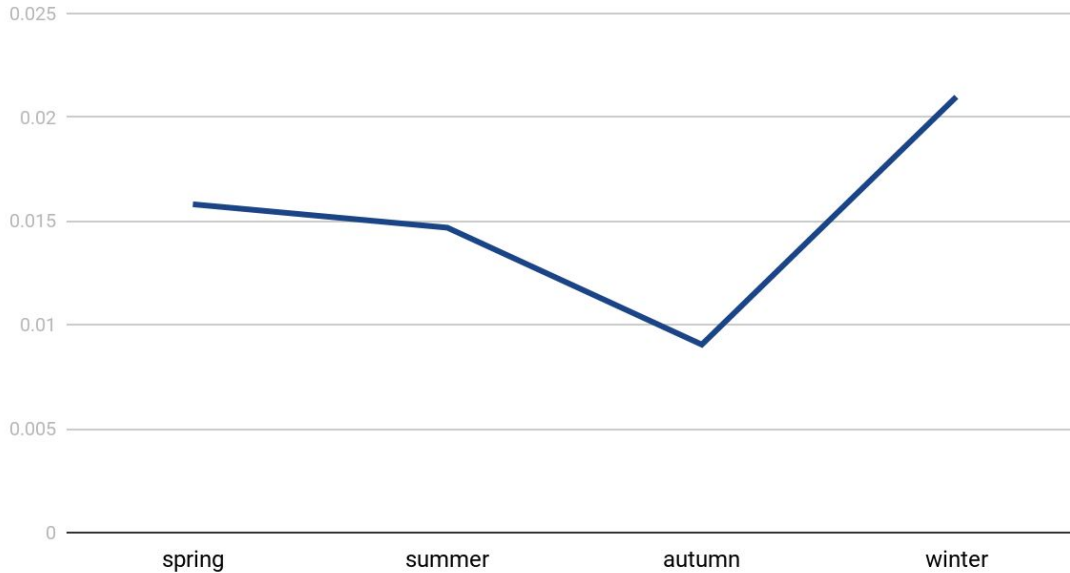


November had the **lowest** percent of cancellation

February had the **highest** percent of cancellation

Percentage of Cancellation by Season

Percentage of Cancellation by Season

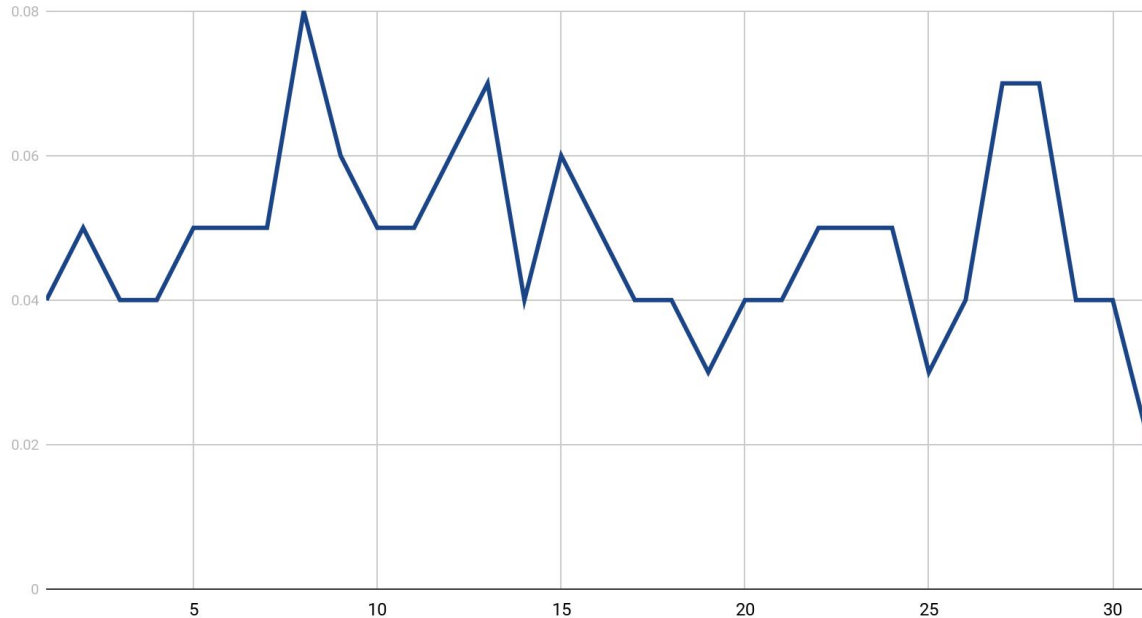


Autumn had the lowest percent of cancellation

Winter had the highest percent of cancellation

Percentage of Cancellation by Day of Month

Percentage of Cancellation by Day of Month



The cancellation tended to be **high** during the **first** and the **last week of months**

And **lower** in the **middle** of months

Predictive Analysis



Prediction Model on R

Model: Call:

```
lm(formula = CANCELLED ~ YEAR + QUARTER + MONTH + DAY_OF_MONTH + OP_CARRIER_FL_NUM + ORIGIN_AIRPORT_ID +  
DEST_AIRPORT_ID + CRS_DEP_TIME + DEP_TIME + DISTANCE_GROUP, data = mydata)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.00453	-0.00088	-0.00061	-0.00036	1.00301

Coefficients: Estimate Std. Error t value Pr(>|t|)

(Intercept)	9.554e-02	5.181e-03	18.443	< 2e-16 ***
YEAR	-4.706e-05	2.573e-06	-18.292	< 2e-16 ***
QUARTER	-2.982e-04	1.580e-05	-18.870	< 2e-16 ***
MONTH	6.219e-05	4.912e-06	12.661	< 2e-16 ***
DAY_OF_MONTH	-2.652e-07	4.186e-07	-0.634	0.526
OP_CARRIER_FL_NUM	1.376e-07	2.184e-09	63.021	< 2e-16 ***
ORIGIN_AIRPORT_ID	-2.060e-08	2.419e-09	-8.518	< 2e-16 ***
DEST_AIRPORT_ID	-1.166e-08	2.418e-09	-4.822	1.42e-06 ***
CRS_DEP_TIME	-1.138e-06	2.798e-08	-40.681	< 2e-16 ***
DEP_TIME	1.470e-06	2.719e-08	54.045	< 2e-16 ***
DISTANCE_GROUP	-2.567e-05	1.631e-06	-15.735	< 2e-16 ***

Linear Regression Equation

The probability of Flight Cancellation =

$$\begin{aligned} &0.09554 - 0.00004706(\text{Year}) - 0.0002982(\text{Quarter}) + \\ &0.00006219(\text{Month}) - 0.0000002652(\text{Day of Month}) + \\ &0.0000001376(\text{Flight Number}) - 0.000000206(\text{Origin Airport ID}) - \\ &0.00000001166(\text{Destination Airport ID}) - 0.000001138(\text{CSR Dep} \\ &\text{Time}) + 0.00000147(\text{Dep Time}) - 0.00002567(\text{Distance Group}) \end{aligned}$$

Model Testing Outcome

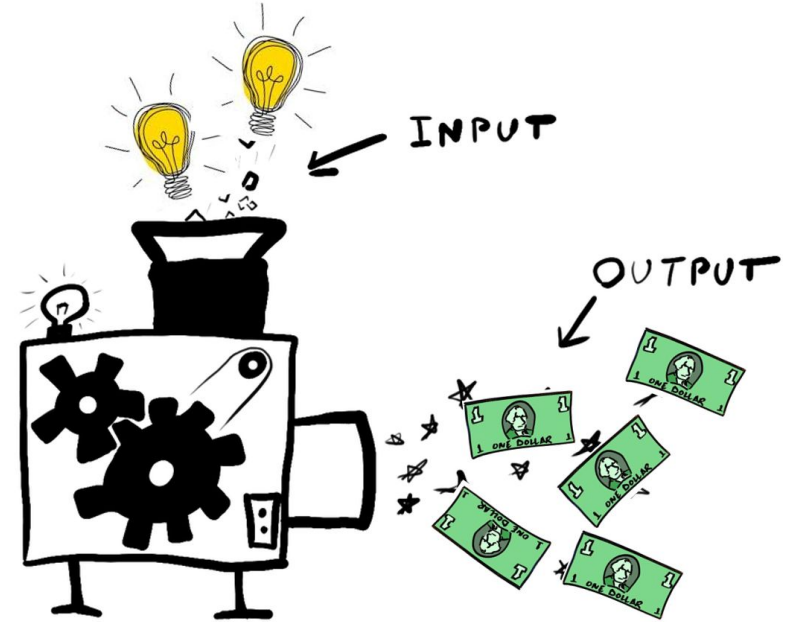
Random sample of 12 months from 2013 - 2017

Run predictive model on this random sample

Calculate diff of probability vs actuals

Mean(diff) = .0000397

Prescriptive Analysis & Recommendations





5:30 PM

JFK

6h 51m

Non-stop



9:21 PM

SFO



4:10 PM

SFO

5h 23m

Non-stop



12:33 AM⁺¹

JFK

1.2% chance of cancellation

2 deals from

\$583

Select →

Get a cancellation insurance with
[Travelex](#)



4:55 PM

JFK

6h 42m

Non-stop



8:37 PM

SFO



9:00 AM

SFO

5h 25m

Non-stop



5:25 PM

JFK

Less than 0.1% chance of cancellation

10 deals from

\$721

Select →

Book now for a full refund guarantee



7:00 AM

JFK

6h 38m

Non-stop



10:38 AM

SFO



6:29 AM

SFO

5h 30m

Non-stop



2:59 PM

JFK

6% chance of cancellation

1 deal

\$550

Select →

Get a cancellation insurance with
[Travelex](#)

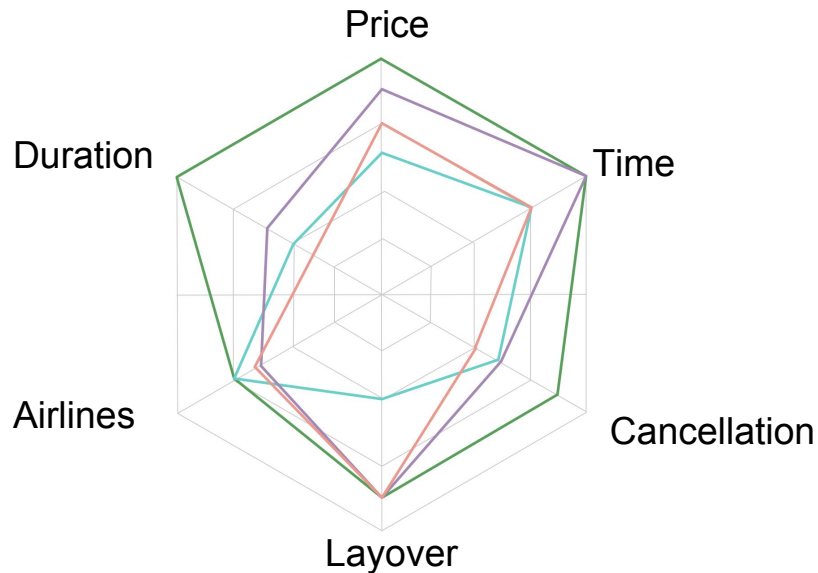
\$\$\$ More
information for
customers

\$\$\$ Affiliation
from Insurance
companies

\$\$\$ Increase of
sales conversion

Looking forward

- Gather customer insights based on click performance
- Build customer profiles based on their preference to cancellation probability
- The prediction model will change according to most up-to-date data. Maintain it by taking out outliers.



Q&A

