Airline Passenger Satisfaction Prediction

ISOM 456 Final Project

Hasong Cho, Anna Choi, Yedda Tao, Will Tung

Agenda

01
Business Understanding

02
Data Understanding

03
Data Preparation

04
Modeling

05
Evaluation

06
Deployment

1.1 Business Understanding: Airline Business Model

Customer Touchpoints
Booking, Check-In, FFP, Baggage Handling, Amenities,
In-Flight Entertainment, Comfort, Timeliness

Travelers

Airlines

Airline Revenues:
Direct from Passengers
From Travel Partners

- In the airline industry:
    - 60% of revenues come direct from passengers (i.e. airfare, fees, and other travel expenses...)
    - 40% come from Travel Partners (i.e. Credit Card Companies, Hotels, Car Rental Companies)
- Out of the 60% of revenues direct from passengers, business travelers are 2x as profitable as other travelers, despite accounting for 12% of all airline passengers

1.2 Business Understanding: Airline Business Model

How can data science improve the financials of an airline?

- Better Customer Satisfaction
  - Better cater to different segments → more satisfied customers → more business -> more revenue from travel partners
  - Better cater to different segments → more satisfied customers → more tolerance for negative experiences -> higher customer loyalty + higher market share
- Lower Costs
  - Classify and understand the impact of strategic decisions → map strategies with customer emotions → eliminate guesswork → impactfully address what matters most → save time and capital investments

1.3 Business Understanding: Airline Business Model

American Airlines

Case Study:
United Airlines vs. American Airlines

- United found that just improving the coffee made customers happier
  American Airlines teamed up with local logistic players to deliver baggage directly to their place, allowing customers to skip the queue.
- *RESULT:* despite their efforts, both these airlines find their place at the bottom of the ACSI (American Consumer Satisfaction Index).

So What?

- *We need data science to find the biggest bottlenecks, what's most important, and what matters least*

1.4 Business Understanding: Predict Customer Satisfaction of Airline Travelers

Statement of problem: Airline travel is consistently rated poorly on the ACSI (American Consumer Satisfaction Index). How can airlines better understand their customers that are most dissatisfied and improve their experiences?

Objective: Build predictive models to predict whether a passenger is satisfied / dissatisfied before on-boarding and post-experience surveys.

Goal (within scope of project):

1. Identify the customer segments most in-need of work using predictive models

Further implementation / For future references:

2. Seek insights on what's most important to customers segments identified in step 1.
3. Turn findings into strategic initiatives

2.1 Data Understanding: Datasets

This data is from

Kaggle (public dataset) and the airline is anonymous.

This data has information on

Survey of airline customers' satisfaction based on customer information, services (internal & external) and flight information.

This data contains

Train set contains 103904 rows & 23 columns and Test set contains 25976 rows & 23 columns.
Two separate train and test dataset are provided, but we will join them and sample for the analysis.

Target Variable is

Customer satisfaction (1: satisfied, 0: dissatisfied or neutral)

2.2 Data Understanding: Types of Features

Categorical

Customer Information
- Gender
- Customer Type
- Type of Travel
- Class

Inflight Services Satisfaction Level
- Inflight Wifi-Service
- Food and Drink
- Seat Comfort
- Inflight Entertainment
- Leg room service
- Baggage Handling
- Inflight Service
- Cleanliness

External Services Satisfaction Level
- Ease of Online booking
- Gate Location
- Online boarding
- Online service
- Check-in service

Flight Information
- Departure/Arrival Time Convenient Satisfaction Level

Customer Satisfaction
- Satisfaction Level

Numerical

Customer Information
- Age

Flight Information
- Flight Distance
- Departure Delay in Minute
- Arrival Delay in Minute

2.3 Data Understanding: Example

- Customer
- Inflight Service
- External Service
- Flight Information
- Satisfaction

- Gender: Male
- Age: 44
- Customer Type: Loyal
- Type of Travel: Business
- Class: Business

- Inflight Wifi-Service: 4
- Food and Drink: 3
- Seat Comfort: 4
- Inflight Entertainment: 3
- Leg room service: 4
- Baggage Handling: 4
- Inflight Service: 5
- Cleanliness: 4

- Ease of Online booking: 3
- Gate Location: 3
- Online boarding: 4
- Online service: 4
- Check-in service: 5

- Flight Distance: 1500
- Departure/Arrival Time Convenient Satisfaction Level: 5
- Departure Delay in Minute: 2
- Arrival Delay in Minute: 1

- Satisfaction Level: satisfied

- Gender: Female
- Age: 30
- Customer Type: Disloyal
- Type of Travel: Personal
- Class: Eco

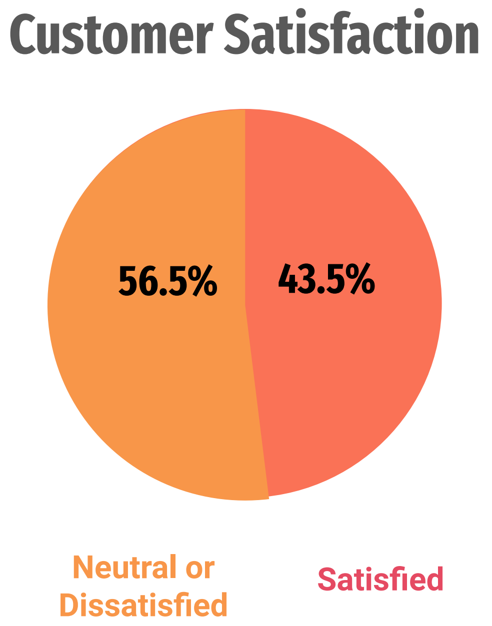
- Inflight Wifi-Service: 0
- Food and Drink: 3
- Seat Comfort: 2
- Inflight Entertainment: 4
- Leg room service: 1
- Baggage Handling: 3
- Inflight Service: 3
- Cleanliness: 4

- Ease of Online booking: 4
- Gate Location: 2
- Online boarding: 3
- Online service: 3
- Check-in service: 4

- Flight Distance: 500
- Departure/Arrival Time Convenient Satisfaction Level: 3
- Departure Delay in Minute: 20
- Arrival Delay in Minute: 18

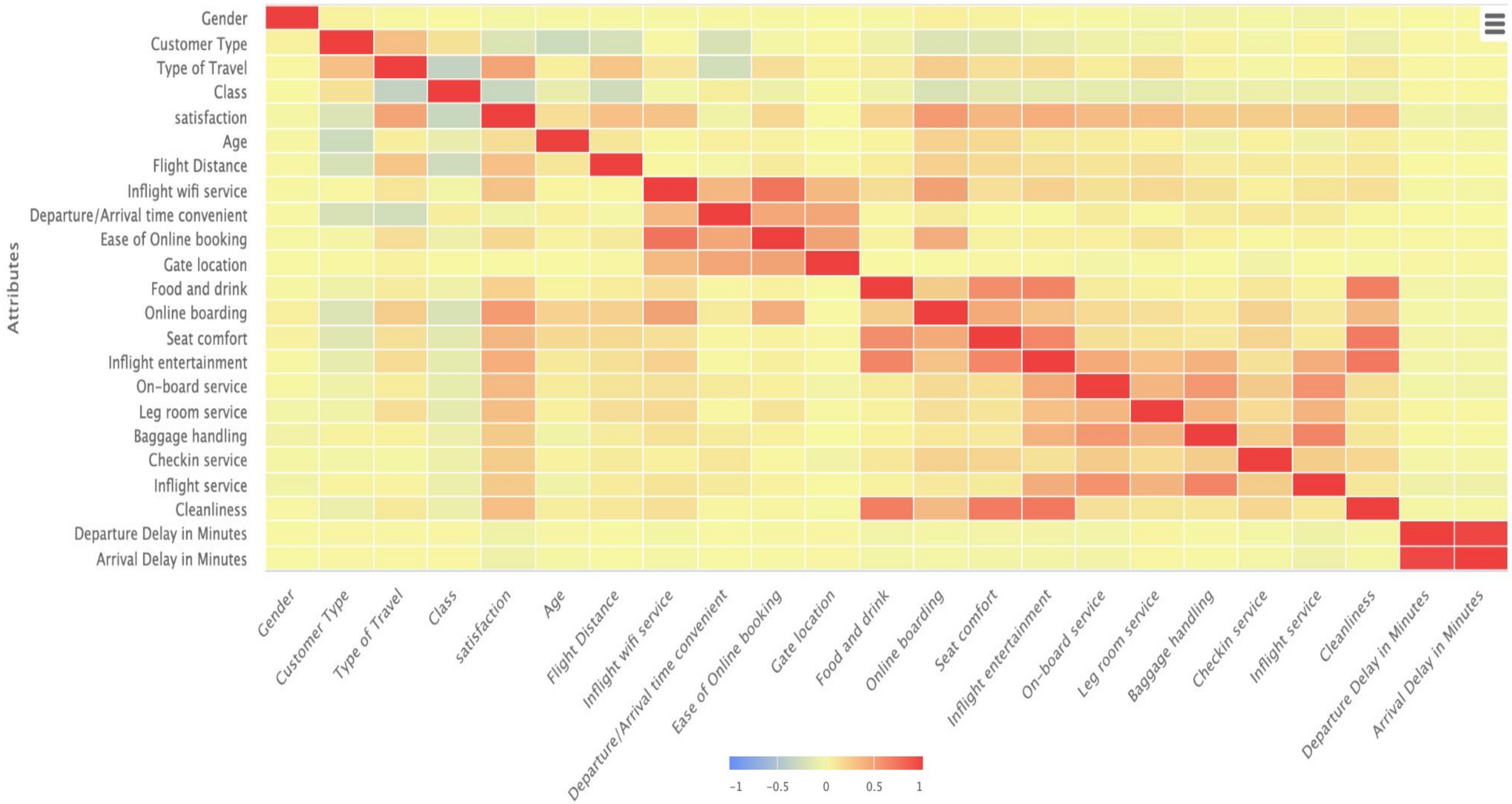
- Satisfaction Level: dissatisfied or neutral

2.4 Data Understanding: Target Variable Distribution



Distribution is about the same for the target variable.

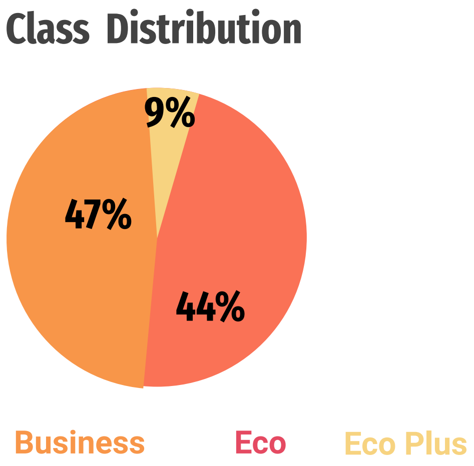
2.5 Data Understanding: Exploring Data

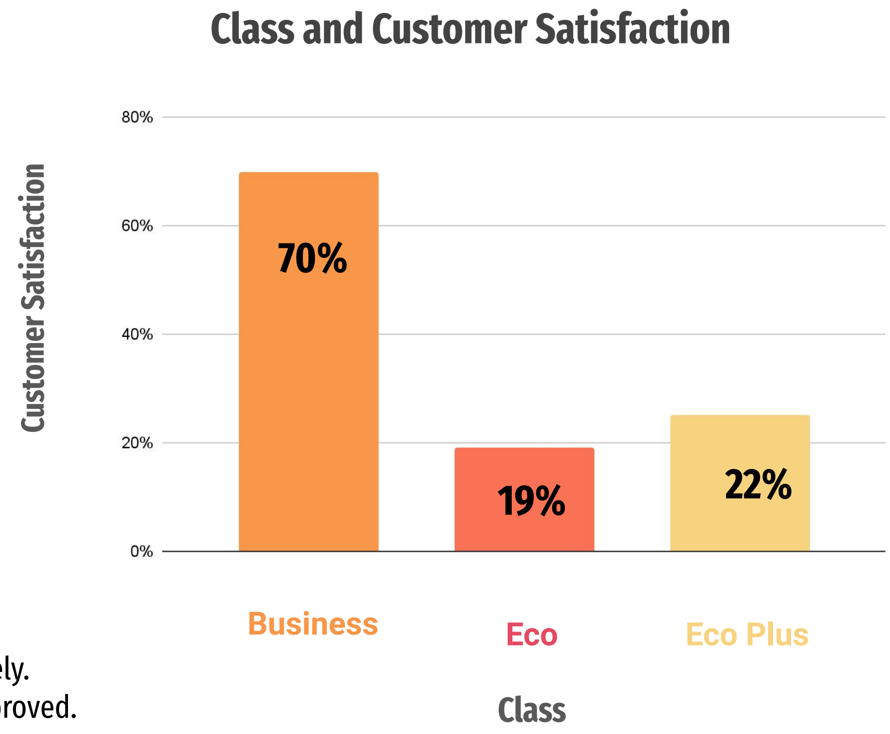


Top3 correlated attributes with satisfaction:

- Online boarding (0.504)
- Type of travel (0.499)
- Inflight entertainment (0.398)

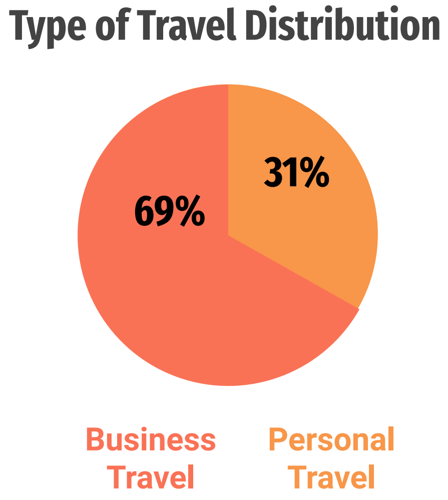
2.6 Data Understanding: Class and Satisfaction

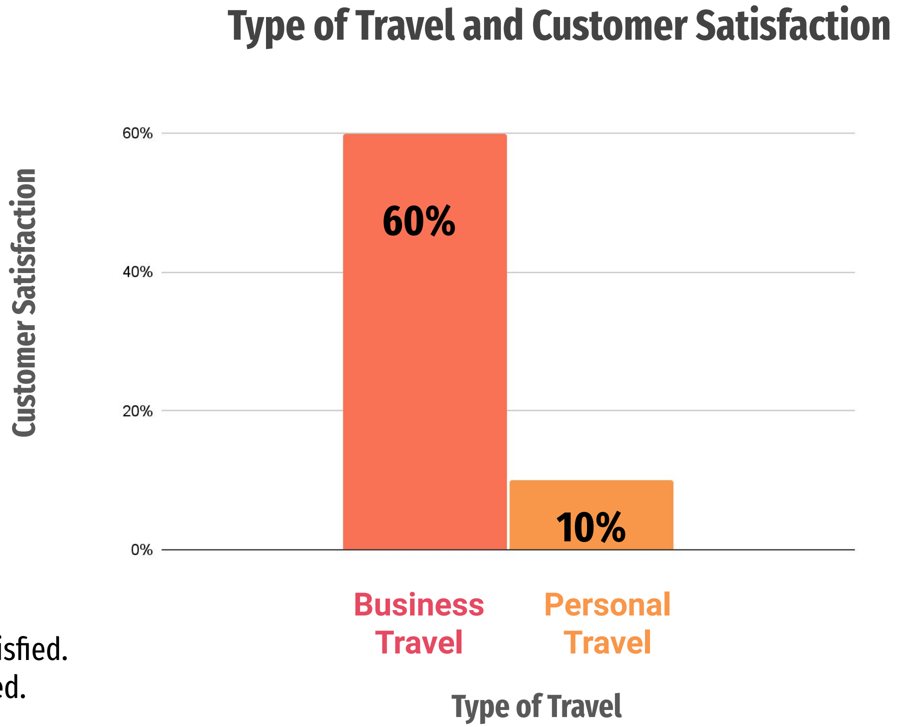




- Most passengers fly with Business and Eco Class.
- 70% of Business class customers are satisfied.
- 19% and 22% of Eco and Eco plus customers are satisfied, respectively.
- Suggest that Eco and Eco plus customer experience needs to be improved.

2.7 Data Understanding: Type of Travel and Satisfaction

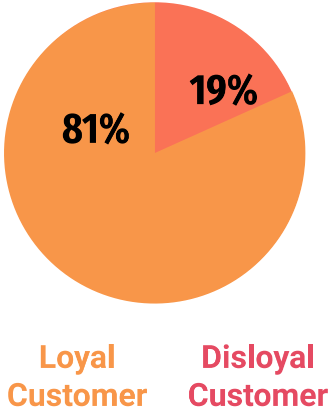




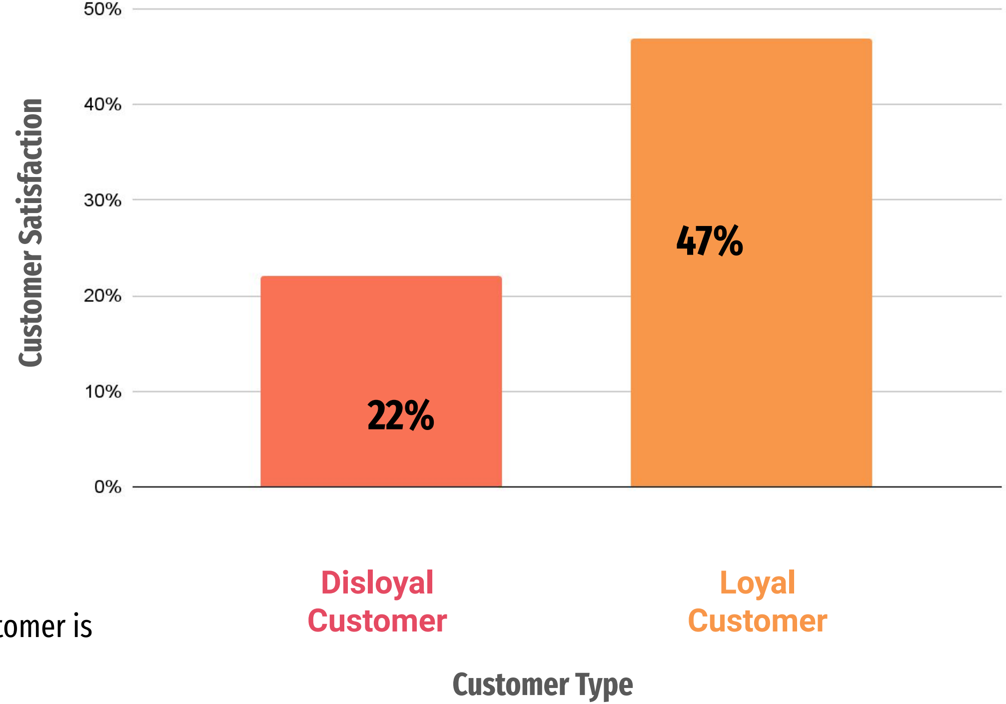
- 69% of customers travel with business purposes.
- 60% of customers traveling with business purpose are satisfied.
- Only 10% of customers with personal purposes are satisfied.

2.8 Data Understanding: Customer Type and Satisfaction

Customer Type Distribution



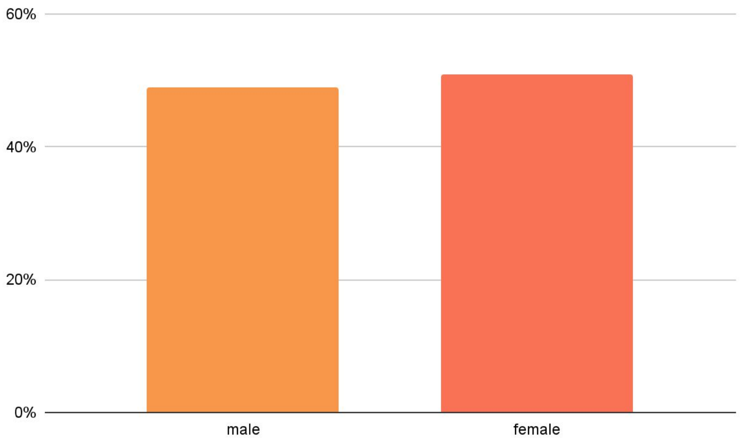
Customer Type and Customer Satisfaction



- Most customers are loyal (81%).
- There is a lower probability of being satisfied if the customer is disloyal.

2.9 Data Understanding: Gender, Age Satisfaction

Gender and Satisfaction

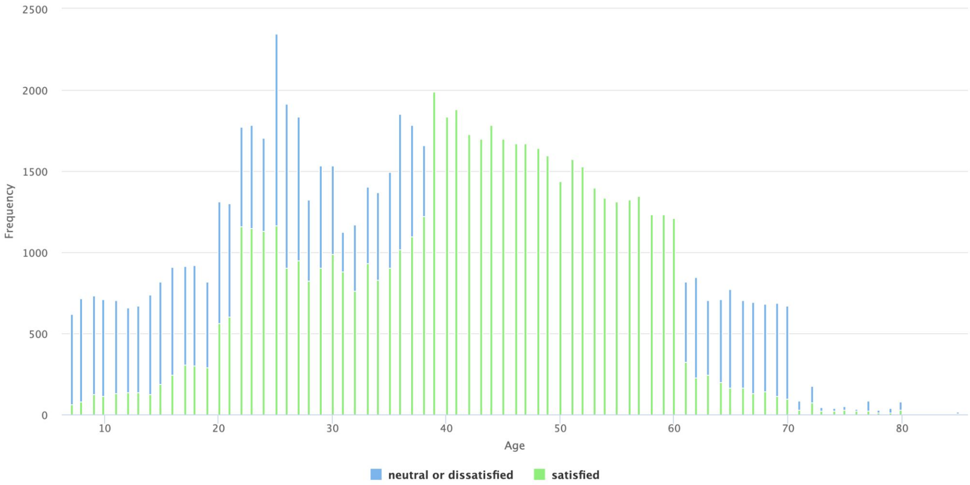


Gender

There is not much satisfaction difference between male and female.

Customer Satisfaction

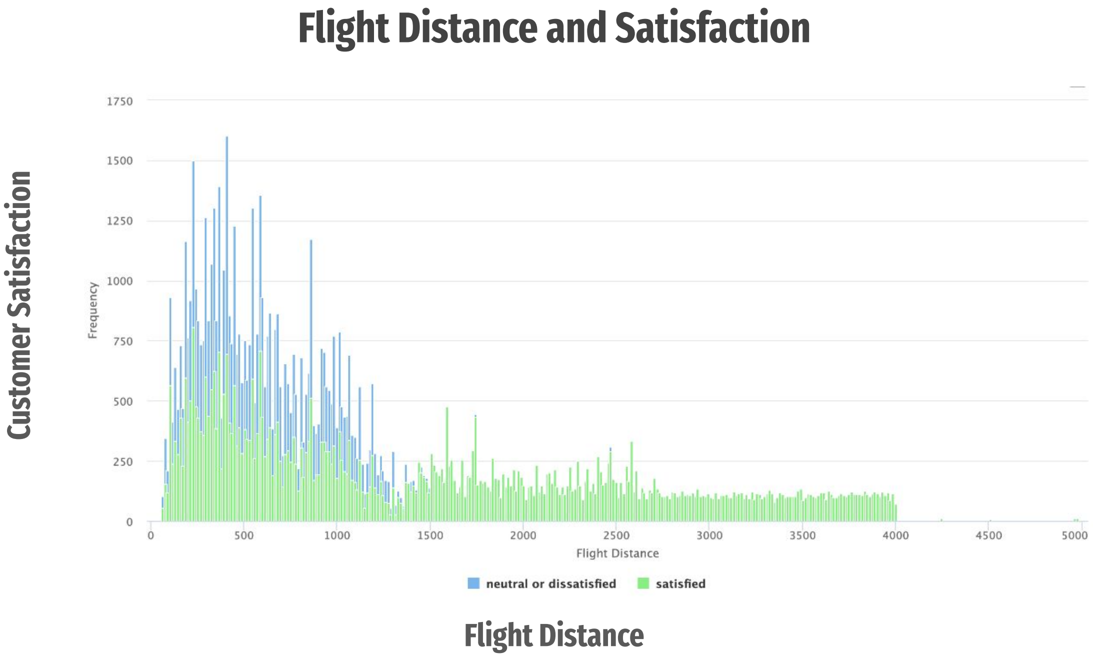
Age and Satisfaction



Age

Customers between age 40 and 60 are more satisfied compared to age 20-38.

2.10 Data Understanding: Flight Distance Satisfaction



Customers flying longer distance have lower satisfaction level.

3.1 Data Preparation: Missing Values

Arrival Delay contains missing values

393 of 129880 are missing.
This is 0.3% of the Arrival Delay containing null values.

Replace with highest correlated attribute

Replace Arrival delay with mean value of Departure Delay.
Departure Delay is the highest correlated column (0.96).

Drop values of 0 from Inflight Services and External Services

0 means not applicable for the services, thus remove the rows.

Correlation between Departure Delay and Arrival Delay

	Departure Delay in Min	Arrival Delay in Min
Departure Delay in Min	1	0.96
Arrival Delay in Min	0.96	1

3.2 Data Preparation: Data Leakage & Normalization

Data Leakage: Exclude variables not predictive of the future at the time of prediction

All inflight services satisfaction level variables
All external services satisfaction level variables
Arrival Delay & Departure Delay times

Data Leakage: 7 Features used

All customer information (gender, age, customer type, type of travel, class), flight distance, satisfaction

Normalization: a necessary step for KNN

KNN assumes that points that are close to one another are similar.
It uses distance to find the similarity; however, scale of measurements influence raw distance measures.
Therefore, normalization transforms variables of different scales into similar scales, so each variable equally contributes to the distance computation.

Normalization: Z score Scaling, Weighted Voting

Z score scaling convert numerical attribute to interval from 0 to 1
Weighted Voting gives more weight on more similar neighbours

4.1 Modeling

- Built four different classification predictive models to determine the best performing one:

  Decision Tree (easy to understand, implement, and use; computationally cheap; simple decision boundary),
  kNN (robust to noise; no assumptions required; computationally expensive),
  Logistic Regression (robust to outliers; fast; linear decision boundary),
  Naive Bayes (fast; independence assumption required).

- Used nested cross-validation for parameter optimization:

  10 folds, Stratified sampling

- Collected five different performance metrics:

  Accuracy, Precision, Recall, F-score, and AUC

➔ Utilize the best performing classification model to predict a passenger's satisfaction and achieve business success (e.g. growth in sales revenue).

4.2 Modeling: Model Results

After optimizing parameters:

Decision Tree

- maximal depth = 16
- criterion = gain ratio
- no pre-pruning/post-pruning

Important attributes:

1. Type of Travel
2. Customer Type
3. Age

Logistic Regression

- no regularization
    - lambda = 0.0
- misclassification costs for FP = 2.0
- misclassification costs for FN = 1.0

Important attributes:

1. Type of Travel
2. Customer Type
3. Class

Decision Tree	**Logistic Regression**
KNN	**Naive Bayes**

kNN

- weighted 5-nearest neighbors
- mixed Euclidean Distance

Naive Bayes

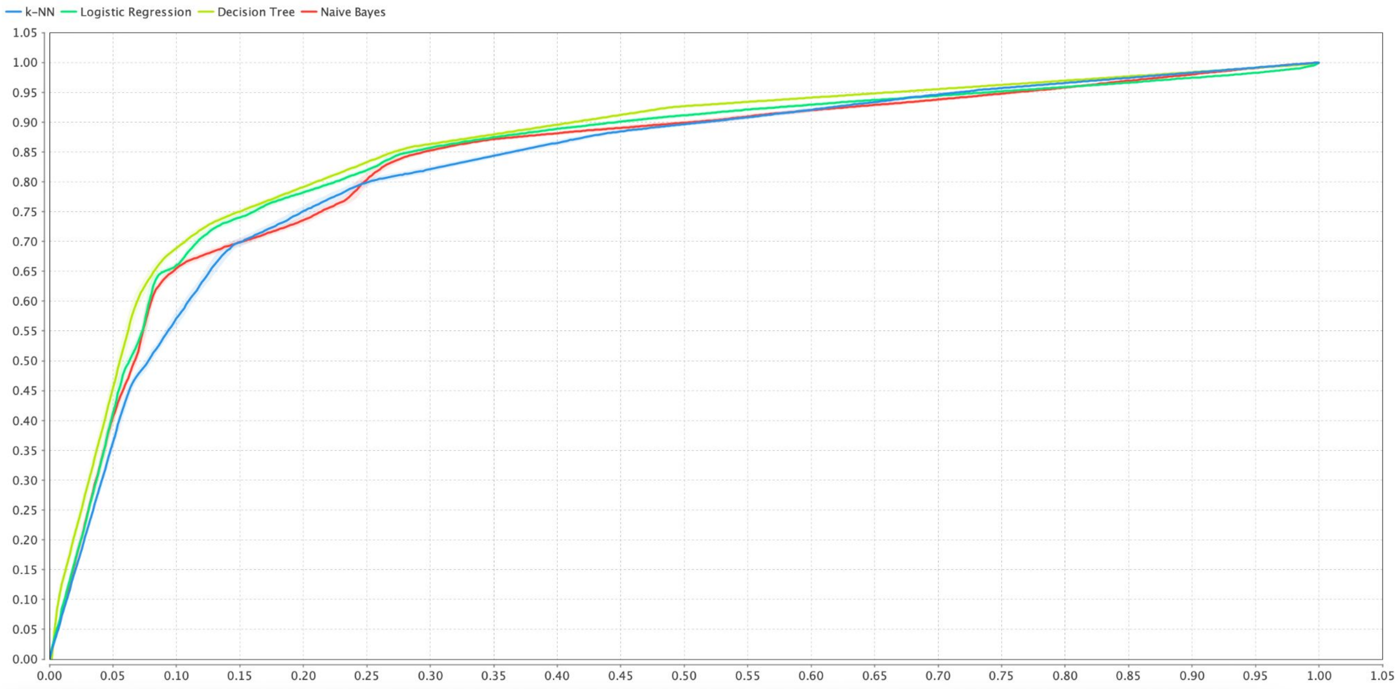
- laplace correction

5.1 Evaluation: Generalization Performance

	Decision Tree	kNN	Logistic Regression	Naive Bayes
Accuracy	0.7984 +/- 0.0038	0.7793 +/- 0.0024	0.7607 +/- 0.0027	0.7875 +/- 0.0032
Precision	0.8093 +/- 0.0043	0.8127 +/- 0.0025	0.9117 +/- 0.0030	0.8087 +/- 0.0029
Recall	0.8480 +/- 0.0030	0.7991 +/- 0.0050	0.6450 +/- 0.0037	0.8242 +/- 0.42%
F-score	0.8282 +/- 0.0038	0.8058 +/- 0.0025	0.7555 +/- 0.0031	0.8164 +/- 0.0029
AUC	0.861 +/- 0.003	0.830 +/- 0.002	0.848 +/- 0.003	0.838 +/- 0.003

- Overall, Decision Tree is the best predictive model that is easy to understand, implement, and use.
- Logistic Regression is the best performing model in the perspective of precision.
  - A high precision is preferred: precision concerns with false positive.
  - Wrongly predicting a passenger who is actually neutral or dissatisfied to be satisfied → undesirable consequences.

5.2 Evaluation: ROC Curves



According to the ROC curves, Decision Tree is the best performing classifier:

- Decision Tree is on the most northwest position.
- Decision Tree has the largest area under the ROC curve.

5.3 Evaluation: Improvement

The predictive model can predict consumer satisfaction levels before they start their journey.

Airline Companies can use the predictive models to:

- *Identify the most important features that impact satisfaction level*
  - **1. Customer Type:** Disloyal customers are 2x more likely to be dissatisfied
    - Build loyalty program for ‘disloyal customer’ by building points system, newsletter with discounts , etc
  - **2. Travel Type:** 90% of personal travelers are dissatisfied (most of which chose Eco/Eco+)
    - Customers is less prestigious class (Eco/Eco+) tend to give lower level of satisfaction for inflight services (food and drink, baggage handling, inflight overall services), so they need to improve service quality. By focusing on providing value-added service which consumers truly-need. For example:
      - Helping them put their luggage in the carrier. Offering each consumers small waste bags so staffs don’t need to walk around to collect waste and disturb passengers.

Benefits: Increase brand reputation and popularity; Stand out from competitors; Retain customers; etc.

6. Deployment

Implementation / future use:

- Find customer satisfaction metrics that are most lacking to disloyal and personal travelers
- Turn findings into strategic initiatives:
    - **Ex.** disloyal and personal travelers rate seat comfort and food-and-drink most poorly.
        - How can you elevate comfort in the economy cabin without incurring high costs and losing seating volume? Could we invest in ergonomics? Partner with brand name chairmakers. Provide free neck pillows or at low cost?
        - What foods/drinks are most sought-after? How can we expand our offerings? Frequency? Presentation? Variety?

Considerations:

- No ethical problems since only basic information is collected for the predictive model
- Risk: Type 1 Error–predict unsatisfied to be satisfied (false positive).
- To reduce the risk: Increase sample size to reach a higher level of statistical significance; Study consumer behavior before/on/after flight.

Questions?

Thank you