Airline Passenger Satisfaction

PRESENTED BY:

Wenyi Zhao & Celine(Xi) Zhang

Overview

Air Travel Consumer Report: Consumer Complaints Against Airlines Rise More Than 300 Percent Above Pre-Pandemic Levels

Thursday, June 23, 2022

OST 20-22
Contact: pressoffice@dot.gov

01

Task

1.train a binary classifier for airline passenger satisfaction using supervised machine learning
2.predict satisfaction:
satisfied OR
neutral satisfaction/dissatisfied
3.airline satisfaction factors evaluation

02

Data

whether a customer is satisfied with the airlines or not after travelling with them.

03

Method: Model

KNN
XGBoost
Random forest

Data

dtypes:**category**(18),float64(1),**int64(3),**object(2)

dtypes: float64(1),
int64(19), object(6)

object: satisfaction& ind(deleted)
int64, float64: numerical variables
category: category variables

https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

	Unnamed: 0	id	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient
0	0	19556	Female	Loyal Customer	52	Business travel	Eco	160	5	4
1	1	90035	Female	Loyal Customer	36	Business travel	Business	2863	1	1
2	2	12360	Male	disloyal Customer	20	Business travel	Eco	192	2	0
3	3	77959	Male	Loyal Customer	44	Business travel	Business	3377	0	0
4	4	36875	Female	Loyal Customer	49	Business travel	Eco	1182	2	3

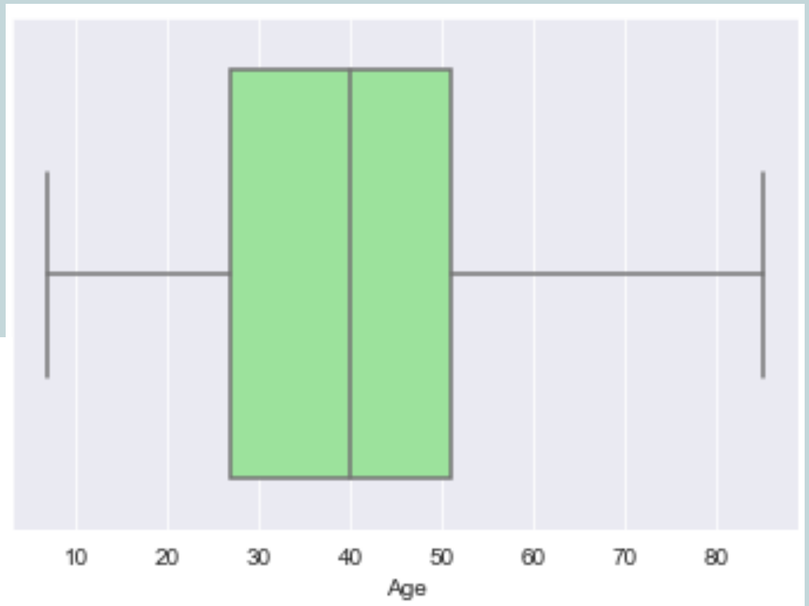
5 rows × 26 columns

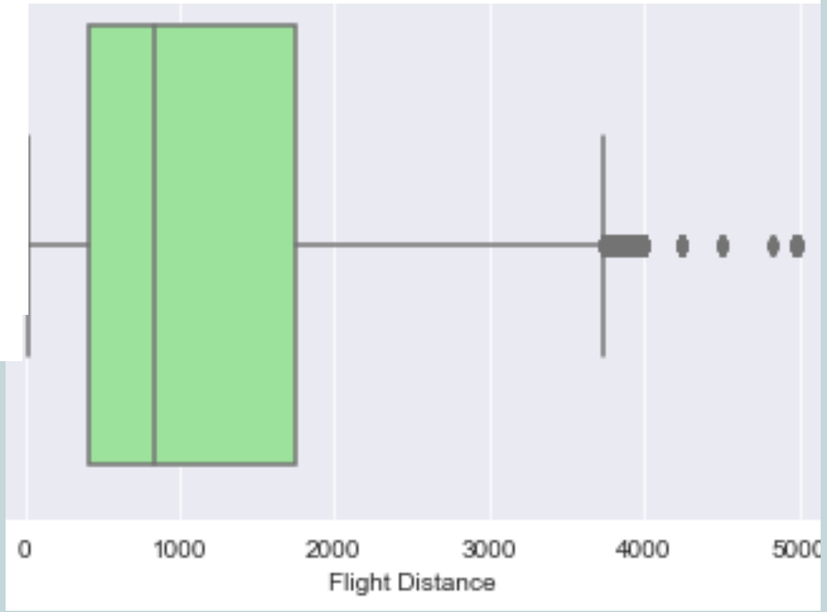
Data Size: neutral or dissatisfied: 73452, satisfied: 56428, already split into train set (80%) and test set (20%)

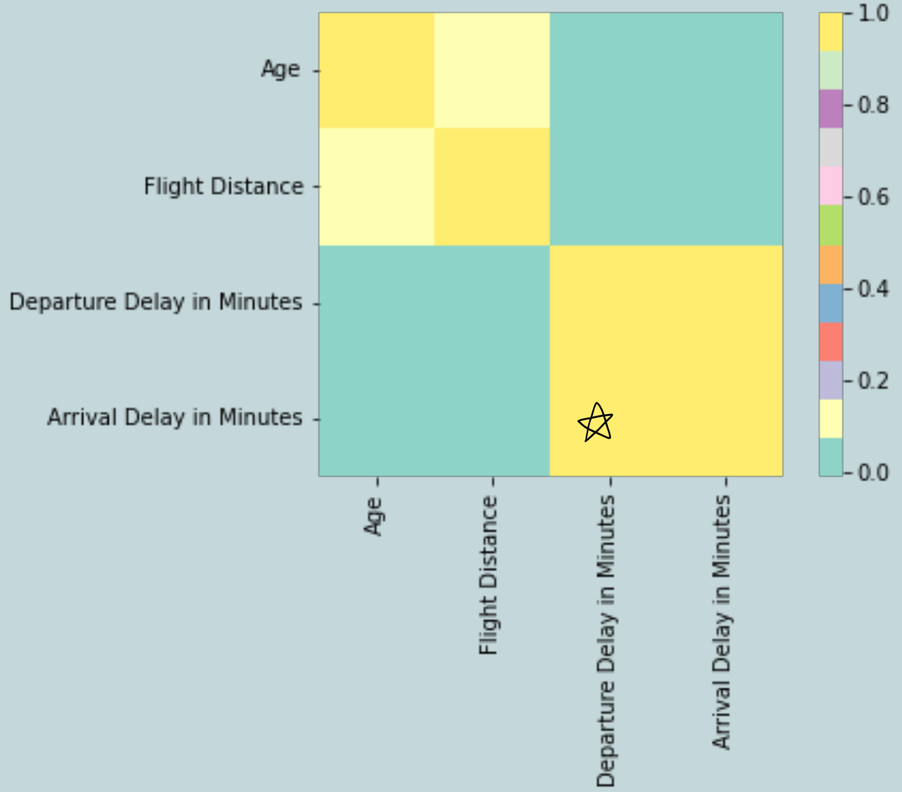
Data

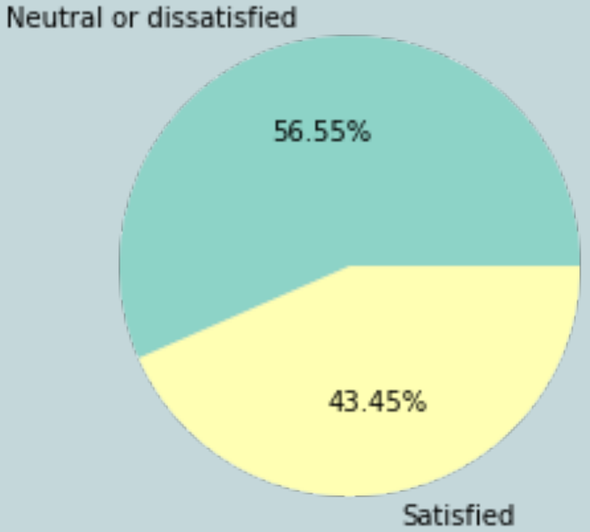
2	Gender	129880 non-null	object
3	Customer Type	129880 non-null	object
4	Age	129880 non-null	int64
5	Type of Travel	129880 non-null	object
6	Class	129880 non-null	object
7	Flight Distance	129880 non-null	int64
8	Inflight wifi service	129880 non-null	int64
9	Departure/Arrival time convenient	129880 non-null	int64
10	Ease of Online booking	129880 non-null	int64
11	Gate location	129880 non-null	int64
12	Food and drink	129880 non-null	int64
13	Online boarding	129880 non-null	int64
14	Seat comfort	129880 non-null	int64
15	Inflight entertainment	129880 non-null	int64
16	On-board service	129880 non-null	int64
17	Leg room service	129880 non-null	int64
18	Baggage handling	129880 non-null	int64
19	Checkin service	129880 non-null	int64

3	Arrival Delay in Minutes	129487 non-null	float64





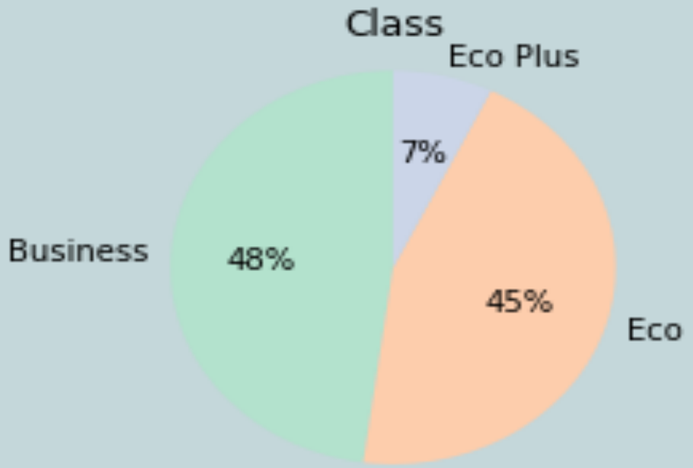


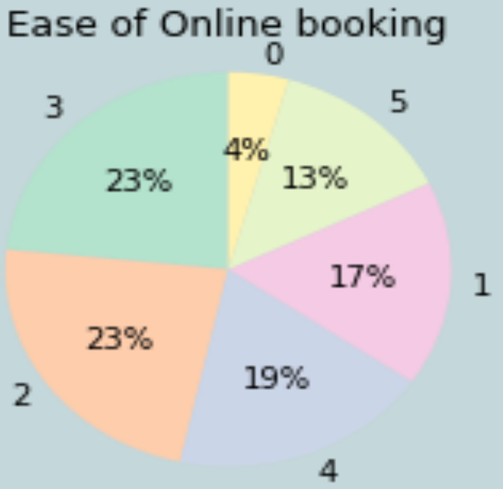


Data

Categorical Data

- Binary: assign 1 and 0
- Non-Binary: assign one binary column per category per categorical feature.
- Example:
- Gender: 0 and 1
- Class: Business: 0 and 1
  Eco Plus: 0 and 1
  Eco: 0 and 1



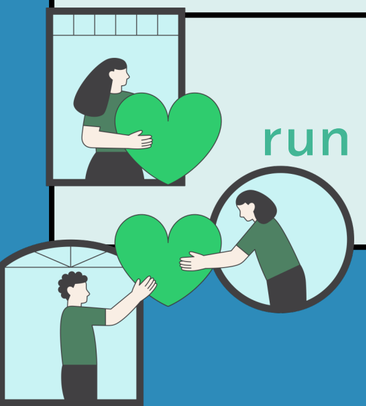


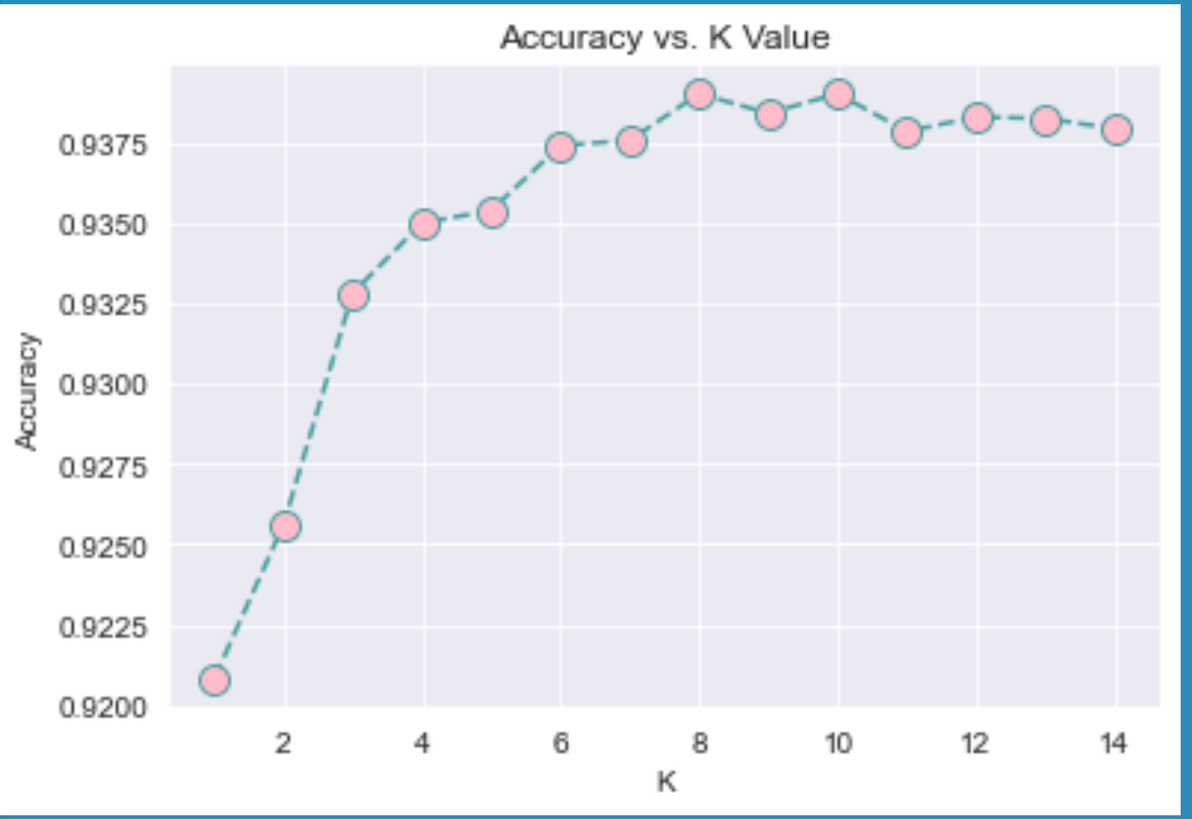
Final Preprocessed Data
Shape: 125899 rows × 92 columns

Age	Flight Distance	Departure Delay in Minutes	Gender	Customer Type	Type of Travel	Class_Business	Class_Eco	Class_Eco Plus	Inflight wifi service_0	...	Inflight service_3	Inflight service_4	Inflight service_5	Cleanliness_0	Cleanliness_1	Cleanliness_2	Cleanliness_3	Cleanliness_4	Cleanliness_5
-0.220161	1.881360	-0.386329	0	1	1	1	0	0	0	...	0	1	0	0	0	0	0	0	1

Model

type of model	KNN(K nearest neighbors)
parameters	n_neighbors = [range(1, 15)]
accuracy	0.939005(n_neighbors=8,10)
run time	43.9s

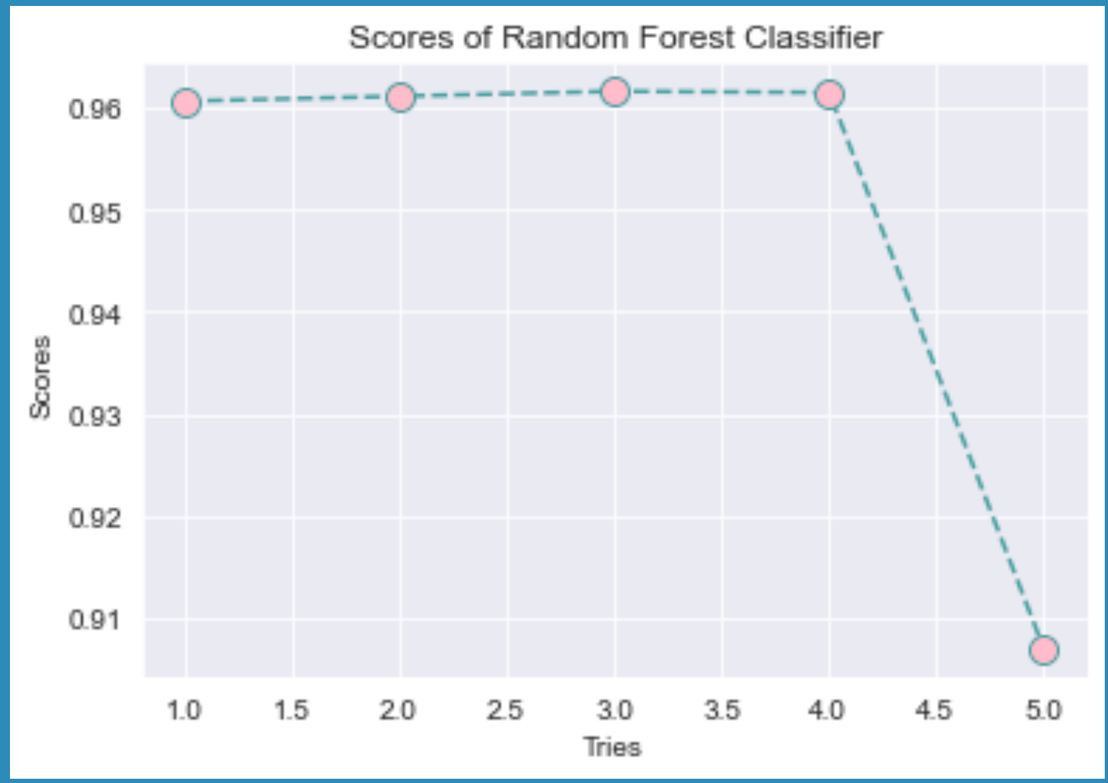




The best K value should be 8 or 10

Model

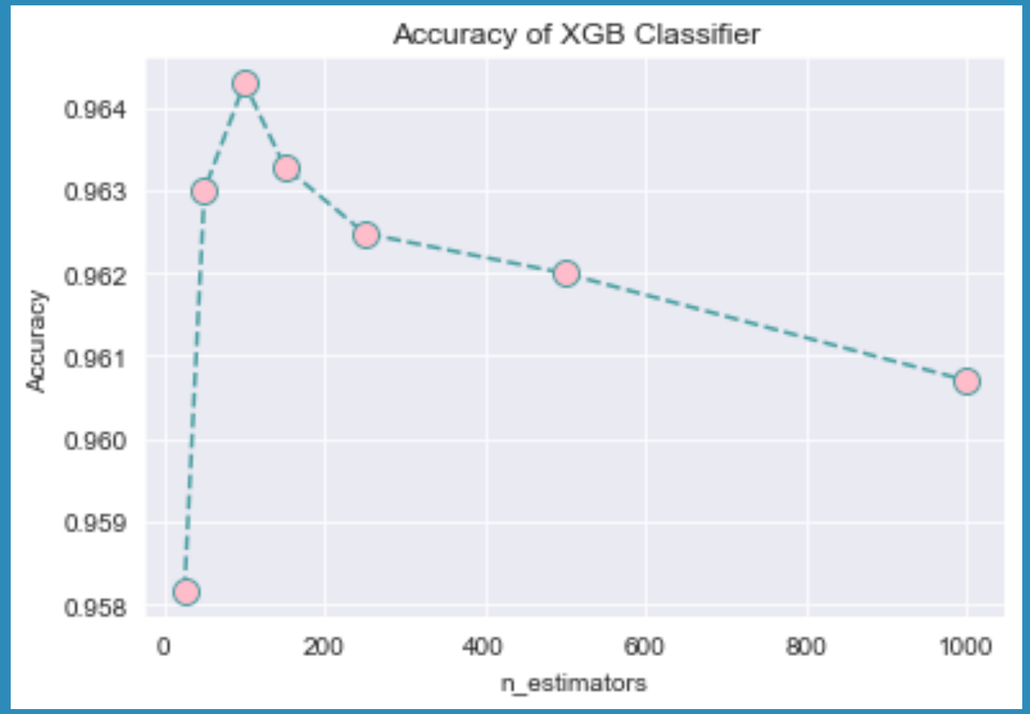
type of model	Random Forest Classifier
parameters	n_estimators = [500,1000] random_state = 42 criterion = "entropy" max_depth = 5
accuracy	0.961631(n_est = 1000, random_state = 42)
feature importance	type of travel
run time	**1m 18.6s**



The fourth try shows the best performance, with n_estimators = 1000

Model

type of model	XGBoost(Extreme Gradient Boosting)classifier
parameters	n_estimators = [25,50,100,150,250,500,1000] eval_metric = ['rmse', 'logloss'] objective=['binary:hinge', 'binary:logistic'] max_depth= 10, learning_rate= 0.1 , gamma= 0.8, reg_lambda= 2 , reg_alpha= 2 , scale_pos_weight= 2 , subsample= 0.8, colsample_bytree= 0.8
accuracy	0.964289(n_estimator=100, default)
feature importance	Online boarding_5
run time	9.9s



when n_estimators >= 100, with the increase of of n_estimators, the accuracy decreases

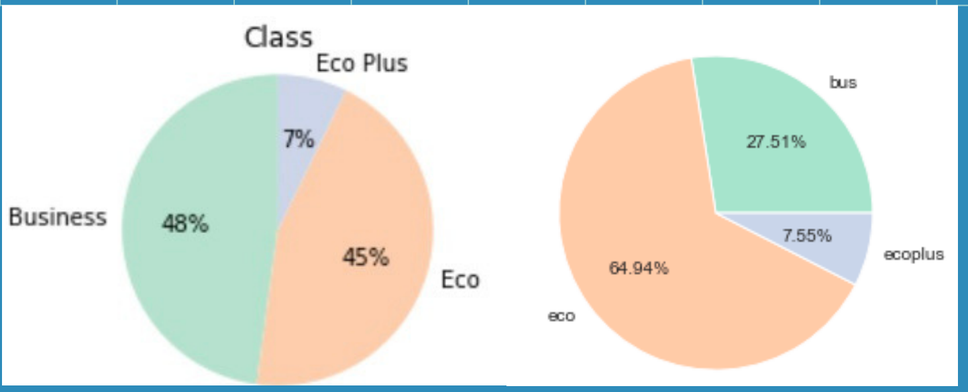
Model Comparison

	Random Forest Classifier	KNN	XGB Classifier
parameter	n_estimators = 1000, random_state = 42	n_neighbors = 8 or 10	Default
accuracy	0.961631	0.939005	0.964289
runtime	1m 18.6s	43.9s	9.9s

Error Analysis -- Examples of where the model fails

- XGB Classifier
- 887 failed predictions out of 24838 test data
- Examples

row	Age	Flight Distance	Departure Delay in Minutes	Gender	Customer Type	Type of Travel	Class_Business	Class_Eco	Class_Eco Plus	Inflight wifi service_0	...	Inflight service_4	Inflight service_5	Cleanliness_0	Cleanliness_1	Cleanliness_2	Cleanliness_3	Cleanliness_4	Cleanliness_5	target class	predict class
4	0.636947	0.056942	-0.386329	0	1	1	0	1	0	0	...	0	0	0	0	0	0	1	0	1	0
25960	-1.077269	-0.740766	-0.386329	0	0	1	0	1	0	0	...	1	0	0	1	0	0	0	0	0	1



whole dataset

failed dataset

1. Systematically fine-tune
2. Data Distribution
3. More advanced model
4. Combination of different models
5. Data preprocessing methods
6. Overfitting? Validation data