Predicting Customer Satisfaction for Airlines with a Binary Classification Model

Sarah Lueling

Assignment Task

Goal

- The dataset contains an airline passenger satisfaction survey.
- Predict satisfaction of airline customer

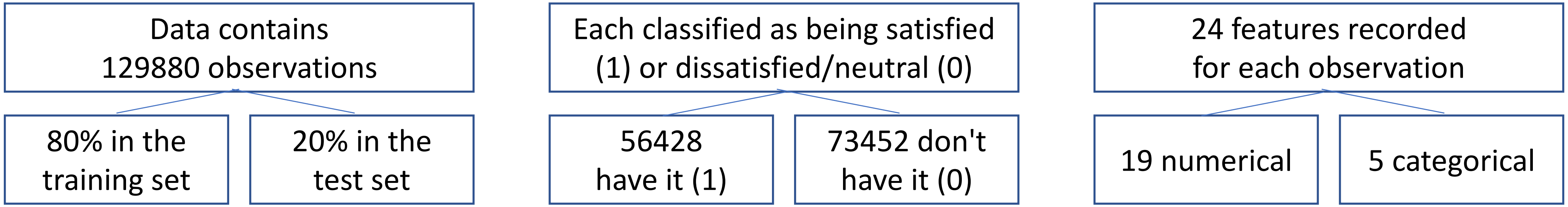
Binary Classification

- Target: Satisfied or not satisfied/neutral
- Use 3 algorithms suitable to binary classification to predict satisfaction

Data

1

Summary of data



Data contains 129880 observations

80% in the training set

20% in the test set

Each classified as being satisfied (1) or dissatisfied/neutral (0)

56428 have it (1)

73452 don't have it (0)

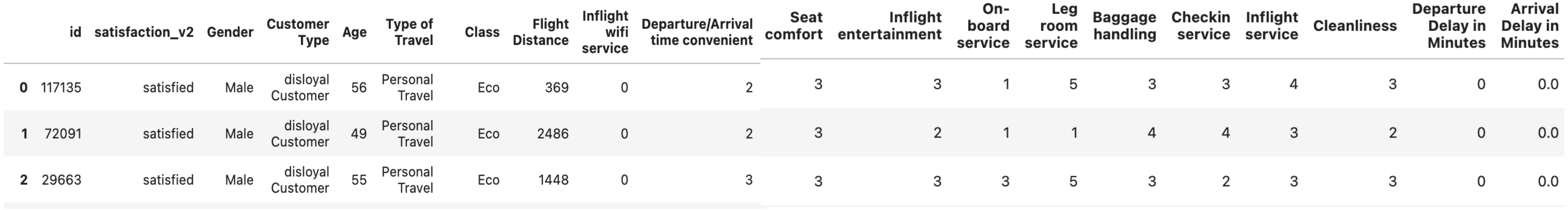
24 features recorded for each observation

19 numerical

5 categorical

2

Appearance of data



	id	satisfaction_v2	Gender	Customer Type	Age	Type of Travel	Class	Flight Distance	Inflight wifi service	Departure/Arrival time convenient	Seat comfort	Inflight entertainment	On-board service	Leg room service	Baggage handling	Checkin service	Inflight service	Cleanliness	Departure Delay in Minutes	Arrival Delay in Minutes
0	117135	satisfied	Male	disloyal Customer	56	Personal Travel	Eco	369	0	2	3	3	1	5	3	3	4	3	0	0.0
1	72091	satisfied	Male	disloyal Customer	49	Personal Travel	Eco	2486	0	2	3	2	1	1	4	4	3	2	0	0.0
2	29663	satisfied	Male	disloyal Customer	55	Personal Travel	Eco	1448	0	3	3	3	3	5	3	2	3	3	0	0.0

3

Pre-processing steps taken

- Dropping unnecessary columns
- Change attribute names
- Standardize scaling: Standardization of the quantitative features
- Label Encoding

Classification Algorithms

- LOGISTIC REGRESSION CLASSIFIER
- DECISION TREE CLASSIFIER
- K NEAREST NEIGHBOUR CLASSIFIER

Logistic Regression Classifier

Explanation	Independent variables are analysed to determine the binary outcome with the results falling into one of two categories.
Accuracy Score	0.874
Tuning	Parameter values : random_state = 42

Decision Tree Classifier

Explanation	Hierarchically, it splits data into subsets which are then split again into the smaller subsets until they become “pure”.
Accuracy Score	0.9472
Tuning	Parameter values : random_state = 42

KNN (K-Nearest Neighbors)

Explanation	Its purpose is to use a database in which the data points are separated into several classes to predict the classification of a new sample point.
Accuracy Score	0.934
Tuning	Parameter values: (n_neighbors=3)

Comparison of Classification Models

	KNN	Logistic Regression	Decision Tree
Pros	Easy to implement Versatility Non-Linear Performance	Simple to implement Less prone to over-fitting with regularization techniques	Requires less effort for data preparation Intuitive and easy to explain
Cons	Slow for large data sets As numbers of variables grow KNN algorithm struggles to predict the output of new data points	Not appropriate for non-linear problems Need to pre-process data	Can lead to overfitting of the data For large dataset it’s can become too complex to interpret and generalize

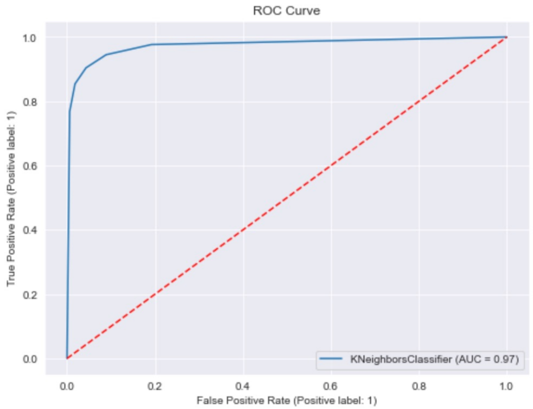
Result Analysis

Results

- Best prediction were given by accuracy of the Decision Tree Classifier: 94%
- Three most important coefficients: “Seat comfort”, “Inflight service”, “Departure/arrival time” convenient”

Error Analysis

- Confusion Matrix:
  - F1 score is KNN model
  - Highest AUC of ROC curve
  - Overfit/underfit
- Analyse raw data
  - Data Collection
  - **Improper splitting of training and test data**



Thanks For Listening