CSci 543 – Data Mining
Project #1

Airline Passenger Satisfaction

Due: April 2, 2021 at 11:59 pm

This project is worth 100 points. Your goal is to explore the Airline Passenger Satisfaction data set, build prediction models, and write report about your findings. The data set is available on Blackboard or can be downloaded from: https://www.kaggle.com/teejmahal20/airline-passenger-satisfaction

The data supposedly comes from a passenger survey, but it isn't clear if it is real or randomly generated. There are two classes: satisfied and (neutral or dissatisfied). Usable features include:

Gender: Gender of the passengers (Female, Male)
Customer Type: The customer type (Loyal customer, disloyal customer)
Age: The actual age of the passengers
Type of Travel: Purpose of the flight of the passengers (Personal Travel, Business Travel)
Class: Travel class in the plane of the passengers (Business, Eco, Eco Plus)
Flight distance: The flight distance of this journey
Inflight wifi service: Satisfaction level of the inflight wifi service (0:Not Applicable;1-5)
Departure/Arrival time convenient: Satisfaction level of Departure/Arrival time convenient
Ease of Online booking: Satisfaction level of online booking
Gate location: Satisfaction level of Gate location
Food and drink: Satisfaction level of Food and drink
Online boarding: Satisfaction level of online boarding
Seat comfort: Satisfaction level of Seat comfort
Inflight entertainment: Satisfaction level of inflight entertainment
On-board service: Satisfaction level of On-board service
Leg room service: Satisfaction level of Leg room service
Baggage handling: Satisfaction level of baggage handling
Check-in service: Satisfaction level of Check-in service
Inflight service: Satisfaction level of inflight service
Cleanliness: Satisfaction level of Cleanliness
Departure Delay in Minutes: Minutes delayed when departure
Arrival Delay in Minutes: Minutes delayed when Arrival

Many of the features have a discrete (Likert) score of 1-5. (or 0-5). As best as I can tell, 0 means missing, 1 is bad and 5 is excellent. There is a training file of almost 104000 instances, and a test file of almost 26000. So it is a nice sized data set.

What you need to do:

- Explore the data to get an idea of the features. Use any of the statistical methods we looked at, and any others you believe will be helpful.
- Develop classification models using a range of algorithms, with effort to tune the associated parameters. Report results on both the training and test sets.
- Write a 2-3 page report with your findings. Think of this as a report to the management of the airline. Include graphs where helpful, and explain the issues you found with the data and challenges you had developing the models (and hopefully what you did to overcome the challenges). You can use the data as-is, or do some preprocessing. If you do any preprocessing though, please explain in your report.
- Submit a zip file with your code (plain code or a notebook) and your report.
- The project will be graded based on the appropriateness and effectiveness of the models developed to analyze the data, the creativity used to extract and present information, and the quality of the report.
- Graduate students must, in addition to all of the above, explore and use at least one learner that we haven't explicitly worked with before (e.g. SVM or an ensemble method), and describe its use in your report, which can be 3-4 pages.