CS 773/873: Data Mining and Security
Summer 2024
Course project
100 Points
Due: August 1, 2024

The course project is an attempt to integrate the techniques you have learned in the class. It may be done in teams of 1 or 2 students. While some guidelines are given as to what is expected, it is more important that the student puts his/her own creativity at work in accomplishing the tasks. There are no restrictions on the tools you may use. From Weka to Excel to your own programs or a combination are all okay.

You will be presenting your results in a report (pdf) with detailed results in the appendix. It is important that the report be well organized with references to the detailed runs in the appendix. It is also important for you to state the tools/techniques you have used to answer any specific task.

This project uses the Airline Passenger Satisfaction Survey data from Kaggle.
https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction

Attribute#	Field	Description
A1	ID	Unique passenger identifier
A2	Gender	Gender of the passenger (Female/Male)
A3	Age	Age of the passenger
A4	Customer Type	Type of airline customer (First-time/Returning)
A5	Type of Travel	Purpose of the flight (Business/Personal)
A6	Class	Travel class in the airplane for the passenger seat
A7	Flight Distance	Flight distance in miles
A8	Departure Delay	Flight departure delay in minutes
A9	Arrival Delay	Flight arrival delay in minutes
A10	Departure and Arrival Tin	Satisfaction level with the convenience of the flight departure and arriva
A11	Ease of Online Booking	Satisfaction level with the online booking experience from 1 (lowest) to
A12	Check-in Service	Satisfaction level with the check-in service from 1 (lowest) to 5 (highest)
A13	Online Boarding	Satisfaction level with the online boarding experience from 1 (lowest) to
A14	Gate Location	Satisfaction level with the gate location in the airport from 1 (lowest) to 5
A15	On-board Service	Satisfaction level with the on-boarding service in the airport from 1 (lowe
A16	Seat Comfort	Satisfaction level with the comfort of the airplane seat from 1 (lowest) to
A17	Leg Room Service	Satisfaction level with the leg room of the airplane seat from 1 (lowest) t
A18	Cleanliness	Satisfaction level with the cleanliness of the airplane from 1 (lowest) to 5
A19	Food and Drink	Satisfaction level with the food and drinks on the airplane from 1 (lowest
A20	In-flight Service	Satisfaction level with the in-flight service from 1 (lowest) to 5 (highest)
A21	In-flight Wifi Service	Satisfaction level with the in-flight Wifi service from 1 (lowest) to 5 (high
A22	In-flight Entertainment	Satisfaction level with the in-flight entertainment from 1 (lowest) to 5 (h
A23	Baggage Handling	Satisfaction level with the baggage handling from the airline from 1 (low
A24	Satisfaction	Overall satisfaction level with the airline (Satisfied/Neutral or unsatisfie

The idea is to make some useful conclusions that could help airline executives improve passenger satisfaction. The project aims to identify crucial factors that influence passenger satisfaction. In order to make it simple for the student, it is organized into steps.

Step 1: Statistics on departure delay (A8) and arrival delay (A9).

Here is a link that talks about useful statistics that a data scientist would like to know about an entity. https://towardsdatascience.com/5-useful-statistics-data-scientists-need-to-know-5b4ac29a7da9

1a. For features A8 and A9, determine

i. Central tendency measures: mean, mode, and median (for each)
ii. The spread: standard deviation (for each)
iii. Percentiles: the 10^{th}, 50^{th}, 75^{th} and 90^{th} percentiles (for each)
iv. 1^{st} quartile, 3^{rd} quartile, and the median
v. The skewness (for each)
vi. The covariance and correlation between A8 and A9
vii. Plot the two delays on a chart to display their distributions https://www.webdatarocks.com/blog/best-charts-for-data-distribution/

1b. Based on the above results in 1a, make useful conclusions. This is an important part of data analysis. It should be an English description that non-mathematically oriented airline executives can understand.

Step 2: Convert numerical values to categorical values

a. Discretize age (A3) to nominal values using the following criteria: 0-15: Child; 16-35: Youth; 36-55 Middle age; 56-70: Old; >70- Senior;
b. Discretize flight distance (A7) to nominal values using the following criteria: 0-500 miles: Short haul; 501-3000 miles: Medium haul; >3000 Long haul
c. Discretize delays (A8 and A9) to nominal values: Small: 0-15; Medium: 16-45; Long: >45
d. Plot the distributions for each of the above using the discretized values.

2b. Based on the distributions in 2d, make some useful observations and write them.

Step 3. Test the following two hypotheses. Show evidence to show whether they are true or false.

You may refer to the following URLs for hypothesis testing procedure.

https://vitalflux.com/linear-regression-hypothesis-testing-examples/

https://vitalflux.com/data-science-how-to-formulate-hypothesis-for-hypothesis-testing/

3H1: Long haul passengers' overall satisfaction is influenced more by the in-flight service quality than by the departure delays.

3H2. Medium haul passengers' overall satisfaction is influenced more by the arrival delays than by the in-flight entertainment.

3H3. Make your own hypothesis, state it, and test it.

Step 4. Find associations between some of the important attributes.

4a. With Gender, Age, Type of travel, Flight distance, Class, Arrival delays, and Overall satisfaction as attributes, determine association rules with a minimum support of 100 and a minimum confidence of 60%.

4b. Provide a plausible English explanation for the associations identified in step 4a

Step 5. Reduce the satisfaction features using PCA. (https://www.youtube.com/watch?v=zuoMjUAPihA https://www.youtube.com/watch?v=THA8zBR64Kk)

5a. Using PCA (Principal Component Analysis), combine features A10-A23 into a single feature. Let us call it PCAS. Now find average, minimum, and maximum of A10-A23 (computed for each passenger record). Let us call them AVES, MINS, and MAXS, respectively. Convert A24 (overall satisfaction) into a numeric value by converting neutral or unsatisfied to 1.0 and satisfied to 4. Let us call it DA24.

5b. Which among PCAS, AVES, MINS, and MAXS could be used as proxy (or proxies) for DA24? You need to make some runs to answer this question.

5c. What benefit, if any, is derived by using PCAS from A10-A23? Is there any additional benefit if we had derived three components from PCA instead of just one? Show some evidence to justify your answer. (Hint: Repeat PCA in 5a with 3 components and then answer the question.)

Step 6. Using linear regression, model the relationship between (i) flight distance (in miles) and arrival delay (in minutes), and (ii) flight distance (in miles) and departure delay (in minutes). Explain the derived models and their accuracy. State whether such a relationship makes sense based on the derived models.

Step 7. Using any of the data mining techniques you are familiar with, answer the following questions. Show your work to justify your conclusions.

7a. Is satisfaction with seat comfort related (or depends on) to passenger Gender?

7b. Is satisfaction with gate location related to passenger age?

7c. Do first time passengers have more or less expectations than returning customers measured in terms of overall satisfaction?

7d. Is there a distinct (statistically significant) difference between business and personal travelers (A5) in terms of their reaction to their flights? (Hint: Use any attribute(s) that you think appropriate to measure their reaction.)

7e. Is there a distinct (statistically significant) difference between business class passengers and economy passengers (A6) in terms of their reaction to satisfaction with food-and-drink?

Step 8.

8a. Determine if any relationship exists between check-in service (A12) and baggage handling (A23) using any data mining technique. Based on this, provide a guidance to airline executives.

8b. Between A10 and A16, which one do you think passengers value the most? Assume that the overall satisfaction (A24) is a good proxy for the value.

Step 9. Executive summary. Having completed the above analysis, provide a summary addressed to a group of executives informing them of what you have learned about air travelers, their likes and dislikes, and their preferences. Provide any other guidance to them. The summary should be backed by your results, and just any baseless statements. Provide references to your specific analysis in making the statements. It always helps to explain the results as graphs or tables instead of some unreadable Weka outputs.

Report organization:

Team members

Executive summary

Steps 1-8: Details of the runs and summaries of the runs with details on conclusions for each step.

Appendix: Showing details of the runs made to address each question.