Université d'Ottawa

Faculté de génie

uOttawa

L'Université canadienne
Canada's university

University of Ottawa

Faculty of Engineering

CSI4142 Introduction to Data Science

Final Examination 2017

Duration: 3 hours

Instructions

a) *This is a closed book examination,* ***subject to the following****.*
b) *You are allowed to bring allow two letter-side note pages (so-called cheat sheets), written or printed on both sides.* ***You are not allowed to use any magnifying glasses.***
c) *No calculators are allowed.*
d) *Please return the examination questions together with your answer booklet(s).*

Name:	
Student Number:	

Notes added on 4 April 2018:
The focus of last year's exam is *very different* than 2018. This is because I covered machine learning is much more detail in 2018.
Please refer to the 2018 exam mark allocation I posted recently. Data marts only constitute 35/100 of your final mark.
We did not cover association analysis (question 10).

FlyAway TRAVEL INSURANCE CASE STUDY

Consider the FlyAway travel insurance company which offers travel insurance for Canadians traveling abroad. The business consists of two main departments, namely the department that sells travel insurance and the one that processes travel claims.

In this business, it is important for FlyAway to be able to analyze their coverage packages and the corresponding claims, in order to determine which ones are the most profitable (and also, on the other hand, the least profitable). These packages are offered to individuals, as well as to families. Usually, many frequent travelers choose to purchase a one year, renewable coverage policy. However, some occasional travelers may only purchase very short term coverage.

FlyAway wishes to determine the profiles of the customers who are most likely to file a claim, in order to adjust their coverage. They are also interested in identifying potential high risk travel destinations, in terms of the frequencies, types and costs associated with typical illnesses and/or accidents. (For example, accidents may refer to mountaineering or scuba diving accidents while typical life-illnesses may refer to yellow fever or malaria.)

FlyAway measures profitability over time by the covered item type (i.e. which kinds of travel coverage), country of destination, demographic profile of the customer, sales region, and event. Events are usually called catastrophes when a large number of travelers are taken ill due to an epidemic or are involved in a major incident such as a land slide.

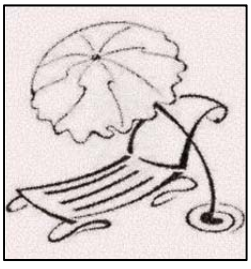
A specific coverage is associated with a travel policy, with a duration ranging from 14 days to one year. A policy is associated with a single traveler or with a family. A family consists of a maximum of two adults and up to 10 children under the age of 18 years old. One of the adults is considered to be the designated contact person associated with the policy. Pensioners between 65 and 75 years old receive a 10% discount.

FlyAway aims to understand what happens during the life-time of a customer, especially those long-term customers who renew their yearly policies. It is also important for the company to identify potential customers who will not renew their coverage, in order to target marketing efforts. This information is important in order to aid FlyAway when estimating the number of claims that will be submitted as well as to determine the overall costs associated with a specific travel coverage packages.

The management of FlyAway interviews you for a newly created position. They wish to hire someone to conduct an initial data science project. Your tasks will include creating a data mart and conducting some data mining.

You are shortlisted for this position and invited to a second interview…

Answer all the following interview questions.



Suppose that the company maintains a Travel-Policy table that contains the following data.

`Travel-Policy(Policy-Number, Policy-Type, Date-Purchased, Date-Start, Data-End, $Amount, Customer-Name, Customer-Age, Customer-Gender, Duration, Marital-Status, Customer-Home-City, Customer-Home-Province, Earlier-claims (y/n), Travel-Destination-City, Travel-Destination-Country, Risk-Level)`

1. Explain how you would determine whether there is a correlation between the Age of a Customer and the Duration of the policy. (4)

2. Suppose that 30% of the tuples do not contain the Travel-Destination-City and Travel-Destination-Country attributes. Explain how you would handle this omission during data preprocessing. (4)

3. Suppose that you decide to construct a data mart in order to manage all the policies, as purchased over the last five years. Draw the dimensional model for this data mart. (16)

4. Suppose that a Customer is a slowly changing dimension on Marital-Status. Explain how you would handle this change during data staging. (4)

5. Explain what the term aggregate refers to, in data mart design terminology, and provide an example of an aggregate that may be constructed against the FlyAway data mart. (8)

6. Recall that your data mart is designed to host the data from the last five years. Consider a query that returns the total amount (i.e. the `sum()`), as paid by Customers from Ontario, who purchased travel policies on 14 February 2017. Show how you would use semi-joins and bitmaps in order to optimize this query. (10)

7. Suppose that TravelAway also maintains a table that tracks the number of claims of travellers at a destination, i.e. `Claims(Destination, Type, Date, Customer-Name, Amount-Claimed)`. Provide the SQL or MDX statement to list, per destination, the amount claimed by a customer, compared to the highest overall claim at that destination. That is, the query should return data as shown in the following sample: (8)

Destination	Type	Date	Name	Amount-Claimed	Highest-Amount-Claimed
Chamonix	Broken leg	1 Dec 2016	Joe	40,000	400,000
Chamonix	Broken arm	4 Dec 2016	Ann	60,000	400,000
Chamonix	Helicopter Evacuation	6 Dec 2016	Sue	400,000	400,000
MT Everest	Broken leg	10 Dec 2016	Joel	500,000	500,000
MT Everest	Broken collar bone	3 Dec 2016	Anne	250,000	500,000
Mt Everest	Sprained ankle	2 Dec 2016	Susie	100,000	500,000

8. Suppose that you preprocessed this data and now wish to use a supervised learning algorithm. The target class is ``risk-level`` which takes one of the two values in the set {low, high}.
    a. Explain what a Bagging ensemble is and discuss when one should consider using this method. (4)
    b. You are asked to construct a classification model and the domain experts inform you that accuracy and interpretability are both important criteria. You consider using a decision tree, a support vector machine or a K-nearest neighbor technique. Which one would be most suitable for this task? Motivate your answer with a discussion of the pros and cons of the three techniques. (12)

9. Currently FlyAway also collects, in addition to the Policy table, a large amount of demographic data about their customers. This information includes details such as income, occupation, travel-interests, previous destinations, travel companions, car rental details, and so on. The current dataset holds the information of 5,000,000 customers, and has 500 attributes.

    The domain experts suggest using the k-means algorithm in order to construct demographic clusters.
    a. Explain how the k-means algorithm calculates the distances between customers. (4)
    b. Suppose that the number of clusters is set to six (6). Explain how the k-means algorithm would proceed to form the clusters associated with the data. (8)
    c. Explain how you would evaluate the quality of the clusters, assuming that there is no "ground-truth". (4)

10. Consider the table that contains ``baskets`` of activities that past Customers combined while on a trip. Show the association rules that were generated after you applied the Apriori algorithm to this dataset. You may assume a *minimum support* of *two*. (10)

TID	**Items**
100	(Cheese-tasting, Scuba-Dive, Hiking, Skydiving)
200	(Parachuting, Hiking, Museums)
300	(Cheese-Tasting, Museums, Hiking, Scuba-Dive)
400	(Parachuting, Scuba-Dive, Hiking)
500	(Parachuting, Hiking, Scuba-Dive)

11. Explain what contextual outliers are, and provide your own example of a contextual outlier one may find in the FlyAway case study. (4)

Fini. Enjoy your summer!