Assignment #1 Intro into Data Mining by Joshua Troup

Q1. If data mining techniques are to be used in the following cases, identify whether the task

required is supervised or unsupervised learning. (8 points)

Also, briefly justify reasons for the task as supervised or unsupervised learning.

a. Deciding whether to issue a loan to an applicant based on demographic and financial data

(with reference to a database of similar data on prior customers).

**Supervised learning. Classification methods are used to distinguish between customers that have been successful in serving the loans or the ones who were not.**

b. In an online bookstore, making recommendations to customers concerning additional

items to buy based on the buying patterns in prior transactions.

**Unsupervised learning. There is no classification or prediction. The goal is to segment data into meaningful segments or patterns**

c. Identifying a network data packet as dangerous (virus, hacker attack) based on comparison

to other packets whose threat status is known.

**Supervised learning. Classification methods used to distinguish between packets that have a virus or not.**

d. Identifying segments of similar customers.

**Unsupervised learning. No classification or prediction. No meaningful segments or patterns.**

e. Predicting whether a company will go bankrupt based on comparing its financial data to

those of similar bankrupt and nonbankrupt firms.

**Supervised learning. Classification methods used to distinguish between bankruptcy based on similar banks.**

f. Estimating the repair time required for an aircraft based on a trouble ticket.

**Unsupervised learning. No classification or prediction. No meaningful segments or patterns.**

g. Automated sorting of mail by zip code scanning.

**Supervised learning. Classifications methods used to distinguish between mailing locations based on zip code scans.**

h. Printing of custom discount coupons at the conclusion of a grocery store checkout based

on what you just bought and what others have bought previously.

**Unsupervised learning. The customer is likely to use the coupon based on the products in the customers cart, not what others bought. No classification or prediction.**

Q2. In fitting a model to classify prospects as purchasers or non-purchasers, a certain company

drew the training data from internal data that include demographic and purchase

information. Future data to be classified will be lists purchased from other sources, with

demographic (but not purchase) data included. It was found that ‘‘refund issued’’ was a

useful predictor in the training data.

Why is this not an appropriate variable to include in the model? (2 points)

**I consider this scenario to be unsupervised learning. Refund issued is not an appropriate variable as the customer was not satisfied with the product therefore returning and losing profit for the company. Demographic is an important variable to include as more people in a certain area are prone to purchase necessities. Example of this would be cold weather gear in very cold climate regions or purchasing surf boards near the coast. Refund issued not relevant to include in a model trying to prospect purchasers.**