# CSCI E 84 Homework 1 Spring 2016

The objective of this homework is to assess your ability to comprehend lessons taught through lecture, labs, require reading covered in Chapter 1 and 2 our text(s):

- Data Science for Business
- Hands-On Programming with R

**Graduate** students will complete all essay question(s) and R Programming exercise.

**Undergraduate** student may skip 1 essay question.

## Essays Question(s) worth 10 points each (50 points)

**Q1** MTC (MegaTelCo) has decided to use supervised learning to address its problem of churn in its wireless phone business. As a consultant to MTC, you realize that a main task in the business understanding/data understanding phases of the data mining process is to define the target variable. In **one or two sentences**, please suggest a definition for the **target variable.** Be as precise as possible—someone else will be implementing your suggestion. (Remember: it should make sense from a business point of view, and it should be reasonable that MTC would have data available to know the value of the target variable for historical customers.)

**Q2** Plumbing Inc. has been selling plumbing supplies for the last 20 years. The owner, Joe, decides that next year it is time to diversify by adding gardening tools to the products. Having had success using customer data to build predictive models to guide direct mail campaigns for special plumbing offers, he considers that data mining could help him to identify a subset of customers who should be good prospects for his new set of products. **Is Joe ready to solve this as a supervised learning problem?**

If yes – what would you suggest as the target variable?

If no - why not?

# CSCI E 84 Homework 1 Spring 2016

**Q3** What is the probability that the sum of two die will be greater than 8, given that the first die is 6?

**Q4** Write a SQL query that returns the total number of sold products? You must use a join and group by function to answer this question. Please show your output in a table format.

**Products Table**

| PRODUCTID | PRODUCTNAME |
|---|---|
| 1 | Apple |
| 2 | Dell |
| 3 | IBM |

**SalesPersonProduct Table**

| SALEPERSONID | PRODUCTID |
|---|---|
| S1 | 1 |
| S1 | 2 |
| S1 | 3 |
| S2 | 1 |
| S3 | 2 |

**You may use w3schools for help if you are rusty in writing SQL Queries. You should write in Oracle PLSQL - http://www.w3schools.com/sql/default.asp**

# CSCI E 84 Homework 1 Spring 2016

**Q5** Choose one of the data technology (Q, H, U, or S) that is most appropriate for each of the following business questions/scenarios.

- Q – SQL Querying
- H – Statistical Hypothesis Testing
- U – Unsupervised Data Mining/Pattern Finding
- S – Supervised Data Mining

**a)** ___ For my on-line advertising the decision tree model yields a response rate of 0.5% and the old, manual targeting model yields 0.3%. Is the decision tree model really better?

**b)** ___ I want to know which of my customers are the most profitable.

**c)** ___ I need to get data on all my on-line customers who were exposed to the special offer, including their registration data, their past purchases, and whether or not they purchased in the 15 days following the exposure.

**d)**___ I would like to segment my customers into groups based on their demographics and prior purchase activity. I am not focusing on improving a particular task, but would like to generate ideas.

**e)** ___ I have a budget to target 10,000 existing customers with a special offer. I would like to identify those customers most likely to respond to the special offer.

**f)** ___ I want to know what characteristics differentiate my most profitable customers.

## Please fill in the blanks with the correct answers!

# CSCI E 84 Homework 1 Spring 2016

**R Programming exercise are required by all students.**

1. **R Programming Exercise worth 50 points:**

Please use the following dataset:

http://www.census.gov/popest/data/index.html

**Please answer the following questions regarding the 2016 President Election. Use R programming to show your answers:**

1 How many qualified voters in the United States – must be a citizen?

2 How many female verses male voters between the age of 18 to 29?

3 How many female verses male voters between age 30 to 44?

4 How many female verses male voters between 45 to 64?

5 How many female verses male voters above the age of 65?

## Submission

Please submit your work as a MS Word Document and R Programming Script in the homework 1 assignment area in Canvas by February 14th at 11:59 pm EST.

You should have 1 MS Document and 1 R Programming script called homework1.R. You will also need to provide your dataset so I can test your answers.

- Homework1.R
- Homework1.docx
- Dataset for R Programming