CIDM/ECON 6308 Exam 1 -Part 2 (100 points)

Due 11:59 PM CDT July 26, 2023

Part 2 of Exam 1 will help you review previous course materials and get ready for the remainder of this course, esp., data mining in Class 07 and 08.

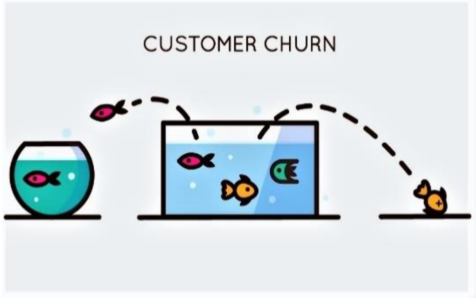
Requirements: This exam is open book, open slides, and open notes, but you are not allowed to collaborate nor discuss with anyone else during the exam period. Any question about the exam should be directed to the instructor.

You are required to follow the instruction to answer all the questions and type your answers via Exam 1 Part 2 Submission on WTClass. This is an individual exam, so sharing your Excel file, screenshots, answers, or anything else about this exam with other students or parties is considered as cheating, which will be reported to the university authority. In addition, it is your responsibility to make your answers meet the required format; otherwise, you would lose points because of wrong format.

Please indicate that you have understood and complied with such requirements in this exam.

Our textbook *Data Science for Business* mentions the example of predicting customer churn (a typical application of data analytics) in multiple chapters such as Chapter 1 and Chapter 14.

Mark just landed a great analytical job with MegaTelCo, one of the largest telecommunication firms in the United States. MegaTelCo provides both wireless and internet services and it has hundreds of millions of customers. They are having a major problem with customer retention in their telecommunication business. Many customers leave, and it is getting increasingly difficult to acquire new customers. Since the telecommunication market is now saturated, the huge growth in the telecommunication market has tapered off. Communications companies are now engaged in battles to attract each other's customers while retaining their own. Customers switching from one company to another is called customer churn, and it is expensive all around: one company must spend on incentives to attract a customer while another company loses revenue when the customer departs. According to a report, annual churn rates for telecommunications companies average between 10% and 67% (Database Marketing Institute, 2008). Customer churn not only increases operation and advertising cost, but also reduces revenue and damages brand image. As Computer Weekly cites, mobile operators spend approximately 15 percent of their revenues on network infrastructure and IT — but a whopping 15 to 20 percent of revenues on the acquisition and retention of customers (Computers Weekly, 2018).



It's long been known retention of existing customers is less expensive than acquisition of new ones. In fact, a Canadian study found it costs nearly 50 times less to retain than acquire (Telecoms, 2018). Therefore, a good deal of marketing budget is allocated to prevent customer churn. The Marketing department is going to designate a special retention offer. Mark's task is to devise a precise, step-by-step plan for how the data science team should use MegaTelCo's vast data resources to predict whether or not a particular customer is going to turn over before s/he actually leave so that MegaTelCo can offer the special retention deal to prevent customer churn. This is even more important to retain high-value customers. In order to predict customer churn, Mark and his data science team need to apply data analytics techniques (esp., data mining). In order to solve this problem, Mark plans to take the data on prior churn and extract patterns, for example, patterns of behavior, that are useful—that can help him to predict those customers who are more likely to leave in the future, or that can help us to design better services.

Considering that it is very time-consuming to download and process millions of records, Mark decides to start with building data mining models via a portion of the data via a random sampling technique. Using the database querying technique, Mark obtains a random sample of about 5000 records (i.e., customers) from the company's data repository.

Next, Mark prepares the data for initial analysis: First, as the dataset has hundreds of attributes, Mark applies feature selection techniques to include a small number of important attributes in his initial models, rather than all the attributes. Next, Mark cleans the data to solve some quality issues of the data such as duplicated and missing values. Finally, Mark obtains a dataset with 13 predictor attributes (or variables) from three categories (see the table as below) and one target attribute (i.e., the attribute of our interest, also called dependent attribute in statistics).

Before running data mining models (to be covered in Classes 7 and 8), Mark plans to first explore this problem in Excel to gain a better understanding of the data and the relationship between other attributes and the target attribute, churn.

Category	Attribute Name	Description	Values
Demographic Information	CustomerID	A Unique ID to identify each customer	Format: NNNN-LLLLL
	Gender	The gender of each customer	Male/Female
	SeniorCitizen	Whether or not this customer was a senior citizen	1=yes, 0=no
	Partner	Whether or not this customer had a partner	Yes/No
	Dependents	Whether or not this customer had dependent(s)	Yes/No
	Tenure	The length of time (in months) this customer had stayed with the company	Whole number
Service Information	PhoneService	Whether or not this customer had phone service	Yes/No
	InternetService	What type of internet service this customer had (code No if a customer had no internet service)	DSL/Fiber/No
Account Information	Contract	The contract type that a customer had with the company	Month-to-Month/ One Year/ Two Year
	PaperlessBilling	Whether or not this customer used paperless billing	Yes/No
	PaymentMethod	The payment method this customer used	Electronic Check Mailed Check credit Card (auto) Bank Transfer (auto)
	MonthlyCharges	The total monthly payment (in dollars) this customer made	Real number
	TotalCharges	The total life-to-date value/revenue (in dollars) this customer contributed	Real number
Target	Churn	Whether or not this customer churned last month	Yes/No

1. **Business Understanding: Describe Analytics Efforts (15 points)**

This case involves multiple domains such as telecom, customer, and churn analytics. Please answer the following two questions to help you better describe how analytics is applied in this case.

1.1. Analytics Domain and Orientation: Which of the following types of analytics is mainly involved in Mark's task of summarizing some patterns and reporting possible relationship in the data? [3 points]

- Descriptive Analytics
- Diagnostic Analytics
- Predictive Analytics
- Prescriptive Analytics

1.2. Analytics Activities: Please indicate whether each of the following statements about analytics activities is true or false by typing T (i.e., True) or F (i.e., False). [12 points and 2 points for each]

□ Sending a special retention deal to those customers who are predicted to churn based on the scoring score computed from the data mining model is an example of evidence generation activity.

□ Retrieving about 5,000 customers' transaction data from MegaTelCo's repository is an example of evidence selection activity.

□ Obtaining about 5,000 customers' demographic information from MegaTelCo's repository is an example of evidence acquisition activity.

□ Generating the probability that each customer is going to turn over from the data mining model is an example of evidence generation activity.

□ Creating a customized service package for high-value customers who are projected to turn over based on the data mining model is an example of evidence emission activity.

□ Producing business rules based on the data mining model and industry standard to determine which customers are our top priority to retain is an example of evidence assimilation activity.

2. Data Understanding & Acquisition (30 points)

SQL Queries: Mark selects the data based on multiple conditions from the data warehouse system using SQL queries. Our class introduces SQL and you have also completed a SQL course on DataCamp, so please answer the following questions. You may refer back to SQL Tutorial or a SQL course on DataCamp.

2.1. Mark wants to select the records in the past three years, i.e., 2020, 2021, and 2022, using a numerical attribute, *year*, from a relational database. His team comes up with ten possible WHERE statements below. Please indicate whether each of them is true or false by typing T (i.e., True) or F (i.e., False) in the front. [10 points and 1 point for each]

- □ WHERE year >=2020 AND year <=2022
- □ WHERE year >=2020 OR year <=2022
- □ WHERE year BETWEEN 2020 AND 2022
- □ WHERE year BETWEEN 2019 AND 2023
- □ WHERE year IN 2020, 2021, 2022
- □ WHERE year IN (2020, 2021, 2022)
- □ WHERE year = 2020 OR year=2021 OR year=2022
- □ WHERE year = 2020 AND year=2021 AND year=2022
- □ WHERE year > 2019 AND year <2023
- □ WHERE year = 2020 OR 2021 OR 2022

2.2. Mark wants to sort the obtained records by the variable *CustomerID* in ascending order. Which of the following statement can be used for this purpose? Choose all that apply [4 points]

- ORDER BY CustomerID
- ORDER BY CustomerID ASC
- ORDER BY CustomerID DESC
- GROUP BY CustomerID
- GROUP BY CustomerID ASC
- GROUP BY CustomerID DESC

2.3. Please indicate whether each of the following statements is true or false by typing T (i.e., True) or F (i.e., False) in the front (16 points: 2 points for each question).

- □ The variable *MonthlyCharges* is a continuous numerical variable.
- □ The variable *SeniorCitizen* is a discrete numerical variable.
- □ The variable *PaymentMethod* is an ordinal variable.
- □ The variable *Tenure* is a discrete numerical variable.
- □ As a type of incidental data, profile data such as *gender* in this case is relatively static.
- □ In this case, the data Mark is using is qualitative.
- □ In this case, Mark uses internal data that is collected by ongoing business activities.
- □ In order to convert the categorical variable, *PaymentMethod*, to a numerical attribute, Mark should create four dummy variables.

3. Data Understanding: Descriptive Analysis of Attributes (20 points)

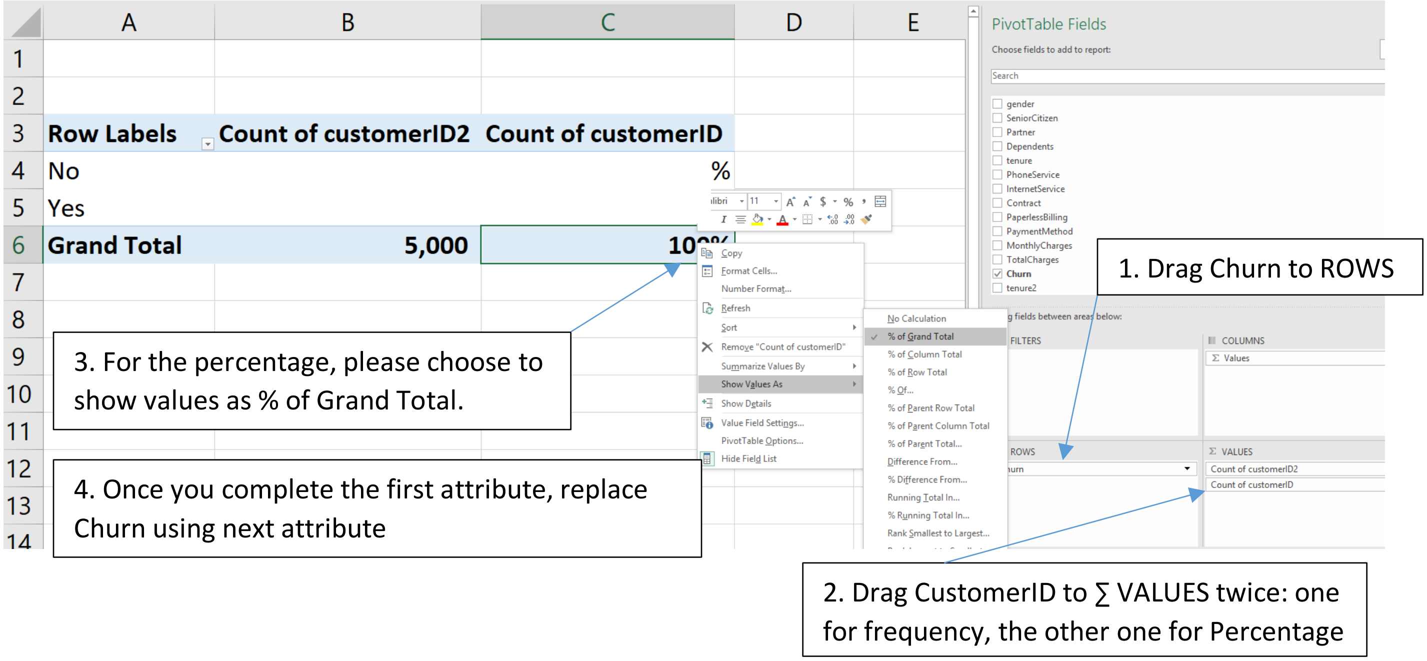
3.1. Describe categorical attributes (10 points: 1 point for each cell)

A summary table[1] tallies the values as frequencies and/or percentages for each category of a variable. A summary table helps us see the differences among the categories in a separate column. Please complete the following summary table for all the categorical attributes (a few examples are already provided). You just need to type the percentage (rounded to the second decimal place such as 74.76), and the frequency is not required.

Attributes	Category or value	Frequency	Percentage (round to the second decimal place such as 12.34%)
Churn	No	3,705	74.10%
	Yes	1,295	25.90%
Gender	Female	2,459	49.18%
	Male	2,541	50.82%
SeniorCitizen	No (0)	-	84.14%
	Yes (1)	-	15.86%
Partner	No	-	1 point
	Yes	-	1 point
Dependents	No	-	1 point
	Yes	-	1 point
PhoneService	No	-	1 point
	Yes	-	1 point
InternetService	DSL	-	34.06%
	Fiber optic	-	43.88%
	No	-	22.06%
Contract	Month-to-month	-	-
	One year	-	-
	Two year	-	-
PaperlessBilling	No	-	40.88%
	Yes	-	59.12%
PaymentMethod	Bank Transfer	-	1 point
	Credit Card	-	1 point
	Electronic Check	-	1 point
	Mailed Check	-	1 point

Note: A quick way to get both frequency and percentage is to use PivotTable for each categorical attribute (you may refer to the instruction below or in LA1).

[1] You are supposed to learn this in your previous statistics course or the leveling course, CIDM 6300 before taking this course. Here is a quick review to help you refresh your memory and get ready for the second module of this course: Technical Application.



	A	B	C	D	E
1					
2					
3	Row Labels	Count of customerID2	Count of customerID		
4	No		%		
5	Yes				
6	Grand Total	5,000	100%		
7					
8					
9					
10					
11					
12					
13					
14					

1. Drag Churn to ROWS

2. Drag CustomerID to ∑ VALUES twice: one for frequency, the other one for Percentage

3. For the percentage, please choose to show values as % of Grand Total.

4. Once you complete the first attribute, replace Churn using next attribute

3.2. Describe Numerical Attributes

Please obtain the minimum, maximum, mean, and sample standard deviation for the three numerical attributes, Tenure, MonthlyCharges, and TotalCharges and then complete the following blanks (2 points for each, 10 points in total). Round your answers to the second decimal place such as 11.11 except the ones already provided.

- The customers' tenure ranges from 0 month (i.e., new customer) to 72 months and their average tenure is 32.34 months, with a sample standard deviation of ____ months.

- The customers' monthly charge ranges from 18.40 dollars to _______ dollars and their average monthly charge is 64.56 dollars, with a sample standard deviation of ____ dollars.

- The customers' total charge ranges from 0.00 dollars to 8,684.80 dollars and their average total charge is _____ dollars, with a sample standard deviation of ____ dollars.

Hint: You can either use the formulas mentioned in LA1 (Step 1.4) or the analysis ToolPak in MS Excel. For loading the analysis ToolPak, please refer to this link; please use it to perform data analysis, please check this link (Data→ Data Analysis → Descriptive Statistics) to find all summary statistics above.

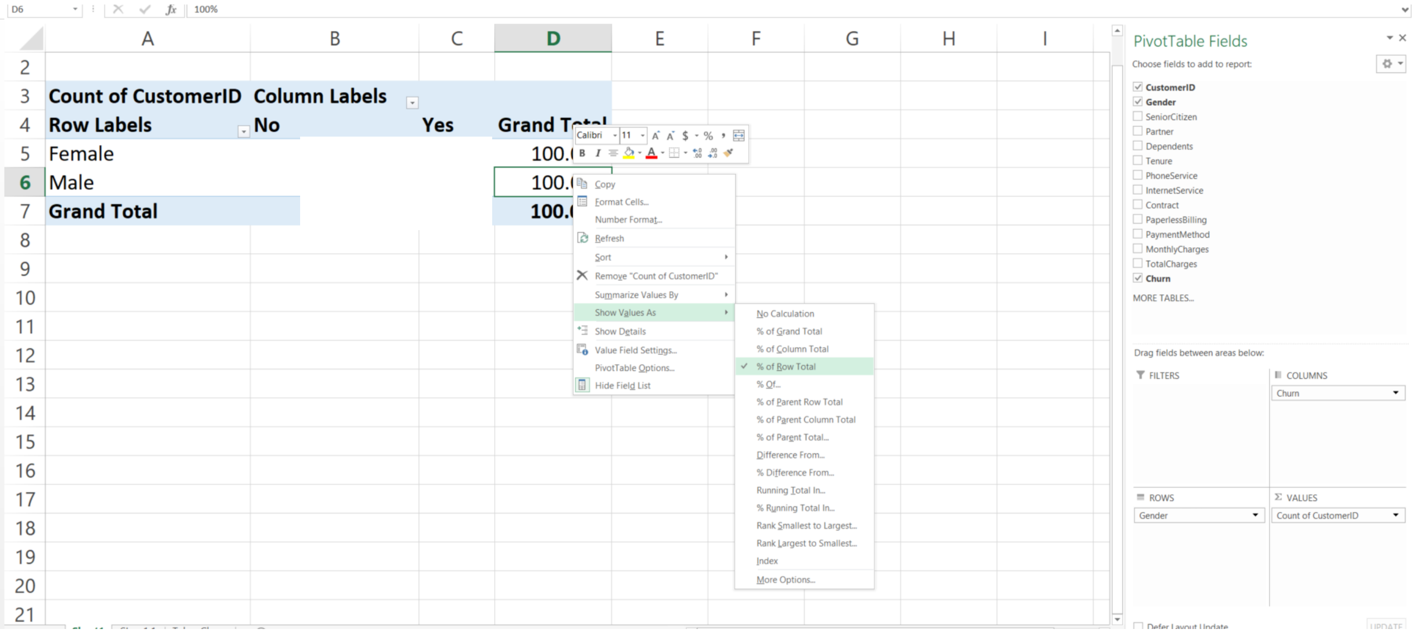
4. Data Exploration: Computing Churn Rate (35 points)

A contingency table[2] cross-tabulates or tallies jointly the value of two or more categorical variables, allowing us to study patterns that may exist between the variables. Tallies can be shown as a frequency, a percentage of the overall total (grand total), a percentage of the row total, or a percentage of the column

[2] You are supposed to learn this in your previous statistics course or the leveling course, CIDM 6300 before taking this course. Here is a quick review to help you refresh your memory and get ready for more advanced analytics techniques.

total. A multi-dimensional contingency table tallies the response of three or more categorical attributes. In Excel, we can construct PivotTable for this purpose. All these efforts are called data discovery, which is defined as the methods enabling us to perform preliminary analysis by manipulating interactive summarizations. Using the contingency table, we can compute churn rate for different groups of customers segmented by each variable and then determine whether or not such variable may be a significant predictor of customer churn before running any data mining models (Note: here, we just explore whether or not such a variable matters, rather than statistically test its significance).

For example, by constructing the following PivotTable, we find that the churn rate of female customers is 26.60%, slightly higher than the churn rate of male customers (25.23%). Because the difference of the two churn rates is quite small, we can conclude that gender may not be a good differentiator for customer churn. In contrast, using the similar way, we find that the churn rate of customers using paperless billing is 32.71% while the churn rate of customers who do not use paperless billing is 16.052%, indicating that the variable paperless billing may be a good differentiator for customer churn.



D6 | 100%

Count of CustomerID	Column Labels		
Row Labels	No	Yes	Grand Total
Female			100.
Male			100.
Grand Total			100.

PivotTable Fields

Choose fields to add to report:

☑ CustomerID
☑ Gender
□ SeniorCitizen
□ Partner
□ Dependents
□ Tenure
□ PhoneService
□ InternetService
□ Contract
□ PaperlessBilling
□ PaymentMethod
□ MonthlyCharges
□ TotalCharges
☑ Churn

MORE TABLES...

Drag fields between areas below:

FILTERS
COLUMNS: Churn
ROWS: Gender
VALUES: Count of CustomerID

□ Defer Layout Update

Show Values As: ✓ % of Row Total

4.1. Please compute the churn rate of each group of customers and round them the second decimal place, as illustrated in the two examples below (1.5 point for each, 27 points in total).

Variable	Customer group	Churn rate	Compare churn rates to determine whether it is a good differentiator
Gender	Female	26.60%	Female customers are slightly more likely to churn than males, indicating gender is not a good differentiator.
	Male	25.23%	
SeniorCitizen	No (0)	1.5 points	
	Yes (1)	1.5 points	
Partner	No	1.5 points	
	Yes	1.5 points	

Dependents	No	1.5 points	
	Yes	1.5 points	
PhoneService	No	1.5 points	
	Yes	1.5 points	
InternetService	DSL	1.5 points	
	Fiber optic	1.5 points	
	No	1.5 points	
Contract	Month-to-month	1.5 points	
	One year	1.5 points	
	Two year	1.5 points	
PaperlessBilling	No	16.05%	Customers using paperless billing are twice more likely to churn than those not, indicating PaperlessBilling is a good differentiator.
	Yes	32.71%	
PaymentMethod	Bank Transfer	1.5 points	
	Credit Card	1.5 points	
	Electronic Check	1.5 points	
	Mailed Check	1.5 points	

4.2. In data mining, one of our primary goals is to identify informative attributes that can effectively reduce uncertainty in prediction, providing valuable insights (for further details on informative attributes, refer to DSB Chapter 3). Here, we will use a simple way to identify informative attributes, which we name as differentiators.

During data exploration, it becomes imperative to pinpoint these differentiators from the numerous variables at hand. To achieve this, we adopt a straightforward rule: when a variable divides the customer base into multiple groups, we consider it a good differentiator if, within these groups, there is at least one group of customers with a churn rate at least twice as high as another group's churn rate. To better understand this concept, consider the two examples provided in the last column of the table above. These examples serve to illustrate whether a variable qualifies as a good differentiator or not. By applying this criterion, we can effectively identify the most influential attributes that significantly impact customer churn rates, aiding us in making more informed predictions and decisions. Using the same rule, which of the seven other categorical variables can be considered as a good differentiator? Choose all that apply. (8 points)

This is the end of Exam 1 Part 2