6-1 Final Project Milestone Three

Joel Meza

Professor Torre Roenne

April 11, 2021

DAT 220 -Fundamentals of Data Mining

Southern New Hampshire University

**Analysis Organization**

I had chosen the cluster analysis to differentiate between the client population into more specific sets of clients with possible like potentials. Gathering this information would be the age, income, marital status, and spending rate. This procedure will be done to differentiate between the clients that are in ads for specified clients that may be useful to aim than everyone else within the turf around the location. Diverse sectors of clients were used within all their available potentials. Clients in sets were made by differentiating the population for the sum spent at restaurants which demonstrate that the more a client is spending within a restaurant, its more possible for them to manage the online spending when clients create an online purchase. In results of sharing this information, the chances are more likely for clients in sets or patterns to highly come back while using advertisements to notify theses clients. Although, one setback from this strategy analysis is that I had only used a single variable to differentiate clients. On behalf of the cluster analysis strategy is used for higher than one variable to be as great when differentiating between customers.

However, when using linear regression to determine which client could possibly make the next online purchase. If this model is accurate to compute the developing spending rate at any given store type, then the model better used to analyze the difference from customers potentials that could affect the total spending. With this experiment or strategy, then Bubba Gump as a company will be able to research more and learn to charm more clients that do not really spend big sums of purchase transactions. As I approach by using all the existing permanent variables to compute the online store spend; however, I found the at website visits and restaurant spend sums are mostly used. When using these two variables within the model, this could not be very effective to use which I found more effective outcomes that demonstrates the spending rates of other stores that create a dissimilar model type for diverse clients in sets from using the cluster analysis. The data used about theses variables in relevance to the linear regression are restricted by the variables comprised into the model.

In reference when using the logistic regression is determined in a client’s that create an online purchase transaction. For a logistic regression it can be computed the possible of other given potentials in which a client makes a purchase in the restaurant or on a precise sum spending. This data could be useful to increase any data about the most significant problems that join with precise type and can be useful to develop ads strategies for growth in sales. As I’ll continue to analyze the information using logistic regression to compute other sets, but it’ll be done also within different types created by the cluster analysis.

**Sources of Error**

Some cases there were some data that was not very rational like such client demonstrating a zero value from the web spend and one value of an online purchase which was created. Although, I was not able to verify the right value, I took the pleasure to eliminate for this information from the information set. However, I will be able to overlook these sets of information before examining these types of issues that I could logically distinguish. Second source of error that was not typed in columns from the JMP data set was the Web Purchase (Yes/No) evaluation which was arranged as ordinal information which is restricted to use the information in certain analyses. Altering the column’s modeling sort to minimal. For the last source of error was the geographic location which only shown the first two digits in the zip code which are incorrect because of the zeros being eliminated from the information set. Altering the zip code digits in the front to be zeros for the first four and altering the information type of each character while computing the first two digits in the potential sources of each error from the client when completing the survey to be applied from the information set. Therefore, a data analyst has to be aware and willing to alter these changes to correct the source errors which is found from our database. As I approach further, I’ll do my best to analyze the information given sets which I will find the cluster analysis to differentiate from clients that trail by other model strategies to compute both spending sums and clients that create a transaction from purchases.

**Meaningful Patterns**

I found the only meaningful patterns to be the clients who spent much at restaurants that manage to spend as much in online purchases. As prior to my research questions this pattern is shown as, “Do clients who spend from third party also divide the equal association with a restaurant rate of spending? What’s a typical amount that a client who spend very much more from a restaurant and online purchasing?”

**Inaccurate Depictions of Data**

For the only imprecise of information which was made by the correlation matrix and regression analyses. Prior to the big size in the sum of zero value information for spending totals. This could be an example to find a solution which is removing zero values from these analytics.

**Alternative Analytic Methods**

Further analysis approaches would be the differences in hierarchical clustering, decision trees, K/means clustering, and regression approaches where these analytic methods have strengths and disadvantages which must be assessed beside the particular analysis being completed. For example, the decision trees don’t require any accounts of linearity and therefore could be useful since many of the clients’ qualities which weren’t found to be linearly associated with our outcome variables. Also, decision trees are simple to understand and associate with business. Another strategy could be parting clients prior with the demographic location information given which is not used from the information set but can be associated to the client’s location. This could be very expensive which could determine the cost of living around the area, which is determined to qualify the income gross, and to located which Bubba Gump Co. restaurant or third-party stores are from distance location.