8-2 Final Project Submission

Joel Meza

Professor Torre Roenne

April 24, 2021

DAT 220 -Fundamentals of Data Mining

Southern New Hampshire University

**Introduction: Business Problem**

Prior to the business channels, why has the sales flattened and diminished from the historical past of two years from the Bubba Gump Shrimp Company?

From the Bubba Gump Shrimp Company, I see it has faced difficulties within sales. Although, I have seen it spike up high in the historical past in sales when the famous name “Bubba Gump” was announced in the movie Forrest Gump and created a great contribution for customers’ interest to come close to the company’s product. Now, it seems that within the past years the company has struggled to make profit in revenue from sales. However, if the sales rate in making profit continues to flatten, then it will lead the company to bankruptcy or into a small business to be ranked throughout nationwide.

**Introduction: Analytical Method**

From this issue, I will apply a method called “Cluster Analysis” along combining it with feedback given from five hundred clients which acquire information that may be beneficial for the Bubba Gump Shrimp Company. Therefore, when using this analytical method would be an effort to attain portions of afforded data about its products which could reflect on gain and loss in sales, statistics, and being aware from the marketing profits from its diverse trades. As I continue using this analytical method, this can offer more data to better help in solving the reduction in sales. There could be a portion that may need more attention which is impacting sales, while this is going, we apply “K-means cluster” which we already know what portions we’ll be in use. As an advantage we would be able to analysis and check for different styles which could lead us to modify the number in clusters that we got, lest we find different discoveries as we proceed with our search.

**Types of Information to be Used**

Accessible Information:

* Retail Partner Sales
* Loyalty Program
* Five Hundred Feedback Surveys
* Online Sales
* Website Channels Performance
* Point of Service System

Non-Accessible Information:

* Client’s Evaluation
* Employee Wages
* Marketing Tactics
* Total Profits
* Third Party Appraising

**Data Mining Tools**

For the data mining tools, I’ll start by using the JMP to guide me as my key tool to strive into data mining. Also, I think it’s very useful and friendly whether information is being used in large- or small-scale amounts, this will organize it to its full potential ability to be used. However, the alternative tool that I would use is Microsoft Excel to organize large or small sums of data mining.

**Data Visuals**

On this kind of data, I’ll start to apply the Cluster Candle Charts which will enable for me to see the Cluster sets to relate them by its equal type. I think this will be my top option to utilize in a visual tool due to the neat and stacks the organized data where its manageable and easy to be understood.

**Research Question**

For the research question I know it should be, “What sets does it generates on its maximum or minimum in sales for the Bubba Gump Company?” Initially, I thought of a different question to reply which was “How come diverse channels of business have diminution on the previous two years?” Although, it was difficult for me to react to this question due to not enough information to track down where there isn’t a time frame. So, I started to focus on Bubba Gump’s Company where the strengths or weakness points in sales are. Which this could help me discern and understand from given feedback on what Bubba Gump company inefficiency to provide good sales and service.

**Research Measurement**

In conclusion, from my research question on reference being categorize to the minimum information trending supplied with the assumption of clusters which represent lowest client sales. Therefore, I’ll have to execute by measuring out the results given to resolve the problems in sales that are set on diverse groups. For measurement purposes, the future upcoming trends in sales will be able to reflect on the diversity and difference to determine how the sales have evolved.

**Follow-Up Question**

Prior to researching and preparing I tried to solve the major issue which is why sales tank and has a high rate in loss over the past two years. Furthermore, I will analyze in detail how the active trending is doing in each group. Either if it’s by age, gender or marital status for each client where I believe this could lead me to know more in depth about the trends in limited visits to the application, websites, and maybe other sub-contracted vendors with purchases. As I continue to analysis, this will enable me to catch up from these clusters the total increased sales as I mentioned in my earlier question and summarize the results.

**Research and Support**

As I start with my research and gather more information, I’ll apply what I get from online sources like articles, press releases, and great catalysts which provide sufficient data that has tanked the sales in the past. As I strive to look in depth, I know I will find more data that can lead me towards evidence and how sales percentage has flown and decreased in such amount of time in the past. The evidence will give me more data towards the diverse number of sales given in retail stores and third-party vendors around the nation. In support, these gathered resources from online websites and articles would give me more insights on a positive note by the famous known film called “Forest Gump” that can provide enough evidence to support my conclusion or hypothesis.

**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 must 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 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.

**Display and Interpretation**

|  | **zip** | **Restaurant** | **RES\_VISITS** | **Webstore\_Spend** | **WEB\_VISITS** | **THIRD\_SPEND** | **THIRD\_VISITS** | **Age** | **Income** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| zip | 1.0000 | -0.0602 | -0.0906 | -0.1043 | -0.0928 | -0.0053 | 0.0465 | -0.0522 | 0.0085 |
| Restaurant | -0.0602 | 1.0000 | 0.5955 | 0.4534 | 0.2078 | -0.0427 | -0.0379 | -0.0033 | -0.0159 |
| RES\_VISITS | -0.0906 | 0.5955 | 1.0000 | 0.2943 | 0.1839 | -0.0801 | -0.0846 | 0.0045 | -0.0307 |
| Webstore\_Spend | -0.1043 | 0.4534 | 0.2943 | 1.0000 | 0.6119 | -0.0059 | -0.0034 | -0.0368 | -0.0299 |
| WEB\_VISITS | -0.0928 | 0.2078 | 0.1839 | 0.6119 | 1.0000 | -0.0409 | -0.0103 | -0.0037 | 0.0301 |
| THIRD\_SPEND | -0.0053 | -0.0427 | -0.0801 | -0.0059 | -0.0409 | 1.0000 | 0.7422 | -0.0827 | -0.0601 |
| THIRD\_VISITS | 0.0465 | -0.0379 | -0.0846 | -0.0034 | -0.0103 | 0.7422 | 1.0000 | -0.0768 | -0.0636 |
| Age | -0.0522 | -0.0033 | 0.0045 | -0.0368 | -0.0037 | -0.0827 | -0.0768 | 1.0000 | 0.1093 |
| Income | 0.0085 | -0.0159 | -0.0307 | -0.0299 | 0.0301 | -0.0601 | -0.0636 | 0.1093 | 1.0000 |

From the first step into my visualization from the given information was to allocate the table shown above while it precisely shows the highlights correlation along with the client’s information. Upon building this table, I found that this data is used for variables that are continuous which make a great fit into the multivariate analysis sort. Also, this table has helped discover which of the variables that are relevant to each other where it clearly leads me to the path to understand how the sales in Bubba Gump company have decrease over the years. Although, I was capable to find the relevant information from each category in restaurant visits, restaurant spend, webstore visits, and webstore spend where I could start to emphasis on the highlighted data given from the chart table above.

**Scatterplot Matrix**

Diagram

Description automatically generated

**Visual Evaluation**

On this table shown above, I could discover in the correlation table to focus down on the variables which are given from data within the scatterplot matrix table. This continues the process of information to gather along with the continuous variables listed as shown above.

****

**Resulting Decision Influence**

From this table above, I am capable to demonstrate a relevant relation from the continuous variables in a particular analysis model. The scatterplot correlation shows the best representation where the regression line specifies how clients are spending more in a Bubba Gump Company location that consumes more by spending online site rather spending inside a restaurant. Although, we know that if a client interest is in restaurant shopping, then it is more likely to spend more sales online shopping due to their use pattern of interest.



The above graph shows the regression logistic model which is supportive tool to calculate the future spending of Bubba Gump’s company existing clients’ information. From the scatter plot table specifies clients which have diverse visits online rather than the restaurant visit that may likely be attracted to purchase online.

**Validity, Reliability, Limitations**

Prior to working on the client information given, I will continue to categorize what the ideal points which I will strive to emphasis on. At first, I will use the correlation from the restaurant spend and the total spent in Bubba Gump web store. Another point would be the relevance in the number of online visits and the possibility of successful purchase transactions. From this correlation its observed from all the diverse methods that are applied on analyzing the client’s information given statistically.

Prior to validity and reliability given the information sets, there are a margin errors to think thoroughly before seeing actual outcome results. The information accumulated was only created upon client’s feedbacks that are given in assessment reviews from there point of interest views. However, there is some type of information that did not have quite enough data to emphasis in, but I continue looking for relevant information to continue my analysis and its validity from this project account.

**Next Steps**

For my next steps I will continue support my growth of sales from the Bubba Gump company to finally achieve the times or number of visits between the restaurant and the web online from its history. Prior to my project account here I most likely pursue the client future transactions in online purchases growth total where they have visited either online or in person at the restaurant. If I could find more relevant feedback, then I would start by extending my voice to our clients who made in the past huge amount cost in purchases to find what the interest was from the product of Bubba Gump company they were attracted. While we know as mentioned previously how clients could spend much in the restaurant, then they will proceed to make more purchases online.

Therefore, I find another possibility in reviews is to count on the geographic location which serves clients most often. In example, if there are big groups in online transactions by clients in a precise location, then Bubba Gump company will count on opening a restaurant close by that location. Within this data given, it’s very important to find potential sales with great sums of purchases to priorly open a new restaurant to receive a return of great investment. Overall, I do appreciate the work with Bubba Gump cliental information given for access to investigate and optimize revenue in sales which have been impacted in the past years while continuing to meticulously analysis, emphasis, and account in this project.