Timothy Kelly

DAT-220 Fundamentals of Data Mining

8-2 Final Project

Updated: July 30, 2021

Originally Created: June 27, 2021

**Business Problem**

The business problem that will be solved which includes customer data collected and provided to us as a main resource will give the Bubba Gump Shrimp Company a better understanding of customer tendencies opposed to an overall generalization. This will be done through visualized graphs depicting patterns and trends.

**Analytic Method**

From an analytic standpoint, we can use variables in areas of interest such as age demographics to target a certain age group or number of visits to determine which restaurants are more frequently visited than others. While it is just a small sample set from a large restaurant company, we can still gain valuable insight into other areas as well. We can use other values such as the amount spent, web store visits, and how much was spent on the website as well. All this data has valuable meaning to the little things that can be achieved from it. Bubba Gump will be able to develop techniques and strategies aimed towards these specific customers. For example, customer ‘x’ has spent a certain amount of money on the website, whereas customer ‘y’ has not spent anything. We can now set up a marketing strategy aimed towards customer ‘x’ to include areas of interest that directly involve the website, whereas customer ‘y’ will receive a different strategy that does not involve the website. This is just one specific area amongst the several items of data that we can extract and develop from this information.

**Analysis Tools**

I chose JMP as my analytic software choice for this research. This is mainly due to my experience and knowledge with this material. I will use the JMP program to analyze the data from the Bubba Gump Shrimp Co. with multiple depictions and visualizations of what we have learned. This program should provide all the necessary results for this particular business insight.

**Data Visualizations**

The types of graphs chosen in this report include a linear regression, logistic regression, and a correlation graph that displays a density ellipse of 95%. I will use these depictions along with the data such as age and restaurant visits to help interpret how to use this data for different marketing strategies for the stakeholders.

**Research Question**

The business question that needs to be answered based on the data would be how the company can more accurately market each customer or each group of customers and make it unique to each one to maximize revenue.

**Research Measurement**

I believe there are several ways in which it can be determined if research was successful. One being user feedback from the company stakeholders directly after delivering the visualizations to them. This will let the company know if the marketing team has a better understanding to utilize this data. Another measurement may not be available till after some time, but it would be the actual results from these marketing strategies to determine if they have been successful based on the results.

**Follow-Up Questions** Some follow up questions to our research may include what other statistics may become available after the implementation of these new marketing strategies based off the data. This data may provide insight into restaurants that may not be doing so well and what can be done to either increase revenue or if other actions need to be taken. Another question that could be asked based on this data is what the customers were purchasing such as gift cards, clothing, or other paraphernalia.

**Research and Support**

There may be other companies who have gone through similar situations that can provide useful information and data to compare against. One being the well-known Target brand store data mining technique used to promote coupons directed towards pregnant females depending on what stage of pregnancy they were in. However, something such as this can only be used in reference since it is all relative since Bubba Gump is a restaurant and Target is a convenient store, both targeting different customer bases. This support can also give valuable insight into pros and cons of the strategies that have and have not worked in the business setting.

**Analysis Organization**

Some of the aspects that are out of my control consist of the information logged in the survey by the customers who chose which fields to complete and which ones to omit. Another factor out of my control would be the types of questions requested in the survey itself. The last error that is out of my control to some degree are the errors in the data, which will be discussed further in the next paragraph.

**Sources of Error**

There were a couple of errors that immediately were questioned while reviewing the customer data. One of these being the incorrect number of people that the data was established from. The scenario called for customer data involving 500 people, but the actual list contained 502 customer entries. Another error would be the possibility of having two different zip code variables that will skew the data.

**Meaningful Patterns**

One of the patterns discovered in the data involved the web store sales related to the number of times customers visited the site. Of the 500 inquiries, most customers only visited the site once for one purchase compared to multiple visits or purchases. This pattern informs the stakeholders that many customers don’t return to the web store and most of the time are only making one purchase.

**Inaccurate Depictions of Data**

Graphical user interface

Description automatically generatedMost of my inaccurate data depictions involved my own human error with trial-and-error working through the different depiction options. However, the two biggest discrepancies I had to overcome were that of the two zip code variables and the two customers with incomplete data that put it over the scenario of 500 customers. After some of the readings and resources, I chose to eliminate the extra two customers since this data was incomplete in the first place. I also chose to omit the zip code variable that involved just two numbers instead of that common five digits. After some quick research that the two-digit zip codes refer to areas of the country and then the five digit refers to the specific location, I chose to also omit the two-digit zip code variable as it seemed rather too generalized in seeking specific data for this business question. Additionally, the screenshot displayed proves what happens when the incorrect variables are used to determine these patterns and trends. The attached logistic chart cannot be used because these variables cannot predict any correlation or trends.

**Alternative Analytic Methods**

After working with the Bubba Gump data and the multitude of options available in JMP to analyze the results, I can think of a few possibilities of alternative methods based off our resources. I believe association analysis could be used to determine which variables occur together more commonly. Even sequence analysis could be used for data mining purposes to determine the time-period of web sales made. For example, if we wanted to see if there was a pattern when customers make purchases at a given time of year such as the holiday season. This would then give the marketing team ideas to use strategies such as promotions to boost sales this time of year, or how they can increase sales during slower times of the year, if possible, as well.

**Display and Interpretation**

Below you’ll find some graphs that depict where most of the revenue is coming from with web store visits. You can see that most customers are only visiting the store once for a one-time purchase. The other obvious observation is that the more visits to the web store by the customers, the more revenue generated. The marketing team can use this information to try to increase visits to generate revenue and return customers. These interpretations are depicted in the followed charts.

**Chart

Description automatically generatedGraphical user interface

Description automatically generated**

**Chart

Description automatically generated**

**Further Statistical Insights**

I added in further charts to enhance and convey the types of patterns and trends we are searching for when data mining. For instance, the first set of charts let the stakeholders have a correlated viewpoint of clustering and the logistic breakdown of these clusters. These clusters are broken down into three categories by number of purchases made online. Again, there is a trend that most people are making one initial purchase from the online store, signaling to the marketing team that they may need to adjust their strategy to increase repeat visits and spending. The last two one-way analysis charts depict the statistics further by how much was spent by state and city respectively. This lets the stakeholders know which restaurants by location are generating the most income and which ones may need further addressing. All these types of charts and statistics help companies determine where to focus their resource and assets.

**Chart

Description automatically generated Graphical user interface, chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Chart, scatter chart

Description automatically generated**

**Validity, Reliability, Limitations**

When it comes to the validity and reliability of the information and the visualizations, I chose to stick with the instructional resources provided to us and constantly compared the Bubba Gump data depictions against example models used in the text and additional resources provided to us throughout the semester. My limitations mainly consisted of the amount of information I have processed and retained from this course to help me decipher the proper data mining visualizations. Another limitation consisted of the data set itself since it had more entries than initially stated which had to be accounted for, or in this case, omitted due to their insignificance.

**Resulting Decision Influence**

To disseminate and facilitate all the data to a potential client or supervisor, I believe visual comparisons will be crucial to really convey the message to interpret the results. By using the generated graphs from the data set, the targeted client or supervisor will be able to visually understand where the sales are being made and the amount of money being generated. By comparing the graphs against the data set, the target audience will be able to understand that the data set is very uninformative without the use of data mining visualizations which provide a better and deeper understanding of patterns and useful information for marketing strategies.

**Visual Evaluation**

I chose these depictions to represent where most of the sales are being generated from as well as the representation to easily convey to the marketing team and stakeholders. This will get the point across on where the teams effort needs to be directed to increase or maintain revenue. My logistic regression graph clearly displays the message on where the drop off is on sales in the online store. My other graphs display the high population of the one-time sales and how staggered it becomes after this.

**Next Steps**

Some of the next steps include the implementation and evaluation of a strategy from the marketing team based on this data. The plan or strategy that can be developed from this data will have to be revisited at another point in time to evaluate if said plan is actually working and up to the teams’ standards. I believe further lines of inquiry include what other patterns can we interpret from the analysis options available to us. Another inquiry should include if there are another line of questions the marketing team wants answers to that can be generated from this data.

**References**

Ahlemeyer-Stubbe & Coleman (2014). *A Practical Guide to Data Mining for Business and Industry*. Retrieved from https://mbsdirect.vitalsource.com/#/