Joseph Silva Jr

10/21/21

SNHU

DAT 220: Final

**Introduction: Business Problem What is the overall business problem you are trying to solve?**

The overall business problem I am trying to discover for Bubba Gump is leading the restaurant to a route which benefits them by showing them the restaurant’s areas which lose revenue and the areas which are benefiting the restaurant. By find these areas, I will help the restaurant find what modifications are needed to prevent loss of revenue from Bubba Gump Shrimp Company and create new routes of services to increase its revenue growth. In order to accomplish this task, I am using a strategy of creating data models by using a small, controlled group of 500 customers’ information from collected data by Bubba Gump Shrimp Company and presenting these models to higher ups of the company.

**Introduction: Analytic Method What is the purpose of the analytic method/approach/strategy you are using? What type of information does it yield?**

The strategy being used is to harvest information from data mining on a group of information from 500 customers. The data will yield information such as purchase history and customer’s responses on their satisfaction with the company. Using the information, I can create histograms, data results, and charts of the results gathered by the 500 customers. These designs will allow me to show the company a visual map on the areas needing modification through improvement and/or creating of new routes to benefit the company through increase growth of revenue.

**What data mining tools will you use to perform the analysis? Why these particular ones?**

The data mining tools I would use to perform the analysis is the JMP Pro data analysis tool. I have already opened the data file using JMP and it showed the information needed. JMP was also able to create graph tables and information using the Bubba Gump data. I also would use this tool system because I have used it for a prior activity in this class and it helped me accomplish the activity. JMP also has many weaknesses such as it is very difficult for new users to learn it. JMP is also very expensive costing approximately $1500 per year in order to use the software (1).

**What data visualizations will you use in your report, and why?**

Using the Bubba Gump Data set, I will use Histograms, quantiles, CDF Plot and summary statistics. Another graph, I would use a pie chart to show a visualization of the revenue brought into the corporation. For the Histograms, quantiles, and summary statistics, I can show the number of purchases has been used to purchase items whether it is from a third party, location, or online purchases. These breakdowns can allow the employees of Bubba Gump a visualization for the statistics of the purchases. I will also these types of data visualization to create a further detail breakdown for the restaurant purchases and the different states. With this information, I can predict which state is first and last for sales, which can be used to help the corporation in showing the areas needing assistance. If I obtain timeline information of revenue, I will use a CDF plot to show an increase or decrease of revenue over time. For a pie chart, I would use to show the breakdown of the total revenue being brought in by percentage using the Restaurant, Webstore, and third-party data columns. I do not believe these visualizations whether they be graphs or numbers have a weakness. The visualizations are created to show executive staff of Bubba Gump, who might not have a background in mathematics, a way to easily explain the data. The plots and numbers can be explained to the executive staff by us when we are giving a breakdown of the information to them during a presentation.

**What is the specific research question that needs to be addressed? What research question will you work from in order to analyze the given data for meaningful patterns?**

A specific research question that needs to be addressed in order to analyze the data to have meaningful patterns would be “What area of sale bring in the most revenue?” and “What area of sale brings in the less amount of revenue?”.

**How will you determine if your research question was answered or if your hypothesis-generation was successful? How will you measure progress?**

Using the data visualization and the data, I can determine the answer to my research question by showing a breakdown of the total amount of revenue from each area of sale. Prior in this milestone, I described how a pie chart can show this breakdown in percentage to show which area of sale needs assistance. I can measure the present progress by using the pie chart and I can measure progress overtime by comparing each area of sales using a CDF plot, if I am given timeline information for each area.

**What are cogent follow-up questions or explorations that should follow from your initial research?**

A cogent follow-up question or exploration I could ask after my initial question is “What assistance is needed to help the area of sale with the lowest amount of revenue being brought to the company?” and another question I could ask would be “How can the company use its resources to benefit each area of sale?”.

**Are there any published sources or other resources that address your line of inquiry? Where do they fall short? How will they help guide your analysis?**

I searched Google and I was unable to find any published sources to assist me with this data information. I would use the resources given to me by the DAT 220 class from SNHU. I would use my book “Practical Guide to Data Mining for Business and Industry” because it gives information on general issues such as missing data information. The book will help guide me when I receive data and fall into issues by comparing examples from the Practical Guidebook to the data that I am receiving from Bubba Gump.

**Analysis Organization:**

The clusters analyzed the data for customers purchases. The web store purchases, and in-restaurant purchases were compared by using Hierarchical Cluster and a K Means Cluster. These clusters showed visual interpretation for the amount of revenue each type of purchase brought to the company from the information provided by the customers.

Next, the linear regression model explained the web channel expenditure with various characteristics of customers such as age and income. The model showed a visualization graph and a linear line comparing web store spending and the characteristics with data points. The graphs were able showed statistics of the RSquare, which compared the variables to each other and if the variables were related to each other.

Lastly, the logistic regression model explained the variables of the WEB\_PURCH\_YN and WEB\_VISITS. This model showed the relationship between customer web store visitation and if they purchased from the web store. The model showed a visual graph showing a web purchase was most likely to happen when the web visit for the customer was 1+.

**Source of Error:**

When reviewing the data, I was unable to find any errors in the raw data which was used to create the graphs. The only issues I found will be discussed in a few paragraphs below under the heading Inaccurate Depictions of Data because it does not show if a purchase was made during each web visit or if it was only one purchase out of multiple visits.

**Meaningful Patterns:**

A Hierarchical Cluster and a K Means Cluster were created using the variables of Web store spend and Restaurant (Spend). Looking at the patterns from the clusters, I was able to determine customers purchased more using the web store for Bubba Gump than in-restaurant purchases. Next, the patterns of a logistic regression using the variables WEB\_PURCH\_YN and WEB\_VISITS showed the data for customers became less when it came to visiting the web store multiple times.

**Inaccurate Depictions of Data:**

An inaccurate area of data would be comparing the variables using WEB\_PURCH\_YN and WEB\_VISITS because we are unable to determine if each visit resulted in a web purchase of an item. This inaccurate data most likely caused an issue with the percentage for the logistic regression model graph because the model showed a 100% when a customer visited the web store 2+ times. These results are more than likely inaccurate due to the 100% results of purchase from the web store from the data mined from the warehouse.

**Alternative Analytic Methods:**

When it came to the graphs, an alternative method would be to break down the data even further into smaller groups than the 500 data points. This method would be most helpful when it came to the linear regression model because the data points were too much to see if there were any patterns within the data points. The graph was unable to show if any characteristics were related when it came to the web store purchase.

**Display and Interpretation:**

Data Set Survey Visualizations:

**Graphical user interface

Description automatically generated with low confidence**

**Diagram

Description automatically generated with medium confidence**

**Chart, waterfall chart

Description automatically generated**

With these graph designs, we can see a visualization of the data taken for the survey. We can see customers tend to go to the webstore to but merchandise instead of using third-parties or visit the restaurants. The webstore can be shown to be the main source of revenue for Bubba Gump. I can also see majority of the customer only visit Bubba Gump restaurant or webstore mostly once instead of being regular customers, which can be a reason for the decrease of revenue over a certain number of years. We can see a breakdown for the age of the customers that will mostly visit Bubba Gump. We can see the age area of 20-45 for the customers have the highest rate of visitations and we can see a decrease of the visitation rate from customers after the age of 55 or below 20. These types of information can show Bubba Gump about which groups their restaurant is being able to affect whether it be age, income, and/or location.

Correlations and Associations:

**Table, calendar

Description automatically generated with medium confidence**

**Graphical user interface, chart, scatter chart

Description automatically generated**

We can see in the correlation that the third-party variables whether it be spending, or visits are very low when comparing their relationship with restaurant purchase / visits and webstore purchase / visits. The third-party correlation degree is in the negatives when comparing its relationship to the other variables. The other variables in the restaurant and webstore categories correlation values have a positive degree which shows they work in comparison with each other when it comes to the customer and revenue of the restaurant.

Diagram, schematic

Description automatically generated

Chart, bubble chart

Description automatically generated

I generated a Hierarchical Cluster and a K Means Cluster using the variables of Web store spend and Restaurant (Spend). The natural amount of clusters developed from these two variables is 19 clusters in the dendogram. While looking at the K Means Cluster data, we can see the web store spends generates the more purchases than the restaurant spends from the data collected by the customers.

Chart, scatter chart

Description automatically generated

I compared the Web store spending variable to two different characteristics of customers. The characteristics were the age and income of the customer. Looking at the graphs, we can see there are not many grouping of data points and they do not appear to line up with the linear fit line. We can also see in the summary of fit the RSquare values for both graphs. The RSquare results show how close the data is related to the fitted regression line. The RSquare is measured from 0% to 100% with 0% being a variability having no explanation or relation to the response data, while 100% means the variability has an explanation or relation to the response data means. Looking at the RSquare values, I believe age and income have no relations with the variable Web store spending.

Chart, histogram

Description automatically generated

Comparing the Web visit and the web purchase variables, we can see the graph visualization. We can see a lot of the customers do not purchase items from the web store because they do not visit the web store. Next, the graph shows there is a 50% chance a customer will purchase an item when they visit the web store for the first time. According to the graph, most customers’ data retrieved for the survey are in the area of visiting the web store either 0 to 1 time. However, looking at the graph we can see that the few customers that visit the web site 2+ times usually have a 100% web store purchase.

**Validity, Reliability, and Limitations:**

The graphs and explanations above allow us to see the validity, reliability, and limitations of these visualizations whether it be the graphs or the numbers. The graphs show how certain characteristics did not affect web store purchases by using the Rsquare numbers. We were also able to see the reliability by comparing web store and in-store purchases revenue. From prior data, I was able to tell the revenue from the webstore was higher than in-store purchase. The graphs and numbers were able to show my hypothesis was correct. Next, the limitations of these graphs were shown when I discovered information was limited by comparing the variables of web store purchases vs the number of visits to the webstore by customers. The data was limited in showing the number of purchases were made during those visits by each customer.

**Resulting Decision Influence:**

I would display the results to a customer or business by showing a comparison of each of their forms of revenue. This comparison will allow a creation of a list showing the lowest and highest forms of revenue. I will then be able to break down which areas need to be cut or modified and which areas need to be pushed forward. These break downs in data will allow companies to reallocate their spending to benefit the company to prevent any further loss and to increase any profit.

**Visual Evaluation:**

While looking at the data graph visualizations, I was able to determine to see if any variables affecting other variables such as the age of customer vs the web store purchases from the customer. I was able to also see the data comparing the web store purchase vs in-restaurant purchases from customers and the data showed the web store purchases brought in more revenue to Bubba Gump over in-restaurant purchases. In the graph with the variable of web store visits, I believe there was an error when it came to a web store purchase from a customer because it showed a 100% purchase after 2+ visits, but the data needs to be able to show if a customer purchased an item on the web store for each visit or a purchase was made for the web store visits in total. A ¼ purchase out of 4 visits is a big difference compared to a purchase being made for each visit of the 4 visits.

**Next Steps:**

The next steps to help Bubba Gump restaurant would be separating the areas that obtain profit from lowest to highest. I would then break down each path of revenue into further detail such as comparing customers’ characteristics with the in-restaurant purchases like it was done with the web store purchases. Breaking down these areas of revenue, the details from the data will be able to show where the company needs to modify, pursue, or dismantle the areas affecting revenue.

**CITATIONS:**

1. *JMP reviews 2021*. Capterra. (n.d.). Retrieved October 25, 2021, from https://www.capterra.com/p/151815/JMP-Statistical-Software/reviews/.