**Food Mart Campaign Cost Prediction**

**A picture containing text, outdoor, building, restaurant

Description automatically generated**

**IST 707: Fall 2022, Wednesday 6:00 PM Class**

**By:**

Matthew Belizaire

Nicholas Nguyen

Sky Ming Li

Tuyen Nguyen

**Introduction**

The Convenience Stores industry has grown in popularity over the five years to 2022, as more Americans have become employed. According to the IBIS World website, the market size which is measured by revenue of the Convenience Stores industry is $37.3 billion in 2022 with the expectation of an increase to 1.8% in market size. Over the past 5 years, the marketing size of this industry in the United States has grown 3.1 % per year. In terms of market size, the Convenience Stores industry is the 27th ranked Retail Trade industry and the 286th largest in the United States. [1]

The industry got affected tremendously in 2019 as the result of the COVID-10 pandemic which forced customers to stay at home due to the health and safety regulations. However, after that, the industry has been growing and in response to growing demand, industry operators have opened additional stores, expanded into new markets, readily adapted to changing customer tastes and with new technology which is more convenience for customers.

With the potential industry which can lead to billions in sales and rising energy prices and eroding fuel margins, a lot of companies have realized how importance of convenience stores. As a mature market, the convenience store segment has become highly competitive, with a battle to gain the market share and share of wallet of each and every consumer. Therefore, each player or firm needs to differentiate itself in order to win over the heart and pocket of individual consumers.

Overall, the purpose of this project which understands more about the prediction of campaign cost that Food Mart spends on its marketing strategy in order to differentiate themselves from other competitors.

**Analysis and Model**

1. **The Data**

The dataset is a csv file (linked [here](https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart)) from Kaggle. It is a comparison of media types, as gathered by different costs, promotion names, stores, food categories, food departments, customers’ information such as income, gender, number of children and other factors.

Since this is a big dataset with more than 60K observations of 40 variables, there are several ways to analyze the campaign cost based on 4 big sections which can be helpful.

* Food Section: Categories, Families, Departments, Brand Names
* Store Section: Sales, Costs, Unit Sales, Cities, States, Countries
* Consumer Section: Occupations, Genders, Martial Statues, Number of Children At Home, Average Cars At Home
* Promotion Section: Promotion Names, Media Types, Costs

Fortunately, there is no missing data in the dataset. However, there are still a few additional steps which need to be done to clean the dataset:

* Remove any countries which are not the United States because the purpose of the project is to predict the campaign cost in the United States market only.
* Remove duplicate columns or variables along with unhelpful information such as: Average Cars At Home column, Gross Weight column, Net Weight column, Education column, etc.

**Analyzing structure data**

Finally, the dataset will look like below and is ready for analysis with less observations and variables. And, this dataset only presents the information in the United States market.

Text

Description automatically generated with low confidence

And, in this report, the main analysis is the correlation between media\_type and cost. Therefore, a few more variables might need to be removed depend on which analysis method is using.

**Data Viz**

The graph below presents the positive relationship between the store sales and costs. The more spent at stores, the more sales the stores will receive. This is a good sign for spending money on marketing campaigns. In addition, the graph was also spotted by unit sales. With the range of unit sales from 1 to 6, the graph was spotted based on the unit sales. As can be seen, the more the unit sales are, the more sales and cost the store will have.

Chart, scatter chart

Description automatically generated

In addition, with food department, the convenience store is doing really well with Produce and Snack Foods. For more details, the Produce department can be any kind of fresh fruits that customers just grab and go when they stop by the convenience store. Perhaps, due to the special location of the store, most of the customers will just buy most of Frozen Foods and Household food department that lead to a high sale for those food departments. The reason for that is because they live close enough to the convenience store but are far away from the grocery store. Therefore, they will buy convenience goods for their house at the convenience store.

Since this is a convenience store, it is understandable that Meat, Canned Products, Seafood department have the lowest sale. Because people stop by the convenience store, they need anything which is quick while waiting on their gas to be filled up; therefore, they rather need something which is already cooked than meat or seafood which requires them to cook.

Chart, bar chart

Description automatically generated

From the graph below, Food Mart has used multiple different media types for their marketing campaign. Sometimes, the company combines several media types as 1 big combination or just use 1 method as the only media type for their campaign. Obviously, the most media type combination with the highest cost is Daily Paper, Radio (approximately $500,000), following by Product Attachment and Daily Paper as only 1 media type for each. Surprisingly, Food Mart has the lowest cost for TV as it seems their target customer group is not the type of spending their time on TV (approximately $150,000). However, with the combination of Daily Paper, Radio and TV, then the company has been spent more on this combination (approximately $350,000).

It is interesting point when it comes to media type for Food Mart. Sometimes, the media type by itself such as TV might not be the best fit media type for the company. However, if it combines with other media types, then it will bring the good result. That’s how the graph below shows.

Chart, bar chart

Description automatically generated

Based on the graph below, there are 5 main groups of customers who go to Food Mart the most divided by membership card. With occupation, the group of customers that come to Food Mart the most is Professional, following by Manual and Skilled Manual group. And, the least group is Clerical. It is completely understandable since the Professional group contains a variety of different jobs and anyone with that kind of job will be considered as a customer who belongs to the Professional group.

Moreover, with membership card type, the most common type is Bronze, following by Normal type. It means that consumers are not considered much about membership card type; or the type of promotional campaign which requires customers to sign up for their membership is not working effectively. Due to the unique characteristic of convenience store that people come to get gas, grab something then leave, they do not have time to sign up for the membership card. Therefore, Food Mart needs to have a better promotional campaign which can attract more customers.

Chart, bar chart

Description automatically generated

The graph below shows the average yearly income group by gender. It is not surprising that the group which makes $10K - $30K and $30K - $50K yearly goes to the convenience store the most. It can be understood that those who make this income can be retired, in the low-income group. They go to the convenience store more often because they might find what they need at the convenience store rather than going to the grocery store. Or, sometimes, people go to spend a certain amount of money on gas rather than filling up their tank. With that being said, the frequency of them going back to the convenience store is higher than other groups. In addition, people who make higher income tend to drive electric cars. Therefore, there is no reason for them to go to the convenience store unless they really need something at that moment. Otherwise, they can wait to go somewhere else.

And, gender does not play an important part in this analysis. Based on the total of 19,200 female and 19,692 male, the graph does not show a big difference on gender no matter what type of income group it is.

Chart, bar chart

Description automatically generated

1. **Data Analysis Method**

**Association Rule Mining**

Association rules are theoretical “if-then” statements and help to show the probability of relationships between data items, within large data sets in various types of databases. Association rules have many applications and are widely used to help discover correlations and co-occurrences between data sets. [2]

Association rules are created by searching data for frequent if-then patterns and using the criteria *support* and *confidence* to identify the most important relationships. [3]

* **Support** – an indication of how frequently the items appear in the data.
* **Confidence** – the number of times the if-then statements are found true.
* **Lift** – A third metric which can be used to compare confidence with expected confidence, or how many times an if-then statement is expected to be found true

In applying the Association Rule Mining method to this project, the outcome is set as RHS (or Right Hand Side) to see how other factors might affect to the outcome.

First, the data set needs to be coerced to transactions for mining. After that, below is the histogram showing the most frequent items, which are helpful for the purpose of better understanding which items might appear in the mined rules.

Chart, bar chart

Description automatically generated

The data set is tested with several supp and conf values in order to narrow down the appropriate number of rules in a way that can be analyzed.

1. **supp = 0.002 and conf = 0.5**

When running the ARM with supp = 0.002 and conf = 0.5, the problem is it comes back with appropriately 1,450,135 rules as the table below. That’s a lot for analyzing purpose.

Text, letter

Description automatically generated

1. **supp = 0.2 and conf = 0.8**

When increasing supp to 0.2 and conf to 0.8, now it only has 2 rules as shown in the table below. This does not have enough rules for analyzing.

Text, letter

Description automatically generated

1. **supp = 0.11 and conf = 0.8**

When decreasing supp to 0.11 and conf to 0.8, it shows that there are 32 rules, which seems enough for analyzing.

Text, letter

Description automatically generated

Out of 32 rules, the table below shows the top 10 rules sorted by decreased lift. The reason for sorting based on lift value is because lift takes into account the support for a rule, but also gives more weight to rules where the LHS and/or the RHS occurs less frequently. The larger the value of lift, the more interesting the rule might be.

Since the purpose of this project is to find out which group of customers will decide to go to Food Mart mostly, then Food Mart should focus on advertising to that group. It is understandable that with Rule 7, 91% of the customers from the data set with the average income from $10K to $30K and normally go to the store type which is Supermarket have Normal membership card.

A picture containing calendar

Description automatically generated

* 1. **RHS: {occupation = Professional}**

To consider the group of customers who are categorized their job as Professional, their average income must be at least above $70K annually. For example, as the table below, with Rule 1, 97% of the customers go to Food Mart has their average yearly income from $130K to $150K and mostly, they are male and single. Same as Rule 1 but when adding the Bronze membership card to the rule, the percentage is only 89% which is still high. With that in mind, the marketing department can target this group of customers when building out their strategy. Since this is the high-end group with high income, the company should focus on satisfying their needs and make them not think they are at convenience stores.

Graphical user interface, text, application

Description automatically generated

Another way to launch an effective marketing strategy is to look at the cost which was spent on marketing for media types and promotion names. Then, the company can evaluate if the company should continue to promote the same campaign or change to a different way.

First, it is needed to set up the cost into 5 categories in the range of the dollar amount: 0-55, 56-80, 81-105, 106-130, 131+.

* 1. **RHS: {cost = 56-80}**

With the cost from $56 to $80, mostly, the company came up with the promotion: I Can’t Believe It Sale and the type of media is Product Attachment (as the table below shows). Since this is the low cost, Product Attachment is the right direction because the target group of customers is those who come to the store and buy any products. It will make them come back since they have a discount or any promotion code for their next purchase.

A picture containing table

Description automatically generated

* 1. **RHS: {cost = 81 – 105}**

When increasing the cost to the range from $81 to $105, as can be seen below, the company also changed the media type and came up with a different promotion name. Since it costs more, the media type also targeted to more customers through Radio, TV or Daily Paper even though they have might or might not never been to Food Mart. And, with those media types, Food Mart also focused on some big cities such as Beverly Hills, Los Angeles where mostly, the high-end customers are located.

StA screenshot of a computer

Description automatically generated with medium confidence

**Decision Tree**

Decision tree method is a commonly used data mining method for establishing classification systems based on multiple covariates or for developing prediction algorithms for a target variable.

Before using this method, the dataset needs to be divided into training and testing test.

Graphical user interface, text

Description automatically generated

With the variety of costs and media types, Food Mart wants to predict which media type the company should allocate the budget to spend more. In order to do that, the company needs to look at how much they spent on previous media types and their cost. And, it is also important to see if a single media type or a combination of several media types is working more effectively.

As can be seen below, the most media type at the top notch is Daily Paper, Radio. This node asks whether the cost of Daily Paper, Radio is larger than or equal to $59. If no, then 12% of the data set shows the media type is Daily Paper, Radio, TV. And, within 88% of the media type: Daily Paper, Radio, it will be 72% of Daily Paper type only if the cost is less than $70. Keep going down to where the cost is larger than or equal to $106, only 28% of it shows media type as Daily Paper only. And, to the same node, 44% shows that the media type is Daily Paper and Radio if the cost is larger than $106.

Diagram

Description automatically generated

While the decision tree is so clear and easy to understand how Food Mart spent their budget on media types, the accuracy of the above decision tree might not be correct due to the calculation below. It only shows 25% accurately. It is understandable since the data set is large but there are only a few nodes shown.



Therefore, when changing the codes to show more nodes as below, it is harder to read since the program must be running through the big data set. However, it improves the accuracy rate up to 73%.

Timeline, scatter chart

Description automatically generated



With that saying, it is really important to build out the training and testing data set and must make sure to understand how many nodes will be enough to analyze the cost prediction. But, the Decision Tree also indicates that the most media type is Daily Paper, Radio which is also aligned with the summary about media types and cost in the Cost By Media Type graph above. And, since Food Mart did not spend much of their budget on TV advertisement, there is no indication of TV in the first Decision Tree.

**K-Means**

One of the most common exploratory data analysis techniques is clustering. And, one of the most used clustering algorithms due to its simplicity is K-means. This algorithm tries to partition the dataset into K pre-defined distinct non-overlapping subgroups (or clusters) where each data point belongs to only one group.[3]

With this particular project, before running the clustering method, the media type needs to be removed out of the dataset for the cluster purposes.

This method will be run with several different k values.

* K: Number of clusters

Since the purpose of this project is to analyze the cost of media types; therefore, there is no need to set a specific amount of k when running the dataset. With that saying, there will be several analyses with different k values.

When running the K-Means analysis with cost as the only variable. Overall, due to the different amount of money spent on media types, there are several media types have the same cluster while others have different depending on K. It is very helpful for Food Mart if they wanted to allocate the campaign budget for other marketing advertisements in the future. And, the higher K is, the more/less common each media type has in compared to each other.

Below is example of showing the media type at different Ks:

1. K = 4

Chart, bar chart

Description automatically generated

There is no cluster 3 in Bulk Mail. That means every media type except Bulk Mail has similar traits and features together.

1. K = 6

Chart, bar chart

Description automatically generated

When k = 6, there is still no cluster 3 for Bulk Mail and there is also no cluster 6 for Daily Paper.

1. K = 8

Chart, bar chart

Description automatically generated

With k = 8, there are so several important differencies among these media types. For example, Bulk Mail does not have cluster 3, 4 and 7. Or, Cash Register Handout does not have cluster 2, 4 and 8. Daily Paper does not have cluster 6.

**Conclusion:**

There are 13 media types in which Food Mart has been used and each of them costs Food Mart differently. And, with those methods above, it is helpful for Food Mart to see how effective their media type is promoting and its cost. From there, they can direct to which media type is suitable for which group of customers.

For example, with Association Rule Mining method, for the cost from $56 to $80, the company should continue to use Product Attachment as the main media type due to the low cost and effectiveness that the media type would bring for the company if the company wants to target to customers who go to Food Mart frequently. With this type of media type, Food Mart can build customer loyalty because the more customers go to Food Mart, the more discount coupons they receive from the Product Attachment media type. Or, with Decision Tree model, Food Mart can decide which media type the company should focus on based on the range of cost. However, one disadvantage of this method is due to the large dataset, the accuracy rate might be affected depending on how many nodes Food Mart wants to show. Last but not least, with K-Means method, Food Mart can see which media type have the same characteristics then the company can allocate their budget accordingly.

Overall, depending on the purpose of the analysis, Food Mart can choose the suitable method which shows some interesting learning. From there, the company will decide which method would be best given the specific range of cost. In addition, this might also help Food Mart understand about their customers and use the correct media type.

**References:**

[1] “Industry Market Research, Reports, and Statistics.” *IBISWorld*, https://www.ibisworld.com/industry-statistics/market-size/convenience-stores-united-states/#:~:text=What%20is%20the%20growth%20rate,to%20increase%201.8%25%20in%20202.

[2], [3] Lutkevich, Ben. “What Are Association Rules in Data Mining (Association Rule Mining)?” *Business Analytics*, TechTarget, 9 Sept. 2020, http://www.techtarget.com/searchbusinessanalytics/definition/association-rules-in-data-mining.

.