BUDT704 Project Proposal:

Business Analytics of Convenient Food Mart

Group 7 (Code Pandas): Dave Anshul, Dhole Deepak, Nadipelli Amuktha Malyada, Shenoy Aditya, Simeen Afia, Zhang Zizheng

# **Introduction**

## Today, data and its application are growing at an exponential rate. Merging with business objectives, it can unravel new opportunities and possibilities in various realms. One such sector, customer acquisition and retention, benefits greatly from diverse and voluminous data. Proposing to analyze and interpret the existing customer base from a chain of convenience stores for this project, we intend to infer the store profitability relative to demographic elements, customer preferences, and effective marketing strategy thereby, contributing significantly to the commercial success of each store.

## Elaborating further, we implement predictive analytics to positively identify prospects who are more responsive to extensive advertising campaigns, promotional offers, or a particular product or service. Additionally, the broad and varied information on the storehouses can be explored to gather crucial trends and drive data-inclined decision-making for condensing or expanding store operations. Above all, marketing managers, store executives, and other relevant departments can largely leverage the insights to prepare scalable incentives to acquire and retain customers while also satisfying business objectives. Transforming this enormous information into quality data, we thereby derive actionable intelligence to enable informed strategies and optimize commercial engagements.

## **Questions of Interest**

Some questions which might be of major interest for the executives to improve their businesses can be

* What proportion of stores in the city/country mostly use recyclable packages?
* What kind of store appears to be most profitable in terms of its features and performance of store at different locations?
* What is the most effective type of media used for promotion?
* Analysing customer behaviour based on their attributes.
* Which product category performs best based on the location and store size?
* Predicting the cost of acquiring potential customer based on customer and store attributes.

# **Data Processing and Analysis**

## **Dataset Description**

Convenient Food Mart is a chain of stores across multiple locations, selling various grocery items, food products and some electronics and hardware. Once one of the largest food mart chains in America, CFM has about 325 stores spread across the country, having many stores outside the US as well.

This dataset describes various attributes of the stores as well as its customers. Attributes for the store range from net sales, costs, square footage for each section of the store to the store’s location, number of amenities offered, etc. Customer attributes cover their age, marital status, income, type of membership with the store, etc.

The dataset also provides information about various marketing campaigns, the media where these were rolled out and the cost for acquiring each customer.

There are a total of 60,000 customers whose data has been stored, in 40 columns. We acquired this dataset from kaggle.com. The publisher on Kaggle has, unfortunately, not linked the original source of the dataset.

## **Data Processing Tasks**

* **Indexing, selection and filtering**

The data doesn’t have a proper primary key, so we will need to make our own unique indices. Moreover, the data has 40 columns, not all of which will be used. Using our best judgement, we may want to select columns that will actually be useful in analysis.

* **Managing missing data**

We will have to check for missing data and see if its presence and proportion affects the quality of analysis. Depending on the column and our judgement, we may have to either impute data or drop the column,

* **Creating dummy variables**

The data has a lot of categorical variables so we may need to create dummy variables for each of them.

## **Data Analysis**

1. Recyclable package – What is the proportion of stores that are using recyclable packages? In which area (city/country) is it used most frequently?

The questions are to be solved by descriptive statistics. There is a categorical column named “recyclable package” in the dataset that tells whether or not a store uses recyclable package. To know the proportion, we count the package-using ones and divide it by the total of stores. For the latter question, we use bar chart and pie chart to visualize the data and draw conclusion. To better understand the geographical distribution of package use, we also consider to present the data on a map.

1. Profitability of store based on store features. In other words, what kind of store appears to be most profitable?

We assume that the profit yielded by a store is determined by other parameters of the store, which can be expressed formally by the following function:

p = f(x1, x2, …, xn)

where every xi is a column value retrieved from or derived from the original dataset, where p is the estimate of profit yielded by a given store. Our goal is to find such a function that predicts the profit of a store based on the other information of the store, for instance, total area, layout, etc.

For visualization, we plan to use bubble plot, line plot and scatter plot to present the regression model. Bubble plot helps present numeric relationships between a dependent variable and a set of independent variables, whereas line plot and scatter plot demonstrate the correlation between the dependent variable and any given independent variable.

1. Which is the most effective media type for promotion?

Depending on the actual situation, we consider either to select a column from the original dataset or to construct a new metric based on the dataset to measure the effectiveness of a promotion. With valid metric for effectiveness, we can then use the sample mean of each promotion category to judge their effectiveness. Bar chart helps figure out what kind of media type is most effective.

1. What is the customer behavior based on gender and marital status?

Statistic tools that help answer this situation include correlation coefficient, variance analysis, and t test. Bar chart best suits our goal.

1. What is the performance of store in different locations?

In this question we are to check if store performance varies across cities and countries. Bar chart combined with map and pie chart best visualize the analysis.

1. Which food category performs best based on the location and store size?

In this question, we probe into two pairs of correlation, one between food category and store location, the other between food category and store size. We want to confirm whether consumers at different locations prefer different food types and whether store size affects the popularity of food types. Variance analysis could be used to check whether it differs significantly as location and store size vary. We consider to visualize the data on a map, which provides direct view of the variation across locations and sizes.

1. Predict cost to acquire the customer based on customer attributes and based on store attributes.

Since we have categorical and numerical variables and we are trying to predict cost to acquire customers, for the predictive tasks, we can use simple ML models like multivariate regression.

# **Expected Findings**

Main focus of our findings will be towards increasing business operations and practices. As dataset is big enough for us to find a relevant solution, we still believe data might prove us wrong with our assumptions to result.

* In recent years, we have seen companies and stores move toward becoming more eco-friendly. However, this dataset was compiled long time ago, a time when being eco-friendly was not necessarily a priority for most firms. We expect that there won't be too many stores which will be using recyclable packages.
* Seeing that the core competency of CFM is selling food products we would expect that stores which allocate more space for this products (such as groceries, snacks, meat) would do better than other stores in terms of net sales.
* Again, considering the time this dataset was compiled we would expect that more traditional firms of media promotions, popular at that time (Newspaper, radio) would work more effectively than “newer” promotional media like TV.
* Intuitively, while looking at customer behavior, we would expect customers that generate most sales to be largely married females with children.
* Knowing that most of the CFM stores are located in the US, we would expect US stores to perform better than those in Mexico and Canada. Furthermore, even within the US since CFM was first started in the state of Ohio, we would expect stores in and around Ohio to have most number of stores and highest sales.
* We would expect locations with more higher population of families to have more sales in groceries. Whereas, other locations with other type of customers may have some different food/ product category.

**Project Timeline**

|  |  |  |
| --- | --- | --- |
| **Task** | **Task Lead** | **Due Date** |
| Data processing | Aditya | 10/19/2022 |
| Exploring data visualization libraries | Deepak | 10/19/2022 |
| Analysis on recyclable package | Anshul | 10/25/2022 |
| Data visualization of recyclable package analysis | Afia | 10/26/2022 |
| Analysis on store profits | Amuktha | 10/25/2022 |
| Visualizing store profits | Zizheng | 10/26/2022 |
| Analysis on customer behavior | Aditya | 10/25/2022 |
| Visualizing analyzed data and setting plotting standards | Deepak | 10/26/2022 |
| Restructuring and formatting with Markdown | Afia | 11/08/2022 |
| Analysis on store promotion | Anshul | 11/16/2022 |
| Store promotion visualization | Zizheng | 11/17/2022 |
| Analysis of store based on location | Afia | 11/16/2022 |
| Visualizing store analysis based on location using plotly | Amuktha | 11/17/2022 |
| Predictive analysis of prices based on customers | Aditya | 11/30/2022 |
| Predictive analysis of prices based on store features | Deepak | 11/30/2022 |
| Analysis on food category and store size | Aditya | 11/16/2022 |
| Data visualization of food category and store size | Anshul | 11/17/2022 |
| Restructuring and formatting with Markdown | Zizheng | 11/24/2022 |
| Project Report | Amuktha | 12/2/2022 |