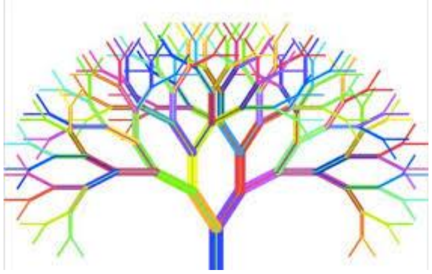

Analysis of consumer preferences and store specific attributes to predict customer loyalty and store profitability.

A Capstone Project

Vaishnavi Kommaraju

BrainStation

Contents

INTRODUCTION

As digital marketing and data science continue to thrive symbiotically, a lot of businesses concentrate on cross selling and upselling. In this scenario, the key challenge is to understand the association between different item sets purchased together and the combination of goods bought most frequently. One such machine learning technique which helps the retailers understand the relationships between products is called “Market Basket Analysis”. In other words, this technique helps the retailers to recognize patterns and intricacies between different products that the individuals purchase. As massive amounts of purchase data and consumer transaction data are stored digitally, it is important for organizations to convert this digital data into business insights to help strategic decision making and marketing. For example, if a retailer has prior knowledge that a consumer who purchases beer is likely to purchase diapers, they might place these products in close proximity to each other, cross sell such products or upsell.

OBJECTIVES

The following are the objectives of this project:

1. To understand and present an overview of the Supermarket business at Foodmart.
2. To identify the most frequently bought item sets.
3. To identify the high value loyal customers for the store.
4. To identify groups of customers with similar attributes.
5. To devise efficient market strategies for conversion of churning out customers into Platinum and Gold customers.
6. To design marketing strategies to improve profit margins for the superstore.

DATA

Food Mart data has been used for the purpose of this project, which is available open source. This dataset consists of above 200,000 transactions and 75 features (before data cleaning). This provides a perfect base to apply different machine learning techniques to get a comprehensive idea of consumer buying behavior in supermarkets.

FEATURES OF DATA

The dataset used for this study consists of the following information over the 1996 to 1998.

1. Product specific features: Product name, packaging, brand names, category, subcategory, department, gross weight, net weight, etc.
2. Store specific attributes: Store sales, store cost, unit sales, store province, store city, store district, store area, meat area, frozen area, coffee bar, salad bar, prepared food, store type, video store, etc.
3. Customer specific attributes: Income, Gender, education, occupation, homeowner, number of cars, number of children, city, province, etc.

The structure and schema of the dataset has been explained in detail in the Jupyter notebook.

THE SUBSEQUENT SECTIONS DISCUSS THE BUSINESS IMPLICATIONS AND INSIGHTS FROM THE TABLEAU DASHBOARDS ATTACHED ALONG WITH THIS DELIVERABLE.

The tableau workbook attached in this deliverable has the following dashboards:

1. **Exploratory data analysis**
2. **Customer analysis**
3. **Cluster analysis**
4. **RFM Model**
5. **Profit Margin model**
6. **Model Accuracy for data scientists**

EXPLORATORY DATA ANALYSIS

The exploratory data analysis is done on Tableau and can be seen under the Dashboard: "Exploratory Data Analysis" in the Tableau workbook.

Key considerations:

- No Null values
- No duplicated rows.
- Store Pin code is an anomaly.

This dashboard contains the following views:

- Countries in the dataset.
- Total sales
- Total cost
- Total number of units
- Average store sales by store type
- Average sales by province
- Top categories of products
- Top products within each category
- Daily sales over each month

- Shelf height
- Shelf width

Key Insights

- Overall, Canned shrimp, Plastic products, canned anchovies, pain relievers, electrical products, drinks, starchy foods etc. are the top revenue generating categories. These categories are grouped by average sales per category.
- Small groceries contribute lowest to the retail sales, while supermarkets contribute the highest.
- There appears to be a peak in the store sales once in five days in a month. This might suggest that most people visit the supermarket on specific days like Mondays, Weekends, etc.
- The dashboard also indicates shelf height and shelf width for specific categories of products. The histograms indicate that the data has many modes which indicates there are many categories/sections within each category.

CUSTOMER ANALYSIS

This dashboard indicates the turnover in terms of number of units. We have the following views in this dashboard:

1. No. of customers each day in a month
2. Distribution of customers by their occupation.
3. No. of customers each day in a week.
4. Distribution of customers by education.
5. Distribution of customers by membership
6. Income distribution of customers
7. Distribution of departments w.r.t. number of customers in each department.

Key Insights:

1. The most popular departments are produce, snack foods, dairy, and meat. However, we saw earlier in the EDA dashboard that these products are not major contributors in terms of dollar value of sales for the supermarket.
2. The number of customers entering the store exhibit specific peaks once in a week, showing busy days during the month.
3. Overall, there are a greater number of customers with bronze membership, but for higher income groups, we can see a higher number of gold and silver memberships.
4. We can see higher instances of Partial High school and high school degree, followed by bachelor’s degree among the customers shopping at the superstore.
5. Likewise, management and professionals seem to be the dominant categories for occupation among the higher income groups.

CLUSTER ANALYSIS (Part of Feature Engineering)

This dashboard presents an in-depth analysis of clusters of customers obtained from K-Means clustering. Overall, we have identified 50 clusters of customers based on the K-Means algorithm.

This dashboard presents:

- The distribution of clusters in terms of brand preferences.
- The distribution of clusters in terms of average profitability and average frequency of purchase.
- The differences in the clusters with respect to geographical location (city).

Key Insights

1. Looking at the box plots for brand preferences, we observe several outliers. Some of the cluster have extreme low preferences for certain brands, while other have very high preferences for the same brand. For the purpose of analysis, I have picked the top brands in terms of average sales.
2. These clusters are highly differentiated with respect to city. We can observe that almost each cluster of customers are part of a different city.
3. Surprisingly, lower frequency clusters of customers contribute slightly higher towards the average profitability, whereas higher frequency customers seem to be less profitable. One of the reasons could be that higher frequency customers normally visit the store for less expensive products like Produce, Meat, etc. while lower frequency customers buy more expensive products. We can also see some overlapping of clusters in this graph, indicating that 2D representation of clusters may not be appropriate since there are several other variables which differentiate these clusters.

RFM ANALYSIS DASHBOARD (Feature Engineering)

The procedure and the feature engineering process of constructing the RFM score has been explained in the Jupyter notebook attached with this deliverable. This dashboard indicates specific/distinct characteristics of Platinum and Gold customers vs. silver and bronze customers. Platinum and Gold customers are more beneficial for the business because they generate higher revenues, visit the store more frequently and are most recent customers.

Key Insights

1. There are higher number of management and skilled professionals among Platinum and Gold categories as compared to others.
2. The average sales are higher for Platinum and Gold customers.
3. A majority of platinum customers have an income between $30k - $50k.
4. Most importantly, the RFM analysis section of this dashboard indicates the position of different clusters of customers in terms of profitability. Each cluster has more than 500 customers. A majority of platinum and gold customers contribute toward an average profitability of 4.0. But we also see some silver and bronze customers at this level of profitability. The goal of the business should be to convert these churning out customers into Platinum and gold.
5. Likewise, we also see some Platinum and gold customers at lower profitability levels. As data scientists, we must develop some effective marketing strategies to push these platinum and gold customers to higher profitability levels by inducing them to spend more on expensive products at the store.

RFM MODEL (Model Evaluation, Selection and Results)

The detailed process of model evaluation and selection has been enumerated in the Jupyter Notebook. The Tableau dashboard presents the model results.

This dashboard is a result of the output generated by the Random Forest Classifier to predict the RFM Loyalty levels of each customer.

The coding system used to measure RFM Loyalty level is as follows:

- ‘0’ indicates Bronze customers (churning customers).
- ‘1’ indicates Silver customers (churning customers).
- ‘2’ indicates Gold customers (high value).
- ‘3’ indicates Platinum customers (highest value).

The dashboard indicates all the variables which are significant in the predictive model, which can be used to devise efficient marketing strategies to increase the loyalty score of customers.

The model predicted the RFM Loyalty level of customers with an accuracy of 78%.

Key Insights

1. Increasing the area for frozen foods induces a higher consumer spending and helps in increasing the monetary value of customers.
2. Likewise, since a majority of customers buy meat and groceries, it is worthwhile to increase meat area and grocery area space in the superstore.
3. Having a coffee bar at the supermarket provides a platform for customers to socialize and spend more time at the store. This would in turn, generate a higher consumer spending because people may tend to buy products from the store which were initially not there on their shopping list.
4. Most of the customers are professionals who are busy at work, so they prefer prepared meals. Hence one should ensure a department is available to handle prepared meals section of the superstore.
5. Finally, there are specific areas like Washington and California where the number of platinum and gold customers is usually very high. It may be a good proposition to open more stores in these geographical areas.

PROFIT MARGIN MODEL

The detailed process of model evaluation and selection has been enumerated in the Jupyter Notebook. The Tableau dashboard presents the model results.

This dashboard is a result of the output generated by the Random Forest Classifier to predict the Profit margin levels. The coding system used to measure profitability is as follows:

- ‘0’ indicates low profitability

- ‘1’ indicates medium profitability
- ‘2’ indicates high profitability.

The dashboard indicates all the variables which are significant in the predictive model, which can be used to devise efficient marketing strategies to increase the profit margin levels of the store.

The model predicted the RFM Loyalty level of customers with an accuracy of 69%.

Key Insights:

1. Cities like San Francisco, Bellingham, etc. have a high impact on store profitability.
2. Specific stores like Store 14, store 5, store 2 and store 22 indicates higher profitability as compared to others.
3. Average SRP known as Shelf Ready Packaging is a key factor in explaining movements along different profit margin levels. SRP is designed to make shelf replenishment more efficient since it helps in optimizing shelf space. Small items, pouched product, odd-shaped items that require time and attention to stock are likely candidates for SRP.
4. Small groceries on an average have higher profitability levels as compared to Supermarkets.

MODEL ACCURACY FOR DATA SCIENTISTS

This dashboard has been designed for data scientists to evaluate model accuracy. This is similar to a confusion matrix. The values of ‘0’, ‘1’, ‘2’ and ‘3’ have been explained earlier in this report. This dashboard indicates the number of cases for predicted values vs. Actual values for both the models that were built in this project.

FUTURE SCOPE

As a next step, we can

1. Develop individual category specific models to improve profitability and customer loyalty.
2. Work with more recent data as and when available.
3. Build a web application to predict customer loyalty and profit margin levels.