Final Report: Group 7

Customer profiling and acquiring cost, Revenue Maximization For Food Mart Retailer

Team Members
Atharva Atre
Supreeth Bannur Sudhakara
Abdul Jawad Mohammed
Keerthi Kumar Reddy Kancham Reddy

Prepared for

Data Analytics in Business using R

Professor: Dr. Adolfo Coronado

Purdue University Fort Wayne

Table of content

Sr.no	Content	Page Number
1	Introduction	3
2	Purpose of our project	3
3	Midterm Report Summary	4
4	I Predictive Model Analysis	11
5	II Best Model Selection Process	12
6	III Dynamic Dashboard	15
7	Conclusion	15

Introduction

We live in a world with many huge monopolies in the retail sector (Walmart and Amazon). Even amidst such staunch competition, other franchises such as Aldi, Dollar General, and 7-Eleven have successfully survived. In this project, we analyzed data on a Convenient Food Mart (CFM), founded in Chicago in 1958.

CFM is a chain of convenience stores in the United States. The private company's headquarters are in Mentor, Ohio, and approximately 325 stores are currently in the US. Convenient Food Mart operates on the franchise system. Convenient Food Mart was the nation's third-largest chain of convenience stores as of 1988. The NASDAQ exchange dropped Convenient Food Mart the same year when the company failed to meet financial reporting requirements. Hence, analysis of the CFM data can also benefit other markets in understanding different parameters where they can do better.

The two main parameters which any retail store focuses on are overall revenue and the number of customers. Firstly, we analyzed the dataset. We would like to understand consumer patterns (how often a consumer buys Bread along with milk) and try presenting an analysis of the item stocking recommendation for the mart using the Apriori Algorithm, as well as the cost borne by CFM for acquiring a new customer (via advertisement). Customer acquisition is the main component in business, so it's imperative to know which form of advertisement suits which type of customers, instead of spending on low-yielding channels.

Purpose of Our Project

● To predict the media 'Cost' incurred by the CFM for acquiring new customers. This includes the effectiveness of different advertisements, customer profiling, shopping preferences, and customer membership programs.

● To organize a better company strategy for a retailer. Recognize your profit margins and work to them while considering the cost of client acquisition.

● Customer acquisition is the main component in business, so it's imperative to know which form of advertisement suits which type of customers, instead of spending on low-yielding channels.

● The project could provide a deeper perspective into possible correlations with customer lifestyles and preferences, yielding more engaging and impactful marketing campaigns for specific customers.

- The project will thoroughly examine the expenditure in different market segments, providing investors with better suggestions on where to make cost reductions and increase yield.

Beneficial Audience

Observing customer trends could generate ideas not only for Food Mart, but potentially other supermarket management teams to create cost-efficient marketing campaigns that are more appealing and catered towards various preferences. The inferences derived from the product sales could also inspire changes in store layout to prioritize popular products and shine a broader spotlight on them. With customer profiling being an integral aspect of our analysis, Food Mart could use customer information to check different advertising methods for increasing revenue. To name a few parts of the organization that will have a direct impact are as follows:

- **Food Mart Management** - for getting new customers and revenue maximization

- **Store Design Team** - to design the store in the most efficient way

- **Food Mart Procurement team** - to decide which items to order with what frequency

Midterm Report Summary

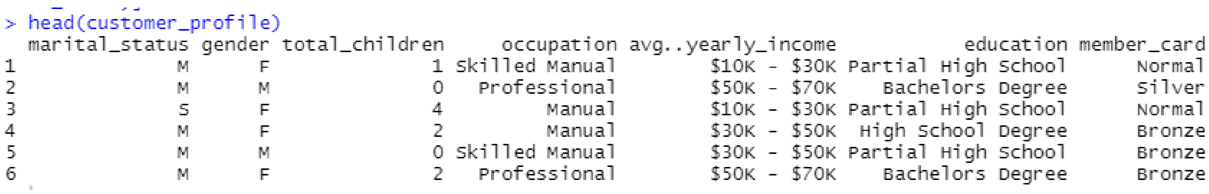
Dataset:

The raw dataset for this is available at GitHub.
https://github.com/jpvelez/foodmart

The raw dataset is divided into eight distinct datasets with a total of about 200K data points with over 75 features. We found a refined version of the same data set at Kaggle.
https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart.

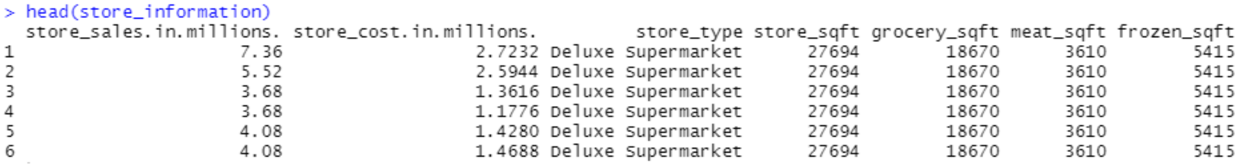
The feature list (columns) for the data is divided mainly into four categories, as mentioned below:

- Customer Profile: Marital Status, gender, total children, occupation, annual income, education level, membership status, etc.



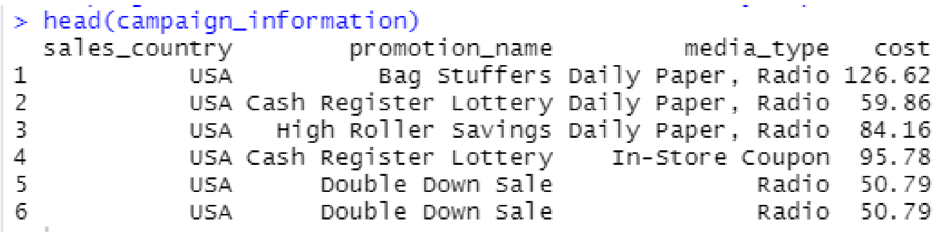
```
> head(customer_profile)
  marital_status gender total_children     occupation avg..yearly_income           education member_card
1              M      F              1 Skilled Manual        $10K - $30K Partial High School      Normal
2              M      M              0   Professional        $50K - $70K     Bachelors Degree      Silver
3              S      F              4         Manual        $10K - $30K Partial High School      Normal
4              M      F              2         Manual        $30K - $50K  High School Degree      Bronze
5              M      M              0 Skilled Manual        $30K - $50K Partial High School      Bronze
6              M      F              2   Professional        $50K - $70K     Bachelors Degree      Bronze
```

- Store Information: Store sales, costs, building type, market area, meat/grocery/frozen food space, city/state location, etc.



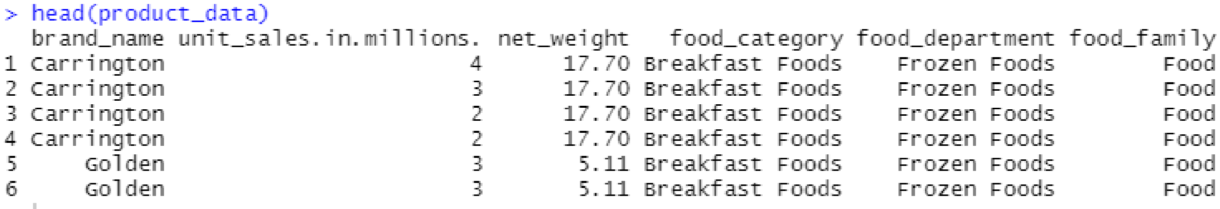
```
> head(store_information)
  store_sales.in.millions. store_cost.in.millions.         store_type store_sqft grocery_sqft meat_sqft frozen_sqft
1                     7.36                  2.7232 Deluxe Supermarket      27694        18670      3610        5415
2                     5.52                  2.5944 Deluxe Supermarket      27694        18670      3610        5415
3                     3.68                  1.3616 Deluxe Supermarket      27694        18670      3610        5415
4                     3.68                  1.1776 Deluxe Supermarket      27694        18670      3610        5415
5                     4.08                  1.4280 Deluxe Supermarket      27694        18670      3610        5415
6                     4.08                  1.4688 Deluxe Supermarket      27694        18670      3610        5415
```

- Campaign Information: Sales country, promotion name, customer acquisition cost, etc.



```
> head(campaign_information)
  sales_country        promotion_name         media_type   cost
1           USA          Bag Stuffers Daily Paper, Radio 126.62
2           USA Cash Register Lottery Daily Paper, Radio  59.86
3           USA   High Roller Savings Daily Paper, Radio  84.16
4           USA Cash Register Lottery    In-Store Coupon  95.78
5           USA      Double Down Sale              Radio  50.79
6           USA      Double Down Sale              Radio  50.79
```

- Product data: Brand, unit sales, weight, product type, food type, etc.



```
> head(product_data)
  brand_name unit_sales.in.millions. net_weight   food_category food_department food_family
1 Carrington                       4      17.70 Breakfast Foods    Frozen Foods        Food
2 Carrington                       3      17.70 Breakfast Foods    Frozen Foods        Food
3 Carrington                       2      17.70 Breakfast Foods    Frozen Foods        Food
4 Carrington                       2      17.70 Breakfast Foods    Frozen Foods        Food
5     Golden                       3       5.11 Breakfast Foods    Frozen Foods        Food
6     Golden                       3       5.11 Breakfast Foods    Frozen Foods        Food
```

Data cleaning:

- The Customer Income feature had values in range format, which we converted into single values. We converted the values into three types, average, the minimum value of the range, and the maximum value of the range.
- We introduced education levels for the education qualification

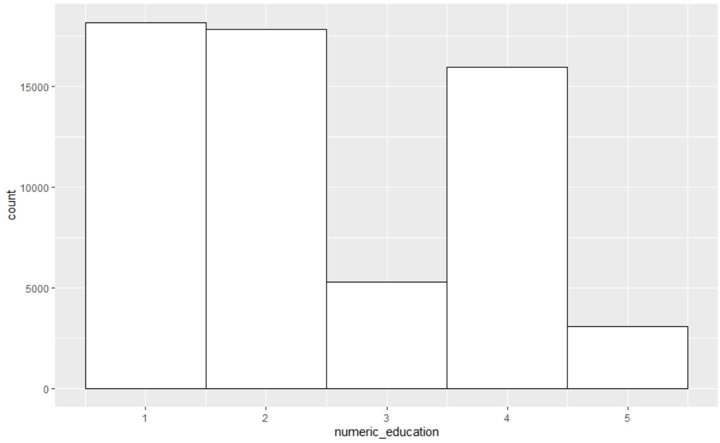
  'Partial High School' = 1

  'High School Degree' = 2

  'Partial College' = 3

  'Bachelor's Degree' = 4

  'Graduate Degree' = 5

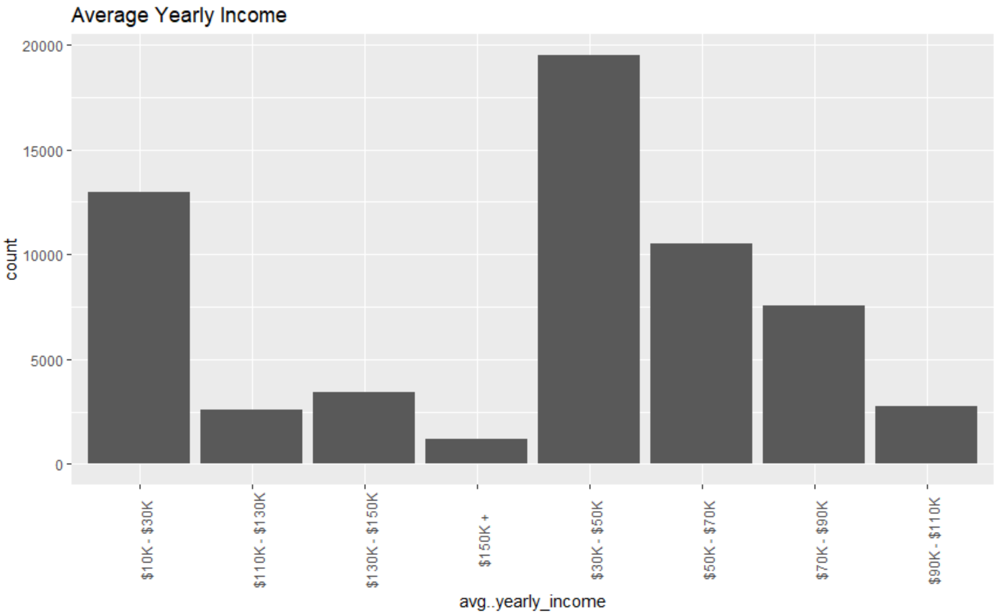


Preliminary Predictive Analytics

Research Questions answered via EDA analysis -

Question 1 - What is the average income of the customers who shop at CFM ?

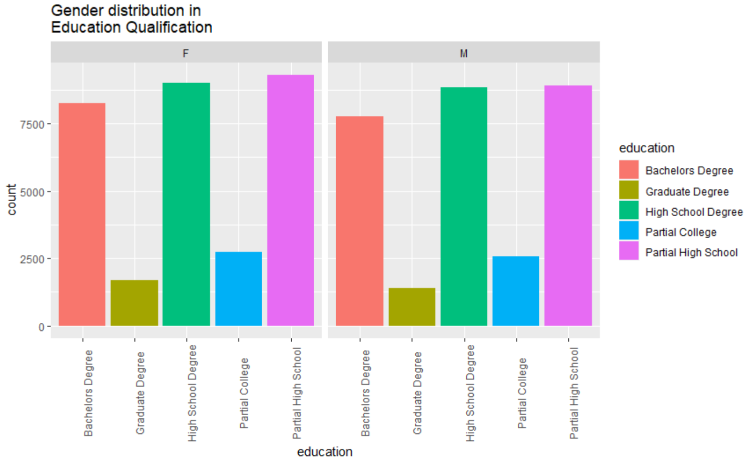
Ans -



From the above bar plot, it is observed that people whose average yearly income is in the bracket of 10k to 90k $ are more significant in number visiting the CFM mart compared to those whose average annual income is more than 90k+.

Question 2 - What is the Education distribution trend between the male and the female customers ?

Ans -

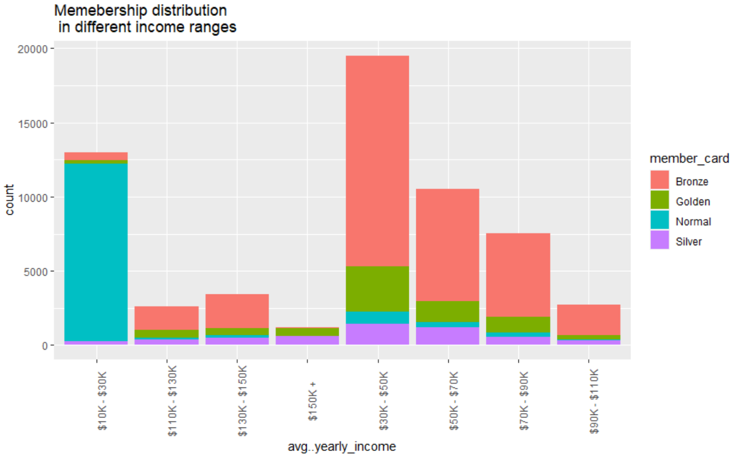


The analysis shows that the gender distribution in education qualification follows the same trend between male and female genders.

We have five levels of education, as shown in the plot above. Both males and females labeled 'M' and 'F' in the above graph have maximum customers (>7500) for the education level of Bachelor's Degree, High School Degree, and Partial High school Degree.
CFM has the least number of customers (<2500) for Graduate Degrees and Partial College education levels.

Question 3 - How effective is the membership program of CFM across various income groups?

Ans -



From the analysis, it is observed that:

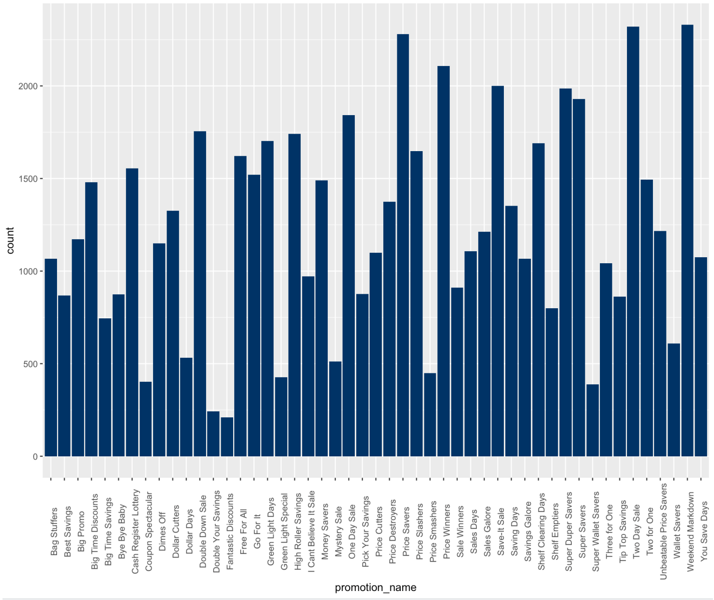
- The bronze card is the most preferred type of card for customers visiting the CFM except for those with an average annual income of 10k-30k$ and 150k+ $.

- The golden card is the second most preferred type of card for customers visiting the CFM except for those whose average annual income lies between 10k-30k$.

- A normal card is the most preferred type of card for people whose average annual income lies between 10k and 30k$.

Question 4 -What are the top 3 promotional taglines that make the most sales for CFM Ans -

Data distribution for different promotions



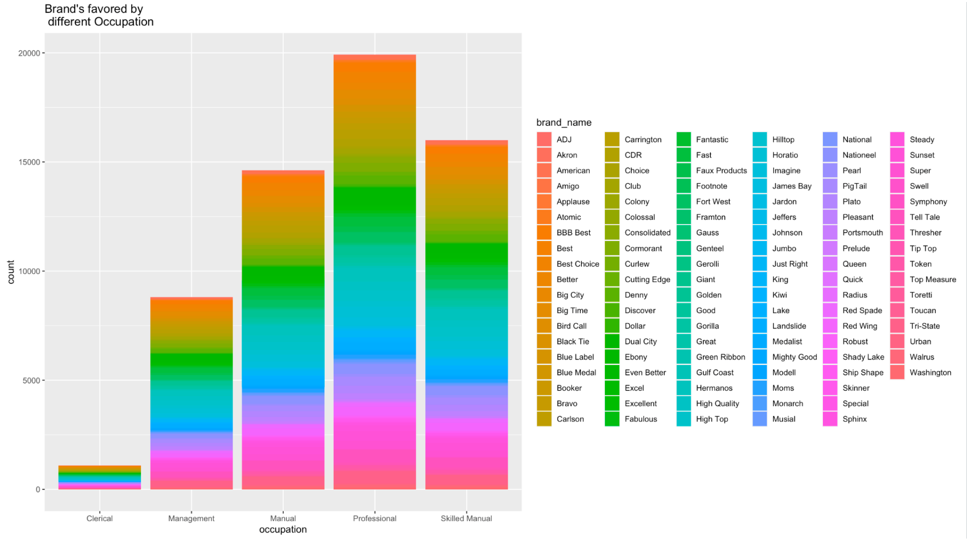
A common tactic used by marts is offering hoardings on merchandise such as 'Huge Discount' or 'daily minimum.' We have analyzed all the different tags used by CFM and ranked them according to how they affect sales.
As shown in the bar graph above, the top 3 promotional tags according to sales volume are-

1. Weekend Markdown
2. Two-day sale
3. Price Saver

The worst-performing sales tags are-

1. Fantastic Discount
2. Double your savings
3. Super wallet saving

Question 5 - Are there any brand preferences for customers with different occupations?
Ans -

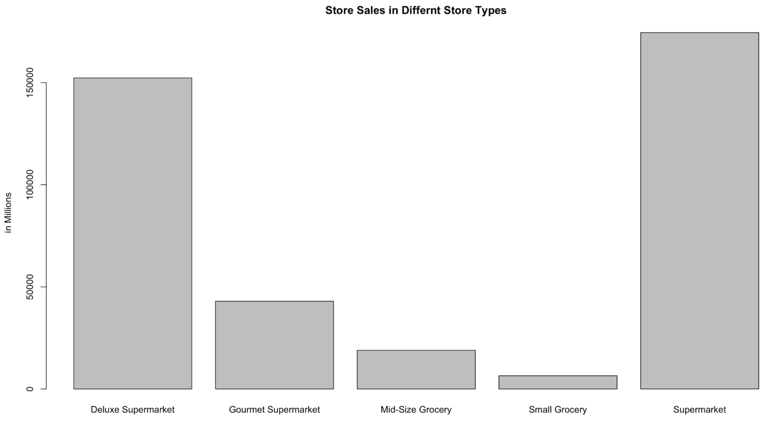


We categorized CFM customers according to their occupation as follows -

- Clerical
- Management
- Manual
- Professional
- Skilled Occupation

Then we plotted the preference of each occupation group for every brand sold at CFM. This data can be helpful for brands to design new products for their customer base.

Question 6 - Which type of store has the most sales for CFM?
Ans -



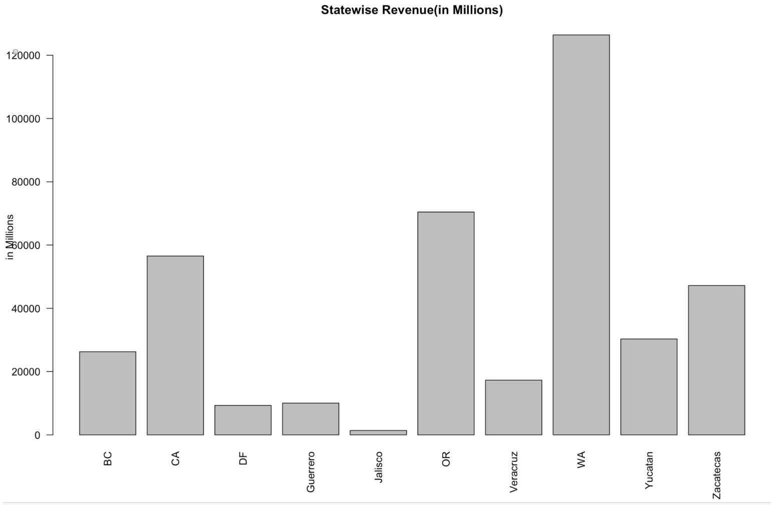
We have five different types of stores, i.e.,

- Deluxe Supermarket
- Gourmet Supermarket
- Mid-Size Grocery
- Small Grocery
- Supermarket.

From this plot of store sales distributed among the different types of stores, we have the following observations:

1. Supermarket type of store is the most profitable among the different store types, followed by marginally lower sales in Deluxe supermarkets.
2. The least profitable stores are small grocery and mid-size grocery stores.

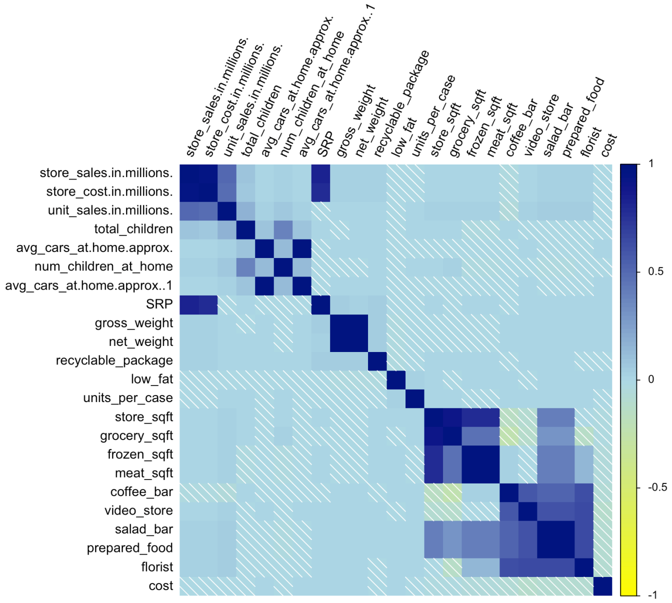
Question 7 - Which state has generated the least and highest revenue for CFM?
Ans -



In the above plot, we can observe that the WA state has the highest sales with sales of around 120000 million, followed by OR with sales of about 70000 million, and CA is the third place with sales of around 60000 million. The state with the lowest sales is Jalisco.

Question 8 - Predicting the media campaign cost for Convenient Food Mart (CFM).
Ans -
The main objective of this project will be to predict the media 'cost' incurred by CFM for acquiring new customers. For this purpose, we first need to select all relevant features for this prediction. Our next steps include finalizing the features we will use for predicting the media cost, model selection, and building.

The feature correlation matrix is shown below:



The importance of features can be estimated from data by building a model. Some methods, like decision trees, have a built-in mechanism to report on variable importance. For other algorithms, the importance can be estimated using a ROC curve analysis conducted for each attribute. We will be doing modeling as our next step in the project.

I. Predictive Model Analysis

In order to predict the incurred cost of FoodMart media campaigns, we initialized and fine-tuned four machine learning algorithms for optimal performance. For each of the models, 70% of the dataset was used for training, 20% for validation, and 10% for out-of-sample testing. Due to the cost variable being a continuous value, the prediction task was regarded as a regression problem.

i. **Random Forest Regressor**: Random Forest models are combinations/ensembles of multiple decision trees which calculate the final prediction based on an average of all the trees' individual outputs. It was primarily chosen for its ability to handle complex datasets with multiple features.

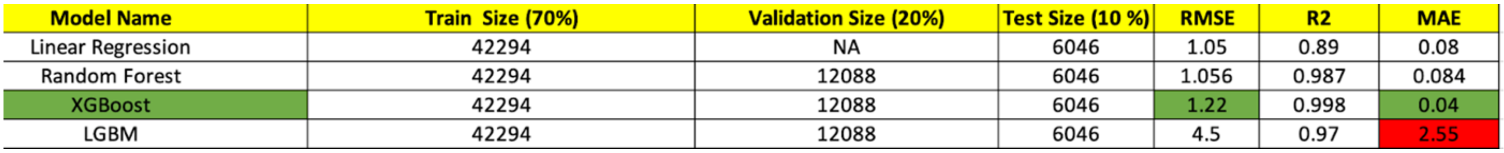
ii. **XGBoost Regressor**: Similar to the Random Forest, the XGBoost is a decision tree-based ensemble model that differentiates itself through the use of a 'Gradient

Boosting' algorithm to minimize loss. Its efficiency and low computation speed made it a great candidate for our problem.

iii. **Light Gradient Boosting Machine (LGBM) Regressor**: The LGBM algorithm is a faster version of the XGBoost regressor due to having a much more optimized Gradient Boosting algorithm.

iv. **Linear Regression**: A linear regressor tries to generate a prediction line that best fits/overlaps with as many samples as possible. The lower the error metrics, the closer the line is to perfectly fit all points.

II. Best Model Selection Process



Model Name	Train Size (70%)	Validation Size (20%)	Test Size (10 %)	RMSE	R2	MAE
Linear Regression	42294	NA	6046	1.05	0.89	0.08
Random Forest	42294	12088	6046	1.056	0.987	0.084
XGBoost	42294	12088	6046	1.22	0.998	0.04
LGBM	42294	12088	6046	4.5	0.97	2.55

Table 1. The resulting error metrics from each of our fine-tuned models

To gauge and compare the performances of each model, we opted to use three metrics:

a. **Root Mean Squared Error (RMSE)**: The RMSE score is a measure of how close the data points are to the predicted line of fit. A lower RMSE value is usually a stronger indicator of good performance. It is calculated as follows -

$$RMSE = \sqrt{(f - o)^2}$$

where *f* represents a predicted value and *o* is an observation:

b. **Mean Absolute Error (MAE)**: The MAE score is the magnitude of the distance between the predicted line and each true observation. Lower values are highly preferred when analyzing model performance. It is calculated as follows -

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n}$$

MAE = mean absolute error

y_i = prediction

x_i = true value

n = total number of data points

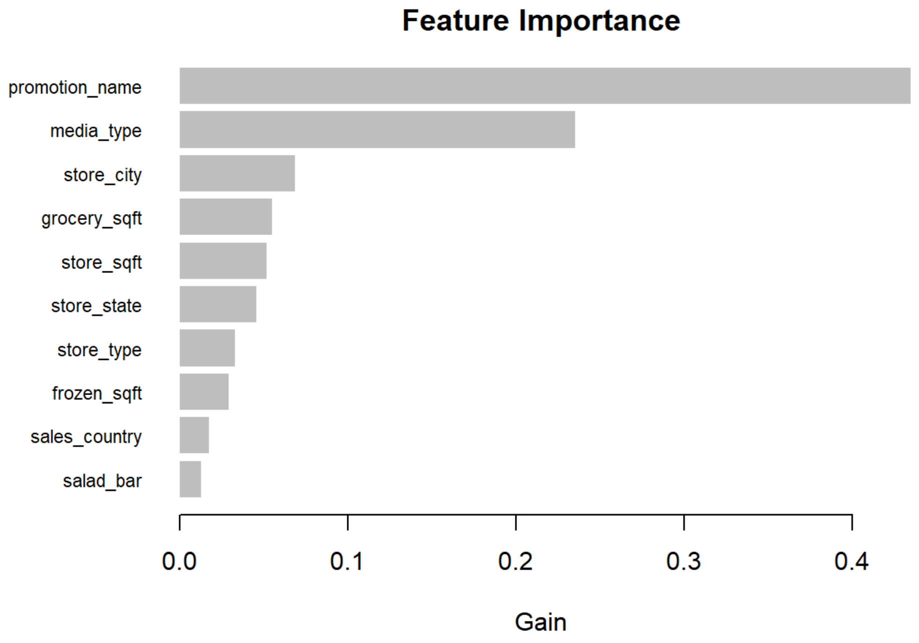
c. **R2 Score**: The R-Squared value is the proportion of the variance that is explained by the model. A higher R2 value indicates a stronger correlation between the variables and better performance from the model. It is calculated as follows, where *yi* is an observation while ŷ is the predicted value:

$$\begin{aligned} R^2 &= 1 - \frac{\text{sum squared regression (SSR)}}{\text{total sum of squares (SST)}}, \\ &= 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - y)^2}. \end{aligned}$$

Performance Analysis: As shown in Table 1, the gradient boosting capabilities of the XGBoost regressor gave it an edge over the other models in terms of R2 score and Mean Absolute Error when predicting the test set. Despite being more efficient, the LGBM regressor achieved the worst RMSE and MAE values among the models. The R2 scores for all the tree-based algorithms were comparable, with the Linear Regression slightly worse. Overall, the XGBoost model proved to be the stronger performer on all performance metrics, slightly worse than linear regression in the RMSE score. However, it is possible that further hyperparameter tuning for the models could potentially have resulted in a different set of outcomes.

Feature Importance

Feature importance is calculated as the decrease in node impurity weighted by the probability of reaching that node. The node probability can be calculated by the number of samples that reach the node, divided by the total number of samples. The higher the value the more important the feature.

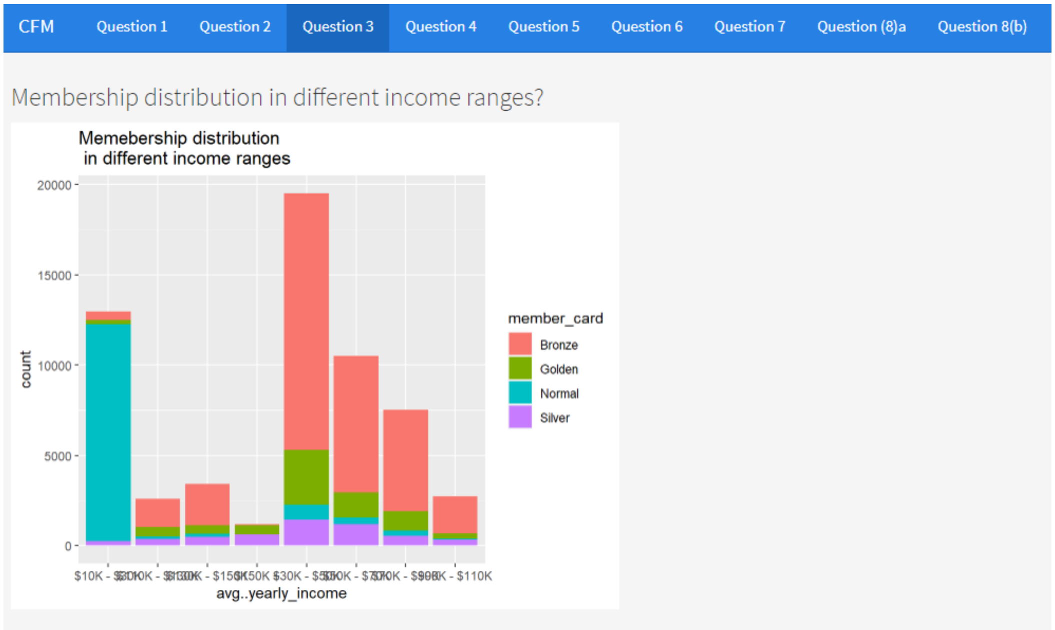


As shown in the above image, tree-based models generated the following feature importance for various features. 'Promotion_name' was the most important feature for promoting the 'cost' of acquiring new customers. As shown in our research question, the following 'promotion_name' were most effective:

1. Weekend Markdown
2. Two-day sale
3. Price Saver

The second most important feature was 'media_type', while the least significant feature was 'salad_bar'.

III. Dynamic Dashboard



We constructed a dashboard examining the major research questions we hoped to get insight from to add an interactive element to our study. For each of the eight questions, an appropriate plot is displayed, showing the findings from the data. The same dashboard can be available as an HTML file on the browser.

Conclusion

This project was an analysis of different aspects of the CFM data. We began with identifying beneficial audiences before performing an EDA analysis. Then, we answered important research questions which could give business insights about demographic and sales information. For predicting the cost of acquiring new customers, we performed a comprehensive comparison between various predictive models, from which we selected the best model based on a set of chosen metrics. Finally, we created a dynamic dashboard to showcase an interactive way to visually explore each of our research questions.

Overall, the CFM cost prediction project served as a launching pad for our team's career in data analytics as well as a learning experience.The knowledge gained related to the research process, machine learning, and analytic procedures will aid us in our future endeavors.