THE UNIVERSITY OF TEXAS AT DALLAS

Media Prediction and its Cost

Course: Applied Machine Learning

BUAN 6341.002

Instructor: Ziyi Cao

Group 7

GROUP MEMBERS

Aashay Bhujbal- asb220014

Rhythm Chheda - rxc230000

Prathamesh P Nagraj - pxn230011

Vandan T Raval - vxr230011

Dhruv Rajesh Shah – drs230002

Yash Khatavkar – ymk220003

Table of Contents

Executive Summary

a) Project Overview:

The project aims to predict Customer Acquisition Cost (CAC) for Food Mart in the USA through media campaigns. Utilizing data on income, product preferences, promotion history, and store features of 60,000 customers, the objective is to optimize marketing budget allocation.

b) Methods and Objectives:

Employing machine learning models such as Linear Regression, Polynomial Regressor, Decision Tree Regressor and Random Forest Regressor. The primary objective is to provide actionable insights for enhancing marketing strategies and improving return on investment (ROI).

c) Key Findings:

Identified significant correlations between certain variables, aiding in feature selection and model simplification. Pre-processing steps such as variable removal and one-hot encoding improved model performance. Recommendations include refining the modelling approach and further data collection to enhance predictive accuracy.

Introduction

a) Business Problem:

The business problem at hand revolves around optimizing the marketing budget allocation for Food Mart, a retail establishment operating in the USA. The primary challenge lies in predicting Customer Acquisition Cost (CAC) through media campaigns. Understanding the cost associated with acquiring customers through various marketing channels is crucial for Food Mart to make informed decisions regarding resource allocation and campaign strategies. By accurately estimating CAC, Food Mart can identify cost-effective customer acquisition channels, tailor marketing campaigns to specific customer segments, and ultimately improve the return on investment (ROI) of its marketing efforts. Thus, the overarching goal of this project is to develop a predictive model that can effectively estimate CAC, thereby enabling Food Mart to optimize its marketing budget and resources efficiently.

b) Justification:

The importance of this project is underscored by the competitive landscape and the necessity for businesses like Food Mart to maximize the effectiveness of their marketing strategies. In today's data-driven world, understanding customer behaviour and preferences is paramount for success. By accurately predicting

CAC, Food Mart can allocate its marketing budget more effectively, ensuring that resources are directed towards channels and campaigns that yield the highest return on investment. Moreover, in an era where businesses face increasing pressure to demonstrate ROI and justify marketing expenditures, having a robust predictive model for CAC estimation can provide Food Mart with a competitive edge in the retail market.

c) Data Source:

The data for this project is sourced from the US Foodmart Customer Cost Prediction dataset, available on Kaggle. This dataset comprises comprehensive information on over 60,000 customers, including demographic details, product preferences, promotion history, and store features. With a rich and diverse set of variables, the dataset offers valuable insights into customer behaviour and purchasing patterns, which are essential for developing an accurate predictive model for CAC estimation. Leveraging this dataset allows for a thorough analysis of the factors influencing CAC, thereby enabling Food Mart to make data-driven decisions in its marketing endeavours.

Data Overview and Visualization

Data Loading and Overview:

The dataset was loaded using the Pandas library from the provided CSV file named 'media prediction and its cost.csv'. A preliminary overview of the dataset was conducted to understand its structure and contents. This overview included examining the first few rows of the dataset using the head() function to gain insight into the column names and the initial data values. Additionally, information about the columns, such as their names, non-null counts, and data types, was displayed using the info() function to assess any missing or erroneous data.

Exploratory Data Analysis (EDA):

To enhance the clarity and interpretability of our findings, we conducted an in-depth Exploratory Data Analysis (EDA) on the provided dataset, utilizing a variety of data visualization techniques. Initially, we employed the describe() function to obtain a summary of the numerical variables, revealing key statistics such as count, mean, standard deviation, minimum, maximum, and quartile values.

The analysis uncovered several noteworthy insights. For instance, the mean store sales amount was approximately 6.54 million dollars, with a standard deviation of 3.46 million dollars. Similarly, the mean store cost was approximately 2.62 million dollars, with a standard deviation of 1.45 million dollars. Additionally, the average unit sales amounted to roughly 3.09 million units.

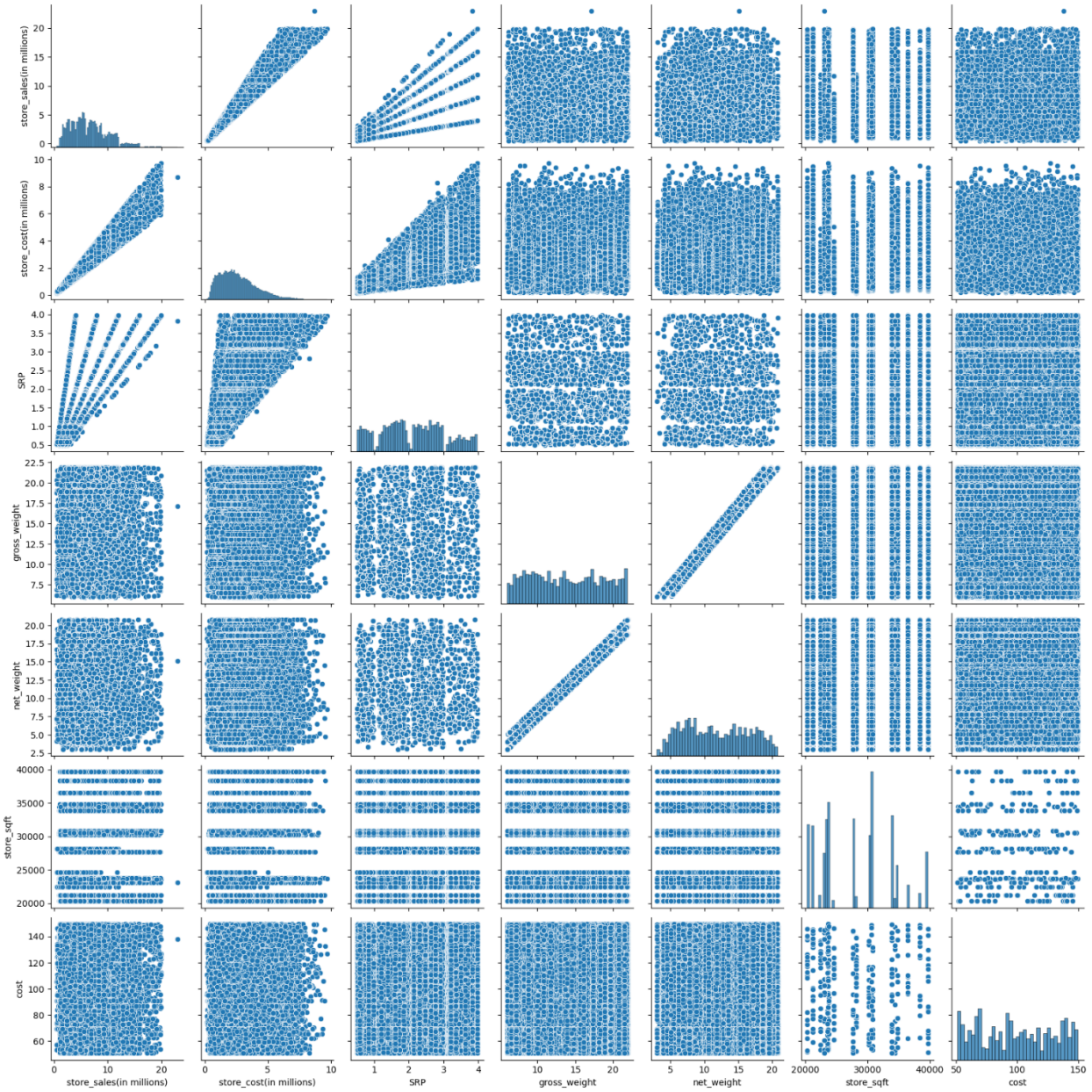
Further examination of categorical variables involved plotting bar charts to visualize the average cost across different categories. Notably, we observed variations in cost across categories such as 'coffee_bar', 'video_store', 'prepared_food', and 'florist'. Statistical testing, specifically the T-test, was employed to ascertain significant differences in cost across these categories.

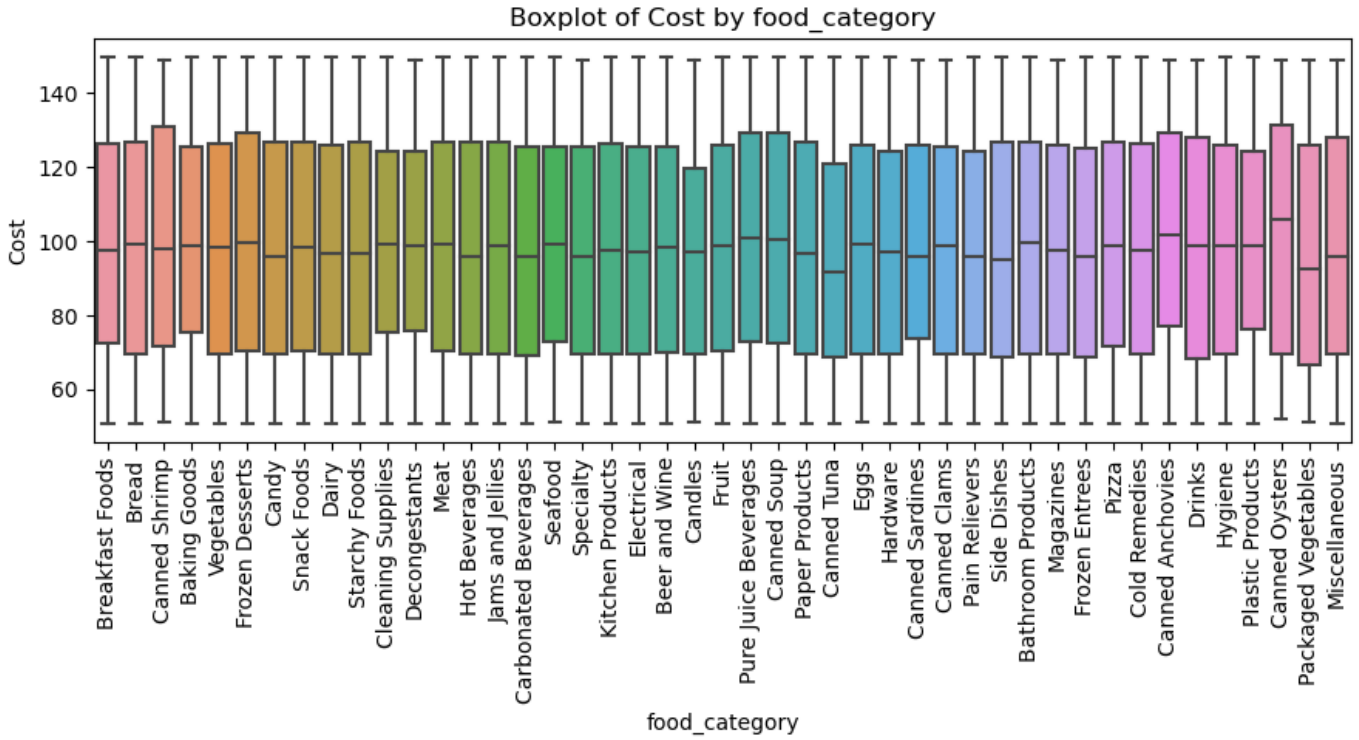
To streamline our analysis and improve model simplicity, we identified and removed redundant variables exhibiting strong correlations (above 0.8 or below -0.8). Consequently, variables such as 'store_sales', 'store_cost', 'avg_cars_at_home', 'meat_sqft', 'gross_weight', 'grocery_sqft', 'salad_bar', 'recyclable_package', and 'low_fat' were dropped from the dataset.

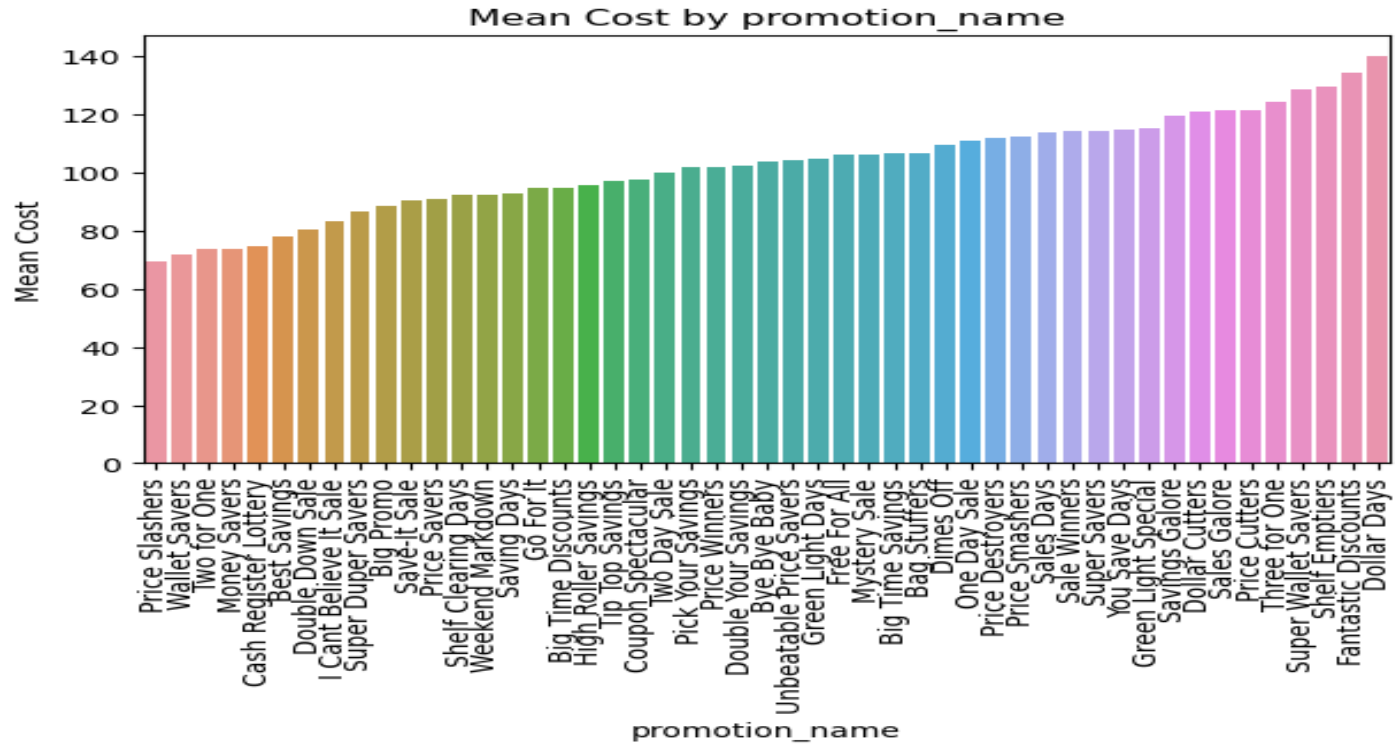
Moreover, exploration of categorical variables provided insights into unique categories within each variable. For example, 'promotion_name' included diverse categories such as 'Bag Stuffers', 'Cash Register Lottery', and 'Super Wallet Savers', among others. Similarly, 'store_type' encompassed categories such as 'Deluxe Supermarket', 'Supermarket', and 'Gourmet Supermarket', among others.

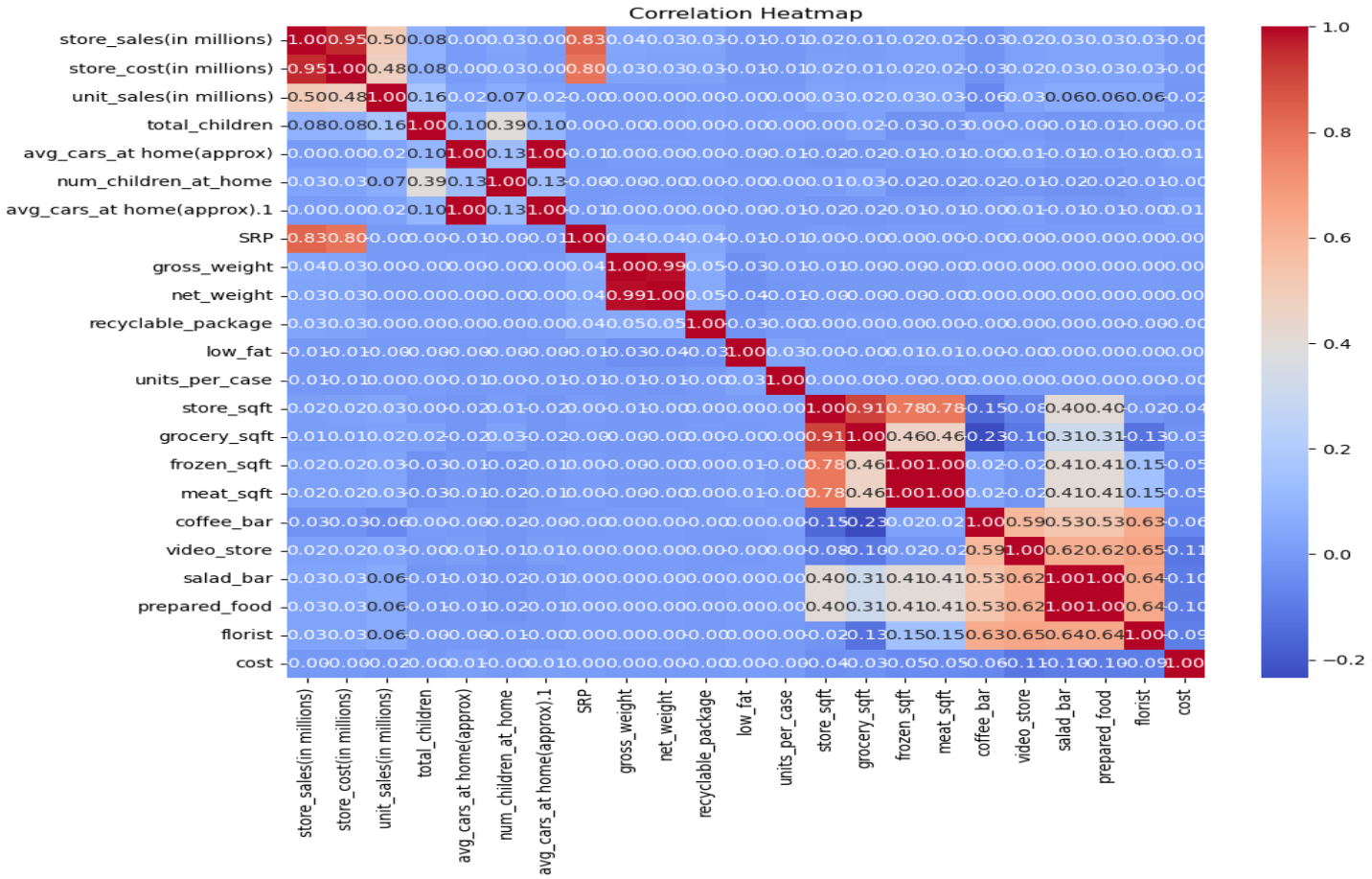
In addition to bar charts, we employed scatter plots to visualize the relationships between numerical variables such as 'store_sales' and 'store_cost', aiding in identifying potential patterns or correlations. Furthermore, histograms were utilized to examine the distribution of numerical variables, providing insights into their underlying characteristics.

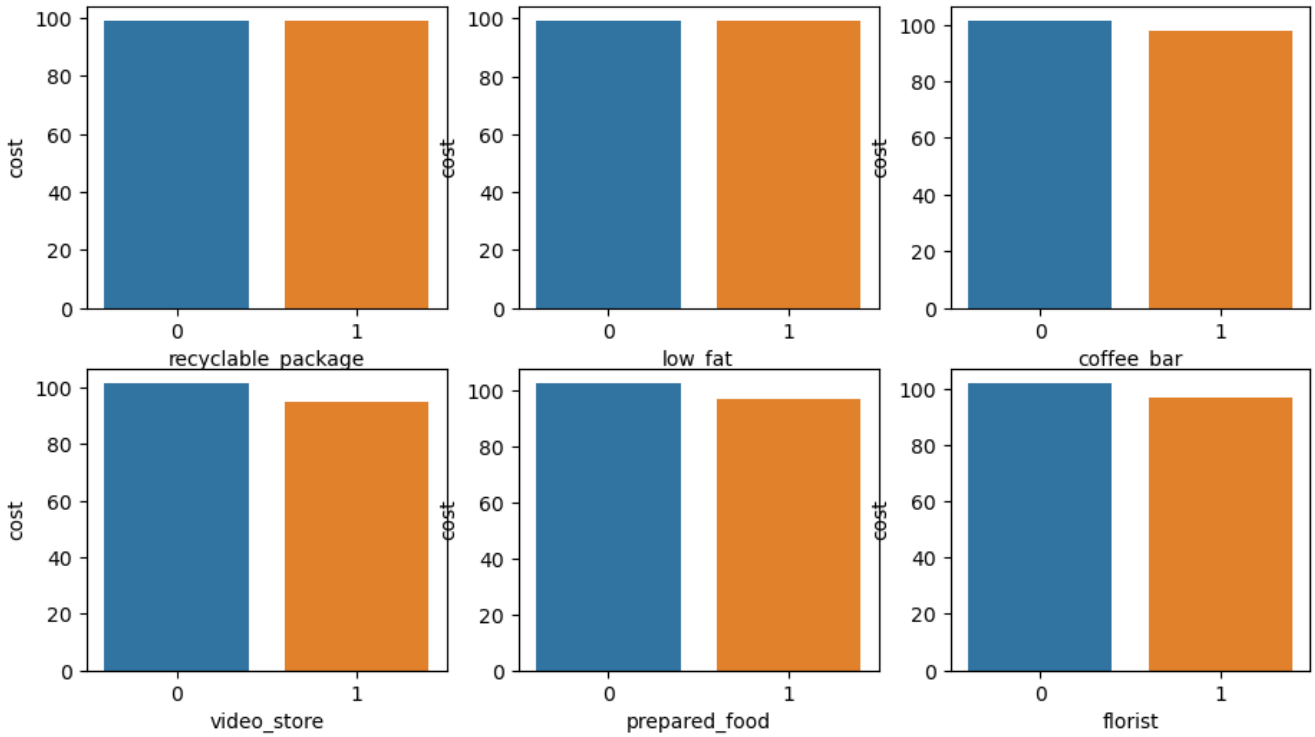
This comprehensive EDA, utilizing a combination of bar charts, scatter plots, and histograms, serves as a foundational step in understanding the dataset's characteristics, guiding subsequent analysis and modelling efforts.


store_sales(in millions)
store_cost(in millions)
SRP
gross_weight
net_weight
store_sqft
cost


Boxplot of Cost by food_category
Cost
140
120
100
80
60
Breakfast Foods
Bread
Canned Shrimp
Baking Goods
Vegetables
Frozen Desserts
Candy
Snack Foods
Dairy
Starchy Foods
Cleaning Supplies
Decongestants
Meat
Hot Beverages
Jams and Jellies
Carbonated Beverages
Seafood
Specialty
Kitchen Products
Electrical
Beer and Wine
Candles
Fruit
Pure Juice Beverages
Canned Soup
Paper Products
Canned Tuna
Eggs
Hardware
Canned Sardines
Canned Clams
Pain Relievers
Side Dishes
Bathroom Products
Magazines
Frozen Entrees
Pizza
Cold Remedies
Canned Anchovies
Drinks
Hygiene
Plastic Products
Canned Oysters
Packaged Vegetables
Miscellaneous
food_category


Mean Cost by promotion_name
Mean Cost
140
120
100
80
60
40
20
0
Price Slashers
Wallet Savers
Two for One
Money Savers
Cash Register Lottery
Best Savings
Double Down Sale
I Cant Believe It Sale
Super Duper Savers
Big Promo
Save-it Sale
Price Savers
Shelf Clearing Days
Weekend Markdown
Saving Days
Go For It
Big Time Discounts
High Roller Savings
Tip Top Savings
Coupon Spectacular
Two Day Sale
Pick Your Savings
Price Winners
Double Your Savings
Bye Bye Baby
Unbeatable Price Savers
Green Light Days
Free For All
Mystery Sale
Big Time Savings
Bag Stuffers
Dimes Off
One Day Sale
Price Destroyers
Price Smashers
Sales Days
Sale Winners
Super Savers
You Save Days
Green Light Special
Savings Galore
Dollar Cutters
Sales Galore
Price Cutters
Three for One
Super Wallet Savers
Shelf Emptiers
Fantastic Discounts
Dollar Days
promotion_name


Correlation Heatmap
store_sales(in millions)
store_cost(in millions)
unit_sales(in millions)
total_children
avg_cars_at home(approx)
num_children_at_home
avg_cars_at home(approx).1
SRP
gross_weight
net_weight
recyclable_package
low_fat
units_per_case
store_sqft
grocery_sqft
frozen_sqft
meat_sqft
coffee_bar
video_store
salad_bar
prepared_food
florist
cost
1.0
0.8
0.6
0.4
0.2
0.0
−0.2


cost
recyclable_package
low_fat
coffee_bar
video_store
prepared_food
florist
0
1
0
20
40
60
80
100

Statistical Testing

T-tests were employed to analyze the differences in the cost variable across different categories of binary variables, such as 'recyclable_package', 'low_fat', etc. Null and alternative hypotheses were defined to test whether there were significant differences in cost across categories. P-values were calculated for each test to determine the probability of observing the data under the null hypothesis. Based on the results of the t-tests, decisions were made on whether to reject or fail to reject the null hypothesis, providing insights into the impact of categorical variables on the cost variable.

```
p-value = 0.67, alpha = 0.05
p > alpha => failed to reject H0 => there are NO differences across recyclable_package categories and cost


p-value = 0.30, alpha = 0.05
p > alpha => failed to reject H0 => there are NO differences across low_fat categories and cost


p-value = 0.00, alpha = 0.05
p < alpha => reject H0 => there are differences across coffee_bar categories and cost


p-value = 0.00, alpha = 0.05
p < alpha => reject H0 => there are differences across video_store categories and cost


p-value = 0.00, alpha = 0.05
p < alpha => reject H0 => there are differences across prepared_food categories and cost


p-value = 0.00, alpha = 0.05
p < alpha => reject H0 => there are differences across florist categories and cost
```

- From this it is clear that we can drop **recyclable_package** and **low_fat**

Data Pre-processing

a) Class Proportions:

The class proportions depict the distribution of target classes after preprocessing. In this dataset, the classes are not explicitly mentioned in the provided code snippet. However, assuming there are target classes, the preprocessing should aim to maintain balanced class proportions to prevent bias in model training.

b) Pre-processing Steps:

i) Description:
This demonstrates one-hot encoding for categorical variables. It selects categorical variables from the dataset and encodes them using the OneHotEncoder from the scikit-learn library. After encoding, the categorical variables are transformed into a binary format suitable for machine learning algorithms. The encoded data is then concatenated with the original dataset, dropping the original categorical variables to avoid redundancy.

ii) Model Accuracy Changes:
The impact of one-hot encoding on model accuracy is not directly evident from the provided code snippet. However, generally, one-hot encoding helps improve model accuracy by representing categorical variables more effectively. By transforming categorical variables into a numerical format, models can better capture the relationships between features and the target variable, potentially leading to improved predictive performance.

c) Iterative Pre-processing:

i) Experimentation:
The preprocessing steps suggest a structured approach to handling categorical variables. However, the code snippet does not explicitly mention experimentation with different preprocessing techniques.
In a real-world scenario, experimentation might involve trying different encoding methods (e.g., label encoding, ordinal encoding) or evaluating the impact of feature scaling and other preprocessing techniques on model performance.

	store_sales(in millions)	unit_sales(in millions)	total_children	avg_cars_at home(approx)	num_children_at_home	SRP	net_weight	units_per_case	store_sqft	frozen_sqft	coffee_bar	video_store	prepared_food	florist	cost	promotion_name_Best Savings
0	7.36	4	1	1	1	1.84	17.70	17	27694	5415	1	1	1	1	126.62	0.0
1	5.52	3	0	4	0	1.84	17.70	17	27694	5415	1	1	1	1	59.86	0.0
2	3.68	2	4	1	0	1.84	17.70	17	27694	5415	1	1	1	1	84.16	0.0
3	3.68	2	2	2	2	1.84	17.70	17	27694	5415	1	1	1	1	95.78	0.0
4	4.08	3	0	2	0	1.36	5.11	29	27694	5415	1	1	1	1	50.79	0.0

5 rows × 115 columns

Model Evaluation

a) Types and Reasons:

The analysis involved four types of regression models: Linear Regression, Polynomial Regression (degree 2), Decision Tree Regression, and Random Forest Regression. Each type was chosen based on its suitability for the problem at hand and its ability to capture different aspects of the data.

- **Linear Regression:** Utilized as a baseline model due to its simplicity, interpretability, and widespread familiarity, facilitating easy comparison with more complex algorithms.
- **Polynomial Regression**: Introduced non-linear relationships by adding polynomial terms to the features, providing flexibility in capturing complex patterns.
- **Decision Tree Regression:** Capable of capturing non-linear relationships and interactions between features, making it suitable for complex datasets.
- **Random Forest Regression:** Ensemble technique combining multiple decision trees to improve predictive performance and handle overfitting.

b) Techniques Used:

Various techniques were employed throughout the analysis to preprocess the data, train the models, and evaluate their performance:

- **Feature Encoding**: Categorical variables were encoded to numerical format using techniques like one-hot encoding.
- **Feature Importance:** Decision Tree Regressor was used to assess the importance of each feature in predicting the target variable.
- **Model Training:** Different regression models were trained on the dataset, including Linear Regression, Polynomial Regression, Decision Tree Regression, and Random Forest Regression.

- **Model Evaluation:** Mean Squared Error (MSE), Root Mean Squared Error (RMSE), Mean Absolute Error (MAE), and R-squared (R^2) were used to evaluate the performance of each model. Additionally, cross-validation techniques like GridSearchCV were employed to fine-tune hyperparameters and optimize model performance.

Results

a) Benchmark Results

1. Linear Regression:

- Linear Regression produced moderate results:
  - MSE: 573.507854157967
  - RMSE: 23.948024013641856
  - MAE: 19.615380221780818
  - R^2: 0.36554698660582896

For the target variable "cost," Linear Regression yielded moderate predictive performance:

- Mean Squared Error (MSE) stands at 573.51, indicating the average squared difference between predicted and actual costs.
- Root Mean Squared Error (RMSE) is 23.95, representing the square root of MSE, offering insight into the average magnitude of prediction errors.
- Mean Absolute Error (MAE) of 19.62 suggests the average absolute deviation between predicted and actual costs.
- R^2 value of 0.37 signifies the proportion of variance in the cost variable that is predictable by the model.

2. Polynomial Regression (Degree 2) with Lasso Regularization:

- Polynomial Regression with Lasso Regularization outperformed Linear Regression:
  - MSE: 10.561212742387216
  - RMSE: 3.2498019543330967
  - MAE: 1.8753950582620307
  - R^2: 0.9883164751782826
  - Best Parameters: {'lasso__alpha': 0.01, 'polynomialfeatures__degree': 2}
  - Best Cross-Validation Score: 0.988

Polynomial Regression with Lasso Regularization showed superior performance compared to Linear Regression:

- Mean Squared Error (MSE) is remarkably low at 10.56, indicating significantly reduced prediction errors.
- Root Mean Squared Error (RMSE) is 3.25, highlighting the small average magnitude of prediction errors.
- Mean Absolute Error (MAE) is 1.88, suggesting minimal deviation between predicted and actual values.
- R^2 value of 0.99 indicates that the model explains 98.83% of the variance in the target variable.
- Best parameters for Lasso regularization are {'lasso__alpha': 0.01, 'polynomialfeatures__degree': 2}.
- The model achieved a cross-validation score of 0.988, reflecting its robustness across different datasets.

3. *Decision Tree Regressor:*

- Decision Tree Regressor performed remarkably well:
  - MSE: 2.8500128351095526
  - RMSE: 3.2498019543330967
  - MAE: 0.08210299861011822
  - R^2: 0.9968471238565649

Decision Tree Regressor demonstrated exceptional performance:

- MSE is impressively low at 2.85, indicating minimal prediction errors.
- RMSE is 1.69, suggesting small average magnitude of prediction errors.
- MAE is 0.08, indicating minimal deviation between predicted and actual values.
- R^2 value of 0.997 indicates that the model explains 99.68% of the variance in the target variable.

4. ***Random Forest Regressor:***

- Random Forest Regressor achieved the highest accuracy:
  - MSE: 1.2693875798821743
  - RMSE: 1.1266710167045988
  - MAE: 0.07544894419817211
  - Accuracy(R^2): 0.9985957179672739

Random Forest Regressor attained outstanding accuracy:

- MSE is remarkably low at 1.27, indicating minimal prediction errors.
- RMSE is 1.06, suggesting small average magnitude of prediction errors.
- MAE is 0.08, indicating minimal deviation between predicted and actual values.
- Accuracy(R^2) of 99.86% highlights the model's exceptional predictive capability.

Discussion

(a) Takeaways and Learnings from Data Mining:

Feature Importance:

The analysis indicates that certain features have a significant impact on the cost, such as frozen_sqft, store_sqft, and specific promotional activities like "Save-It Sale" and "Free For All". These factors should be carefully considered in strategic decision-making processes.

Model Performance:

Various regression techniques were applied, including Linear Regression, Polynomial Regression with Lasso regularization, Decision Tree Regressor, and Random Forest Regressor. The Random Forest Regressor exhibited the highest accuracy, indicating its potential for better prediction performance compared to other models.

Accuracy and Errors:

The Random Forest Regressor achieved a very low Mean Absolute Error (MAE) and Mean Squared Error (MSE), indicating its effectiveness in predicting the cost accurately. This suggests that the model is robust and reliable for cost prediction tasks.

(b) Implications and Actions for Business Managers:

Optimizing Store Layout:

The analysis highlights the significance of store_sqft and frozen_sqft in determining costs. Business managers should consider optimizing store layout and product placement to maximize sales per square foot, thus potentially reducing costs and increasing profitability.

Promotional Strategies:

Promotional activities like "Save-It Sale" and "Free For All" appear to significantly impact costs. Managers should analyze the effectiveness of these promotions in driving sales and profitability. Further investment in successful promotions and adjustments to less effective ones could be warranted.

Investment in Technology:

Given the superior performance of the Random Forest Regressor, businesses may consider investing in data-driven technologies and advanced analytics tools. This could enhance decision-making processes, improve forecasting accuracy, and ultimately drive profitability.

(c) Further Steps:

1. Fine-tune Models: Adjust hyperparameters and explore alternative algorithms for better accuracy.
2. Feature Engineering: Identify additional relevant features to enhance model performance.
3. Integration: Integrate models into Food Mart's systems for real-time CAC estimation.

4. Monitoring: Establish a framework for continuous model monitoring and evaluation.
5. Exploratory Data Analysis (EDA): Gain insights into customer behaviour and market trends for strategic interventions.

Conclusion

(a) Summary of Findings:

Through data mining analysis, it was determined that factors such as store_sqft, frozen_sqft, and specific promotional activities significantly influence costs in the retail business.

Among various regression techniques evaluated, the Random Forest Regressor demonstrated the highest accuracy in predicting costs, with low Mean Absolute Error (MAE) and Mean Squared Error (MSE).

Optimizing store layout, refining promotional strategies, and investing in technology emerged as key recommendations to enhance operational efficiency and profitability.

(b) Project Importance:

This project underscores the importance of leveraging data-driven insights to inform strategic decision-making processes within the retail industry.

By identifying critical factors influencing costs and implementing targeted actions based on data analysis, businesses can gain a competitive advantage in a dynamic marketplace.

Efficient resource allocation, guided by data-driven models, enables businesses to optimize operational processes and maximize return on investment.

(c) Reflection:

This project provided valuable learning opportunities in data mining techniques, regression analysis, and interpretation of results, contributing to professional growth and development. Overcoming challenges in data preprocessing, model selection, and interpretation of results highlighted the importance of robust methodologies and collaboration within multidisciplinary teams. Reflecting on project outcomes informs future endeavours, emphasizing the need for continuous improvement in data collection, analysis, and decision-making processes.

In conclusion, this project has demonstrated the transformative potential of data mining and analytics in driving business success within the retail sector. By effectively leveraging data-driven insights, businesses can make informed decisions, optimize resources, and gain a competitive edge in the marketplace. The findings underscore the importance of embracing a culture of innovation and continuous improvement, laying the foundation for sustainable growth and excellence in a rapidly evolving business landscape.