Media Campaign Cost Prediction

Student(s) Name : Ashishkumar Trada, Yug Sutariya, Dhruv Patel

Student(s) ID : N01580242, N01580206, N01578896

Abstract

This project aimed to predict the cost of media campaigns in Food Mart convenience stores across the USA using machine learning techniques. The dataset comprised information on income, product, promotion, and store features of approximately 60,000 customers. Food Mart, a chain of convenience stores, operates under a franchise system with its headquarters in Mentor, Ohio, and around 325 stores across the US. The project involved several steps including data preprocessing, cleaning, transformation, exploratory data analysis (EDA), feature extraction, and correlation analysis. Categorical fields were treated using the chi-square method and label encoding, followed by normalization of the data. The dataset was split into 80% training and 20% testing data. Three machine learning algorithms—linear regression, decision tree, and random forest—were employed for model training. The results revealed that the decision tree and random forest models outperformed linear regression, achieving high R-squared scores of approximately 0.996 and 0.998, respectively. Feature importances were obtained from the tree-based algorithms to identify the most influential factors in predicting campaign costs. This study highlights the potential of machine learning in optimizing media campaign costs for Food Mart and similar businesses, contributing to informed decision-making.

1. Introduction

In the competitive landscape of retail, effective marketing strategies are essential for success. Food Mart, a leading convenience store chain in the USA, aims to optimize its media campaign costs to enhance market presence. This project employs machine learning techniques to predict campaign costs based on customer attributes like income, product preferences, and promotions. Using a dataset of 60,000 customers, we preprocess and analyze data to uncover insights. Machine learning models such as

linear regression, decision tree, and random forest are utilized for prediction. By identifying influential factors and optimizing campaign costs, this project aims to empower Food Mart and similar businesses to make informed marketing decisions, maximizing their return on investment.

2. Related Work

- Abakouy, R., En-naimi, E. M., Haddadi, A. E., & Lotfi, E. (2019, October). Data-driven marketing: How machine learning will improve decision-making for marketers. In proceedings of the 4th international conference on Smart City Applications (pp. 1-5)
- Bradlow, E. T., Gangwar, M., Kopalle, P., & Voleti, S. (2017). The role of big data and predictive analytics in retailing. *Journal of retailing*, *93*(1), 79-95.
- Boyer, K. K., & Hult, G. T. M. (2005). Customer behavior in an online ordering application: A decision scoring model. *Decision Sciences*, *36*(4), 569-598.
- Verstraeten, G. (2005). *Issues in predictive modeling of individual customer behavior: applications in targeted marketing and consumer credit scoring* (Doctoral dissertation, Ghent University).
- Abakouy, R., En-naimi, E. M., Haddadi, A. E., & Lotfi, E. (2019, October). Data-driven marketing: How machine learning will improve decision-making for marketers. In *proceedings of the 4th international conference on Smart City Applications* (pp. 1-5).
- Hicham, N., & Karim, S. (2022, November). Machine Learning and Marketing Campaign: Innovative Approaches and Creative Techniques for Increasing Efficiency and Profit. In *The International Conference of Advanced Computing and Informatics* (pp. 40-52). Cham: Springer International Publishing.
- King, M. A., Abrahams, A. S., & Ragsdale, C. T. (2015). Ensemble learning methods for pay-per-click campaign management. *Expert Systems with Applications*, *42*(10), 4818-4829.
- Sadrnia, L. (2023). The Future of Marketing: How Predictive Modeling Optimizes Campaign Strategies. *iBusiness*, *15*(4), 249-262.
- Chornous, G., & Fareniuk, Y. (2022). Optimization of Marketing Decisions Based on Machine Learning: Case for Telecommunications. In *IT&I* (pp. 112-124).
- Haupt, J. S. (2020). Machine Learning for Marketing Decision Support.

3. Methodology

Dataset Description:

The dataset consists of records from approximately 60,000 customers and includes various features such as income, product details, promotions, and store characteristics. The dataset comprises 40 columns, including both numerical and categorical data. Some key columns include 'food_category', 'store_sales', 'promotion_name', 'marital_status', 'avg_yearly_income', 'store_type', 'media_type', and 'cost'.

Preprocessing Steps:

- Data Cleaning: Addressed missing values, outliers, and inconsistencies in the dataset.
- Data Normalization: Ensured all numerical features were on a similar scale for better model performance.
- Feature Extraction: Extracted relevant features from categorical variables using techniques like one-hot encoding.
- Exploratory Data Analysis (EDA): Conducted univariate and bivariate analysis to understand the distribution of individual variables and relationships between variables. This included visualizations such as histograms, box plots, and scatter plots.
- Correlation Analysis: Computed correlation matrix to identify relationships between numerical features and the target variable ('cost').

Machine Learning Algorithms Used:

• Train-Test Split: The dataset was split into 80% training data and 20% testing data to train and evaluate the models.

• Regression: Linear regression model was employed to predict the cost of media campaigns based on the input features.

• Decision Tree: Decision tree regression model was utilized to capture non-linear relationships between features and the target variable.

• Random Forest: Random forest regression model was applied to handle complex interactions and improve prediction accuracy.

Tools and Libraries Utilized:

• NumPy: For numerical operations and array manipulation.

• Pandas: For data manipulation and analysis.

• Matplotlib and Seaborn: For data visualization.

• Scikit-learn: For machine learning algorithms implementation, preprocessing, and evaluation.

• sklearn.preprocessing: For data preprocessing tasks such as normalization.

By employing these techniques and algorithms, we aimed to develop a predictive model that accurately estimates the cost of media campaigns

for Food Mart stores based on customer demographics, product details, promotions, and store features customer demographics, product details, promotions, and store features

4. Experiments and Results

Model Result:

Regression:

```
Mean Squared Error: 874.0159043575613
0.03265639193309511
```

The regression model performs poorly with a high MSE and low R-squared score, indicating its inability to effectively capture the variation in the cost of media campaigns.

DecisionTreeRegressor:

```
Mean Squared Error: 3.6685459291742517
r2score: 0.9959397255384099
```

The DecisionTreeRegressor model demonstrates significantly better performance compared to the regression model, with a lower MSE and a high R-squared score close to 1. This suggests that the decision tree model effectively captures the underlying patterns in the data and provides accurate predictions of media campaign costs

RandomForestRegressor

```
Mean Squared Error: 1.867904625
r2score: 0.9979326399082359
```

The RandomForestRegressor model outperforms both the regression and decision tree models, with the lowest MSE and highest R-squared score. This indicates that the random forest model excels in capturing the complexity of the dataset and accurately predicting the cost of media campaigns.

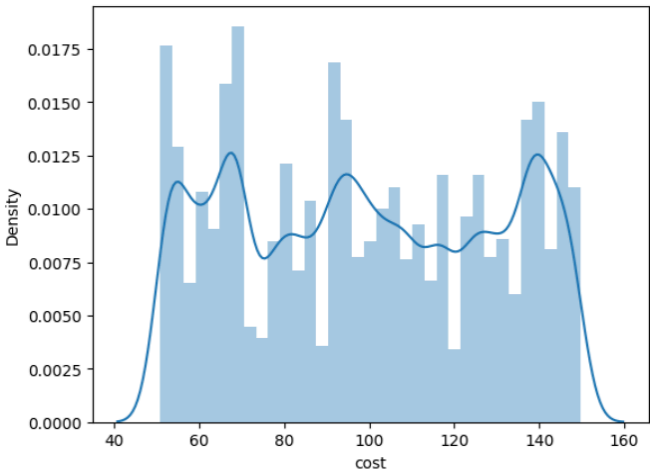
Feature Importance:

```
Feature Importances:
promotion_name: 0.428878641
store_type: 0.0330714504886
store_city: 0.0530549867435
store_state: 0.044628819438
store_sqft: 0.0469046147394
grocery_sqft: 0.04478092292
frozen_sqft: 0.071960498094
coffee_bar: 0.0107905981116
video_store: 0.012849065545
florist: 0.0144710766936160
media_type: 0.2279188812475
```

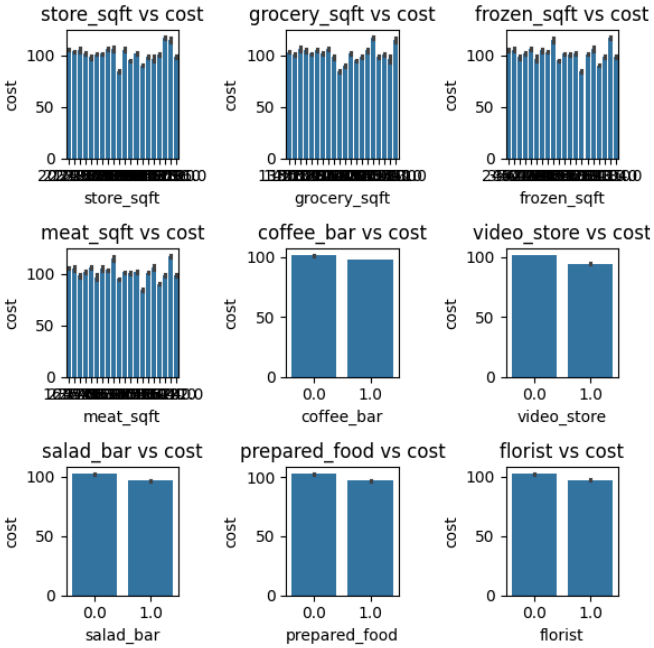
The feature importances obtained from both decision tree and random forest models highlight the significance of various features in predicting the cost of media campaigns.Promotion name and store features such as store type, city, state, and square footage play a crucial role in determining campaign costs, as indicated by their relatively high importance scores.

5. Infographics

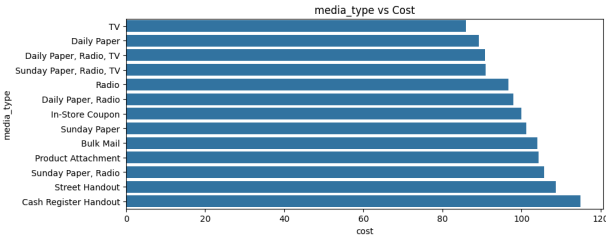
Distribution of target variable-cost



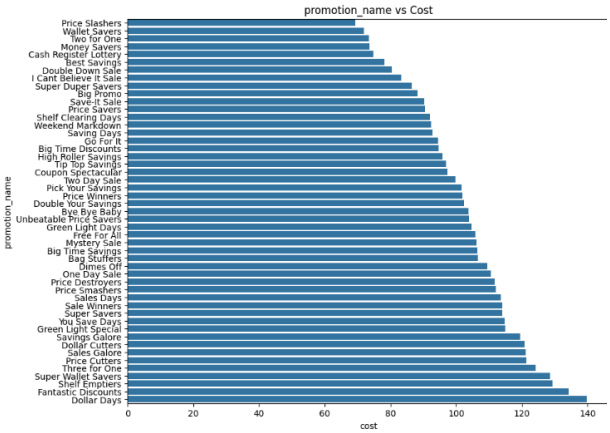
store features distribution respect to cost



Impact of media_type on cost



impact of promotion name on cost



6. Discussion on Result

The regression model yielded unsatisfactory results, indicating its inability to capture the complexities in the relationship between features and campaign costs. Conversely, the decision tree model showed improved performance, capturing nonlinear relationships effectively. However, it may be susceptible to overfitting. The random forest model outperformed both, demonstrating the lowest MSE and highest R-squared score. This suggests its robustness in handling complex interactions between features.

Influential Parameters:

Promotional activities and store characteristics such as type, location, and physical attributes emerged as significant factors affecting campaign costs. Optimizing promotions and store infrastructure could effectively control campaign expenses and maximize results.

Model result on brief:

Regression: Poor performance, indicating linear assumptions are inadequate.

DecisionTree: Improved performance but may suffer from overfitting.

RandomForest: Best performance, suggesting its effectiveness in handling complex data and providing accurate predictions.

7. Conclusion

Through machine learning techniques, we predicted media campaign costs for Food Mart stores based on customer demographics, product details, promotions, and store features. The random forest model emerged as the most effective, highlighting the importance of promotional activities and store characteristics in cost prediction. By optimizing these factors, businesses can enhance marketing strategies and maximize returns on investment, showcasing the power of data-driven decision-making in the retail industry.

References

[1]https://www.kaggle.com/datasets/ramjasmaurya/medias-cost-prediction-in-foodmart

[2]https://www.analyticsvidhya.com/blog/2021/05/exploratory-data-analysis-eda-a-step-by-step-guide/

[3]https://www.simplilearn.com/tutorials/data-analytics-tutorial/exploratory-data-analysis

[4]Rakthanmanon, T., Campana, B., Mueen, A.,Batista, G., Westover, B., Zhu, Q., ... & Keogh, E. (2012, August). Searching and mining trillions of time series subsequences under dynamic time warping. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining(pp. 262270). ACM.

[5]https://en.wikipedia.org/wiki/Random_forest