The Future of Grocery Shopping: Analyzing and Predicting Marketing Campaign Success

Crystal Luo '24 Wellesley College Data Science Major Capstone

Background and Research Question

This project utilizes data from iFood, a leading Brazilian online food ordering and delivery platform with more than 80% of the market share in Brazil's grocery market sector.

The dataset is collected from 2,240 customers from a prior marketing campaign. Leveraging this dataset, this project seeks to build a predictive model to address the following research questions:

1) **How do customer demographics and historical purchasing behaviors forecast the success of marketing campaigns in grocery retail?**
2) **Which predictive model is most effective in maximizing profits for future marketing campaigns?**

Data

Description

Utilizes data from 2,240 customers targeted in a previous marketing campaign for a product offering. Customers were randomly selected and reached out to by phone.

- Outcome Variable: customers' responses to the product offer (purchase vs. no purchase) there was a large class imbalance.

Response = 0	Response = 1
1906	334

- Categorical Variables (2):
  - *Marital_Status*
  - *Education* Level
- Date Variable (1)
- Numeric Variables (25):
  - Income, Product category spending, frequency of purchases, and historical engagement with marketing campaigns, etc.

Data Cleaning:

Income. Imputed 24 NA values using MICE function.

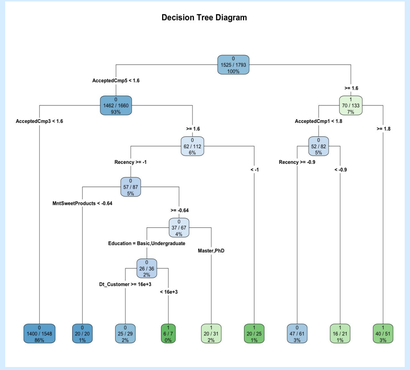
Marital Status. "Alone" was combined with "Single", and less conventional categories like "Absurd" and "YOLO" were grouped under "Other".

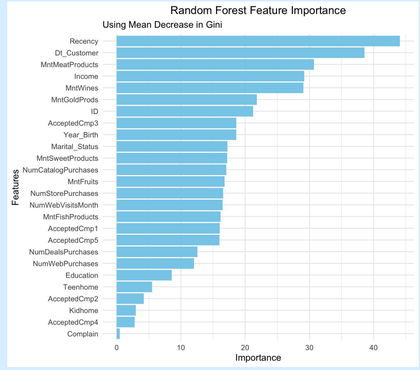
Education Level. "2n Cycle" was combined with "Master", as they refer to the same level of education. "Graduation" was renamed to "Undergraduate".

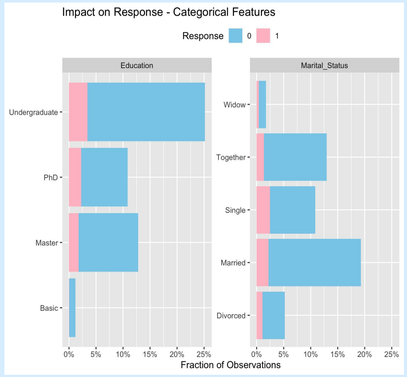
Methodology and Models

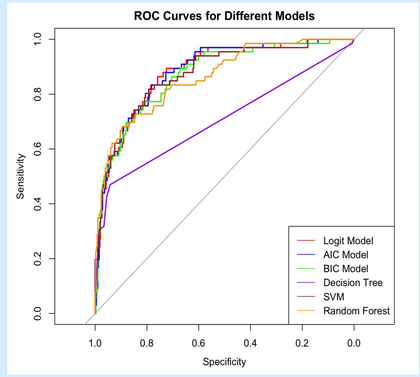
The data was first standardized, then randomly split into 20% testing and 80% training. Logistic Regression was carried out first, followed by a Variance Inflation Factor (VIF) analysis to ensure no multicollinearity influenced the predictive variables. Since no multicollinearity was detected, we proceeded with a stepwise regression based on both Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) to streamline the model by selecting significant predictors while reducing complexity.

Given the large feature space of this dataset, it is not clear whether the data points have linear or non-linear decision boundaries. Therefore, multiple models that are suitable for a multiclass classification problem were employed: Support Vector Machines (SVM), Decision Tree, and Random Forest. For each model, I fitted the model on the training set, and evaluated model performance on the testing set.









Results and Evaluation

This project focused on the following metrics to evaluate each model:

Model	Accuracy	Precision	Recall	F1 Score	AUC
Logistic Reg.	88.81%	91.48%	95.80%	93.59%	0.8806
AIC Model	88.59%	91.46%	95.54%	93.45%	0.8813
BIC Model	89.71%	91.56%	96.85%	94.13%	0.8703
Decision Tree	87.02%	89.10%	96.59%	92.70%	0.7056
SVM	89.04%	91.29%	96.33%	93.74%	0.8747
Random Forest	89.26%	89.93%	98.43%	93.98%	0.8618

All models performed pretty well across all metrics. AUC and F1 will be the main metrics for evaluation as they can deal with class imbalance, and consider both minimizing wasted effort (high precision) or maximizing campaign reach (high recall). The Logistic Regression model with the AIC criteria has the largest ROC AUC value. The model is as follows:

$$\log\left(\frac{p}{1-p}\right) = 58.870441 + 1.933812 \cdot \text{EducationMaster} + 2.623675 \cdot \text{EducationPhD} + 1.659186 \cdot \text{EducationUndergraduate} - 1.169866 \cdot \text{Marital\_StatusMarried} + 0.285563 \cdot \text{Marital\_StatusSingle} - 1.152211 \cdot \text{Marital\_StatusTogether} + 0.294965 \cdot \text{Marital\_StatusWidow} - 0.665504 \cdot \text{Teenhome} - 0.003963 \cdot \text{Dt\_Customer} - 0.975670 \cdot \text{Recency} + 0.417033 \cdot \text{MntMeatProducts} + 0.183161 \cdot \text{MntGoldProds} + 0.267369 \cdot \text{NumDealsPurchases} + 0.175180 \cdot \text{NumWebPurchases} + 0.280796 \cdot \text{NumCatalogPurchases} - 0.592911 \cdot \text{NumStorePurchases} + 0.350194 \cdot \text{NumWebVisitsMonth} + 0.537772 \cdot \text{AcceptedCmp3} + 0.205763 \cdot \text{AcceptedCmp4} + 0.408576 \cdot \text{AcceptedCmp5} + 0.403341 \cdot \text{AcceptedCmp1} + 0.206161 \cdot \text{AcceptedCmp2}$$

Variables of significance:
Education (AIC), Marital Status (AIC), Recency (RF), Dt_Customer (RF), AcceptedCampaign5 (DT)

Discussion and Conclusions

Each model suggested different variables of significance. Logistic Regression model yields highest AUC value and is most interpretable. Given the strong performance of the Logistic Regression model, the underlying relationships between the features and the response seem more-or-less linear. It might not have been completely necessary to employ such robust ML techniques. However, the different models identified customer characteristics that could help maximize profits and reduce costs.

For future improvements:
1) it is not completely clear when the data was collected
2) the class imbalance may not have been accurately represented when splitting the dataset into training and test data.
3) tuning hyperparameters for ML models