Technical Report on Promotion Response Predictive Modeling

Introduction

This project revolves around predicting customer responses to promotional activities in the retail sector, a critical task for optimizing marketing strategies and enhancing customer engagement. The aim is to identify which customers are likely to be 'active', meaning they will response positively to promotions. Accurately anticipating these reactions is key to delivering personalized experiences and building loyalty, which in turn can lead to increased sales and a stronger market position.

The analysis is based on a dataset containing 20,000 instances and 17 features, sources from Kaggle, reflecting customers' past interactions with various promotions and transactios. Key attribute such as transaction frequency, promotion types, and transaction monetary values have been considered to capture the diverse factors that may influence a customer's decision to engage with promotional offers. This dataset serves as a foundation for developing a model that can not only predict response with high accuracy but also provide insights into the effectiveness of different promotional strategies.

Approach

To effectively predict customer responses to promotional activities, I adopted a multi-faceted approach focusing on advanced feature engineering and strategic model development. Initially, I enriched our feature set by incorporating additional data elements related to customer transactions, specifically 'total spent', 'total transactions', and 'unique categories'. These features were selected for their relevance to customer purchasing behavior and their potential impact on promotional responsiveness.

Feature Engineering

In the feature engineering phase, a correlation analysis (shown as figure 1) was crucial in identifying redundancies within our dataset. This analysis led to the elimination of highly correlated variables, thus reducing multicollinearity, and enhancing the predictive power of our model. Specific features (such as 'total spent', 'total transactions' and 'promo') are dropped due to high correlation included transaction-related attributes, which overlapped significantly with total spent and total transactions.

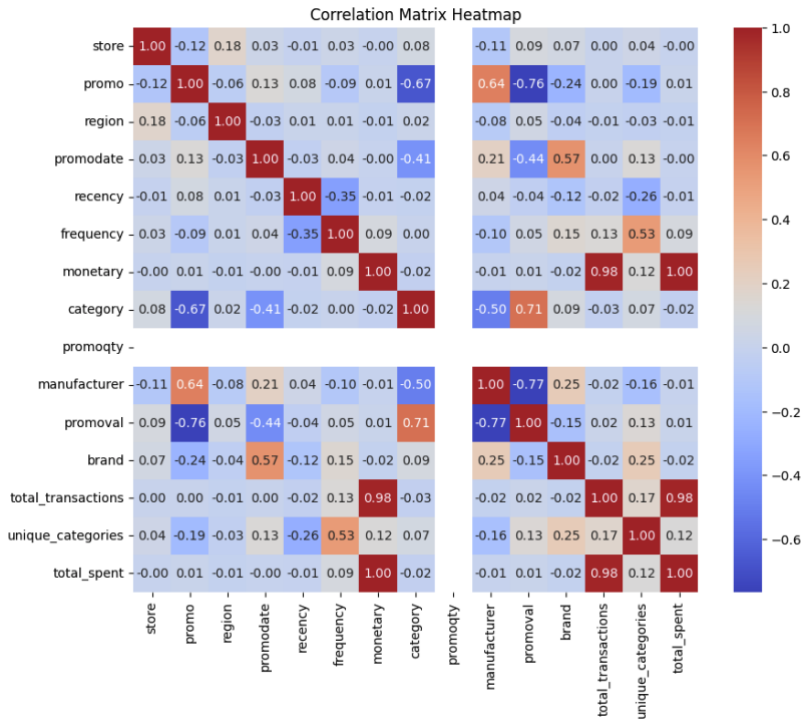

Fig [1] Correlation Matrix Heatmap

To accommodate categorical variables such as 'brand, 'store', 'manufacturer', and 'region', I converted these into dummy variables. This transformation was essential for integrating categorical data into our model, allowing it to process a broader spectrum of information effectively.

Prior the model training, I addressed the challenge of missing data through strategic imputation, ensuring no valuable information was lost. This preprocessing step was vital for maintaining the integrity and completeness of our dataset.

Model Development

For the model development phase, I chose the Random Forest algorithm due to its robustness and ability to handle diverse data types and complex interaction effects among features. I engaged in rigorous hyperparameter tuning using RandomizedSearchCV, which explored a range of potential settings to optimize model parameters such as the number of estimators, maximum depth, minimum sample split, and maximum features. The parameter's tuning process was guided by the goal of maximizing the ROC-AUC score, a critical measure for assessing the model's ability to discriminate between classes effectively.

Non-obvious Implementation Details

In the process of developing our predictive model using the Random Forest algorithm, a critical implementation detail that may not be immediately apparent is the methodical handling of missing data. The Random Forest algorithm, despite its robustness and

versatility, requires a fully populated dataset for training and predictions, as it cannot inherently process missing values.

To address this requirement, a strategic approach was adopted to manage missing entries in the dataset. Recognizing the potential impact of missing data on the model's performance and the integrity of the analysis, I implemented a mean imputation strategy using the 'SimpleImputer' from scikit-learn's preprocessing tools. This choice of imputation was driven by the mean's ability to preserve the overall statistical characteristics of the dataset without introducing significant biases, which is crucial given our model's sensitivity to the input data distribution. The imputation was applied consistently across testing datasets to maintain uniformity in data treatment, which is vital for valid model evaluation and performance testing.

Experimental Setup

The experimental setup for our predictive model was meticulously designed to ensure the robustness and reliability of our findings. Our goal was to develop a model that not only performs well on historical data but also generalizes effectively to new, unseen data.

Model Configuration

For the model training, I utilized a RandomForestClassifier due to its effectiveness in handling both categorical and numerical data, its inherent feature selection capabilities, and its robustness to overfitting with default settings. The configuration of the RandomForest was determined through a systematic search for optimal hyperparameters using RandomizedSearchCV, which examined combinations of several key parameters such as n_iter, cv and etc.

Validation Technique

To ensure the model's generalizability, I employed stratified k-fold cross validation with 'k=5'. This method is particularly suited to our dataset as it preserves the percentage of samples for each class, which is crucial given the imbalanced nature of our response variable. This approach helps mitigate any potential biases that could arise from imbalanced data and ensures that each fold is a good representative of the overall dataset.

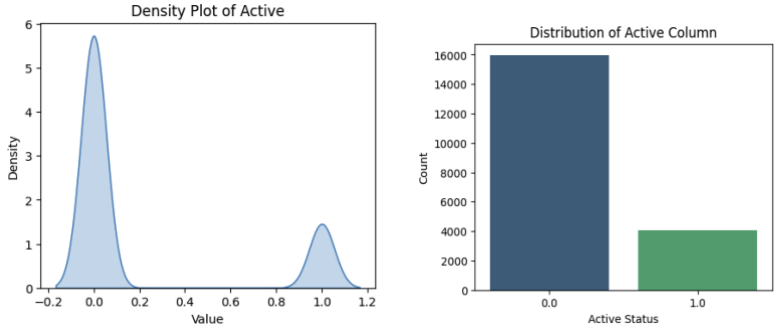

Fig [2] Balanced check for the target variable – 'active'

Performance Metrics

The primary metrics used to evaluate the model's performance were 'Test Accuracy' and the 'Area Under the Receiver Operating Characteristic Curve (AUC-ROC). Accuracy provides a straightforward measure of overall correctness, while AUC-ROC offers a more nuanced view by measuring the ability of the model to distinguish between the classes at various threshold settings. The model achieved an impressive 81% accuracy on the testing set and an AUC of 69% on the test set, demonstrating its efficacy and the success of our feature engineering and model tuning efforts.

Future Directions

The predictive model developed in this project has shown promising results, yet there remains significant potential for enhancement and refinement to further increase its efficacy and application scope. Looking ahead, several key areas could be explored to push the boundaries of our current solution:

Advanced Feature Engineering

I plan to delve deeper into more sophisticated feature engineering techniques. This includes leveraging interaction terms that could uncover complex relationships between features and exploring time-series for promotions that are time-bound. The integration of text analytics to parse and utilized unstructured data from customer feedback could also provide deeper insights into customer preferences and behavior patterns.

Adoption of Advanced Modeling Techniques

While RandomForest has provided robust results, the exploration of advanced models such as deep learning could unveil new possibilities, especially capturing non-linear relationships and interactions at scale. Implementing neural networks could particularly be transformative in handling high-dimensional data and enhancing the model's ability to learn from a large volume of transactional data.

Expanding Data Sources

To enrich our model's understanding and predictive accuracy, expanding our data sources will be crucial. Incorporating external data such as economic indicators, market trends, and demographic data could provide a more holistic view of the factors influencing customer behaviors. Additionally, partnerships with other platforms could be explored to gather broader behavioral data across different shopping channels.

References

- CleverTap. (n.d.). RFM Analysis. Retrieved April 12, 2024, from https://clevertap.com/blog/rfm-analysis/

- Brian Keng. (2024). [Assignment 3 - Example]. Colab Research. from https://colab.research.google.com/drive/1lCSchHT4yQaqzMdnGaL9-xvalLCjKcTK#scrollTo=5_yNT76IIaSZ.

Rotman

Final Page

Grade: _____