Changsong Yang 1001416149

RSM8521 Assignment 3

Introduction

In recent years, direct marketing has garnered increased attention due to the growth of digital channels. Rather than relying on mass media channels, direct marketing emphasizes delivering targeted promotional content directly to consumers. Determining the likelihood of a customer responding to a specific promotion can be invaluable for designing and managing marketing campaigns, as well as enhancing a company's sales performance.

The objective of this project is to develop a predictive model that can accurately estimate the probability of a customer responding to a promotion, using historical transaction data, promotion details, and customers' past responses to promotions. The model's output is a probability value ranging from zero to one. Feature engineering was employed, and various techniques were implemented to enhance the model's predictive capabilities. The model's performance was assessed using the ROC AUC score.

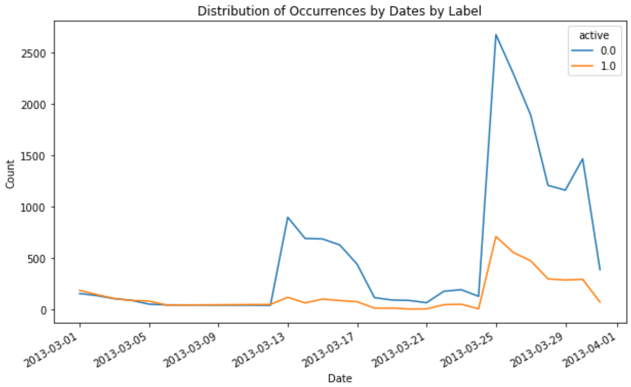
A training dataset was employed to train, tune, and evaluate the model. The best model, after hyperparameter tuning, was utilized to make predictions on a test dataset.

Exploratory Data Analysis

There are three datasets available for predicting customers' responses.

- **transactions**: Transactions for all customers between 2012-03 and 2013-02
- **promos**: Metadata about each promotion
- **train_history**: Promotions given to a subset of the customers during 2013-03 and whether or not they responded

A preliminary exploratory data analysis was carried out to examine the dataset and identify possible connections between customers' responses and various features. Figure 1 illustrates the distribution of customer responses by promotion date, while Figure 2 displays the distribution of customer transactions by transaction date. The target variable - whether the customer responded - is represented by different colours in both figures. As observed, the counts of customers who responded and those who did not follow a similar trend and exhibit comparable seasonality. It is important to note that the datasets are imbalanced, with customers who responded constituting only about 25% of the training dataset.



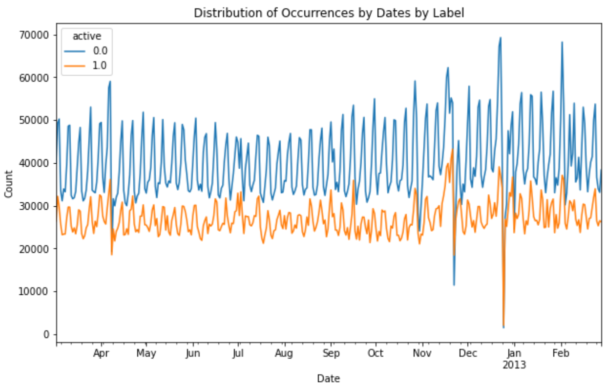


Figure 1: distribution of customer responses by promotion date from training history dataset; Figure 2: distribution of customer transactions by transaction date from the transactions dataset.

Feature Engineering

Feature engineering was performed to improve the predictive power of the model. Extra columns were created by aggregating on existing data.

Recency, Frequency, and Monetary Value

The recency, frequency, and monetary value (RFM) of each customer were determined using the transaction data. The methodologies employed for the calculations and the underlying rationales are described below:

- **recency**: The number of days between the most recent date in the transactions table and the latest transaction record for a customer. The hypothesis associated with this feature is that customers who have made purchases more recently may be more interested in promotions.
- **frequency**: The total number of days on which a customer made purchases, divided by the total number of days between March 2012 and February 2013. The hypothesis associated with this feature is that customers who make purchases more frequently may be more interested in promotions.
- **monetary** value: The cumulative amount spent by a customer between March 2012 and February 2013. The hypothesis associated with this feature is that customers who spend more on purchases may be more interested in promotions.

These RFM metrics provide valuable insights into customer behavior, enabling us to create a more accurate and effective predictive model.

Three-Month Recency, Frequency, and Monetary Value

In addition to the RFM features derived from the entire duration of the transaction dataset, RFM features for the most recent three months leading up to the latest date in the transaction dataset were also extracted using similar methodologies. These features might prove beneficial as they represent more recent customer behavior, which could be more pertinent when predicting their responses to promotions. The names of these additional features are provided below.

- **recency_3_month**
- **frequency_3_month**
- **monetary_3_month**

Brand and Category Counts

A customer may be more inclined to respond to a promotion involving a product or brand they have previously purchased. As a result, the number of purchases a customer made for the promoted brand within their past transaction history was calculated using the training dataset. Similarly, the number of purchases a customer made within the product category was also computed. To calculate these two features, aggregations were performed on the transaction dataset, and the resulting table was joined with the training dataset.

- **category_count**: Group the transaction dataset by customer ID and product category to count the total number of purchases within each category. Perform a left join between the training table and the transaction table on the customer ID and product category to include only the purchase count of the promoted category in the resulting table.
- **brand_count**: Group the transaction dataset by customer ID and product brand to count the total number of purchases within each brand. Perform a left join between the training table and the transaction table on the customer ID and brand category to include only the purchase count of the promoted brand in the resulting table.

By incorporating these features, we can better understand the customer's past behaviour and preferences, potentially improving the predictive power of the model.

One-Hot Encoding

The model used to predict the probability of a customer responding to a promotion is a tree-based model, which requires categorical features to be numeric variables. One-hot encoding was used to convert the original categorical variables unique binary value. This method was used on the training dataset, and the same model was used to transform the test dataset.

Joining Tables

The engineered tables were joined as illustrated in Figure 3.

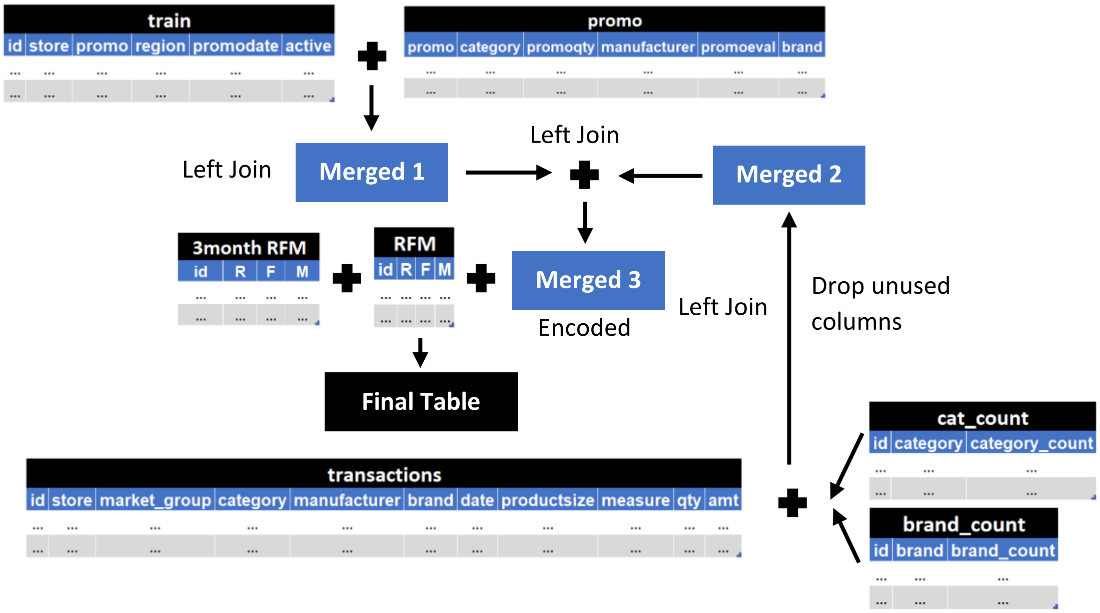
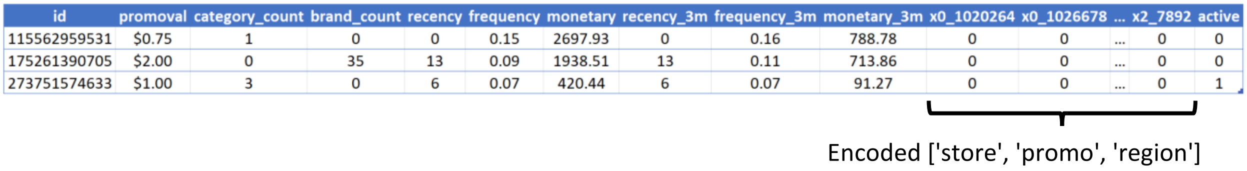

train
id store promo region promodate active
promo
promo category promoqty manufacturer promoeval brand
Left Join
Merged 1
Left Join
Merged 2
3month RFM
id R F M
RFM
id R F M
Merged 3
Encoded
Left Join
Drop unused columns
Final Table
cat_count
id category category_count
transactions
id store market_group category manufacturer brand date productsize measure qty amt
brand_count
id brand brand_count

Figure 3: tables were joined, and features were engineered to create the final ADV.

Feature Summary

The final analytical data view (ADV) comprises 163 features, customer IDs, and the target variable, with store, promotion ID, and region information one-hot encoded. Variables such as manufacturer, brand, promodate, promoqty, and category were excluded due to the impracticality of encoding them and their weak association with the target variable.

Table 1: the final analytical data view that will be used as the inputs of the model



id	promoval	category_count	brand_count	recency	frequency	monetary	recency_3m	frequency_3m	monetary_3m	x0_1020264	x0_1026678	...	x2_7892	active
115562959531	$0.75	1	0	0	0.15	2697.93	0	0.16	788.78	0	0	...	0	0
175261390705	$2.00	0	35	13	0.09	1938.51	13	0.11	713.86	0	0	...	0	0
273751574633	$1.00	3	0	6	0.07	420.44	6	0.07	91.27	0	0	...	0	1

Encoded ['store', 'promo', 'region']

Building the Model

To develop a model that accomplishes the project's objective, the engineered dataset was divided into training and testing sets using an 85:15 ratio. The training set was employed to train and fine-tune the model, while the test set was used to evaluate its performance.

Random Forest Classifier

As the aim is to predict whether a customer will respond to a promotion, a classification model is required. Considering the presence of categorical features with a significant number of unique values, a model that can effectively handle categorical inputs and is not heavily impacted by the curse of dimensionality is preferred. As a result, a random forest classifier was chosen as an initial approach to predict customer responses.

Imbalanced Dataset

The provided training dataset is imbalanced, with only a quarter of the customers having responded to promotions. To mitigate the impact of this issue, the RandomOverSampler technique was tested to oversample the training dataset. However, since no significant improvement in the ROC AUC score was observed, oversampling was not implemented.

Hyperparameter Tuning

The hyperparameters used for tuning the random forest model are shown Table 2. A k-fold cross-validation with k=5 was employed to assess model performance. The ROC AUC score served as the evaluation metric. The optimal hyperparameters, which yielded the highest AUC score, were identified and subsequently used to generate predictions.

Table 2: hyperparameters used for model tuning; Table 3: top five most important features from the model.

HYPERPARAMETERS	SEARCHED	BEST
N_ESTIMATORS	100, 200, 500, 800	800
MAX_DEPTH	5, 10, 20, 30, None	20
MIN_SAMPLES_SPLIT	2, 5, 10	10
MIN_SAMPLES_LEAF:	1, 2, 4, 8	4

	Importance
monetary	0.098367
MONETARY_3_MONTH	0.097044
X1_209524	0.094442
FREQUENCY	0.08645
FREQUENCY_3_MONTH	0.083279

Model Performance

The best-performing model achieved an ROC AUC score of 0.69, signifying a 69% probability of accurately distinguishing between positive and negative classes. This result represents a substantial improvement compared to random guessing. Table 3 lists the top five most important features along with their respective importance values. The monetary values, frequency, and the presence of promotion ID 209524 exhibit considerable predictive power in the model.

Future Work

While the current model demonstrates reasonable performance in predicting customer responses to promotions, there are several avenues for future work to improve the model's accuracy and applicability in real-world scenarios.

- **Feature Engineering**: Many features were not used in the current model. Additional features could be engineered to capture more customer information to improve the model's predictions.
- **Algorithm Exploration**: Other classification algorithms can be explored, such as GBMs or DL models.
- **Hyperparameter Tuning**: More extensive and fine-grained hyperparameter tuning can be performed.

Conclusion

In this project, a random forest predictive model was developed to estimate customer responses to promotions using historical transaction data, promotion details, and past responses. The model achieved a 0.69 AUC score. Key factors contributing to the likelihood of response included monetary values and frequency. This model can help companies target customers more effectively, optimize marketing resources, and enhance sales performance.

Appendix

ChatGPT

In this project, ChatGPT was employed exclusively to enhance the linguistic quality of the report.

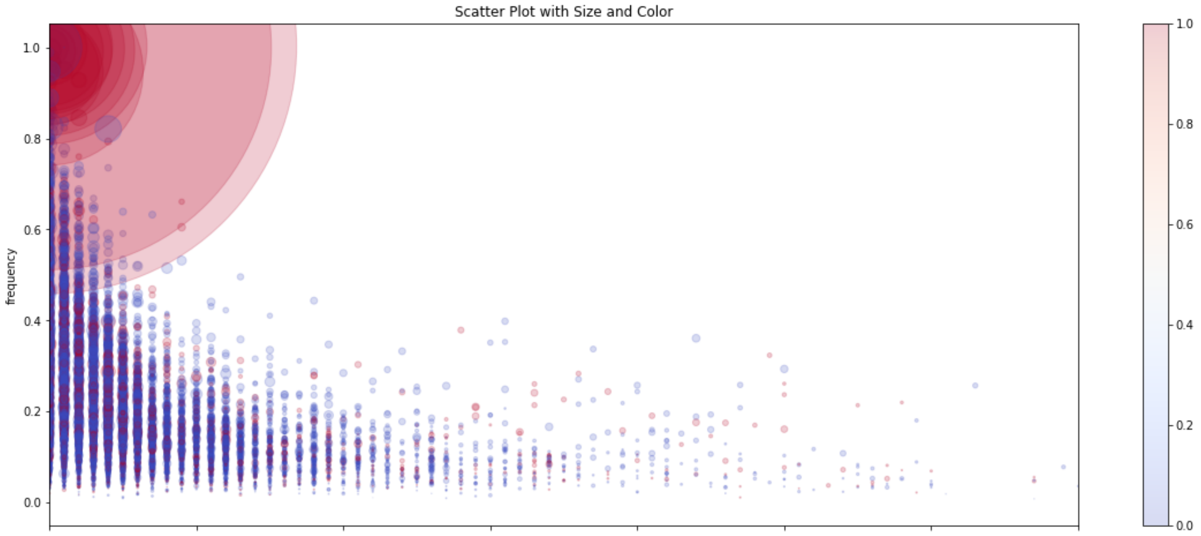
Feature Exploration

X axis: Recency

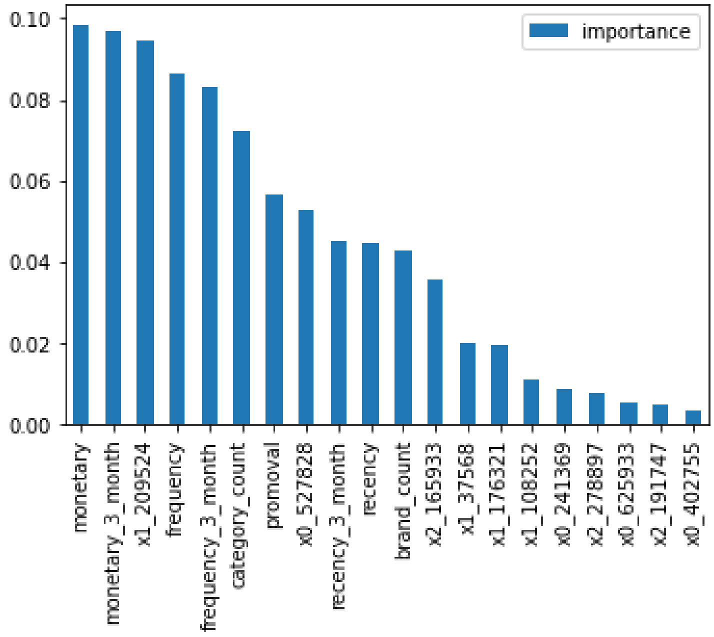
Y axis: Frequency

Size: Monetary values

Colour: Target variable



Importance of Features


importance
monetary
monetary_3_month
x1_209524
frequency
frequency_3_month
category_count
promoval
x0_527828
recency_3_month
recency
brand_count
x2_165933
x1_37568
x1_176321
x1_108252
x0_241369
x2_278897
x0_625933
x2_191747
x0_402755