Business Intelligence Case Study

DEVELOPING UPLIFT MODEL TO OPTIMIZE MARKETING PROMOTION CAMPAIGN

Presented by Anniza Mega

Overview

About Me

- Hello everyone, introducing my name Anniza Mega, a student in Dibimbing.id Business Intelligence batch 9. I am happy to share my project.
- In today's competitive market, understanding customer behavior and optimizing customer relationships are crucial for business success. This Business Intelligence (BI) project focuses on two essential aspects: Customer Churn Analysis and Customer Lifetime Value (CLV) Optimization. The project aims to leverage data analytics to identify at-risk customers and develop strategies to retain them while maximizing their long-term value to the company.

INTRODUCTION

What is Uplift Modeling?

Uplift modeling, also known as incremental modeling, differential response modeling, or true lift modeling, is a predictive analytics technique used to estimate the causal effect of a treatment or intervention on an individual outcome. Unlike traditional response modeling, which predicts the likelihood of an outcome irrespective of any treatment, uplift modeling aims to measure the difference in outcomes between a treated group and a control group. This approach helps in identifying individuals who are most likely to be positively influenced by the treatment, allowing for more targeted and efficient interventions.

This approach divides the people who are eligible to be contacted into four behavior segments:

Sure things: the folks who would do the thing regardless of whether or not you contact them.

Persuadables: the folks who will only do the thing if contacted.

Lost Causes: the folks who will never do the thing regardless of whether you contact them.

Do Not Disturbs: the folks who will be dissuaded from doing the thing because you contacted them. Also called Sleeping Dogs (as in “don’t wake a sleeping dog”).

Business Problem

The company had a dataset of 64,000 customers with brief information, such as historical promotion usage (Discount or Buy One Get One), offers that had been extended, and conversion results (purchase or not). The company wants to optimize Marketing Promotion Campaigns to target promotions at potential customers.

Objective

The main objective of this project is to build an uplift model to predict the effectiveness of promotional strategies and optimize marketing campaigns to maximize conversions and reduce campaign costs

DATA OVERVIEW

```
[120] import pandas as pd

      df = pd.read_csv('/content/data.csv')

      df.head()  # Display first few rows
```

	recency	history	used_discount	used_bogo	zip_code	is_referral	channel	offer	conversion
0	10	142.44	1	0	Surburban	0	Phone	Buy One Get One	0
1	6	329.08	1	1	Rural	1	Web	No Offer	0
2	7	180.65	0	1	Surburban	1	Web	Buy One Get One	0
3	9	675.83	1	0	Rural	1	Web	Discount	0
4	2	45.34	1	0	Urban	0	Web	Buy One Get One	0

Next steps: View recommended plots

This dataset contains Marketing Promotion Campaign data conducted by a company. This dataset contains a total of 64.000 customer data. 06

Content:

- recency: months since last purchase history: value of the historical purchases
- used_discount: indicates if the customer used a discount before
- used_bogo: indicates if the customer used a buy one get one before
- zip_code: class of the zip code as Suburban/Urban/Rural
- is_referral: indicates if the customer was acquired from referral channel
- channel: channels that the customer using, Phone/Web/Multichannel
- offer: the offers sent to the customers, Discount/But One Get One/No Offer
- conversion: customer conversion(buy or not)

DATA PROFILING

```
df.info()  # Data types and non-null counts
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 64000 entries, 0 to 63999
Data columns (total 9 columns):
 #   Column         Non-Null Count  Dtype
---  ------         --------------  -----
 0   recency        64000 non-null  int64
 1   history        64000 non-null  float64
 2   used_discount  64000 non-null  int64
 3   used_bogo      64000 non-null  int64
 4   zip_code       64000 non-null  object
 5   is_referral    64000 non-null  int64
 6   channel        64000 non-null  object
 7   offer          64000 non-null  object
 8   conversion     64000 non-null  int64
dtypes: float64(1), int64(5), object(3)
memory usage: 4.4+ MB
```

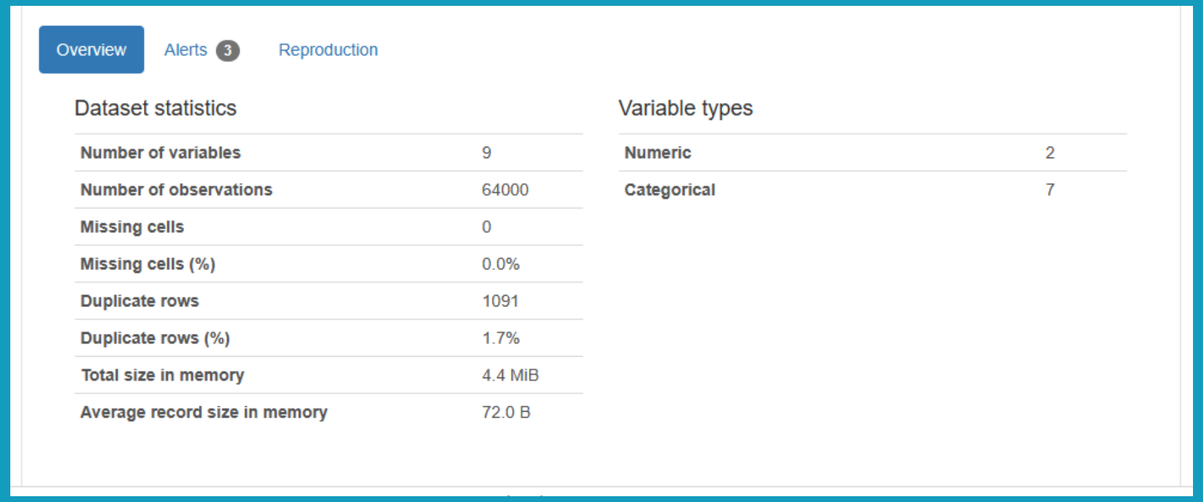
There are 9 columns and 64000 rows in this data which consists of 3 catagorical data types and 6 numerical.

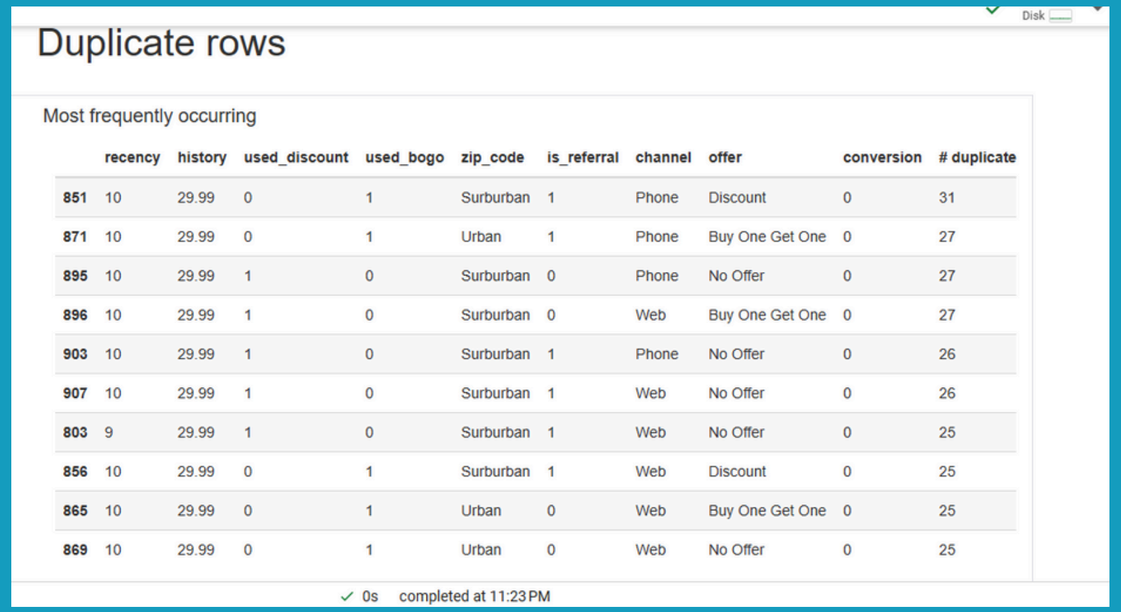
```
[123] df.describe()  # Summary statistics
```

	recency	history	used_discount	used_bogo	is_referral	conversion
count	64000.000000	64000.000000	64000.000000	64000.000000	64000.000000	64000.000000
mean	5.763734	242.085656	0.551031	0.549719	0.502250	0.146781
std	3.507592	256.158608	0.497393	0.497526	0.499999	0.353890
min	1.000000	29.990000	0.000000	0.000000	0.000000	0.000000
25%	2.000000	64.660000	0.000000	0.000000	0.000000	0.000000
50%	6.000000	158.110000	1.000000	1.000000	1.000000	0.000000
75%	9.000000	325.657500	1.000000	1.000000	1.000000	0.000000
max	12.000000	3345.930000	1.000000	1.000000	1.000000	1.000000

Descriptive Statistics:
Contains basic descriptive statistics for each column in the data frame that has a numeric data type or, with certain parameters,for columns with other data type.

DATA PROFILING



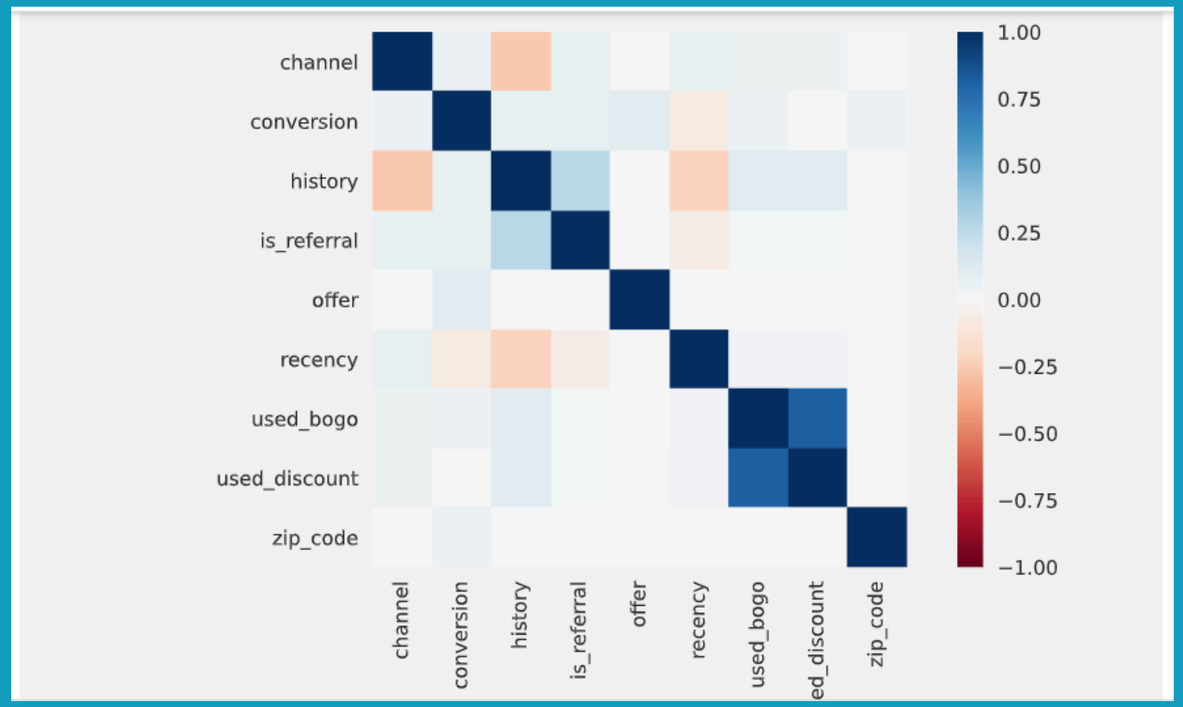


	recency	history	used_discount	used_bogo	zip_code	is_referral	channel	offer	conversion	# duplicate
851	10	29.99	0	1	Surburban	1	Phone	Discount	0	31
871	10	29.99	0	1	Urban	1	Phone	Buy One Get One	0	27
895	10	29.99	1	0	Surburban	0	Phone	No Offer	0	27
896	10	29.99	1	0	Surburban	0	Web	Buy One Get One	0	27
903	10	29.99	1	0	Surburban	1	Phone	No Offer	0	26
907	10	29.99	1	0	Surburban	1	Web	No Offer	0	26
803	9	29.99	1	0	Surburban	1	Web	No Offer	0	25
856	10	29.99	0	1	Surburban	1	Web	Discount	0	25
865	10	29.99	0	1	Urban	0	Web	Buy One Get One	0	25
869	10	29.99	0	1	Urban	0	Web	No Offer	0	25

The dataset contains 9 variables and 64,000 rows. While there are no missing values, 1.7% of the rows are duplicates. It needs to be determined whether these duplicate rows should be removed.

After checking the duplicate row data, there is duplicate data that has the potential to affect the model to be created, therefore at the next stage it is necessary to drop the duplicate row

DATA PROFILING



Heatmap | Table

	channel	conversion	history	is_referral	offer	recency	used_bogo	used_discount	zip_code
channel	1.000	0.050	-0.259	0.070	0.000	0.067	0.055	0.061	0.003
conversion	0.050	1.000	0.070	0.074	0.089	-0.076	0.052	0.005	0.049
history	-0.259	0.070	1.000	0.266	0.000	-0.226	0.100	0.101	0.000
is_referral	0.070	0.074	0.266	1.000	0.000	-0.053	0.021	0.020	0.000
offer	0.000	0.089	0.000	0.000	1.000	-0.002	0.000	0.000	0.000
recency	0.067	-0.076	-0.226	-0.053	-0.002	1.000	0.026	0.032	0.000
used_bogo	0.055	0.052	0.100	0.021	0.000	0.026	1.000	0.817	0.000
used_discount	0.061	0.005	0.101	0.020	0.000	0.032	0.817	1.000	0.000
zip_code	0.003	0.049	0.000	0.000	0.000	0.000	0.000	0.000	1.000

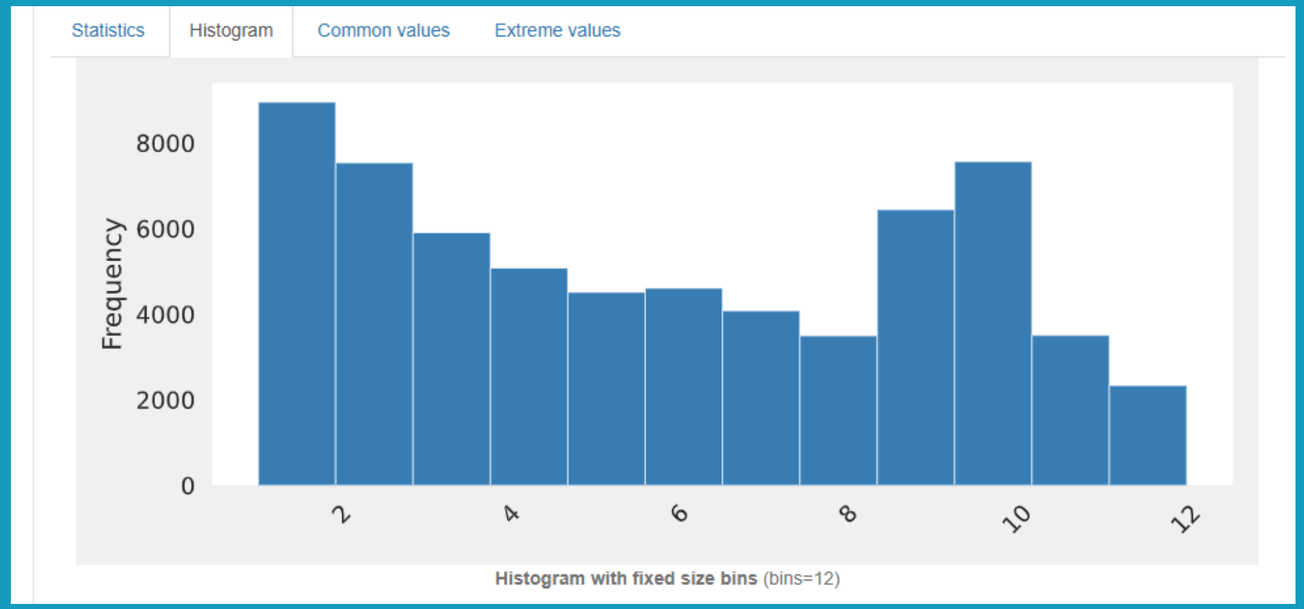
Correlation:

The correlation analysis revealed a high positive correlation between the `used_bogo` and `used_discount` variables, with a value of 0.817. Therefore, it is recommended to exercise caution when interpreting the model.

DATA EXPORATION

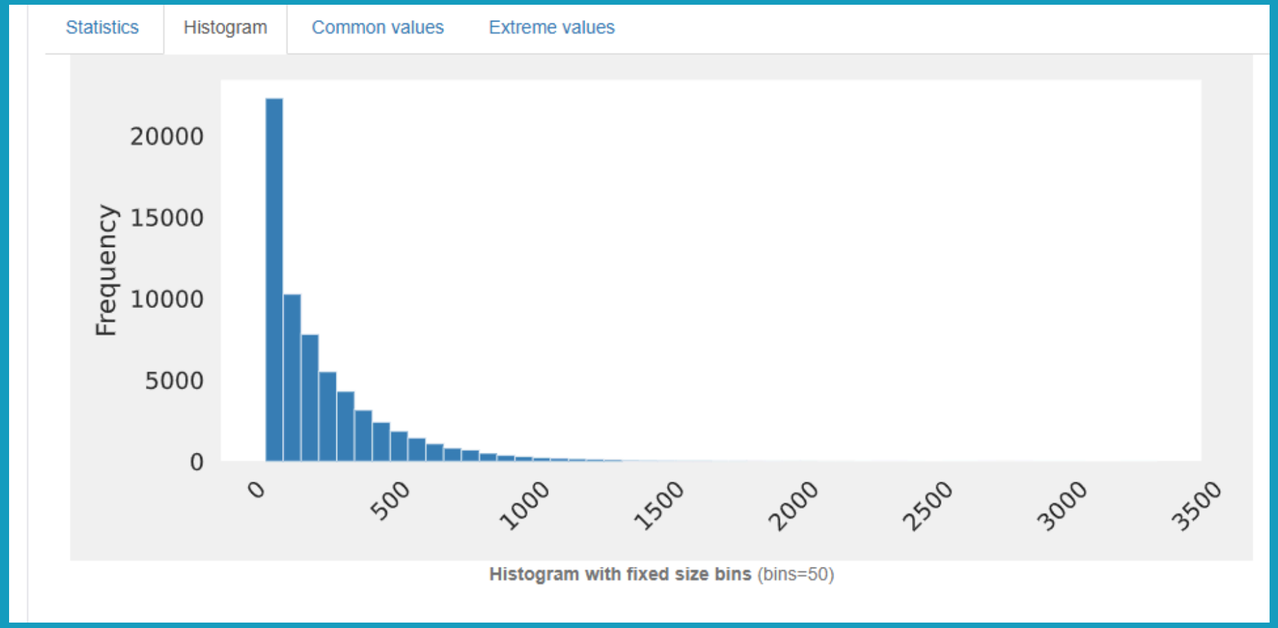
Uplift Modelling .ipynb

RECENCY DISTRIBUTION



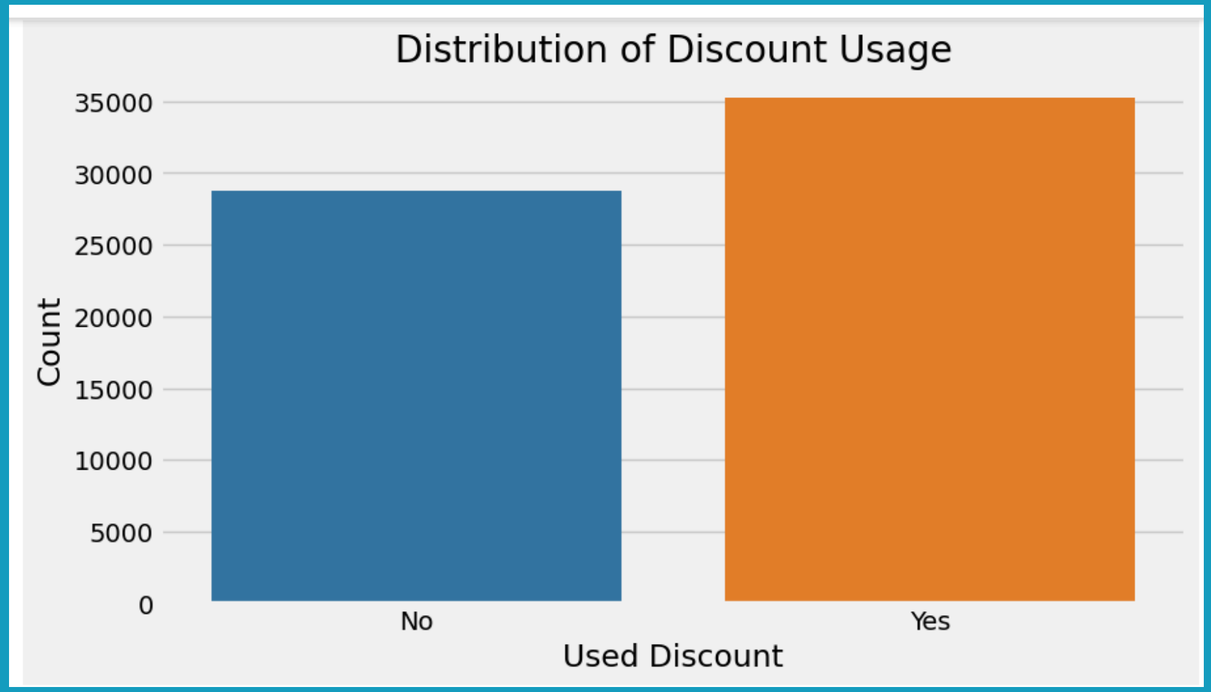
'Recency' refers to the time elapsed since a customer last made a purchase, with smaller numbers indicating more recent activity. The distribution data shows that the largest number of customers (8,952) made their last purchase just one day ago, indicating high engagement. The number of customers decreases steadily up to day 8, reflecting a decline in activity over time. However, there are notable increases on day 9 (6,441) and day 10 (7,565), which could be attributed to successful promotions or reminders. Beyond day 10, the number of customers drops again, suggesting a need for retention strategies to maintain customer engagement.

TRANSACTION HISTORY



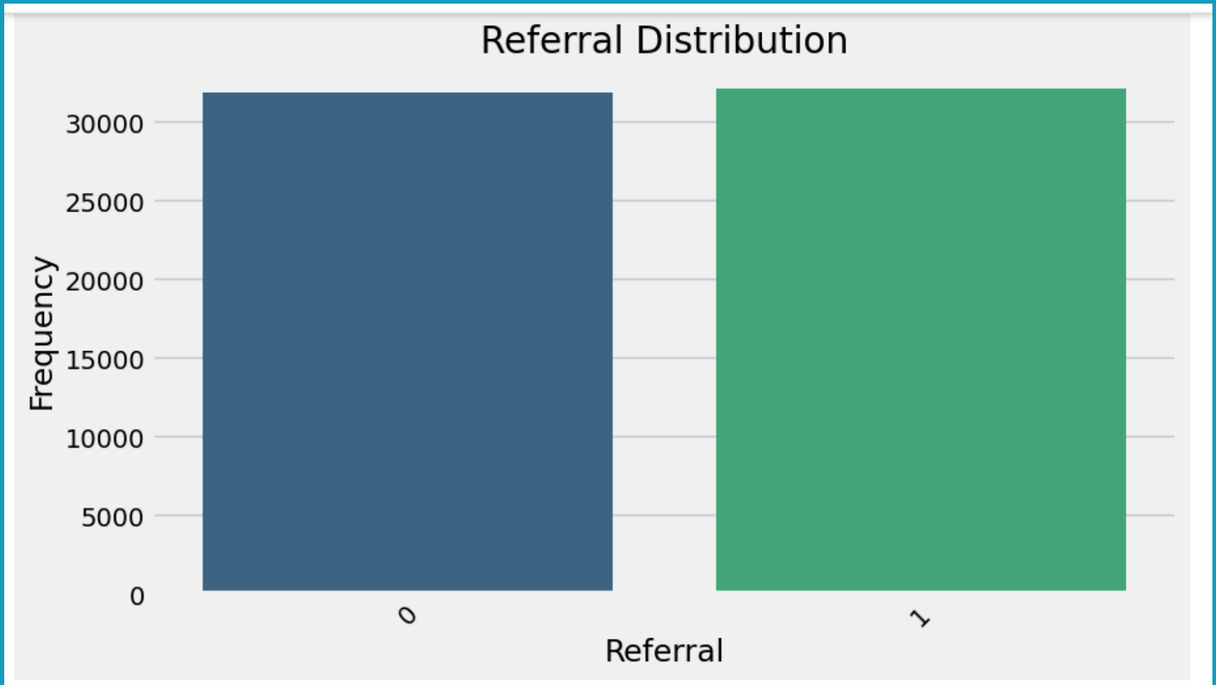
The 'History' data set shows substantial variability and is heavily right-skewed, with a few very high values significantly influencing the mean. The high kurtosis indicates that there are more outliers than would be expected in a normal distribution. Given these characteristics, it's essential to consider the impact of these extreme values when analyzing and interpreting the data.

DISCOUNT DISTRIBUTION



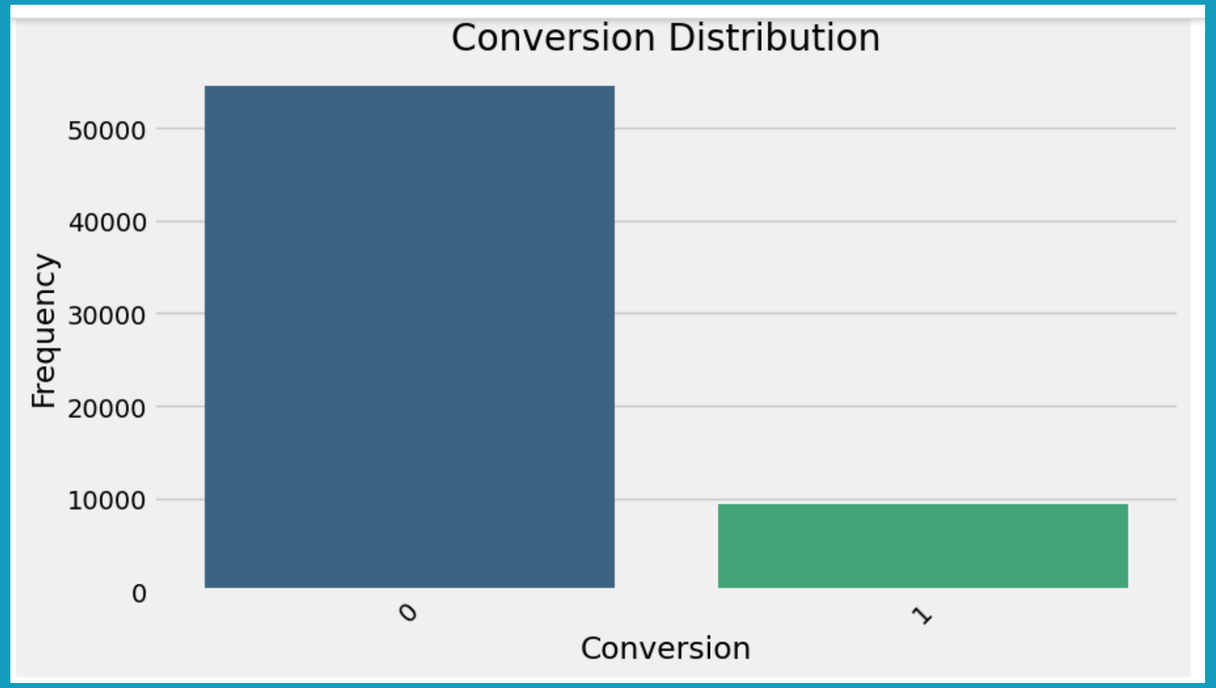
The data shows that more customers were given discounts compared to those who were not. Specifically, 35,266 customers received discounts, while 28,734 did not. This higher number of discounted customers suggests that customers tend to seek more value and prefer to save costs.

REFERRAL DISTRIBUTION



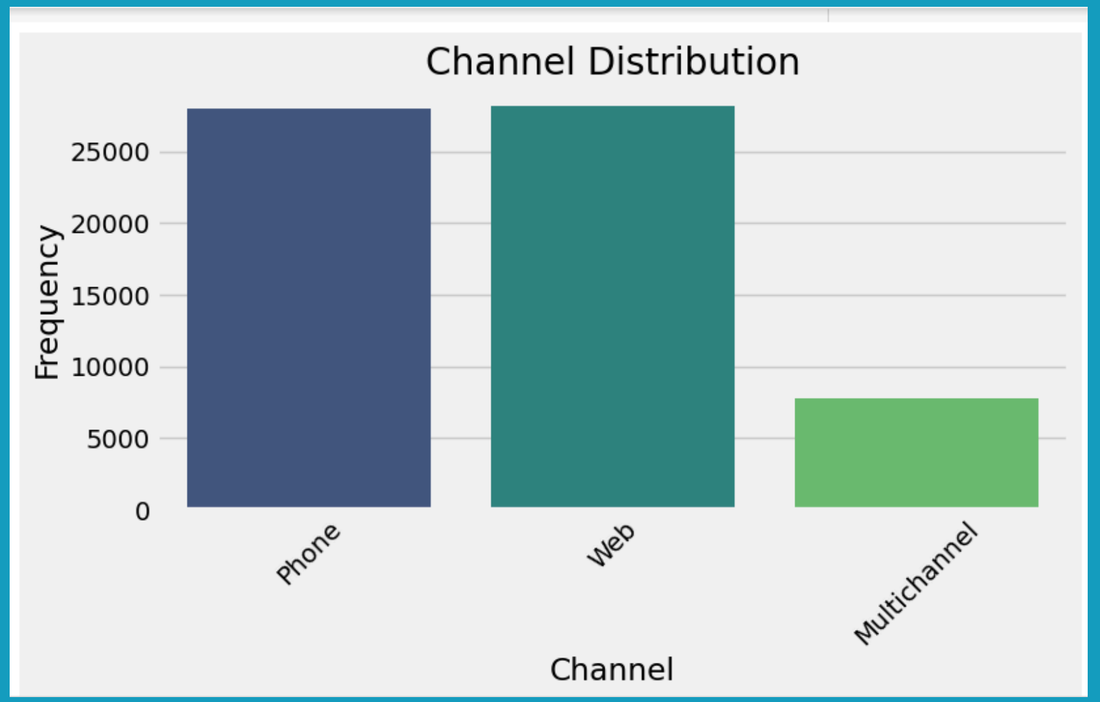
The data reveals that the number of customers acquired through referral channels is nearly equal to those acquired through non-referral channels, with 32,144 customers coming from referrals and 31,856 from other sources. This indicates that referral channels are highly effective in attracting customers. Additionally, customers who come through referrals tend to be more loyal and have a higher Customer Lifetime Value (CLV).

CONVERSION DISTRIBUTION



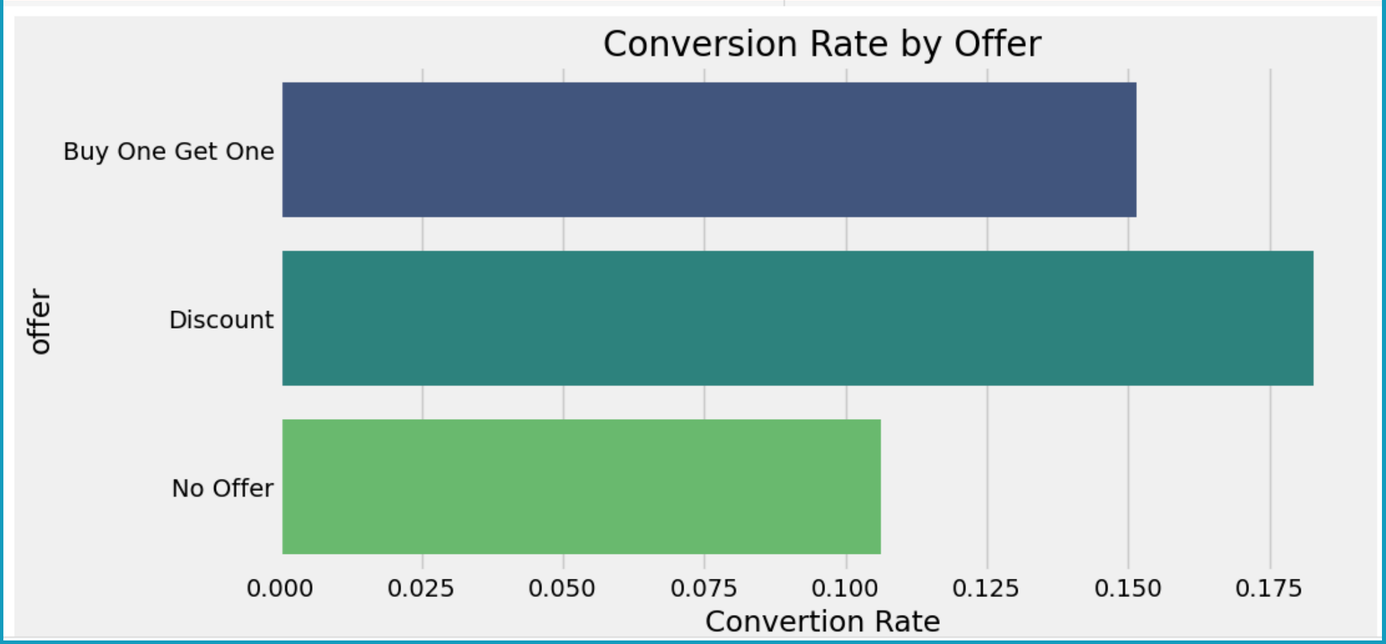
The number of customers who successfully converted to purchases after receiving promotional treatments, such as discounts and BOGO (Buy One Get One) offers, was relatively low, with only 9,394 customers making purchases. In contrast, there were 54,606 customers who did not make a purchase. These findings suggest that companies need to develop effective strategies to better target potential customers for discount or BOGO promotions.

CHANNEL DISTRIBUTION



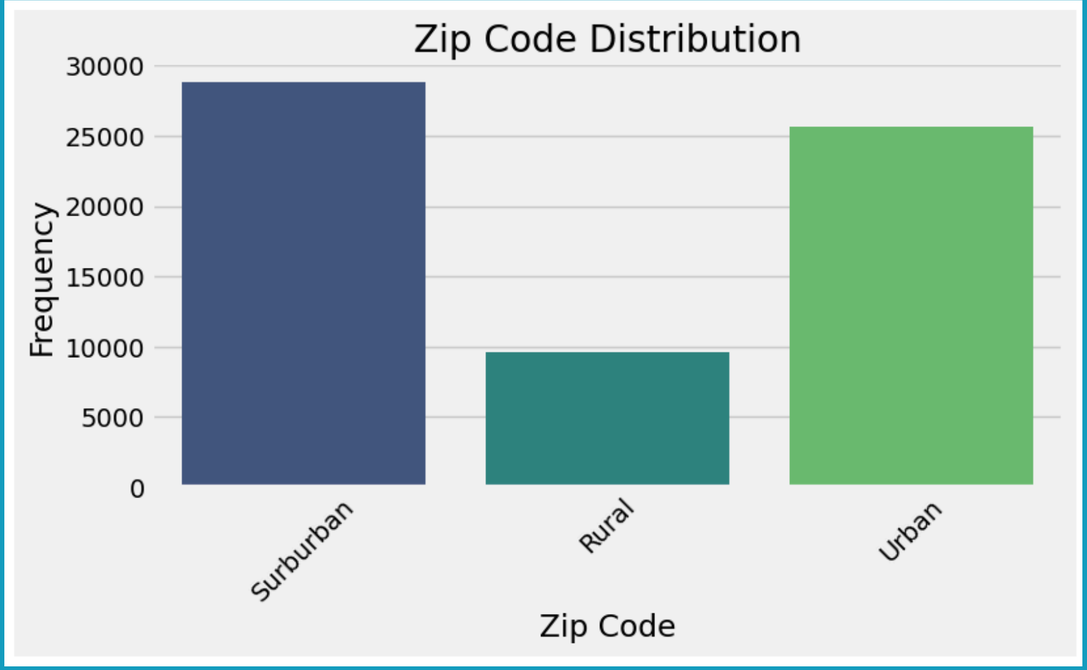
Customers primarily utilize web and phone channels, with only a minority engaging in multichannel interactions. This suggests that companies could prioritize efforts on web and phone channels for customer attraction, though further examination is needed to determine whether multichannel interactions have a significant impact on conversion rates.

CHART CONVERSION RATE BY OFFER



The bar chart shows the effectiveness of different promotional offers on customer conversion rates:
Buy One Get One (BOGO): Highest conversion rate at 15%.
Discount: Conversion rate is 17,5%.
No Offer: Lowest conversion rate at 10%.
Promotional offers, especially Discount, significantly increase the likelihood of customers making a purchase compared to no promotions.

ZIP CODE DISTRIBUTION



With regards to zip code distribution, there are 28,776 customers categorized as Suburban, followed by 25,661 customers classified as Urban, and 9,563 customers classified as Rural. These findings suggest that marketing efforts could be targeted towards the two classes with the highest customer counts, namely Suburban and Rural areas.

SPLITING DATA

Split Data

```
[ ] X_train, X_test = train_test_split(
        df.drop(variable_to_drop, axis = 1),
        test_size = 0.5,
        random_state = 42
    )
```

```
X_train.head()
```

	recency	history	zip_code	channel	offer	conversion
36841	10	29.99	Urban	Web	Discount	0
42427	10	662.03	Surburban	Web	Discount	1
35161	10	197.71	Surburban	Web	Discount	0
13578	7	29.99	Urban	Web	Discount	0
14909	8	299.38	Surburban	Web	No Offer	0

```
[ ] X_test.head()
```

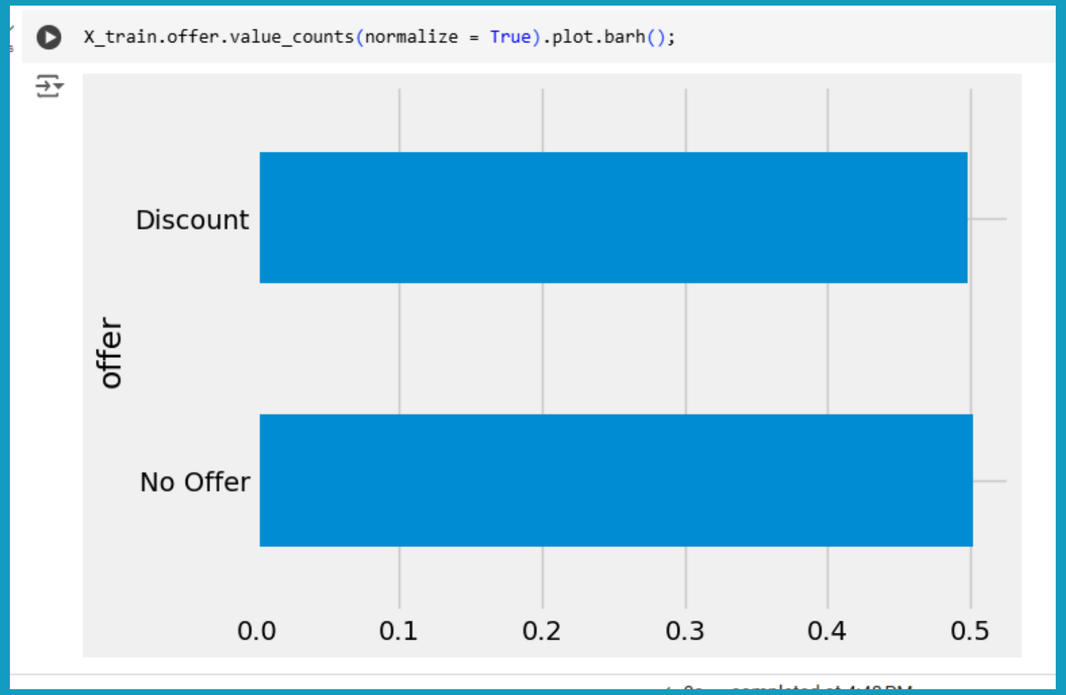
	recency	history	zip_code	channel	offer	conversion
322	4	121.87	Rural	Web	No Offer	0
14339	3	184.83	Urban	Phone	Discount	0
2348	2	400.48	Urban	Multichannel	No Offer	0
33454	10	175.25	Rural	Web	Discount	1
25505	2	1640.42	Rural	Multichannel	Discount	0

This code is splitting a dataset `df` into training and testing sets using the `train_test_split` function from a machine learning library.

- **`X_train`, `X_test`**: These variables will hold the features (independent variables) of the training and testing sets, respectively.
- **`train_test_split`**: This function splits the dataset into random train and test subsets.
- **`df.drop(variable_to_drop, axis=1)`**: This drops the specified variable (or column) from the dataset along the columns axis.
- **`test_size=0.5`**: This parameter specifies that 50% of the data will be used for testing, and the rest for training.
- **`random_state=42`**: This sets a seed for random number generation to ensure reproducibility.

Splitting data before preprocessing helps to prevent data leakage, where information from the testing set unintentionally influences preprocessing decisions. By separating the data first, preprocessing steps are applied only to the training data to avoid biasing the model evaluation on the testing set.

DATA PREPROCESSING



This horizontal bar chart visually represents the distribution of values within the "offer" column in the training dataset (X_train). It offers clarity on the proportional representation of each category within the "offer" column, aiding in understanding the relative frequencies of each category.

DATA PREPROCESSING

```
[46] # Encode categorical variables on X train data
     dummies = pd.get_dummies(X_train[categorical_columns], drop_first=True)
     X_train = pd.concat([X_train.drop(categorical_columns, axis=1), dummies], axis=1)

     # Encode categorical variables on X test data
     dummies = pd.get_dummies(X_test[categorical_columns], drop_first=True)
     X_test = pd.concat([X_test.drop(categorical_columns, axis=1), dummies], axis=1)
```

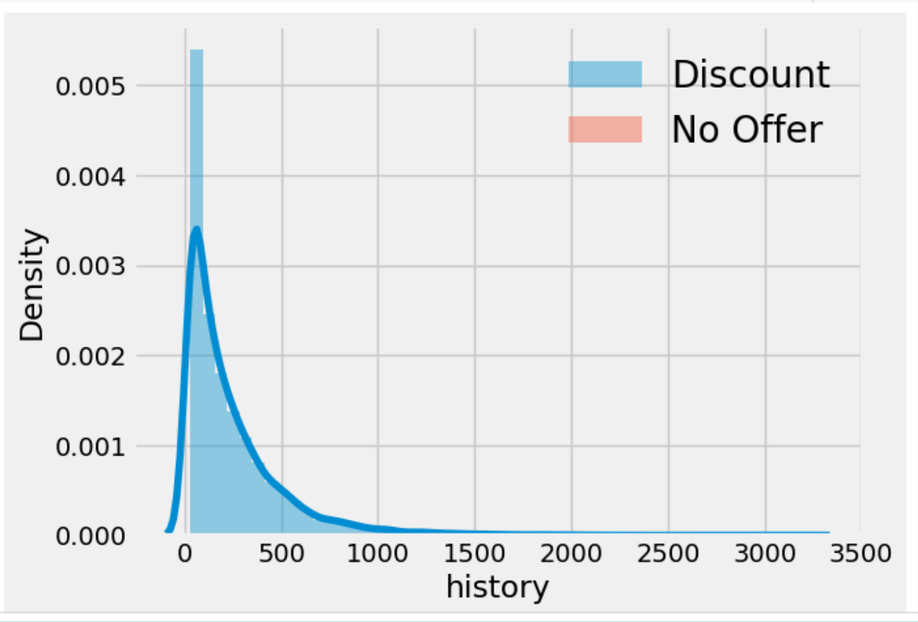
```
X_train
```

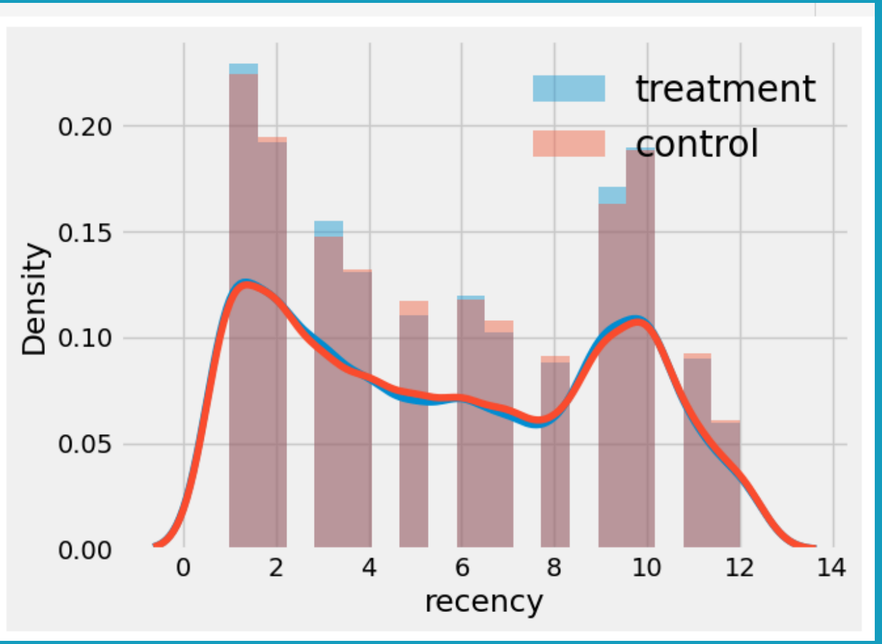
	recency	history	offer	conversion	zip_code_Surburban	zip_code_Urban	channel_Phone	channel_Web
36841	10	29.99	Discount	0	False	True	False	True
42427	10	662.03	Discount	1	True	False	False	True
35161	10	197.71	Discount	0	True	False	False	True
13578	7	29.99	Discount	0	False	True	False	True
14909	8	299.38	No Offer	0	True	False	False	True
...	...	...	...	...	...	...	...	...
6265	2	205.12	No Offer	0	False	True	False	True
11284	10	37.04	No Offer	0	False	True	False	True
38158	4	80.23	Discount	1	False	False	False	True

Encoding:

One-hot encoding is necessary because many machine learning algorithms cannot directly handle categorical data. By converting categorical variables into a binary format (0s and 1s), one-hot encoding allows these algorithms to interpret categorical variables as numerical features, improving model performance and accuracy.

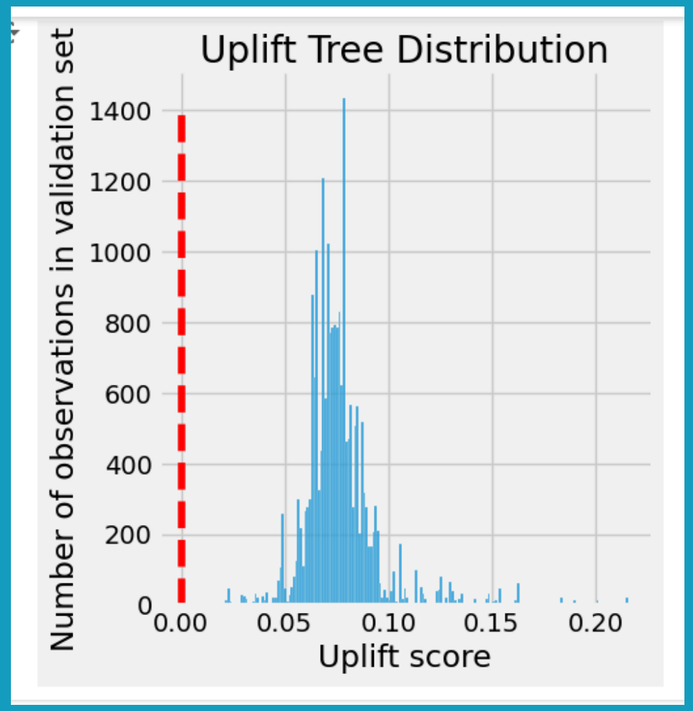
DATA PREPROCESSING





Visualize the distribution of purchase history between 'Discount' and 'No Offer' types. Additionally, visualize how the time since last purchase (recency) is distributed among the treatment and control groups.

DEVELOPING UPLIFT MODEL

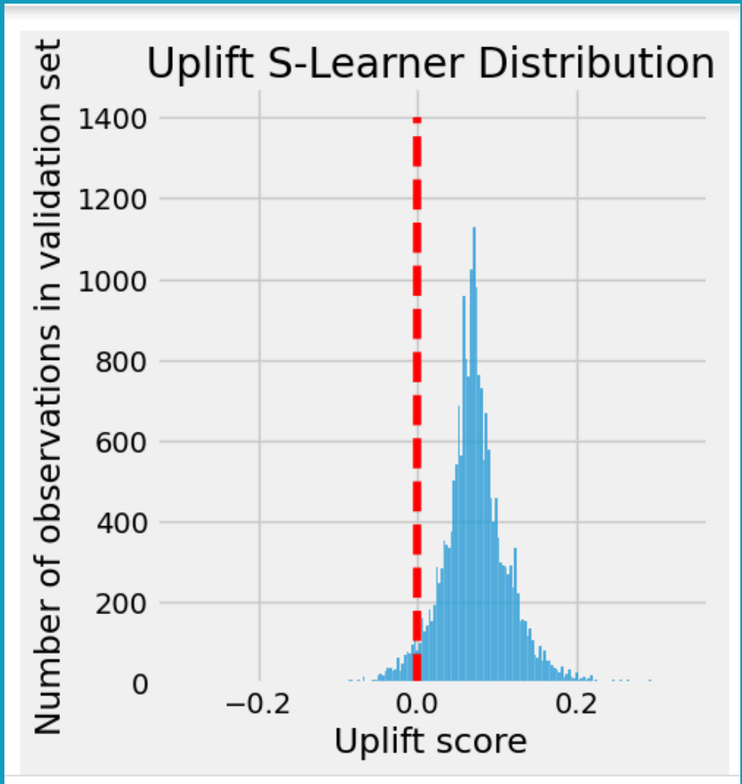


Analysis

1. Distribution Shape: The distribution of uplift scores is centered around a value slightly above 0, with most of the scores falling between 0.05 and 0.1. This suggests that the model generally predicts small positive treatment effects for most instances.
2. Concentration of Scores: There is a high concentration of scores around 0.07 to 0.09, indicating that a significant number of observations are predicted to have a modest positive uplift.
3. Few Negative Scores: There are very few observations with negative uplift scores (left of the red line), implying that the model predicts that most individuals will benefit, to some degree, from the treatment.
4. Spread of Scores: While the majority of scores are tightly clustered, there are some scores spread out up to 0.2, suggesting that for a small subset of the population, the model predicts a higher positive treatment effect.

In summary, the plot indicates that the uplift forest model predicts small but positive treatment effects for most individuals, with a strategy of targeting individuals with scores in the higher end of the distribution likely to yield better outcomes. Further model tuning or additional features might be needed to enhance the differentiation of treatment effects across the population.

DEVELOPING UPLIFT MODEL



Analysis

1. **Centering Around Zero**: The fact that most uplift scores are around 0 suggests that for a large number of observations, the treatment effect is minimal or neutral. This means that the intervention does not significantly change the outcome for these individuals.
2. **Positive Skew**: The right skewness indicates that there are more individuals with positive uplift scores than negative ones. This implies that the treatment has a beneficial effect for a larger portion of the population.
3. **Implications of Zero Uplift Score**: The red dashed line at 0 serves as a reference point. Observations to the right of this line (positive uplift scores) are those for which the treatment is predicted to have a positive impact. Conversely, observations to the left (negative uplift scores) are predicted to be negatively impacted by the treatment or to benefit more from no treatment at all.

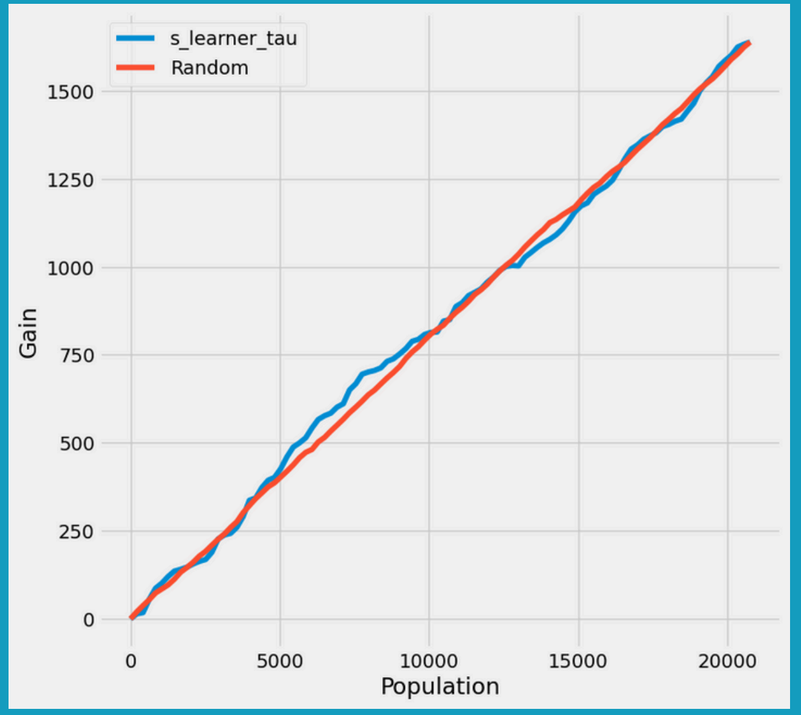
The distribution of uplift scores from the S-Learner model indicates that the treatment has a varying impact across the population, with a central tendency around zero and a positive skew. This information can be leveraged to target interventions more effectively, focusing on individuals who are most likely to benefit, thereby optimizing the overall effectiveness of the treatment. Further refinement and analysis of the model can enhance these insights and improve decision-making processes.

MODEL EVALUATION

Cumulative Gain Plot

Interpretation of the Graph

- **Blue Line (s_learner_tau)**: This line represents the gain curve for the S-learner model. It shows how much gain is achieved as you target more of the population according to the model's predictions.
- **Red Line (Random)**: This line represents the gain curve for a random targeting strategy. It serves as a baseline to compare the model against. If the model is effective, the gain curve should be above this random line.



Short Explanation of the Graphic Result

- **Initial Gain**: At the beginning (left side of the graph), the blue line (S-learner model) rises more sharply than the red line, indicating that the model is initially targeting individuals who are more likely to benefit from the treatment.
- **Middle and End Gain**: As the population increases, the blue line continues to stay slightly above the red line, indicating that the model consistently performs better than random targeting across the population.
- **Overall Performance**: The S-learner model (blue line) shows a higher cumulative gain compared to random targeting (red line), demonstrating that the model effectively identifies and targets individuals who are more positively impacted by the treatment.

In summary, the gain curve indicates that the S-learner uplift model is effective in identifying the right individuals for the treatment, leading to higher cumulative gains compared to a random strategy.

MODEL EVALUATION

Cumulative Gain Plot

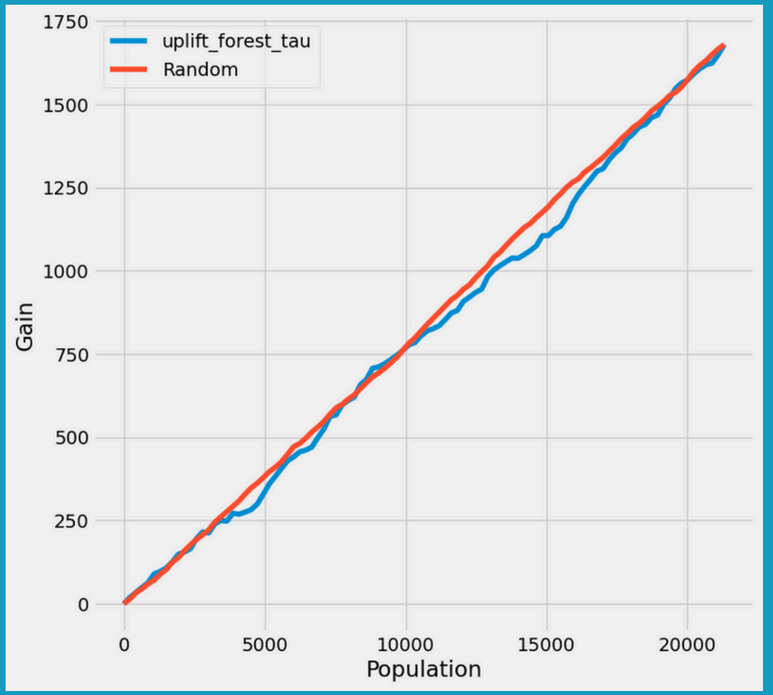
Interpretation of the Graph

- **Blue Line (uplift_forest_tau)**: This line represents the gain curve for the uplift forest model. It shows the cumulative gain achieved as more of the population is targeted according to the model's predictions.
- **Red Line (Random)**: This line represents the gain curve for a random targeting strategy, serving as a baseline for comparison.

Short Explanation of the Graphic Result

- **Initial Gain**: At the beginning (left side of the graph), the blue line (uplift forest model) initially stays very close to the red line, indicating that the model is not significantly better than random targeting for the first part of the population.
- **Middle and End Gain**: As the population increases, the blue line shows slight fluctuations but stays mostly close to or slightly above the red line. This suggests that the uplift forest model performs only marginally better than random targeting across the population.
- **Overall Performance**: The uplift forest model (blue line) shows a cumulative gain that is only slightly higher than the random targeting (red line). This indicates that the uplift forest model has limited effectiveness in identifying and targeting individuals who are more positively impacted by the treatment, compared to random selection.

In summary, the gain curve indicates that the uplift forest model provides only a slight improvement over random targeting. This suggests that the model may need further tuning or that the uplift signal in the data is weak, leading to performance close to random.



MODEL EVALUATION

Model	AUUC Scorre	Random Value AUUC	Qini Score	Random Value Qini Score
S-Learner	0.506463	0.501969	0.019298	0.000000
Uplift Tree	0.48274	0.49702	-0.0097	0.0000

AUUC & Qini Score

The AUUC (Area Under the Uplift Curve) score measures the performance of the model in predicting uplift, with higher values indicating better performance. The Qini score measures the difference in cumulative uplift between the treatment and control groups, with positive values indicating the model's effectiveness in targeting the treatment group.

MODEL EVALUATION

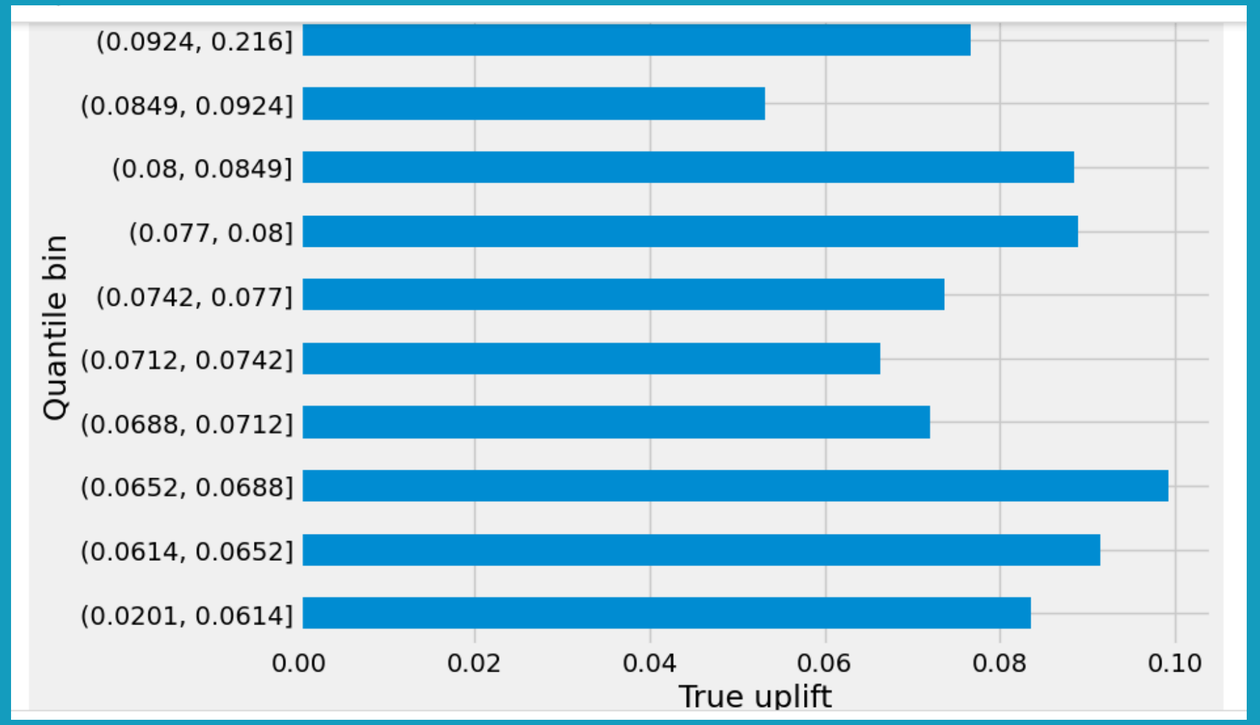
Interpretation Of Uplift Quantil Chart

- **Higher True Uplift in Higher Quantiles**:

Generally, the bars are longer for higher quantile bins, indicating that the higher the predicted uplift score, the higher the actual measured uplift. This suggests that the model is effective in identifying individuals who are more likely to benefit from the treatment.

- **Model Validation**:

This chart validates the model's predictions, as there is a positive correlation between the predicted uplift scores and the true uplift. The bins with higher predicted uplift scores show higher true uplift values.

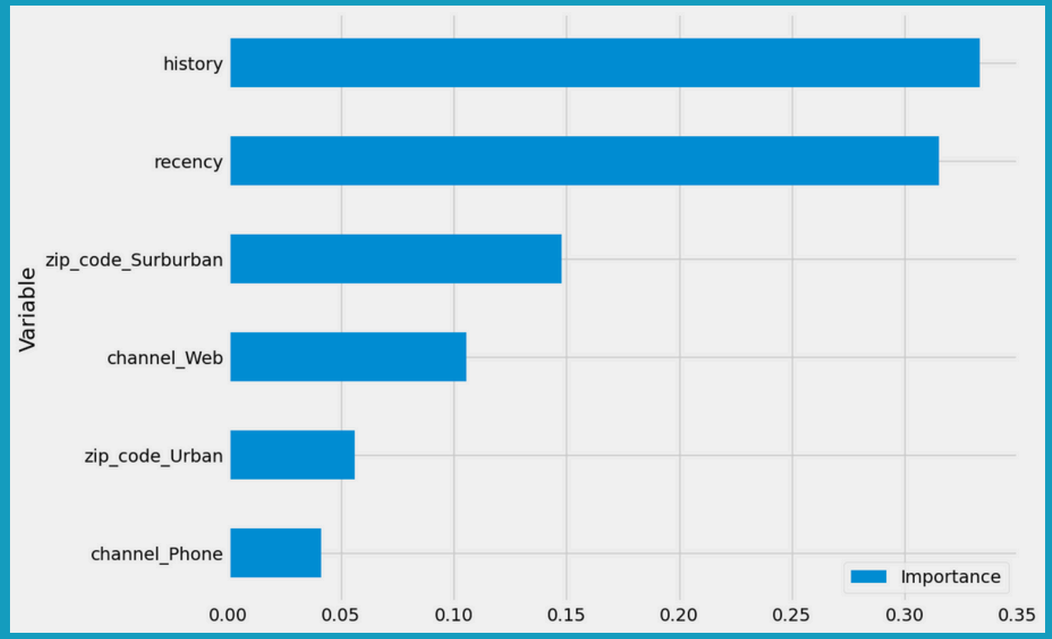


Effectiveness of Targeting:

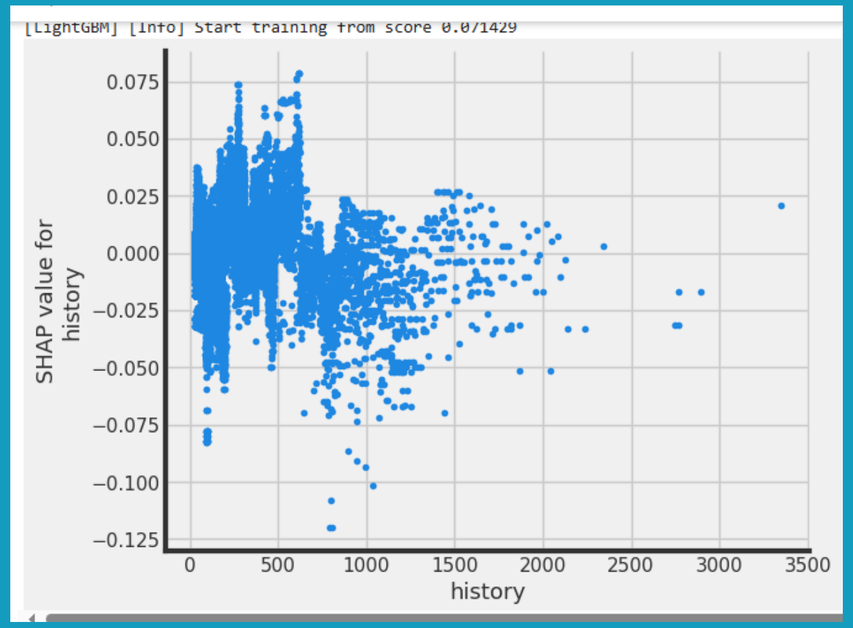
- The model's with Discount treatment ability to correctly predict higher true uplift for higher quantiles indicates it can be used effectively for targeting. Resources can be allocated to individuals in the higher quantile bins to maximize the treatment's impact.

MODEL EVALUATION

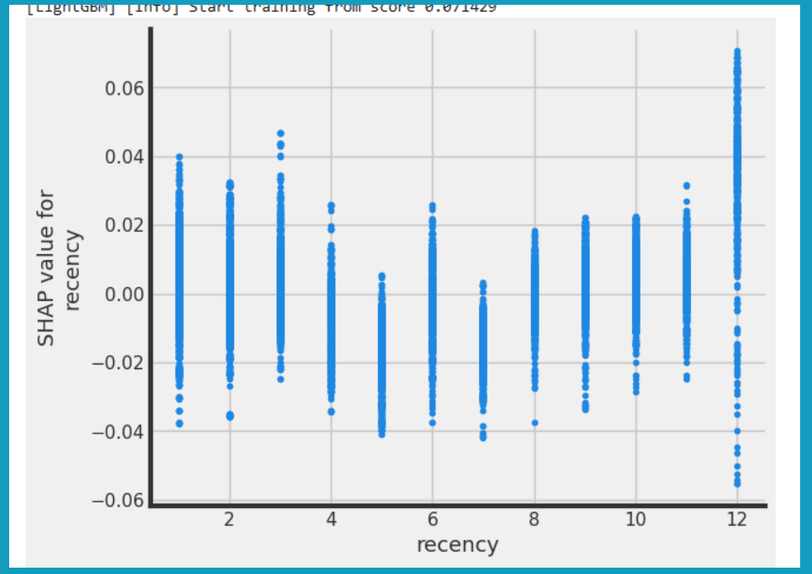
1. **Strategic Focus**:
   - **History and Recency**: The high importance of these features suggests focusing on customers' past behaviors and recent interactions can significantly improve targeting strategies.
   - **Geographical Targeting**: The importance of suburban and urban zip codes indicates that geographical segmentation is useful for predicting uplift.
   - **Channel Preferences**: The significance of interaction channels (web and phone) highlights the need to consider how customers engage with the service or product.
2. **Model Optimization**:
   - The model can be further optimized by ensuring that the most important features are accurately captured and utilized in the analysis.
   - Lesser important features might be refined or combined with other features to see if their importance increases.



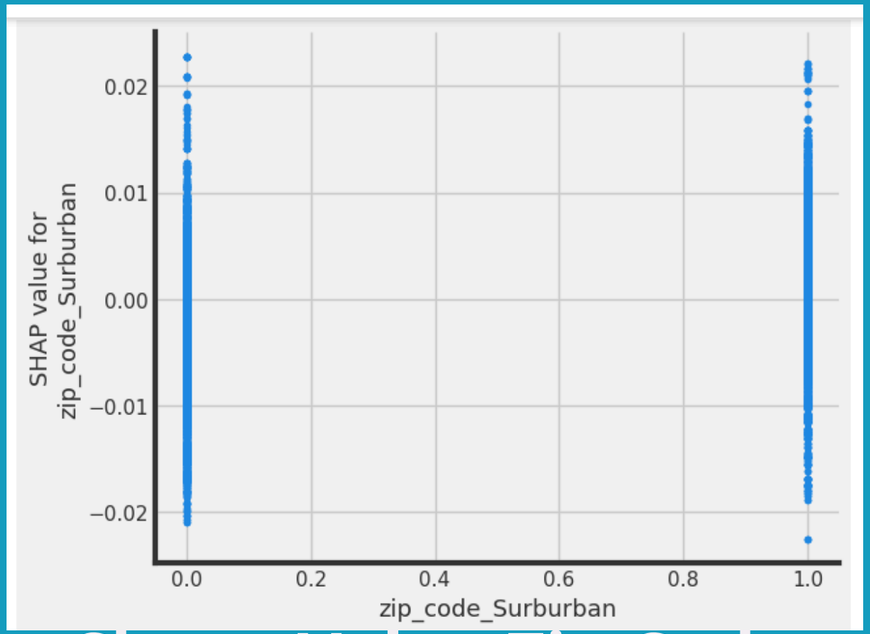
MODEL EVALUATION



Shape Value History



Shape Value Recency



Shape Value Zip Code

CONCLUSION

The uplift modeling project successfully addressed the business problem by enabling more strategic and effective targeting of marketing campaigns. Key findings include:

1. **Effective Targeting**:
   - By focusing on customers with higher uplift scores, the company can significantly enhance the efficiency of its marketing efforts. This targeted approach ensures that resources are directed toward customers who are most likely to respond positively, thereby maximizing ROI.
2. **Insightful Segmentation**:
   - The identification of crucial features such as purchase history and recency provides actionable insights for customer segmentation. This allows the company to refine its marketing strategies and develop more personalized and impactful campaigns.
3. **Optimization Potential**:
   - The model's validation through true uplift analysis confirms its reliability and provides a solid foundation for future optimization. Continuous refinement of the model and features can further enhance its predictive power and business value.

Overall, the project demonstrates the potential of uplift modeling to revolutionize marketing strategies by providing a deeper understanding of customer behavior and enabling more precise and effective targeting of interventions.

THANK YOU

Uplift Modelling .ipynb