**Project: Super store marketing campaign**

**Team-Group18**

**Team Names:**

**Anil Nimma (11597397)**

**Hari Krishna Thunga (11616021)**

**Sowmya Kondeti (11596944)**

**Lawrence Abhinav Gunkonda (11528114)**

**Table of Contents:**

1. Introduction
2. Abstract
3. Methodology
4. Results and Analysis
5. Discussion and Implications
6. Conclusion
7. References

**Introduction**

Introduction: The superstore industry is highly competitive, and businesses are constantly seeking new ways to retain their existing customers while attracting new ones. One such method is by offering exclusive discounts and membership programs. In this case, the superstore is planning to launch a gold membership offer that gives a 20% discount on all purchases, exclusively for existing customers. To promote this offer, the management plans to conduct a phone call campaign to their customers. However, they also want to reduce the cost of this campaign by developing a predictive model to identify which customers are likely to purchase the offer.

**Abstract**

This project aims to develop a predictive model for a superstore's year-end sale gold membership offer. The objective is to predict the likelihood of a customer giving a positive response and identify the factors that affect the customer's response. The data provided will be analyzed to identify these factors, and a prediction model will be built to estimate the probability of a customer giving a positive response. By accurately predicting which customers are more likely to purchase the offer, the superstore can target its marketing campaign more efficiently, ultimately increasing its sales and revenue.

**Methodology**

**Data Plan and Solution:**

1. **Data collection:**

We will collect the data from the superstore's previous year's campaign (Kaggle Dataset). The data contains information about the customers and demographics, their purchases, and their response to the campaign. We will use this data to build a predictive model for the upcoming campaign.

1. **Data Cleaning and Preparation:**

We will clean the data by removing missing values and outliers. We will also transform the data by converting categorical variables into numerical variables using one-hot encoding. We will split the data into training and testing sets to evaluate the performance of the model.

1. **Exploratory Data Analysis:**

We will conduct exploratory data analysis to understand the relationships between the variables and the target variable. We will use statistical techniques such as correlation analysis and hypothesis testing to identify the variables that are most strongly associated with the target variable.

1. **Feature Selection:**

We will use feature selection techniques such as Principal Component Analysis (PCA) to identify the most important variables that are likely to influence the customer's response to the offer.

1. **Model Selection and Evaluation :**

Once we have identified the most relevant features, we will build a predictive model using machine learning algorithms. We will test multiple classification algorithms, such as logistic regression, decision tree, random forest, and gradient boosting, to identify the best model based on performance metrics like accuracy, precision, recall, and F1 score.

1. **Evaluation:**

We will evaluate the performance of the predictive model using various metrics like accuracy, precision, recall, and F1 score. We will also use confusion matrix, ROC curve, and AUC score to visualize the model's performance and identify any issues like over fitting or under fitting. We will use cross-validation to ensure that the model's performance is robust and unbiased.

1. **Conclusion:**

The final deliverable of this project will be a predictive model that estimates the probability of a customer giving a positive response to the gold membership offer. The superstore can use this model to identify the customers who are likely to purchase the offer and target them with the phone campaign. This will reduce the cost of the campaign.

**Data collection**

Data has been collected from the Kaggle website. We collected customers transaction data, demographic data for this predictive analysis.

**Dependent variable/Target variable:** Response (target) - 1 if customer accepted the offer in the last campaign, 0 otherwise

**Independent variables:**ID - Unique ID of each customer  
Year\_Birth - Age of the customer  
Complain- 1 if the customer complained in the last 2 years  
Dt\_Customer - date of customer's enrolment with the company  
Education - customer's level of education  
Marital - customer's marital status whether the customer is married or not   
Kidhome - number of small children in customer's household  
Teenhome - number of teenagers in customer's household  
Income - customer's yearly household income  
MntFishProducts - the amount spent on fish products in the last 2 years  
MntMeatProducts - the amount spent on meat products in the last 2 years  
MntFruits - the amount spent on fruits products in the last 2 years  
MntSweetProducts - amount spent on sweet products in the last 2 years  
MntWines - the amount spent on wine products in the last 2 years  
MntGoldProds - the amount spent on gold products in the last 2 years  
NumDealsPurchases - number of purchases made with discount  
NumCatalogPurchases - number of purchases made using catalog (buying goods to be shipped through the mail)  
NumStorePurchases - number of purchases made directly in stores  
NumWebPurchases - number of purchases made through the company's website  
NumWebVisitsMonth - number of visits to company's website in the last month  
Recency - number of days since the last purchase

**Data cleaning and Preparation**

In this project, the data was collected from the customer database of the superstore. The data includes 22 variables and 2240 observations. The target variable in this analysis is the customer's response to the offer, which is categorized as 0 for no response and 1 for a positive response.

The data was checked for quality using the CRISP-DM process. The initial shape of the data was (2240 X 22), which was then modified as follows:

Age column was extracted, resulting in (2240 X 23) shape

NA values were removed, resulting in (2216 X 23) shape

Outliers in the income variable were treated, resulting in (2208 X 23) shape.

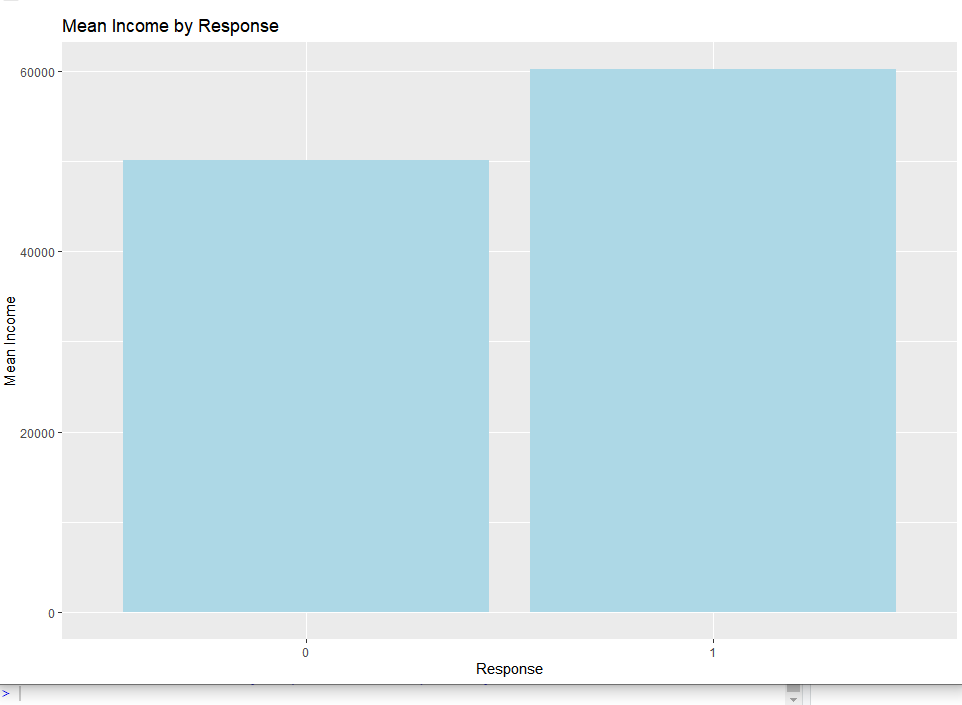
The final dataset used for analysis contained 2208 observations and 23 variables, including the target variable.

**Exploratory data analysis**

1. 15 % of the customers are target group and 85 % customers are from control group.
2. Most of the Customers are Graduates.
3. Average Purchase of Control group is 14.4 which is less compared to target group 17.7

Note: Clearly we can see the Campaign is effective in above case Average purchase.

**Mean Income of Target Group and Controlled Group**



The mean Income of control group is 50110.53

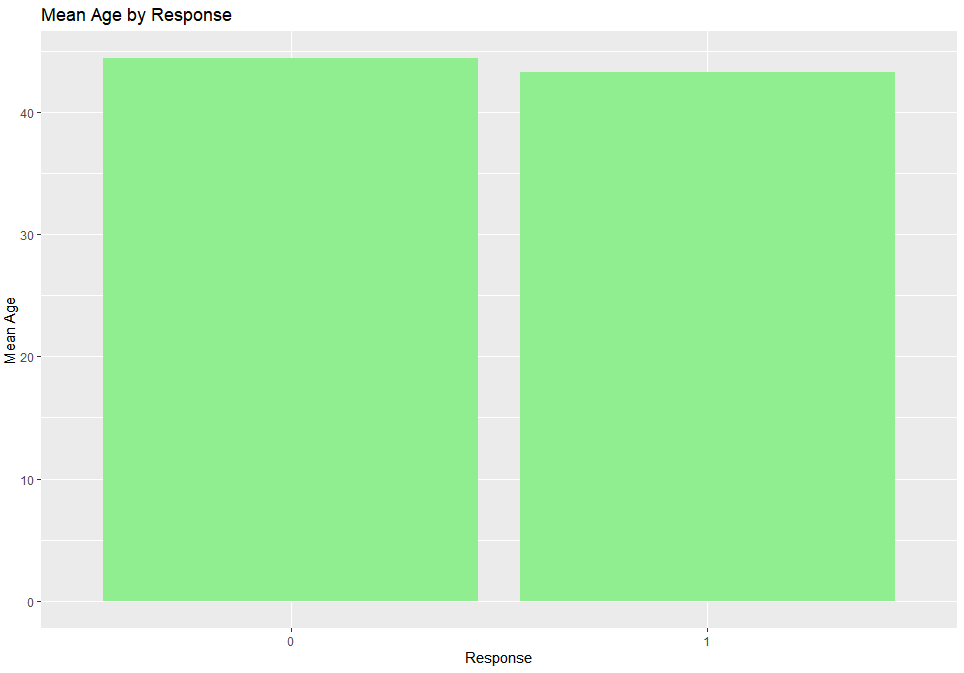
The mean Income of Target group is 60209.68

The mean income of the control group is lower than the mean income of the target group. This indicates that customers who responded to the offer have a higher income level than those who did not respond. This information is crucial for the superstore as it suggests that the campaign is making an impact on revenue. Customers who responded to the offer and purchased products at a discounted price have a higher income, which implies that they have the purchasing power to buy more products in the future. This could lead to an increase in revenue for the superstore.

**Mean Age by Response:**

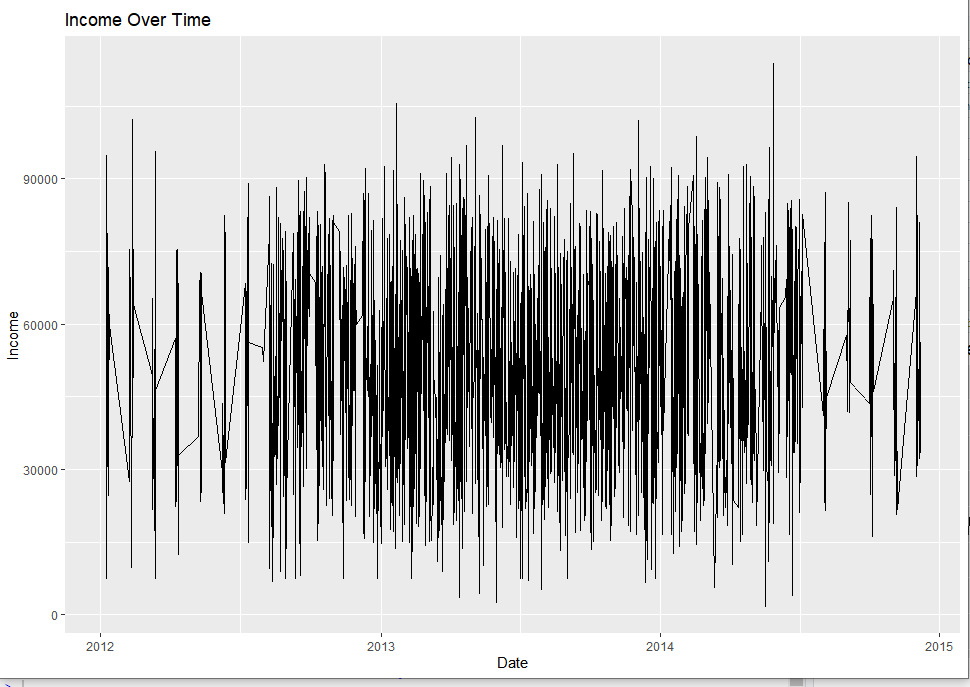
The mean age of controlled group is 44 Years

The mean age of Target Group is 43 Years

****

The mean age of the controlled group (customers who did not respond to the offer) is 44 years and the mean age of the target group (customers who responded to the offer) is 43 years. This difference in mean age is relatively small and may not be significant. However, it is important to study the age distribution of the two groups more closely to see if there are any particular age ranges that are more likely to respond to the offer.

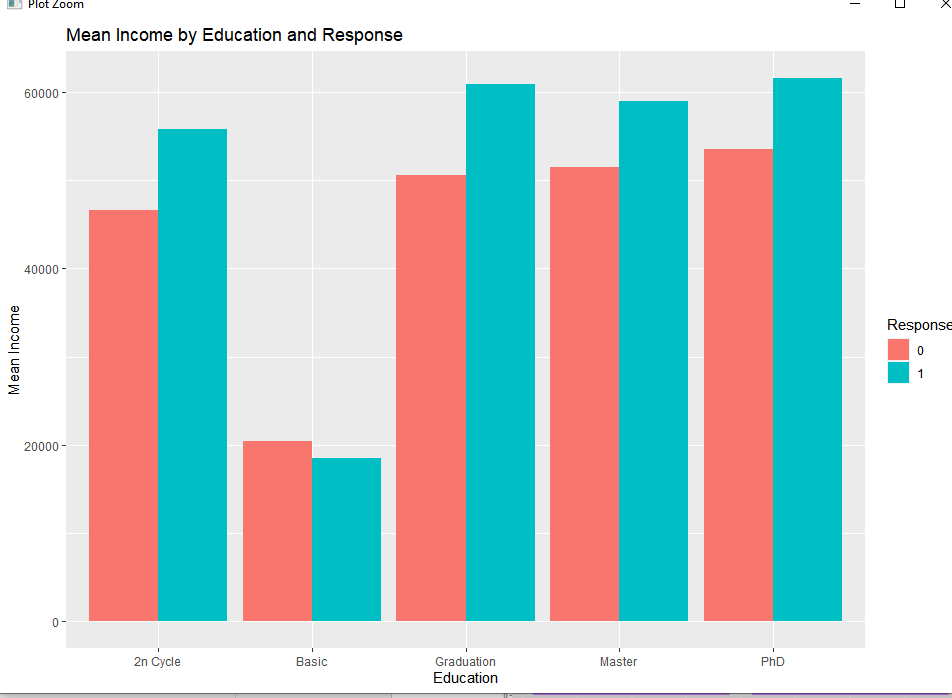
**The line plot for income at different tran\_dates(dt\_customer):**

****

Most of the Transactions are take place in the year 2013 and 2014.

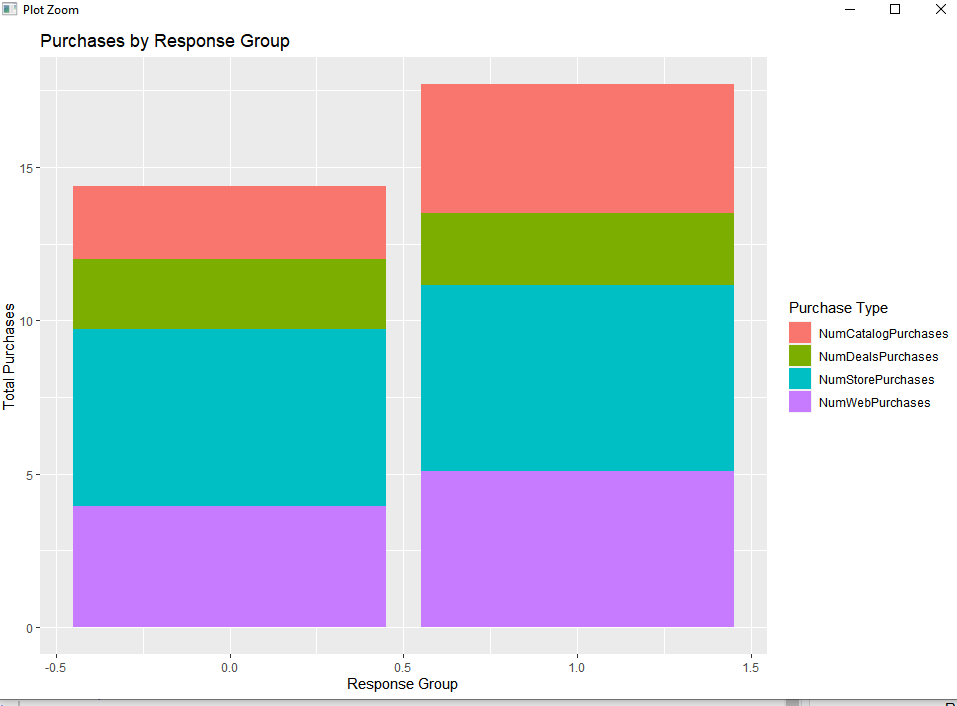
And also we can see there is volatility in revenue or Income.

**Mean Income based on education:**



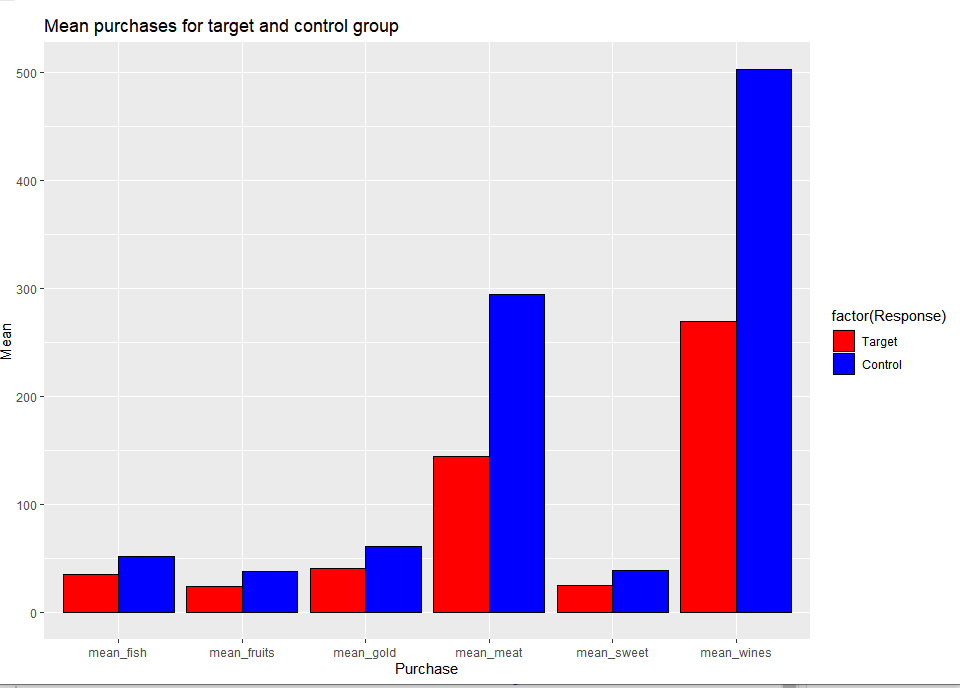
The Graduates and PhD having higher income or revenue.

**Purchase by Response:**

****

The mean Purchase were Consider for different purchase type clearly we can see the customers are more interested to buy from stores and next to that is web.

**Mean Purchase of different item types based on response variable:**

****

We can conclude that customers are more interested in purchasing wines and meat products. This can be seen in the fact that these two categories have the highest number of purchases and revenue generated. Additionally, the target group, which received the year-end sale gold membership offer, has a higher average revenue generated than the control group, despite purchasing fewer items overall.

These findings suggest that the marketing campaign has been successful in generating interest among customers and encouraging them to purchase more high-value products. However, it is important to note that the control group has made more purchases overall, indicating that there is still room for improvement in targeting potential customers who may be interested in purchasing these high-value products.

Based on analysis ,the graduates are purchasing more and the mean income of graduates and PhD are similar and more likely to respond to the campaign ,The age band 40-45 are more likely to respond to the campaign . The purchases in store and web are more so we can target these audience to gain more and increase revenue.

**Feature Engineering**

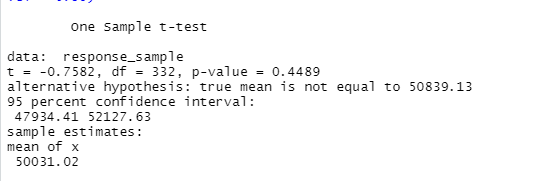
Extracted the age column from the year.

**Data Imbalance:** The original dataset was imbalanced, with only 333 observations in the target group and 1875 observations in the control group. To address this issue, we decided to sample the data. After sampling, both the target and control groups had 333 observations each. This allowed us to have balanced data for further analysis**.**

**Hypothesis Testing:** To check whether the sample data mean is deviating from the population mean, we conducted a hypothesis test.

The null hypothesis was that the sample mean was not deviating from the population mean, while the alternative hypothesis was that the sample mean was deviating from the population mean.

We used a two-sample t-test to test the hypothesis. We set the significance level at 0.05. The t-test resulted in a p-value of 0.4489, which was greater than the significance level. Therefore, we failed to reject the null hypothesis, which means that the sample mean and population mean were not significantly different.



As the p value is 0.4489 which is greater than 0.05 we are not able to reject null hypothesis.

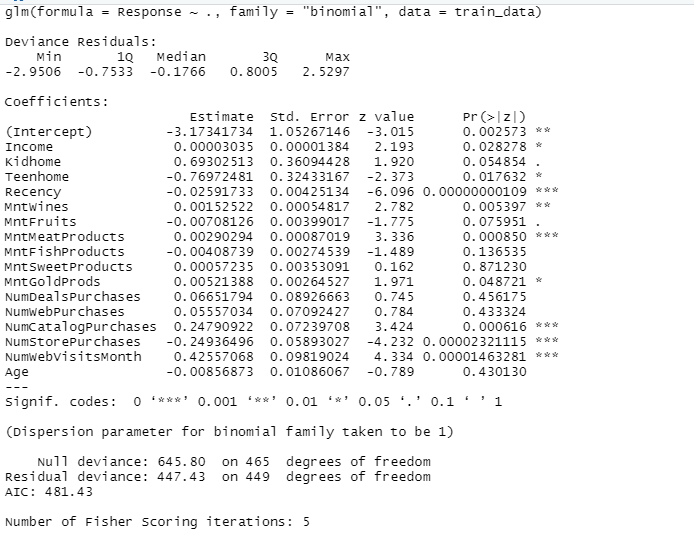
So clearly sample mean and population mean are equal.

**Model**

We have built a logistic regression model to predict whether a customer will fall into a target group or control group based on the given set of variables. We used a train-test split of 70:30 and evaluated the model on both train and validation data.

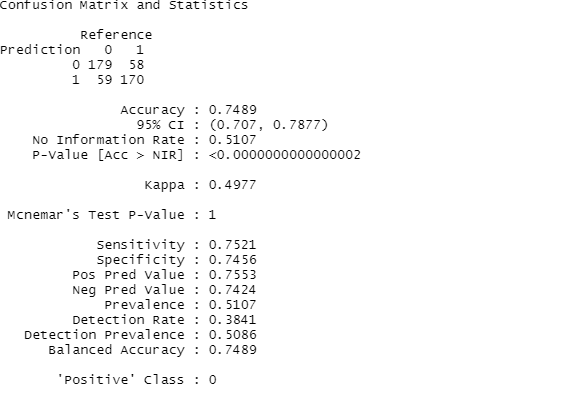
Model Building: We chose logistic regression as our modeling technique since the dependent variable is categorical in nature with two possible outcomes (0 and 1). We built the model using the train dataset and identified the most significant variables from the model summary.

Logistic Regression results:

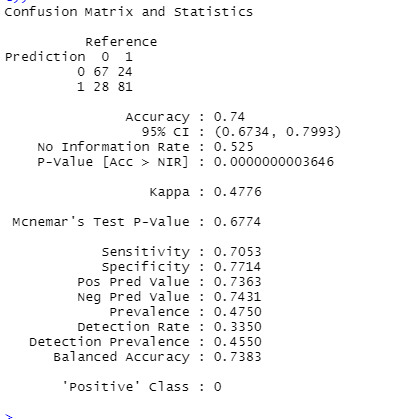


We can see the most significant variables from above list.12

On Train\_data:



On Valid\_data:



Model Evaluation: We evaluated the performance of the model on both the train and validation datasets. The model achieved an accuracy of 0.7489 on the train dataset and 0.74 on the validation dataset. The accuracy score indicates that the model has performed well in predicting the target and control groups.

|  |  |
| --- | --- |
| Data | Accuracy |
| Train\_data | 0.7489 |
| Valid\_data | 0.74 |

Model Performance: The model was able to predict 24 customers from the control group to fall into the target group, and 28 customers from the target group to fall into the control group. This could potentially lead to missed opportunities or incorrect targeting of customers.

Recommendations: To improve the accuracy of the model, we recommend increasing the amount of data used to train the model. Additionally, we suggest investigating the misclassification of customers from the control and target groups to understand the factors contributing to these misclassifications.

**Conclusion**

Overall, the logistic regression model performed well in predicting whether a customer will fall into the target or control group. However, there is room for improvement, and we recommend additional data and analysis to refine the model further.

Potentially out of 237 control group customers 58 (25%) customers may turn to target group if we aim marketing at the customers.

Out of 229 target customers 59 customers may churn and fall into control group we can aim marketing campaign to not to lose these customers.

**References**

<https://www.kaggle.com/datasets/ahsan81/superstore-marketing-campaign-dataset>

---------------------------------------- Thank You-------------------------------------------