INDIAN INSTITUTE OF TECHNOLOGY, KANPUR

MTH416A: REGRESSION ANALYSIS

A PROJECT REPORT ON

Marketing Campaign

Submitted by

PRITHWIJIT GHOSH - 211349
SHIVAM - 211381
SAPRATIVA BHOWAL - 211371
PRATIBHA VISHWAKARMA - 211346

Under the Guidance of

Dr. Sharmishtha Mitra

Department Of Mathematics And Statistics, IIT Kanpur

ABSTRACT

Making a huge success of a marketing campaign is one of the most important thing to the Marketing Company as this is the only way to have higher incentives among the employees. For the same reason the marketing campaign plays a major role in the marketing field. The importance of the area is due in part to the relevance for employees and investors of the company in evaluating the likelihood that the number of customers may decline in a regular interval. Therefore, in this study, we explore, build, and compare the different classification models. We begin by carrying out data preprocessing and exploratory analysis where we impute the missing data values after treating the outliers. To address the data imbalance issue, we apply Synthetic Minority Oversampling Technique (SMOTE) to oversample the minority class labels. Then, we build an appropriate model using Logistic Regression on our cleaned dataset and finally, analyze and evaluate the performance of the models on the validation datasets using several metrics such as accuracy, precision, recall, etc., and rank the models accordingly.

ACKNOWLEDGEMENT

As it is rightly said that the real learning comes from a practical work.

The success and final outcome of this assignment required a lot of guidance and assistance from many people and we are extremely fortunate to have got this all along with the completion of our project work. Whatever we have done in this project is only due to the wonderful guidance and assistance of our instructor Dr. Sharmishtha Mitra, Department of Statistics, IIT Kanpur for giving us this great opportunity to do the project on 'Marketing Campaign' and providing us all support and guidance which made us complete the project work on time. Without her valuable guidance and motivation, it was nearly impossible to work on this project as a team and understand the practical aspect of the course "MTH 416A: Regression Analysis".

Last but not the least we are grateful to all the faculty members and the seniors who constantly remained in touch with us and supported us at many stages.

Prithwijit Ghosh (211349)
Shivam (211381)
Saprativa Bhowal (211371)
Pratibha Vishwakarma (211346)

Contents

List of Tables

List of Figures

1 Introduction

1.1 Inroduction

The official definition of marketing is it is a philosophy whose main focus is providing customer satisfaction. Marketing campaigns promote products through different types of media, such as television, radio, print, and online platforms. Defining a campaign's goal usually dictates how much marketing is needed and what media are most effective for reaching a specific segment of consumers. Companies that lose sales due to major negative press often use marketing campaigns to rehabilitate their image.

Marketing campaigns can be designed with different goals in mind, including building a brand image, introducing a new product, increasing sales of a product already on the market, or even reducing the impact of negative news. Defining a campaign's goal usually dictates how much marketing is needed and what media are most effective for reaching a specific segment of the population.

There are many ways to market products and services to customers, from mailing brochures to coordinating a social media blitz. Small companies can email invitations to a special sale and offer a free product to every customer who brings the invitation. Larger companies can use paid advertising and professional agencies to reach a wider audience.

Whatever the size of the company, it's important that someone is dedicated to handling the influx of traffic a marketing campaign generates. If you are prompting customers to sign up for your email list, you must make sure that the list is managed well and that new customers receive welcoming messages. If visits to your website increase, you must continually update your content to convert this traffic to profitable sales.

Companies that lose sales due to major negative press often use marketing campaigns to rehabilitate their image. One example is Chipotle Mexican Grill, which was investigated by the Centers for Disease Control and Prevention after dozens of customers became sick in 2015 from food safety issues related to E. coli and norovirus.Chipotle's sales dropped 30 percent in the first quarter of 2016, and to regain customer interest, the company offered coupons for free food via direct mail and texts. Chipotle also used online video to announce a 10 million dollar grant to support local farmers.

Here in our data, we have the customer ID, date of birth, income, martial status, education level, number of children etc. so that we can give our customer the best possible outcome and can predict their needs.

The long-running Aflac duck campaign is one example of a campaign that significantly raised brand recognition. The company's brand-recognition rate was just 12 percent when it launched the campaign in 2000, and more than a decade of advertising boosted recognition to 90 percent.

1.2 Goal Of Our Study

By using our response variable i.e. the response given by the customer whether he/she accept or reject the offer in the last campaign, we want to fit a logistic model that actually predicts the the profit of the company for the next marketing campaign. In our model we follow the next steps respectively—

1. Encoding the categorical columns into corresponding **dummy variables**.
2. checking for **outliers** and if it exists then replace them by their suitable estimates.
3. Checking for **missing values** and replace them by their corresponding estimates
4. Examine whether the data is imbalanced or not.
5. Fit our **logistic model** with full set of regressors.

6. Scrutinize the model for multicollinearity issue in the continuous predictor.

7. Finally, select the variables which are necessary and drop all other.

2 Methodology

We already discussed about the data-set (**i.e. Marketing Campaign**) in somewhat subjectively. Now we will discuss about our plannings of the necessary steps from the very beginning to the bottom end. At first we will encode our categorical variables to dummy variables as necessary. Then we will check whether our data contains missing valus or not and also visualizing it by the sparsity matrix. Outliers will be detected (if any) for each of the regressors in the model. This outliers will be considered as missing values and estimated by the corresponding sample estimates. We will check whether our data is imbalanced or not and will take the necessary actions based on that. After fitting the model, multicollinearity issue will be detected very seriously.Then from the fitted model **Model Adequacy**(i.e.Specificity, Sensitivity,Independence,Chi-Squared value etc) will be checked by the confusion matrix and finally using some Machine Learning Models we will compare the quality of our logistic model for this given scenario.

2.1 Data Description

Our data contain the full information of a marketing campaign strategy.In this data-set the column response is our response or target variable,actually represents the whether the customer accept the offer(**i.e. 1**) or reject the offer(**i.e. 0**) in the last campaign.

Now Our regressor variables are both types i.e. categorical and continuous. In categorical columns Customer's education,marital status,complain and in which trial they accept the offer are included and simultaneously in the continuous columns number of children,number of teenager,yearly income,recency and money spend on the items like fruit ,sweet,wines,golds are included.

Now it is a very fruitful question that based on the information of customers can we predict something whether in the last campaign customers successfully grab the offer or we have to think in a different way keeping in mind of the price constraint i.e. more specifically can we fit a suitable model based on the given information to predict the response variable (customer's response in the last campaign)? We try to answer this question by our statistical analytical procedure.

So the complete description of our regressors and response variable is given below –

- **AcceptedCmp1** - 1 if customer accepted the offer in the 1st campaign, 0 otherwise
- **AcceptedCmp2** - 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
- **AcceptedCmp3** - 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
- **AcceptedCmp4** - 1 if customer accepted the offer in the 4th campaign, 0 otherwise
- **AcceptedCmp5** - 1 if customer accepted the offer in the 5th campaign, 0 otherwise
- **Complain** - 1 if customer complained in the last 2 years
- **DtCustomer** - date of customer's enrolment with the company
- **Education** - customer's level of education
- **Marital** - customer's marital status
- **Kidhome** - number of small children in customer's household
- **Teenhome** - number of teenagers in customer's household
- **Income** - customer's yearly household income
- **MntFishProducts** - amount spent on fish products in the last 2 years
- **MntMeatProducts** - amount spent on meat products in the last 2 years

- **MntFruits** - amount spent on fruits products in the last 2 years
- **MntSweetProducts** - amount spent on sweet products in the last 2 years
- **MntWines** - amount spent on wine products in the last 2 years
- **MntGoldProds** - amount spent on gold products in the last 2 years
- **NumDealsPurchases** - number of purchases made with discount
- **NumCatalogPurchases** - number of purchases made using catalogue
- **NumStorePurchases** - number of purchases made directly in stores
- **NumWebPurchases** - number of purchases made through company's web site
- **NumWebVisitsMonth** - number of visits to company's web site in the last month
- **Recency** - number of days since the last purchase
- ***Response (target)* - 1 if customer accepted the offer in the last campaign, 0 otherwise**

This is the full description of the data-set, more specifically name of different columns and their characteristic features. We have to go through the columns and their values for the further analysis.

	ID	Year_Birth	Education	Marital_Status	Income	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds	NumDealsPurchases	Complain	Z_CostContact	Z_Revenue	Response
0	5524	1957	Graduation	Single	58138.0	88	546	172	88	88	3	0	3	11	1
1	2174	1954	Graduation	Single	46344.0	1	6	2	1	6	2	0	3	11	0
2	4141	1965	Graduation	Together	71613.0	49	127	111	21	42	1	0	3	11	0
3	6182	1984	Graduation	Together	26646.0	4	20	10	3	5	2	0	3	11	0
4	5324	1981	PhD	Married	58293.0	43	118	46	27	15	5	0	3	11	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
2235	10870	1967	Graduation	Married	61223.0	43	182	42	118	247	2	0	3	11	0
2236	4001	1946	PhD	Together	64014.0	0	30	0	0	8	7	0	3	11	0
2237	7270	1981	Graduation	Divorced	56981.0	48	217	32	12	24	1	0	3	11	0
2238	8235	1956	Master	Together	69245.0	30	214	80	30	61	2	0	3	11	0
2239	9405	1954	PhD	Married	52869.0	3	61	2	1	21	3	0	3	11	1

Figure 1: Overview of Marketing Campaign data-set

3 Exploratory Data Analysis(EDA)

For every data-set we have to perform the some amount of exploratory data analysis. At first we have to visualize the pattern of the data for necessary columns and also their interrelations via some sort of critical analysis and then some great deal of visualizing tools.

3.1 EDA on response variable

Since our response variable (response) is a binary variable i.e. a categorical variable,so it is reasonable to start with some plot that deals with the the categorical nature of the data. For easy understanding, here we prefer the pie chart.

Here in the given figure the *orange* part shows the *"Not Accepted"* part of the response and the *"accepted"* part is marked by *"blue"* color. Now the ordinary visualization shows that the overall response by the customer is actually not accept the offer. In statistically we can say that **85.1%** people have not accepted the offer and only **14.9%** accepted the offer.

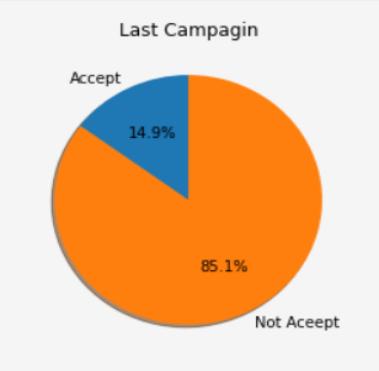

Figure 2: Pie Chart corresponding to the Response Variable

So it is very important to scrutinize the defects, customer's overall choice, their personal information e.g.their marital,education,their response in the earlier campaigns. For this we have to build a model that can predict the customers choice and in the mean time we can also detect the defects and follow a nice budget estimated campaign successfully.

3.2 EDA on different regressor variables

Here we will mainly focus on the different characteristics for different columns mainly through graphical analysis. This graphical tools may help our analysis in proper direction.

- **Information of Regressors on the Last Campaign -**

  For the last campaign let us see that the nature or the change in the values of different columns via mainly bar-plot.

- **Analyzing the Education Details of the Customers**

  The main aim of the campaign is to get the response from the customers. In order to do that we should have a handful knowledge about our customers.

  In this process firstly we are interested in knowing the education qualifications of our customers. In order to know about the same, we plot a graph to get a count of people having basic knowledge or are graduated, PhD scholars, etc.

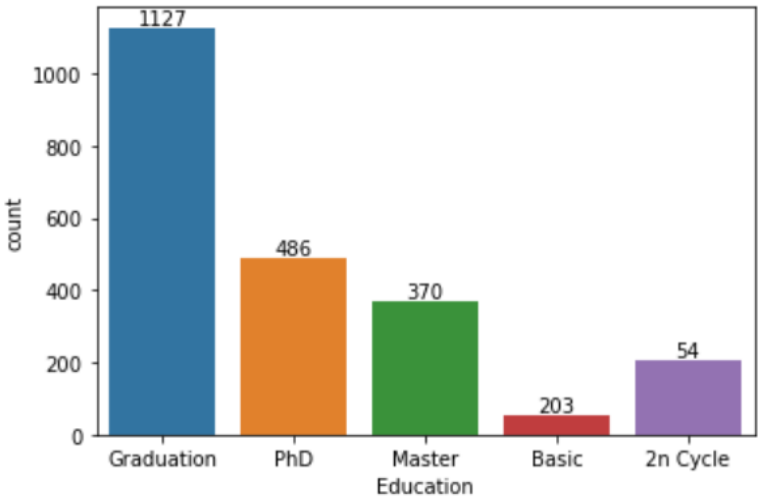

Figure 3: Bar Plot of the education level

Then we plot a graph to analyze the marital status of the customers

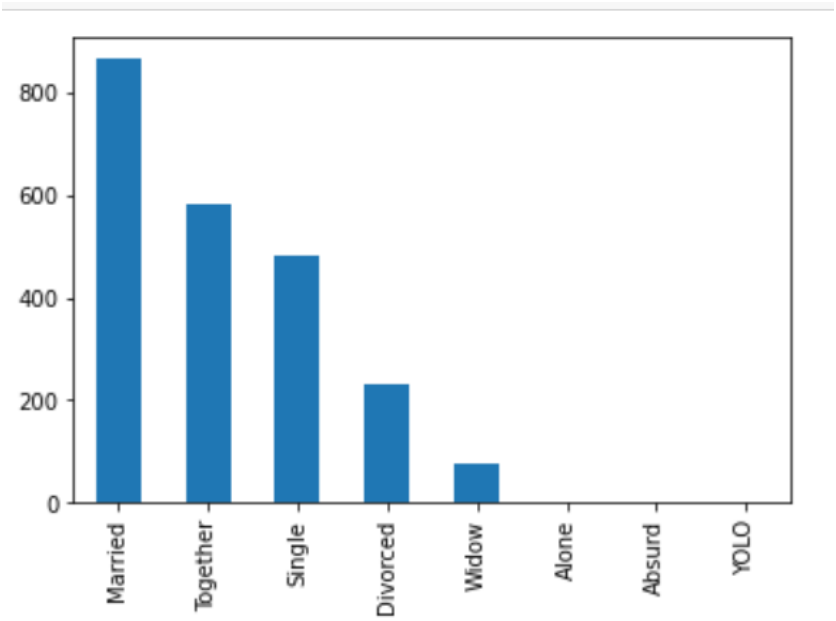

Figure 4: Plot of the marital status

For a new campaign to be successful we must clearly analyze about the last campaign, i.e. how the customer reacts to our previous campaign if it was more attracted to the family with teens or kids. From these graphs we get to know the answer what was the maximum qualification of the customers attracted, i.e. our campaign attracted graduated people or people with low qualifications or even higher than it. To answer all these questions we take the help of some bargraphs.

First bar graph is about the education details of the customers accepting the last campaign.

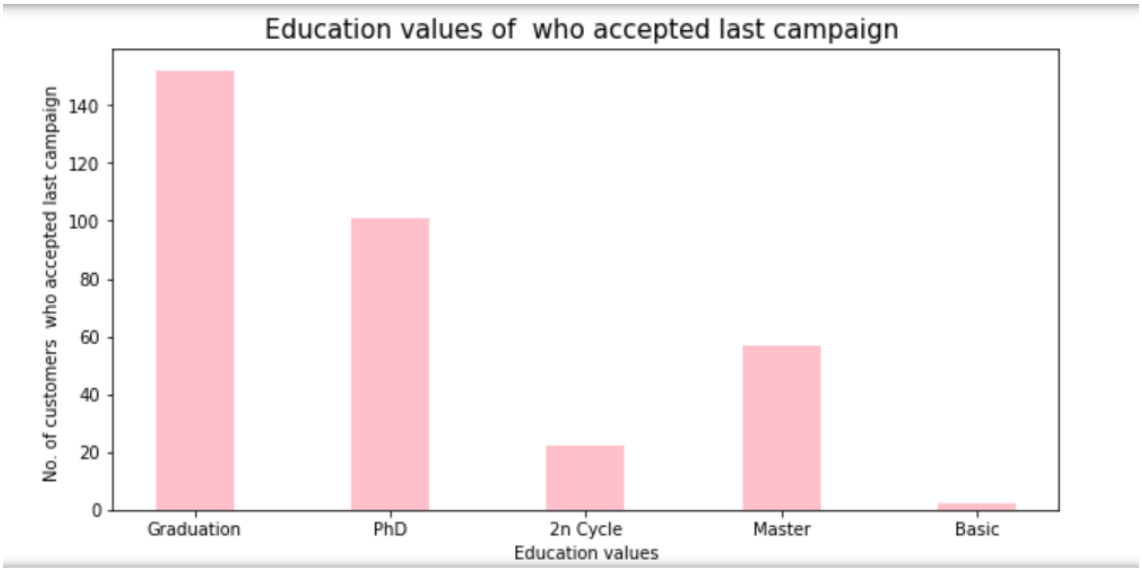

Figure 5: Education values of who accepted the last campaign

Here we see that the last campaign of our marketing team attracted the people pursuing the higher education and wasn't able to attract much with basic knowledge.

Next, we analyze the response of last campaign over the marital status of the customers.

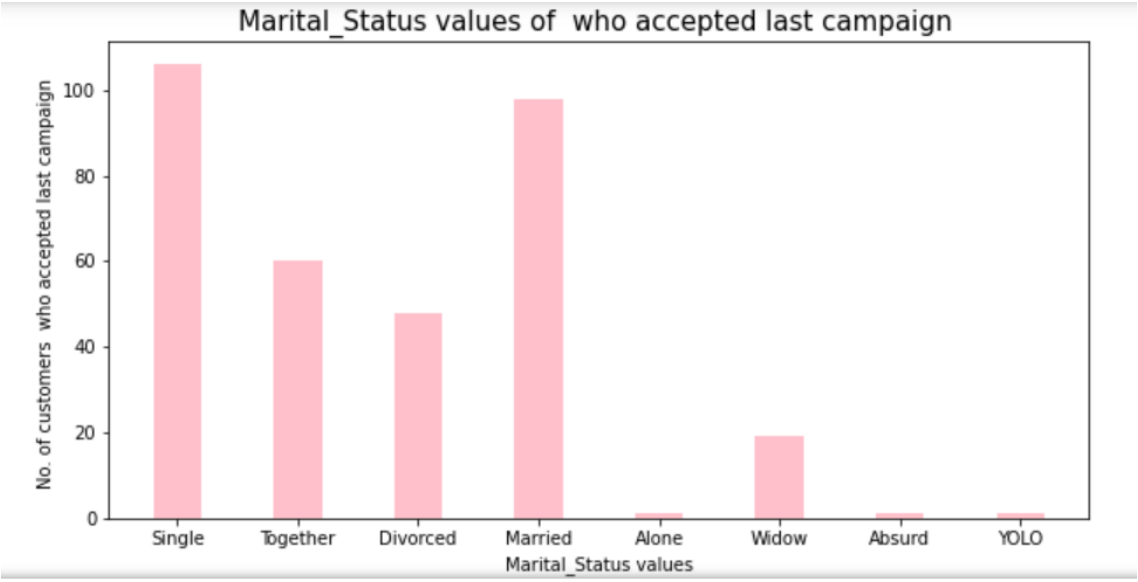

Figure 6: Marital status of who accepted the last campaign

The above bar graph shows that the last campaign was completely fine attracting the Single and Married customers but marketing team could have emphasize more on attracting Together and Divorced category of Marital Status.

Recency values shows the number of days the customers last purchased a product from our company. So, it is important to see whether we are lacking behind to attract the same group of customer. Therefore, we form another bar graph on checking the recency days of our customers.

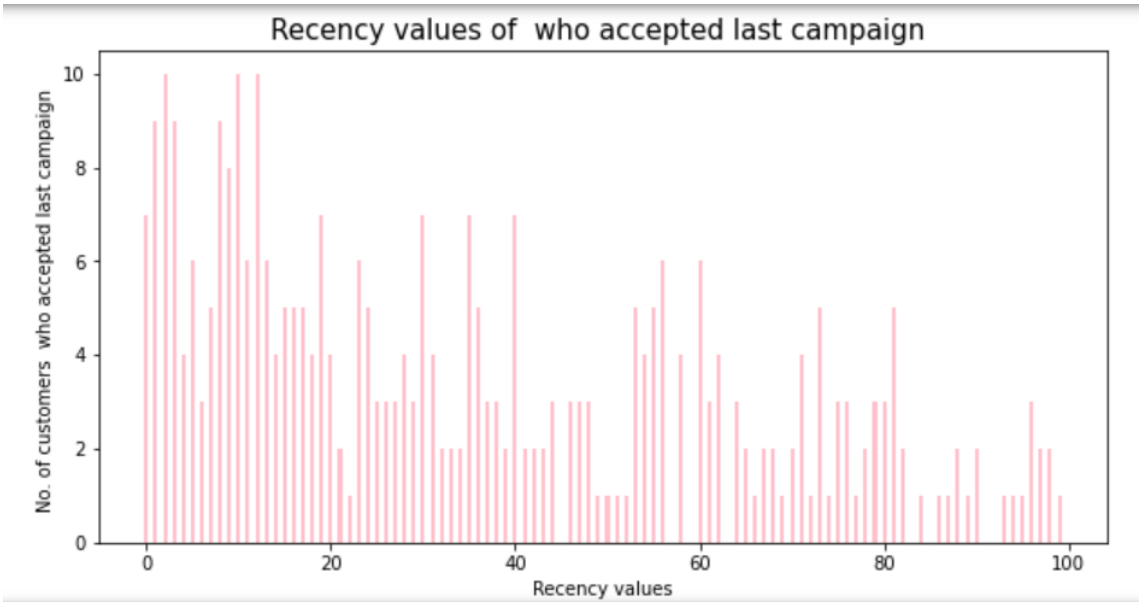

Figure 7: Recency values of who accepted the last campaign

Here, we see that most of the people recently purchased some item from our company. If the sky-scrappers started forming at the right end of the plot then it would be a real concern to check on it.

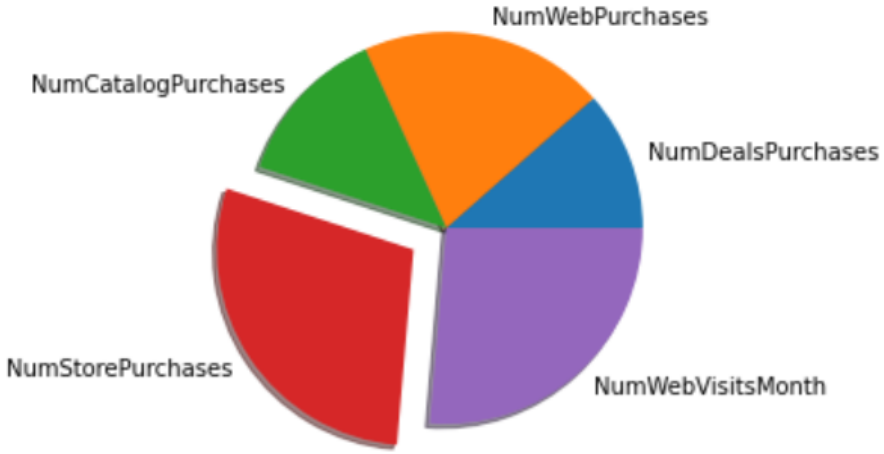

Figure 8: Pie chart for the Number of Purchases from the Different layout

From the pie graph, we can analyse that more than 50% of the purchases are from web visits or from the store purchases.

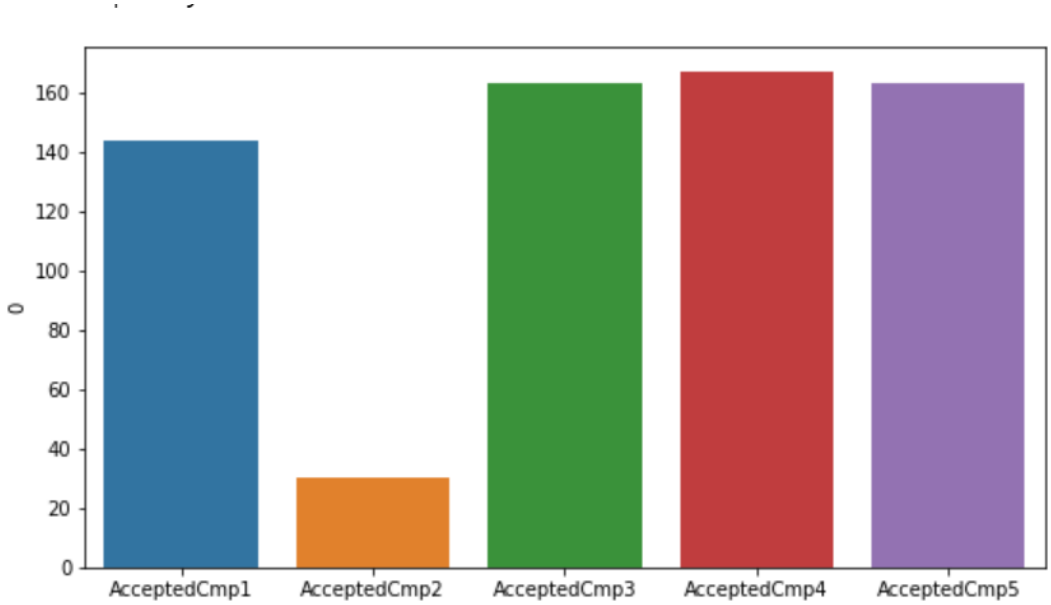

Figure 9: Bar plot analyzing the success of different campaigns

We can interpret from the bar plot that the third, forth and fifth campaigns were equally successful attracting approximately 160 customers but the second campaign was not upto the mark as it attracted just 30 customers.

4 Data Encoding

Data encoding is a principle step for every analysis of categorical data. Here, every categorical columns have to be encoded i.e. turned them into numerical values appropriately. During encoding –

1. We first have to keep in mind that the design matrix must not be singular.
2. There is no restriction on the model.
3. Ordinality have to be explained properly.

Now our marital status column is encoded as follows –

- Total number of categories in the marital status column is 7 So as by the basic rule of encoding we can encode it into 6 numerical variables that takes on the values 0 or 1.So the categories are – **'Single', 'Together', 'Married', 'Divorced', 'Widow', 'Alone','Absurd', 'YOLO'**

  Here using **PYTHON**, we actually encode the variables by the "level hot encoding". Now the following table gives the summarization of what actually we want to say–

Table 1: **Encoded Marital Status Column**

	Encoded Variables						
Raw Categorical variables	Alone	Divorced	Married	Single	Together	Widow	YOLO
Alone	1	0	0	0	0	0	0
Divorced	0	1	0	0	0	0	0
Married	0	0	1	0	0	0	0
Single	0	0	0	1	0	0	0
Together	0	0	0	0	1	0	0
Widow	0	0	0	0	0	1	0
YOLO	0	0	0	0	0	0	1
Absurd	0	0	0	0	0	0	0

Now in our model we use this encoded variables throughout and the categorical variables are now just the 7-dimensional indicator variables, slightly in different ways.

- We have another categorical column "Education" but this column is a ordinal. So we can encode this column into the natural order i.e. based on the ordinality we place the values 1,2,...., upto final order.

So the data encoding for this column is as follows –

$$['\textbf{2n Cycle}' \to 0]$$

$$['\textbf{Basic}' \to 1]$$

$$['\textbf{Graduation}' \to 2]$$

$$['\textbf{Master}' \to 3]$$

$$['\textbf{PhD}' \to 4]$$

Here the left hand side is the categories of the original categorical variable and the right hand side is the corresponding level encoded values in actual order.

5 Outlier Detection

Outlier in the data plays a very important role when we came into the estimation purpose for the model parameters. So we have to remove or estimate the outliers based on the given scenario.

For our model let us draw the box plots to see whether heavy amount of outliers present in our data or not. If the number of outliers are quite small we can simply delete the corresponding rows, otherwise we proceed into further analysis.

Let us first see that the box-plots of the corresponding columns–

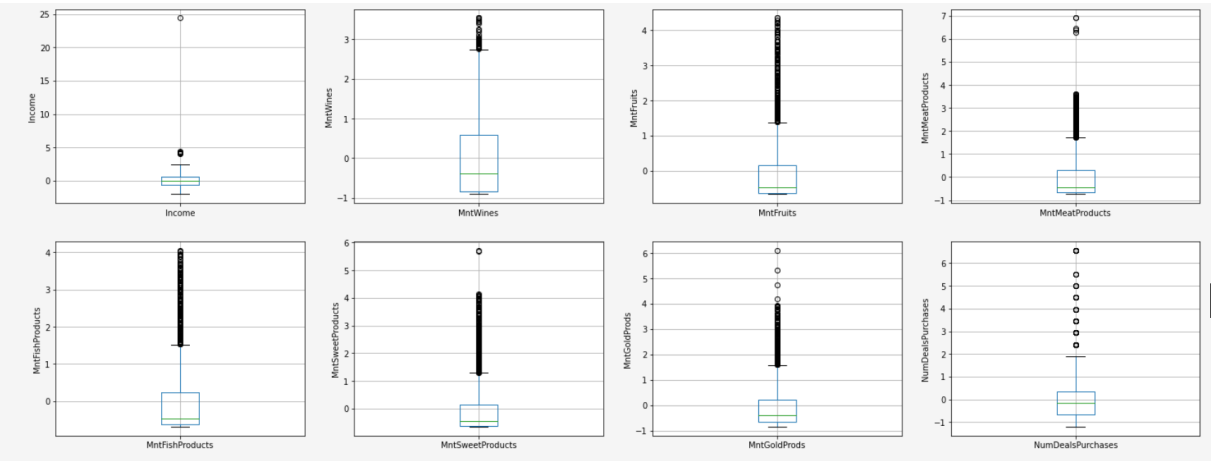

Figure 10: Box-plot for detecting outliers in the present data

Since too much outliers are present in the data, we have to estimate them. Here we apply Interquartile Range method (IQR) for outlier detection and removal of outliers. Observing the boxplot, it is almost sure that the variables are skewed. So, IQR method suits well for such kind of data.

Let, X denotes the variable
Then the IQR =(q_3(X)-q_1(X)) Where, q_3(X)= 3^{rd} quartile ; q_1(X)=1^{st} quartile
Now, let us define,
lower(X)= q_1(X) – IQR (X)*1.5
Upper(X)= q_3(X)+IQR (X)*1.5
If any values of X, say i-th value of X, X[i], say Now we will estimate all the outliers outside the inter quartile range by the corresponding median values of that particular regressors.

In this way we can estimate all the outliers present in the data and hence we can say that, now our data is quite structured than the previous one.

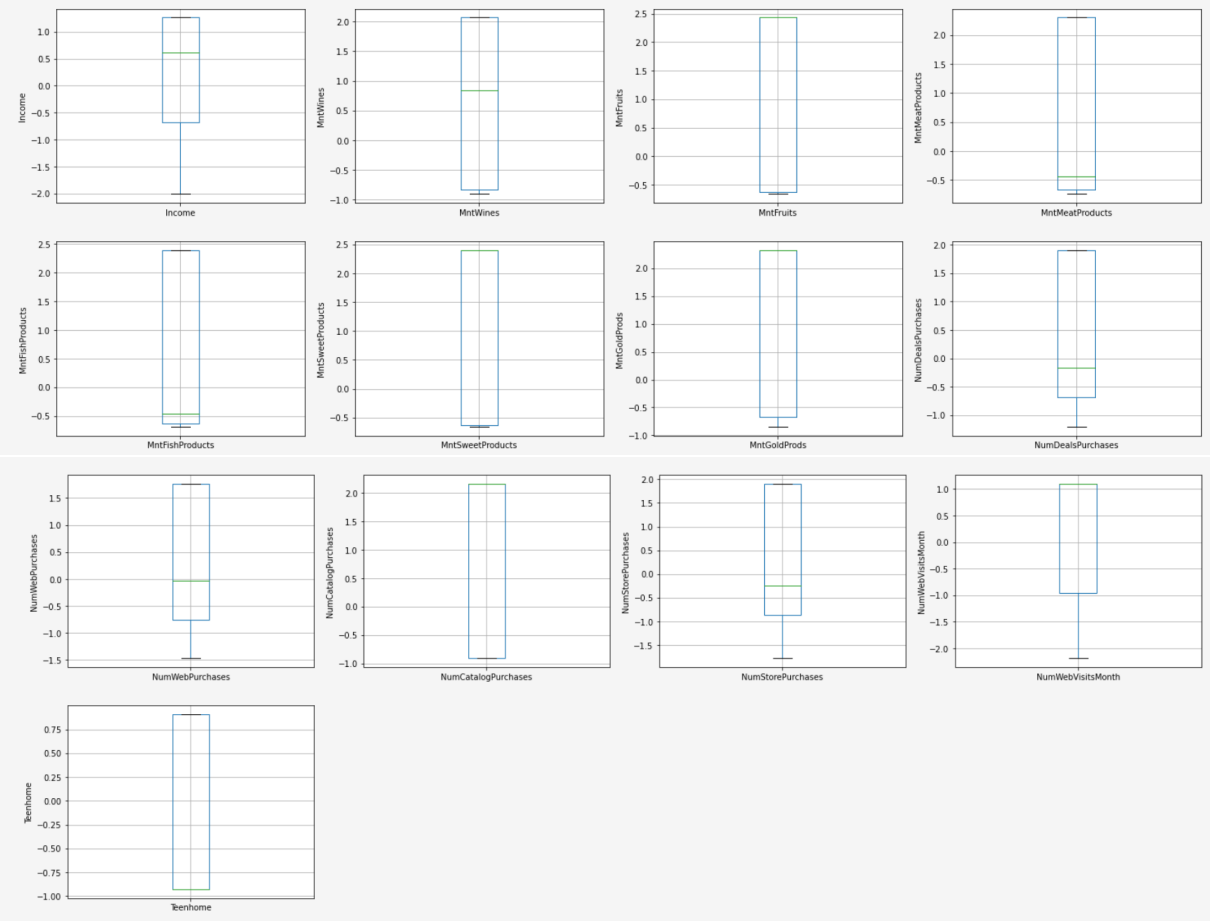

Figure 11: Box-plot after estimating outliers in the present data

6 Missing Data

Missing values are quiet important for data cleaning step in every data analysis project work. If some of the observations are not recorded, totally lost or inappropriately stated then those values are known as the missing values.Usually soft wares or spreadsheets refer them as 'NA'.

If the missing values are quite small then we can remove them from the dataset, i.e. we can delete the whole row containing the missing values. But if the number of missing values are quiet large then they have to be estimated by some rigorous process.

In our data, in the 'Income' column, only 24 observations are missing. So we can delete them. Since we estimate the outliers and balanced our data, so we apply the mean-imputation method for estimating the missing values.

Mean-imputation method is largely affected if some outliers present in the data. Because mean can shift towards the outliers and we may have tedious estimates via this method. But we first lean the outliers, then we apply the mean-imputation method.

Here missing values in every column has been estimated by the corresponding mean values of that column. Now, there are only 24 missing values presented in the 'Income' column. So,we estimate them by the mean of the 'Income' column.

After applying the the missing value imputation technique we finally present the PYTHON output corresponding to each of the columns and that represent the number of missing values present in our data.

```
Education             0
Income                0
Kidhome               0
Teenhome              0
Recency               0
MntWines              0
MntFruits             0
MntMeatProducts       0
MntFishProducts       0
MntSweetProducts      0
MntGoldProds          0
NumDealsPurchases     0
NumWebPurchases       0
NumCatalogPurchases   0
NumStorePurchases     0
NumWebVisitsMonth     0
AcceptedCmp3          0
AcceptedCmp4          0
AcceptedCmp5          0
AcceptedCmp1          0
AcceptedCmp2          0
Complain              0
Z_CostContact         0
Z_Revenue             0
Response              0
Alone                 0
Divorced              0
Married               0
Single                0
Together              0
Widow                 0
YOLO                  0
dtype: int64
```

Figure 12: Number of missing values in the data-set

7 Multicollinearity

Multicollinearity is a silent killer of any regression model. If the multicollinearity is present in the data then all the coefficient magnitudes becomes very large, but the predicted value of the response variables may be highly satisfactory. But if we delete one row or one column then the model change drastically, the negative coefficient becomes positive, the largest coefficient may be smaller in the newer model and finally the prediction may be reversed the previous one.

Let us consider an example: based on our present data we fit a model and estimate response variable, i.e. we give a mathematical logic for predicting some near future observations. The model gives well data fit, all the model accuracy say that this model is a good one. But after one or two week we again collect some data and add to the data-set and the model

changes drastically. If may possible that it's prediction based on the previous data also changes dramatically and we finally conclude that the model is completely worthless.

Let us consider our logistic model –

Now the least square estimators of the model parameters i.e. β's are explained as

Now the matrix (X'X) must be non-singular for the inverse to exist.Since the elements of (X'X) are continuous real values it has a very high probability that it is non-singular.Now the original problem occurs when the matrix is non-singular but it is actually near singular i.e the determinant of (X'X) is ≈ 0. Then this (X'X) matrix is said to be the ill-conditioned matrix.

7.1 Simple Correlation Check

Let us first look into the simple correlation matrix for the whole data-set. Now the correlation matrix is not only hard to interpret but also very difficult to visualize. So what we want to say that instead of correlation matrix we try to visualize it by the correlation heat map. Here light colors represent the high positive correlation and the dark color represent the high negative correlation, in between the color changes with the from light to dark gradient as the value of the correlation changes from 1.0 to -1.0.

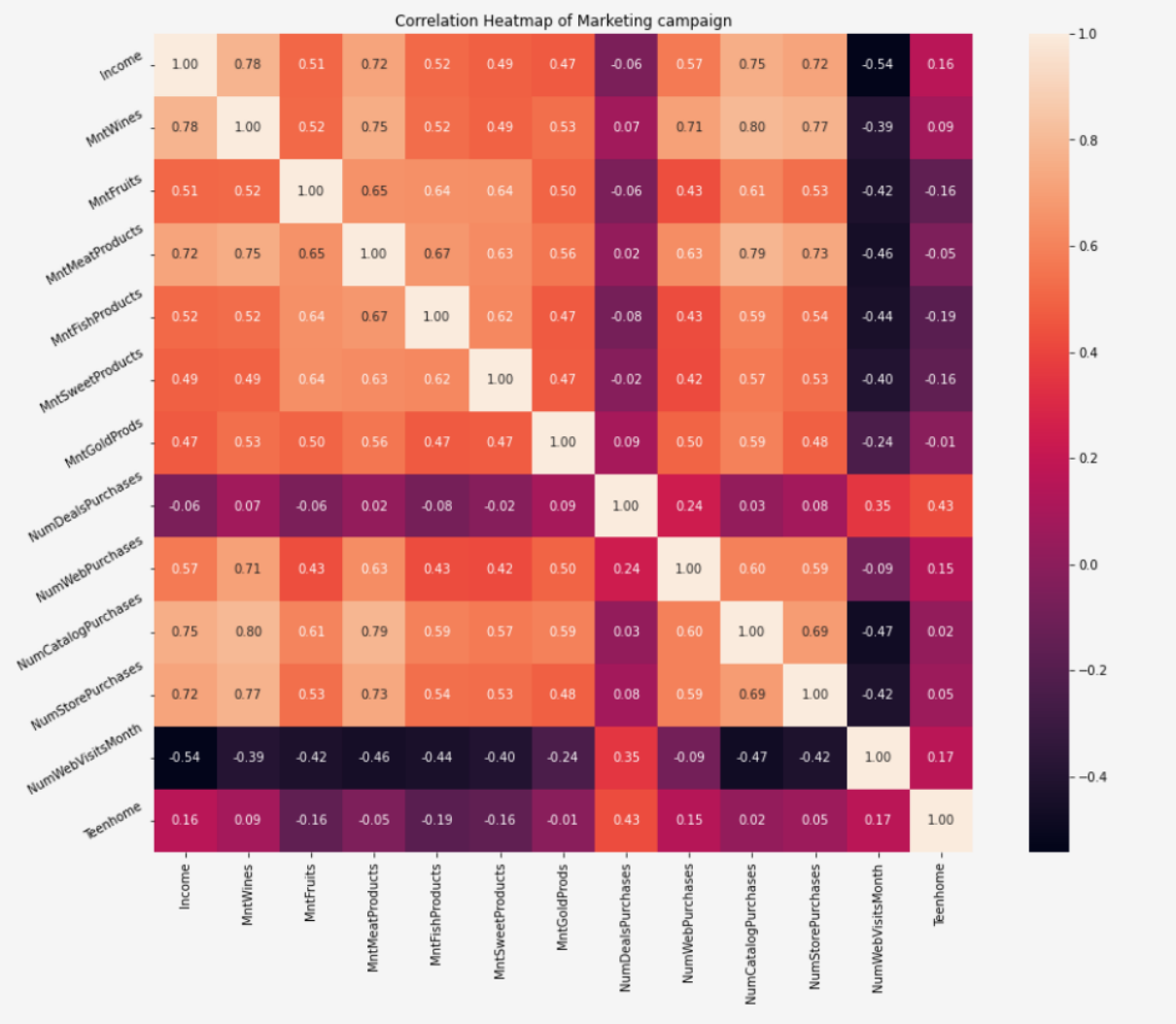

Figure 13: The correlation heat-map of our data-set

7.2 Variance Inflation Factor

Since the correlation matrix is not very informative for our data, we choose our second alternative procedure for finding multicollinearity in the data,if presents. Now multicollinearity exists only if the continuous columns are nearly linearly dependent. so we first standardize the continuous columns and try to apply the VIF method to completely get rid of the multicollinearity problem.

Something—-

We first check the VIF values for our model for each of the regressor variables and as a rule of thumb if

- If VIF value is greater then 5, we can say that multicollinearity is present in the model.
- Else multicollinearity is not present in our model.

	feature	VIF
1	MntWines	5.693785
3	MntMeatProducts	5.529949
9	NumCatalogPurchases	5.257185
0	Income	4.306189
10	NumStorePurchases	3.312221
2	MntFruits	3.140082
4	MntFishProducts	3.017878
5	MntSweetProducts	2.893736
8	NumWebPurchases	2.443291
6	MntGoldProds	2.258905
11	NumWebVisitsMonth	1.765098
7	NumDealsPurchases	1.620554
12	Teenhome	1.532493

Figure 14: Variance Inflation Factor for all the of our data-set

Now for three columns "MntWines", "MntMeatProducts", "NumCatalogPurchases" VIF values are 5.693785, 5.529949, 5.257185 respectively. So here we drop the variable with highest VIF i.e. "MntWines" and as by the rule we have to find the VIF for all the variables present in the model except the one that is excluded last time.

	feature	VIF
7	NumCatalogPurchases	4.195212
0	Income	3.952539
1	MntFruits	3.065607
8	NumStorePurchases	2.877751
2	MntFishProducts	2.874222
3	MntSweetProducts	2.833840
4	MntGoldProds	2.249110
6	NumWebPurchases	2.146884
9	NumWebVisitsMonth	1.753698
5	NumDealsPurchases	1.615131
10	Teenhome	1.515180

Figure 15: Variance Inflation Factor for the remaining variables of our data-set

Since none of the variables have VIF greater than 5, so we can conclude that the multi-collinearity problem is completely resolved from our model.

8 Model Fitting : Logistic Model

Finally based on the given regressors, we fit the logistic model with these new set of regressors and based on that we try to visualize the quality of fitting for our model.

Logistic Model Since the outliers and the missing values are estimated in our data-set, we are now able to set our model. As our response variable 'Y ' is a categorical variable with two categories, that is,
0 which indicates that the customer does not accept the offer and
1 which indicates that the customer accept the offer. Now the logistic regression is a generalized linear regression, it is very important to note it's link function $ln\frac{p_i}{1-p_i}$.

$$Y_i \sim Binomial(N_i, p_i)$$

$$ln(\frac{p_i}{1-p_i}) = \beta_o + \sum_{j=1}^{k} \beta_j x_{ij} = \mathbf{X}\beta \quad for\ i = 1, 2, ..., n$$

where x_{ij} is the element in the i^{th} row and j^{th} column of the model matrix X. To evaluate the accuracy of the classification or the logistic model, we use a confusion matrix.

8.1 Confusion Matrix

confusion matrix is actually the most informative part of the logistic regression model.The confusion matrix shows the effective and the defective part of the model simultaneously.Here at first we have to categorized our predicted response variable into 0 and 1. Then we will check that how much of the response is accurately explained by the predictive model and how much it doesn't. Now there four categories in confusion matrix –

1. **TRUE Negative:**If the predicted model says FALSE and the corresponding response is also FALSE.

2. **TRUE Positive:**If the predicted model says TRUE and the corresponding response is also TRUE.

3. **FALSE Positive:**If the predicted model says TRUE and the corresponding response is also FALSE.

4. **FALSE Negative:**If the predicted model says FALSE and the corresponding response is also TRUE.

8.2 Model Diagnostics

- **Accuracy Score :** Generally our model posses some good properties if it's TRUE positive and TRUE negative part is nearly equal to the total response and the FALSE positive and the FALSE negative part is nearly equal to zero. Then a simple measure of accuracy is the total number of TRUE cases i.e. number of TRUE positives + number of TRUE negatives divided by total number of responses. So mathematically we can write,

$$Accuracy = \frac{Number\ of\ correct\ predictions}{Total\ number\ of\ predictions\ made}$$

- **Precision Score :** It is the number of correct positive results divided by the number of positive results predicted by the classifier. Precision talks about how precise/accurate your model is out of those predicted positive, how many of them are actual positive.Precision is a good measure to determine, when the costs of False Positive is high. So mathematically,

$$Precison = \frac{True\ Positive}{True\ Positive\ +\ False\ Positive}$$

- **Specificity :** Specificity, also known as the true negative rate, measures the proportion of actual negatives that are correctly identified as such. It is the opposite of the recall.So, mathematically we can write,

$$Specificity = \frac{True\ Negative}{True\ Negative\ +\ False\ Positive}$$

- **F1 Score :** The F1 Score is a measure of a test's accuracy, that is, it is the harmonic mean of precision and recall. It can have a maximum score of 1 and a minimum of 0. Overall, it is a measure of the preciseness and robustness of the model.

$$F1\ Score = \frac{2(precision\ X\ recall)}{precision+recall} = \frac{2True\ Positive}{2Ture\ Positive+False\ Positive+False\ Negative}$$

- **ROC Curve :** An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds. This curve plots two parameters: True Positive Rate. False Positive Rate. The ROC curve is produced by calculating and plotting the true positive rate against the false positive rate for a single classifier at a variety of thresholds. For example, in logistic regression, the threshold would be the predicted probability of an observation belonging to the positive class.

9 Train-Test and Split

Now to check whether we are on the perfect direction or on the wrong direction we split the total data into two parts. At first we build our model with 2016 rows and all 29 columns. Based on that model we then test the the remaining 224 observation and also we will see the model accuracy and the overall summary of the model.

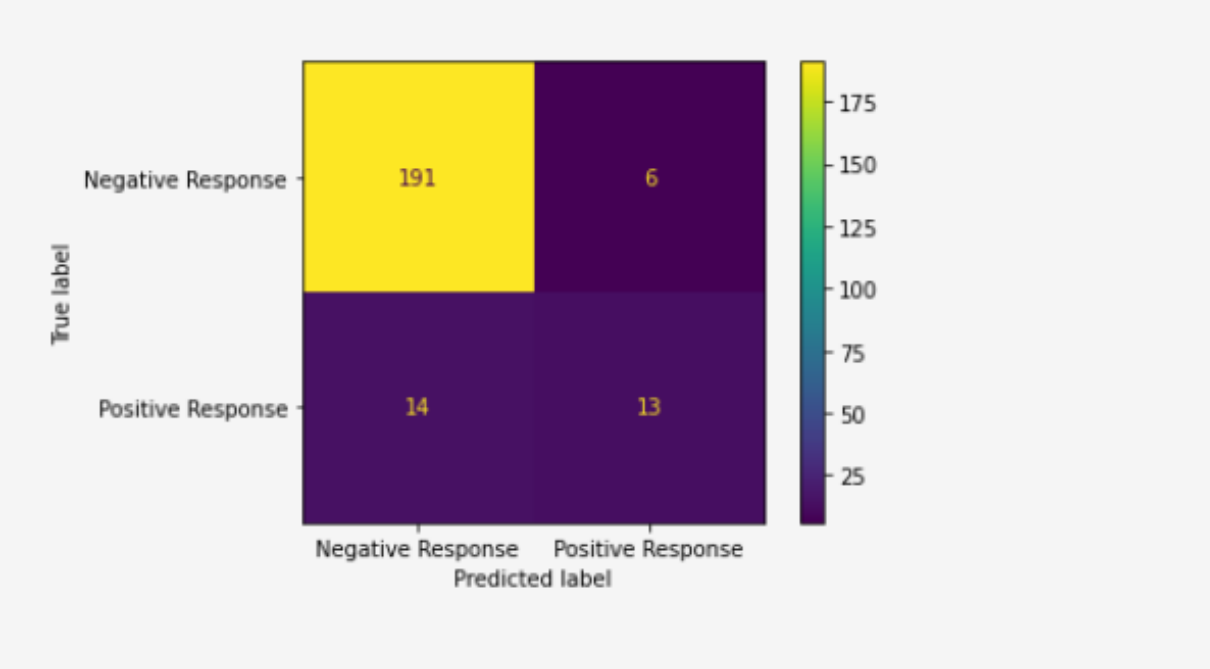

Figure 16: Confusion matrix corresponding to train test

```
0.9107142857142857
              precision    recall  f1-score   support

           0       0.93      0.97      0.95       197
           1       0.68      0.48      0.57        27

    accuracy                           0.91       224
   macro avg       0.81      0.73      0.76       224
weighted avg       0.90      0.91      0.90       224
```

Figure 17: Model accuracy and summary corresponding to train test

Since the model accuracy is nearly 90% the train model fits the data well. Now from the confusion matrix we observe that the TRUE negative part is 191 out of 224 observations and FALSE positive and FALSE negative part are very less and finally TRUE positive is also very tiny amount.
Now the precision for 0 is 0.93 and for 1 is 0.68 and the f1 score is 0.95 for 0 and 0.57 for 1. Everything is quite fine except the fact that the model is extremely imbalanced.

10 Data Imbalance

Imbalanced data are those type of data sets where the target class have an unequal distribution of observations i.e., one class have high number of observations and the other class have low observations. This data imbalance is observed in the data set because sometimes some data are high cost or sometimes it become impracticable to collect the required quantity of data or some times imbalanced data came due to some error.This bias in the training data-set can influence many machine learning algorithms, leading some to ignore the minority class entirely. This is a problem as it is typically the minority class on which predictions are most important.

Data imbalance is also present in our data with huge percentage.Let us see the bar graph that corresponds the visualization of data imbalance–

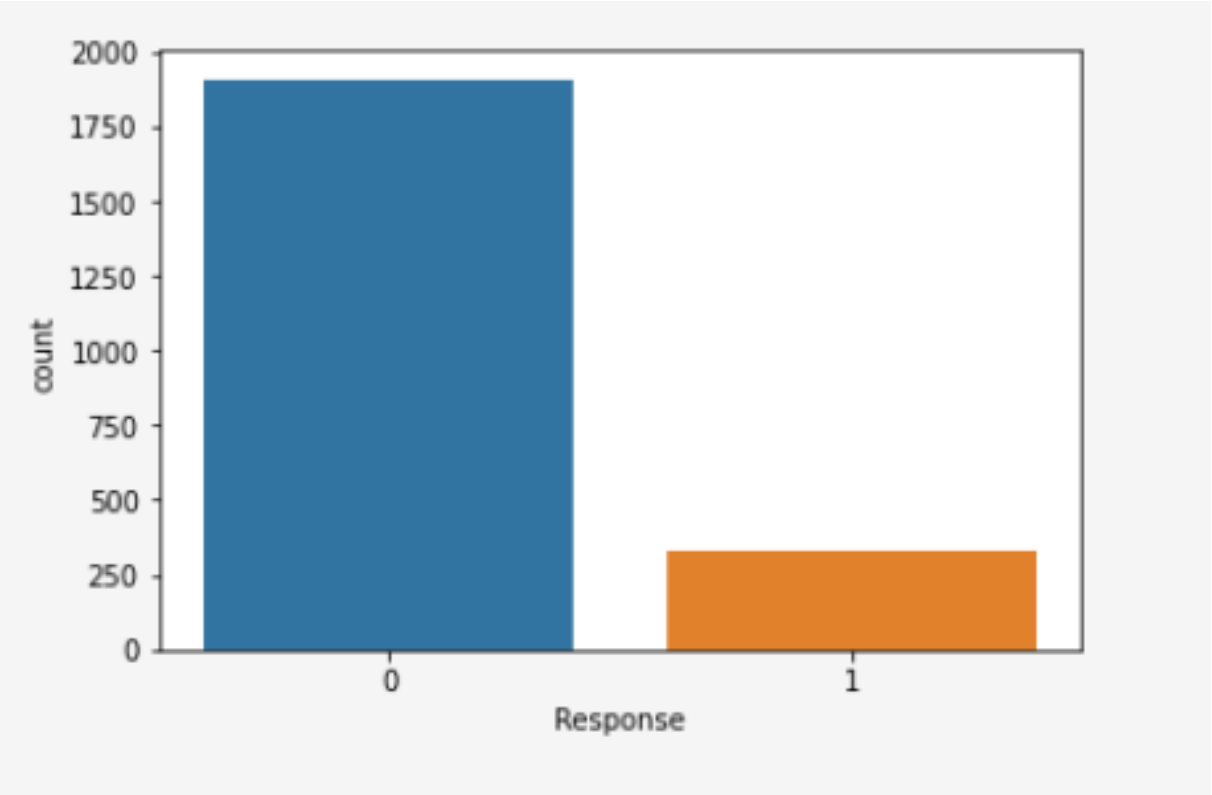

Figure 18: Bar Graph corresponding to the response variance

To over come this problem of data imbalance, we have used some methods. Those methods are as follows -

Oversampling

Oversampling or Random oversampling is a method to randomly allocate data to the minority class. Here the data is randomly allocated to the minority class by the method of simple random sampling with replacement (SRSWR) from the available data of that minority class.

This type of random sampling or data imbalance method is highly effective to those type of data sets which have a skewed distribution. This type of imbalance is also used for the data which have huge outliers in order to cut the outliers and replace them by balanced data.

This technique can be effective for those machine learning algorithms that are affected by a skewed distribution and where multiple duplicate examples for a given class can influence the fit of the model. This might include algorithms that iteratively learn coefficients, like artificial neural networks that use stochastic gradient descent. It can also affect models that seek good splits of the data, such as support vector machines and decision trees.

Now for our data we use this over samping technique and the result is given below —

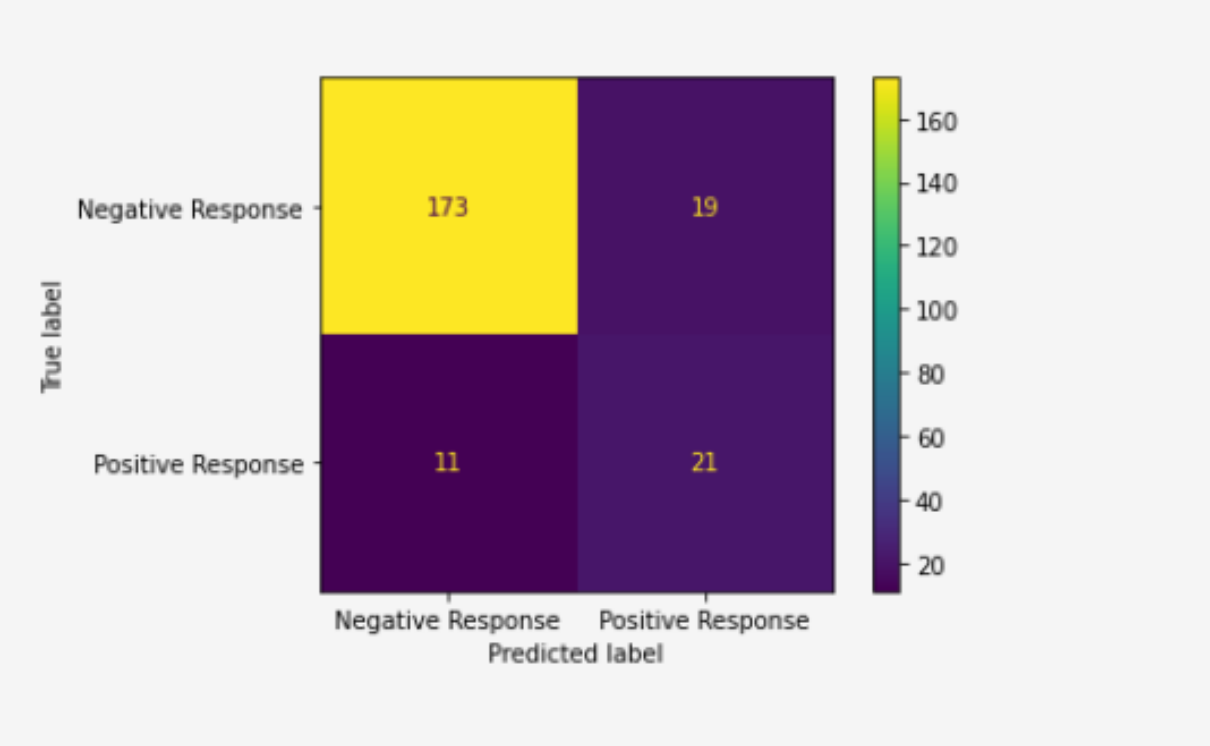

Figure 19: Heat Map for Over sampling method

0.8616071428571429

	precision	recall	f1-score	support
0	0.94	0.89	0.91	186
1	0.57	0.71	0.64	38
accuracy			0.86	224
macro avg	0.76	0.80	0.77	224
weighted avg	0.88	0.86	0.87	224

Figure 20: Summery corresponding to the oversampling method

Since the model accuracy is nearly 86%, so the model fits the data well. Now from the confusion matrix we observe that the TRUE negative part is 173 out of 224 observations and FALSE positive and FALSE negative part are very less and but finally TRUE positive increases slightly.

Now the precision for 0 is 0.94 and for 1 is 0.57 and the f1 score is 0.91 for 0 and 0.64 for 1. Now data is quite balanced than the previous one.

SMOTE Method

A problem with imbalanced classification is that there are too few examples of the minority class for a model to effectively learn the decision boundary.

One way to solve this problem is to oversample the examples in the minority class. This can be achieved by simply duplicating examples from the minority class in the training dataset prior to fitting a model. This can balance the class distribution but does not provide any additional information to the model.

Synthetic Minority Oversampling Technique or SMOTE works by selecting examples that are close in the feature space, drawing a line between the examples in the feature space and drawing a new sample at a point along that line.

At first the total no. of oversampling observations, N is set up. Generally, it is selected such that the binary class distribution is 1:1. But that could be tuned down based on need. Then the iteration starts by first selecting a positive class instance at random. Next, the KNN's (by default 5) for that instance is obtained. At last, N of these K instances is chosen to interpolate new synthetic instances. To do that, using any distance metric the difference in distance between the feature vector and its neighbors is calculated. Now, this difference is multiplied by any random value in (0,1] and is added to the previous feature vector. This is pictorially represented below:

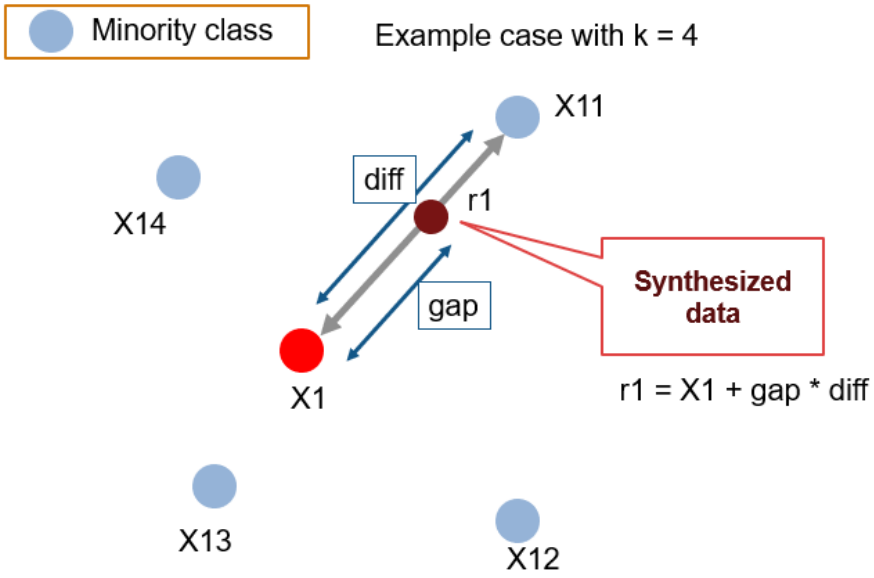

Figure 21: SMOTE

Now for our data we use this over samping technique and the result is given below —

Since the model accuracy is nearly 84%, so the model fits the data well. Now from the confusion matrix we observe that the TRUE negative part is 164 out of 224 observations and FALSE positive and FALSE negative part are very less and but finally TRUE positive increases slightly more than that of oversampling method.

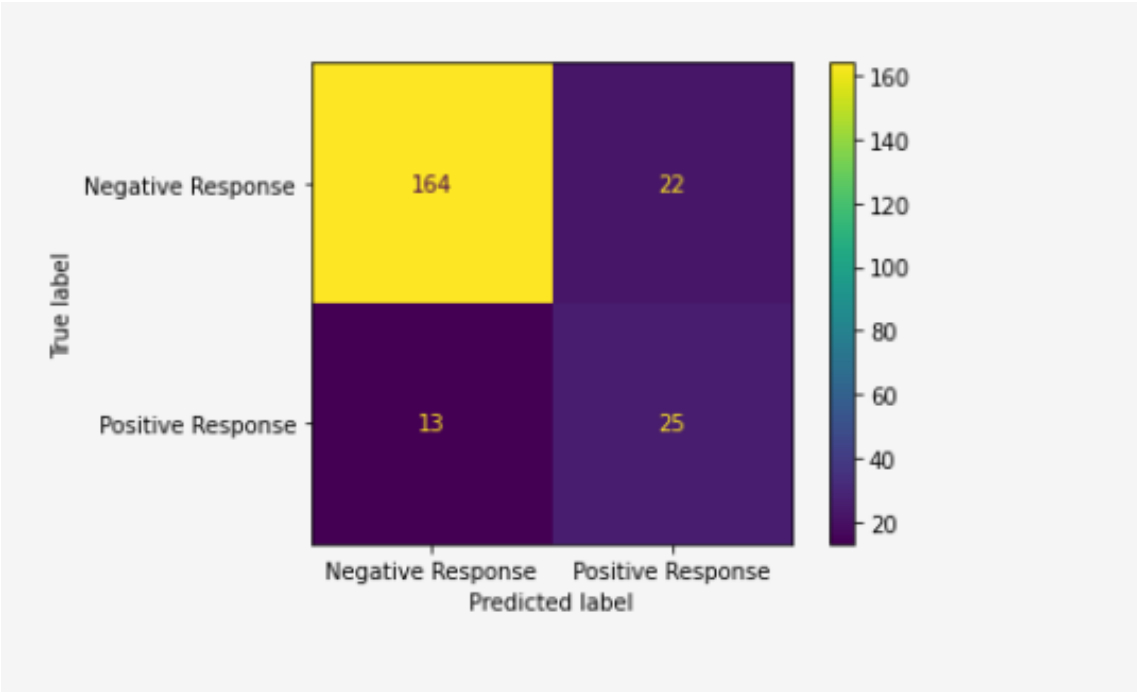

Figure 22: SMOTE Heat map

0.84375

	precision	recall	f1-score	support
0	0.93	0.88	0.90	186
1	0.53	0.66	0.59	38
accuracy			0.84	224
macro avg	0.73	0.77	0.75	224
weighted avg	0.86	0.84	0.85	224

Figure 23: Summary of SMOTE Method

Now the precision for 0 is 0.94 and for 1 is 0.57 and the f1 score is 0.91 for 0 and 0.64 for 1. Now data is quite balanced than the previous one.

11 Final model fitting and diagnostic checking

11.1 Fitting with full model

Now our data becomes balanced and also let us try to fit the logistic model considering all the regressors at the same time. So our model becomes as follows –

```
logmodel1=LogisticRegression()
logmodel1.fit(x_train_os,y_train_os)
y_pred_logistic=logmodel.predict(x_test)
print(accuracy_score(y_test,y_pred_logistic))
```

✓ 0.3s

```
0.84375
```

Figure 24: Accuracy for the full model

Since the accuracy score of the full model is 0.84, we can conclude that this is a better model for our data-set.

```
auc = roc_auc_score(y_test,y_pred_logistic)
print('AUC for Logistic Regression: %.2f' % auc)
```

✓ 0.1s

```
AUC for Logistic Regression: 0.77
```

Figure 25: AUC score for the full model

Now the AUC score of our model is nearly 0.77 which is not too worse and hence there is no serious with our data-set.

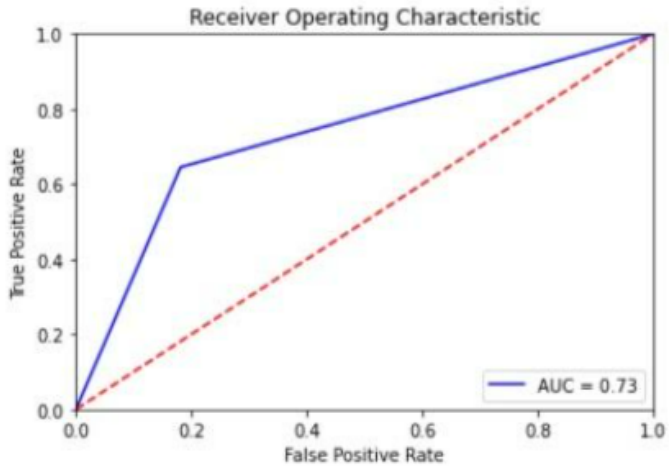

Figure 26: ROC curve for full model

Now the ROC curve along with the AUC score shows that our final model is quite well featured model with high satisfaction for detecting the TRUE negatives.

12 Variable Selection

Since our model contains large number of regressors,it is a better idea to choose some of the regressor variables that can interpret the total model with almost same accuracy as in the full model.

12.1 Stepwise Selection

A combination of forward selection and backward elimination procedure is the stepwise regression for selecting a basket of good variables. It is a not only a modification of forward selection procedure but also the backward elimination one and has the following steps.

- We start with the intercept model and compute the AIC for the model.
- We then compute AIC for all the possibilities of adding one more variable in our intercept only model. We select the variable with the smallest AIC if it has a lower AIC than intercept only model.
- We then again compute AIC for adding one more variable in the model along with the AIC for removing the already added variable. We sort the values by ascending order and variables are either added or the existing variable is subtracted depending on the value of AIC.
- We continue performing these steps until any further action - addition or subtraction, results in increase of AIC of the model.

```
Call:
glm(formula = Response ~ Education + Teenhome + Recency + MntFruits +
    MntGoldProds + NumDealsPurchases + NumWebPurchases + NumCatalogPurchases +
    NumStorePurchases + NumWebVisitsMonth + AcceptedCmp3 + AcceptedCmp4 +
    AcceptedCmp5 + AcceptedCmp1 + AcceptedCmp2 + Married + Together,
    family = binomial, data = D)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5737  -0.4452  -0.2614  -0.1253   3.0823

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)         -2.369946   0.266085  -8.907  < 2e-16 ***
Education            0.387914   0.072249   5.369 7.91e-08 ***
Teenhome            -0.776159   0.105774  -7.338 2.17e-13 ***
Recency             -0.029005   0.002854 -10.163  < 2e-16 ***
MntFruits            0.143391   0.065645   2.184 0.028936 *
MntGoldProds         0.113490   0.064841   1.750 0.080070 .
NumDealsPurchases    0.275796   0.079387   3.474 0.000513 ***
NumWebPurchases      0.192658   0.081806   2.355 0.018520 *
NumCatalogPurchases  0.322175   0.085891   3.751 0.000176 ***
NumStorePurchases   -0.316011   0.077892  -4.057 4.97e-05 ***
NumWebVisitsMonth    0.436115   0.097738   4.462 8.12e-06 ***
AcceptedCmp3         1.713846   0.219096   7.822 5.18e-15 ***
AcceptedCmp4         0.762308   0.262128   2.908 0.003636 **
AcceptedCmp5         1.850898   0.262654   7.047 1.83e-12 ***
AcceptedCmp1         1.422352   0.256225   5.551 2.84e-08 ***
AcceptedCmp2         1.271471   0.542236   2.345 0.019034 *
Married             -1.174607   0.174266  -6.740 1.58e-11 ***
Together            -1.157189   0.198387  -5.833 5.44e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1886.8  on 2239  degrees of freedom
Residual deviance: 1234.7  on 2222  degrees of freedom
AIC: 1270.7

Number of Fisher Scoring iterations: 6
```

Figure 27: Stepwise Variable selection for reducing the model

12.2 Final Model

```
logmodel1=LogisticRegression()
logmodel1.fit(x_train_os,y_train_os)
y_pred_logistic=logmodel.predict(x_test)
print(accuracy_score(y_test,y_pred_logistic))
```

```
0.7946428571428571
```

Figure 28: Accuracy for the final model

Since the accuracy score of the full model is 0.79, we can conclude that this is a better model for our data-set.

```
auc = roc_auc_score(y_test,y_pred_logistic)
print('AUC for Logistic Regression: %.2f' % auc)
```

```
AUC for Logistic Regression: 0.73
```

Figure 29: AUC score for the final model

Now the AUC score of our model is nearly 0.73 which is not too worse and hence there is no serious with our data-set.

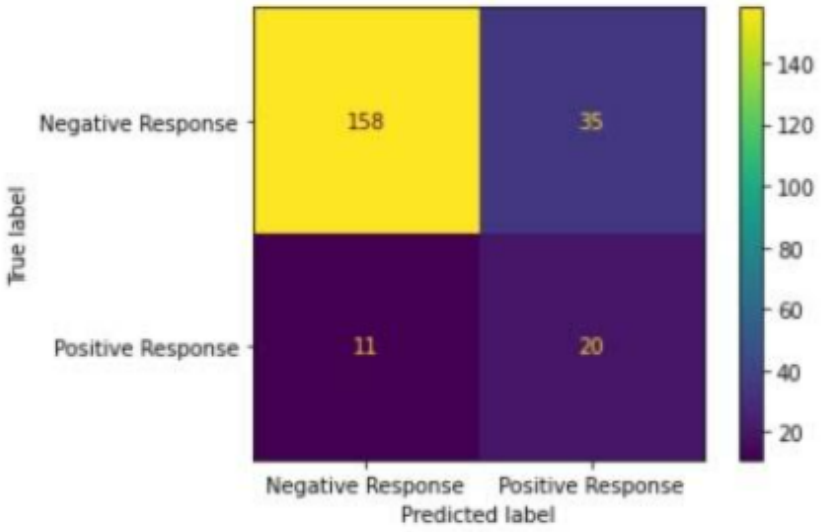

Figure 30: Confusion Matrix for final model

Now from this confusion matrix it is readily obvious that the precision is high for this model and specification is also quite good for our model.

	precision	recall	f1-score	support
0	0.93	0.82	0.87	193
1	0.36	0.65	0.47	31
accuracy			0.79	224
macro avg	0.65	0.73	0.67	224
weighted avg	0.86	0.79	0.82	224

Figure 31: Summary for the final model

Now from the model summary we can conclude that the precision is 0.93 and 0.36 for 0 and 1 respectively. Also the f1 score is 0.87 and 0.47 respectively for 0 and 1.

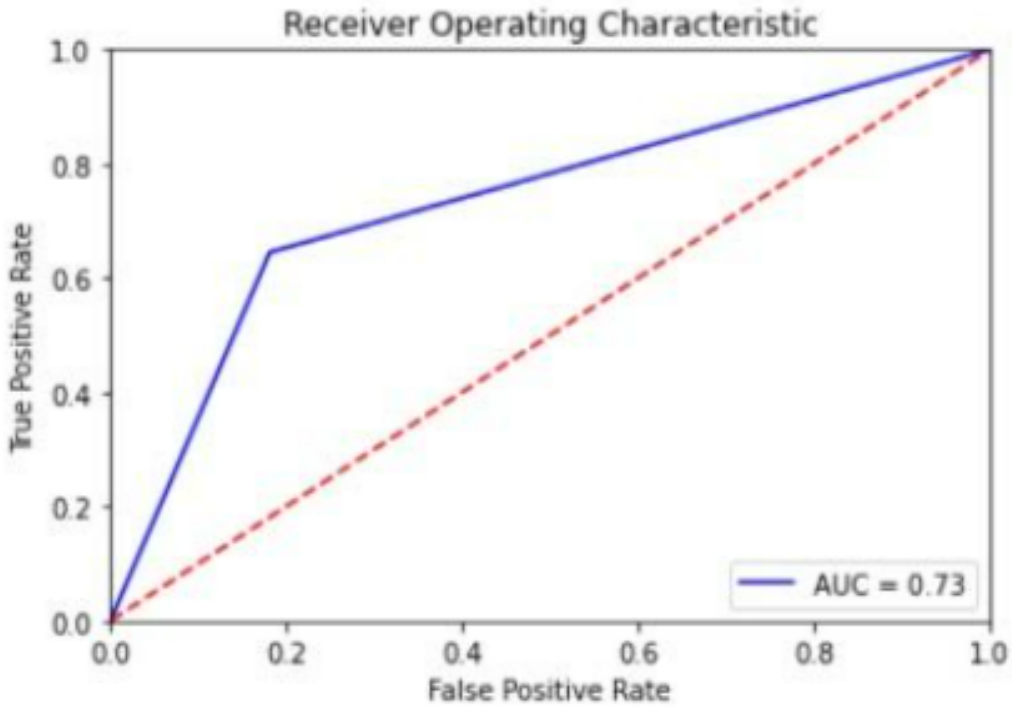

Figure 32: ROC curve for final model

Now the ROC curve along with the AUC score shows that our final model is quite well featured model with high satisfaction for detecting the TRUE negatives.

13 Conclusion

In our study, we have successfully modelled the Marketing Campaign using different variants of Logistic Regression.

Initially, we started with the data pre processing steps such as dealing with

Missing Data and Imbalanced Data. We found during these steps that the Marketing Campaign

Dataset suffered from various serious problems and necessary actions were needed

before modelling the data. We therefore used multiple data imputation techniques

and created synthetic samples to deal with issues of missing and imbalance data.

After obtaining a complete dataset, we found that there is severe multicollinearity in our dataset. This is due to the nature of the Variables in dataset which are

synthetic combinations of common financial Variables. To tackle the problem of

multicollinearity, we used VIF iterative algorithm for eliminating variables from our model.

Post elimination of Multicollinearity,we switched to the problem of Variable

Selection and implemented Stepwise Regression to reduce dimensionality of our

model.

14 References

- Introduction to Linear Regression Analysis, 5th Edition
  Douglas C. Montgomery, Elizabeth A. Peck, G. Geoffrey Vining
- MTH 416A:- Lecture Notes
- https://home.iitk.ac.in/ shalab/regression/Chapter14-Regression-LogisticRegressionModels.pdf
- Applied Regression Analysis, 3rd Edition by
  Norman R. Draper, Harry Smith