Optimizing retail marketing costs by classifying customers into unique segments using machine learning

By Vibodh Jadhav and Sneha Chandorkar

Every retail company runs marketing campaigns to provide exposure to their products. Significant amount of capital is spent in running these campaigns. In this project, we aim to reduce this capital to the minimum by predicting which customer is most likely to respond to the campaigns.

In this case, the company runs 5 consecutive campaigns. We look at the response to the campaigns. We take our focus off from the type of customers who don’t respond to all the campaigns consistently which reduces the spending on marketing. The ML algorithm decides this ‘type’ of customer from the training dataset.

Approach: Classification. We use binary feature of Campaign response as the target feature/variable(dependent variable) and select the most important features using EDA(Exploratory Data Analysis)

Features	Details
ID	Customer's unique identifier
Year_Birth	Customer's birth year
Education	Customer's education level
Marital_Status	Customer's marital status
Income	Customer's yearly household income
Kidhome	Number of children in customer's household
Teenhome	Number of teenagers in customer's household
Dt_Customer	Date of customer's enrollment with the company

Features	Details
MntFruits	Amount spent on fruits in last 2 years
MntMeatProducts	Amount spent on meat in last 2 years
MntFishProducts	Amount spent on fish in last 2 years
MntSweetProducts	Amount spent on sweets in last 2 years
MntGoldProds	Amount spent on gold in last 2 years
NumDealsPurchases	Number of purchases made with a discount
Recency	Number of days since customer's last purchase
Complain	1 if the customer complained in the last 2 years, 0 otherwise
MntWines	Amount spent on wine in last 2 years

Data Preprocessing

Data Cleaning is the first step of Preprocessing. Without good data, the model will not give good results.
We get rid of-

- Missing values (we impute them with mean, here only Income feature has missing values)
- Outliers (we treat outliers by using the IQR method)

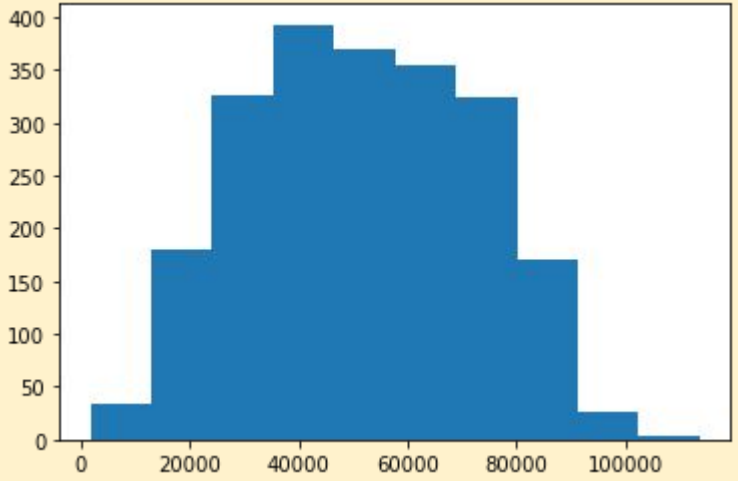
Feature Selection

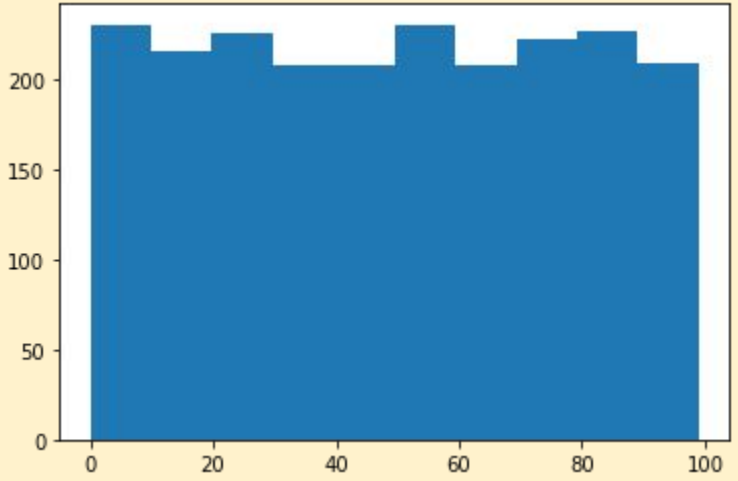
Feature selection is one of the most important steps of the workflow. It significantly boosts the predictive power of the ML algorithms.
We do a three fold analysis for selecting the features-

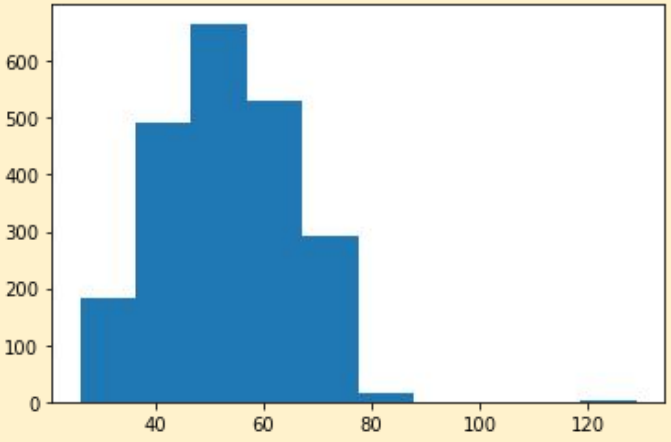
- Univariate analysis by Visualization
- Bivariate analysis by Visualization
- Bivariate analysis using hypothesis testing(Z test)

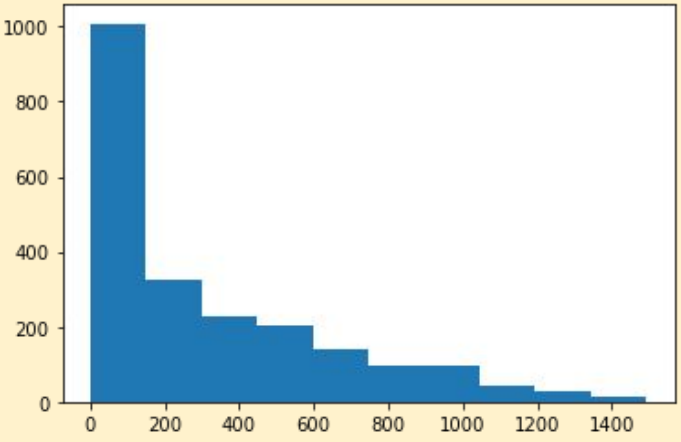
Univariate analysis

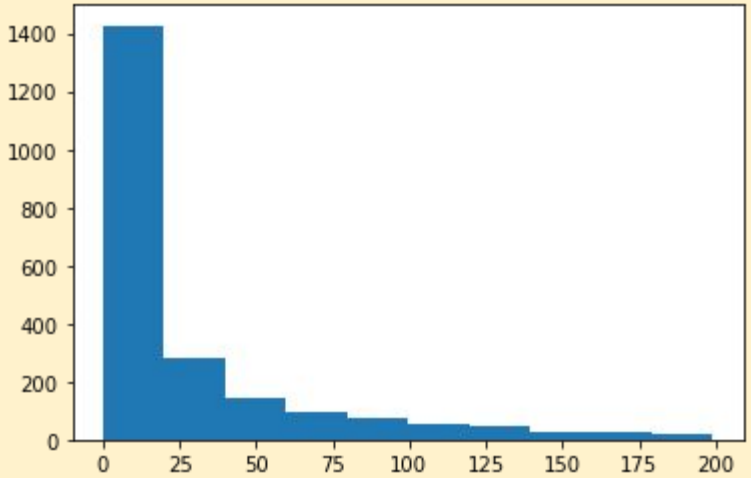
We do univariate analysis to see whether is feature is normally distributed or not.
Let us look at the graphs-

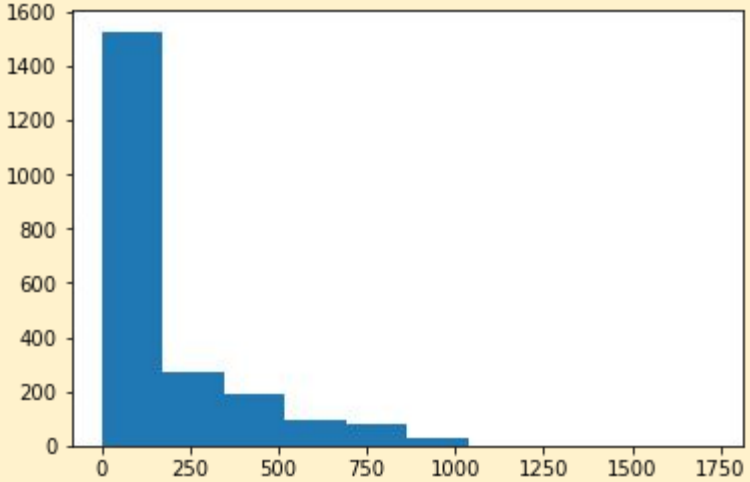

400
350
300
250
200
150
100
50
0
0
20000
40000
60000
80000
100000

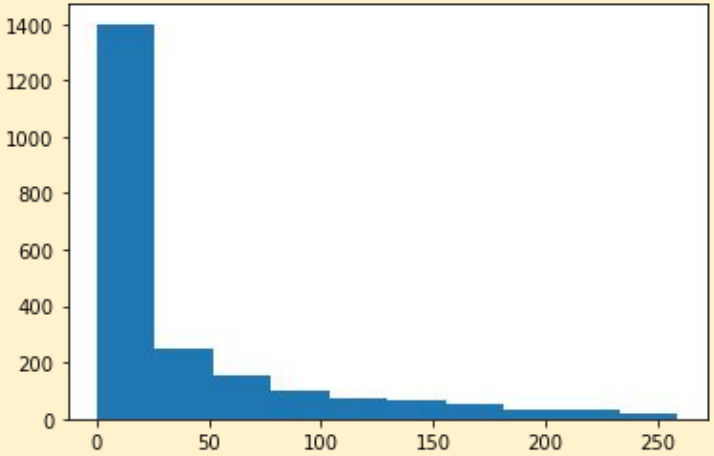

200
150
100
50
0
0
20
40
60
80
100

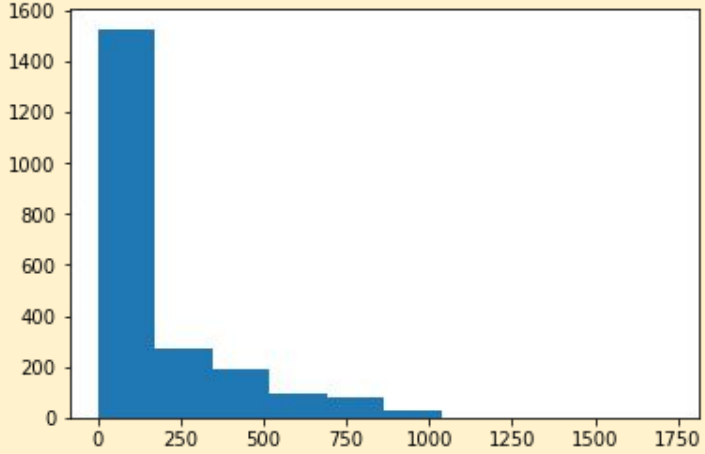

600
500
400
300
200
100
0
40
60
80
100
120


1000
800
600
400
200
0
0
200
400
600
800
1000
1200
1400


1400
1200
1000
800
600
400
200
0
0
25
50
75
100
125
150
175
200

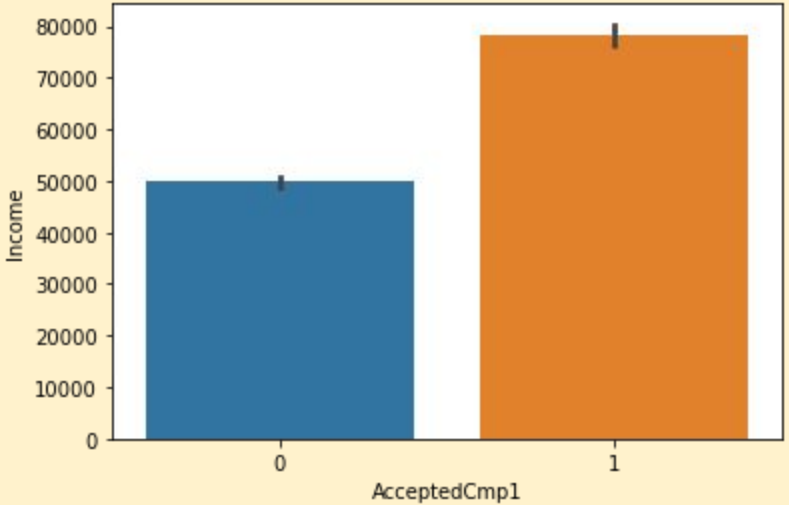

1600
1400
1200
1000
800
600
400
200
0
0
250
500
750
1000
1250
1500
1750

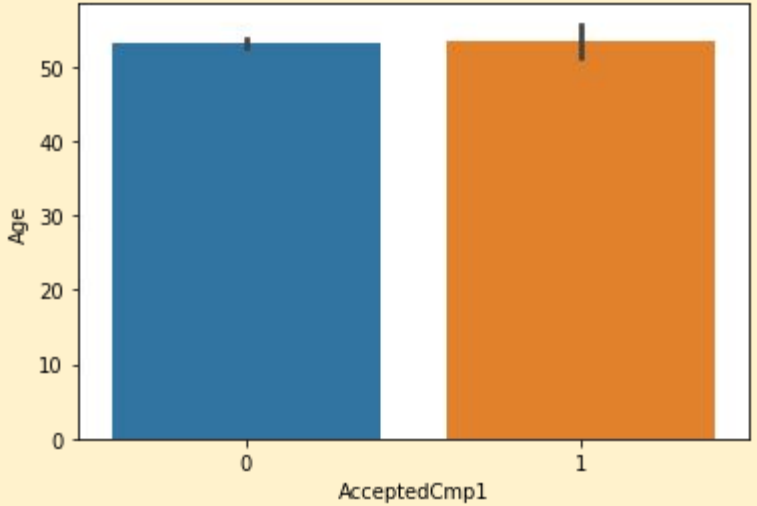

1400
1200
1000
800
600
400
200
0
0
50
100
150
200
250

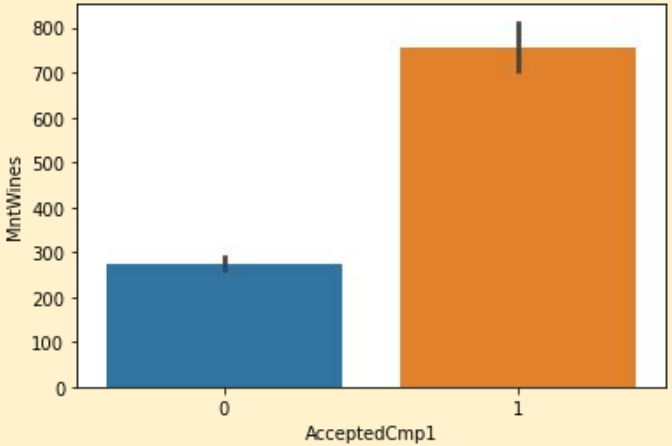

1600
1400
1200
1000
800
600
400
200
0
0
250
500
750
1000
1250
1500
1750

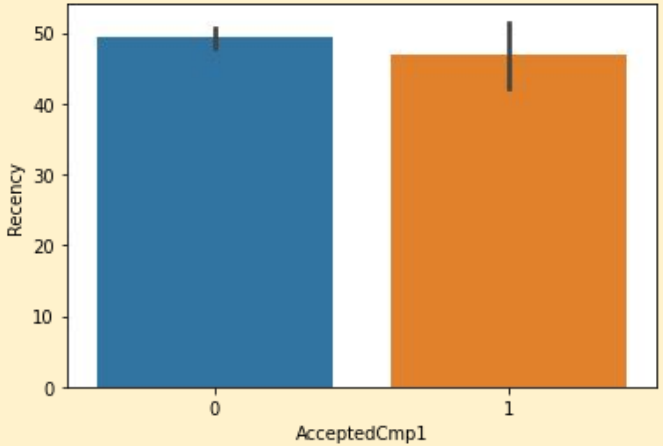
Bivariate analysis using Visualization

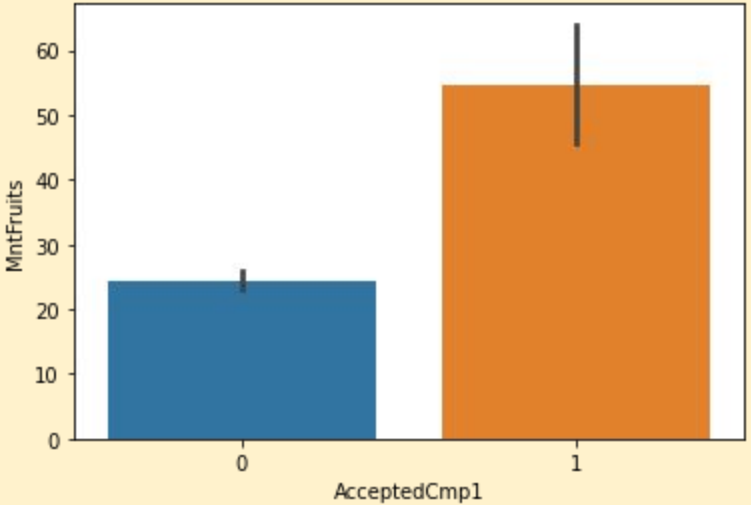
Bivariate analysis is used to understand the relationship between the binary target variable of Campaign response and the all the other features. This helps in feature selection. Let us look at the graphs-

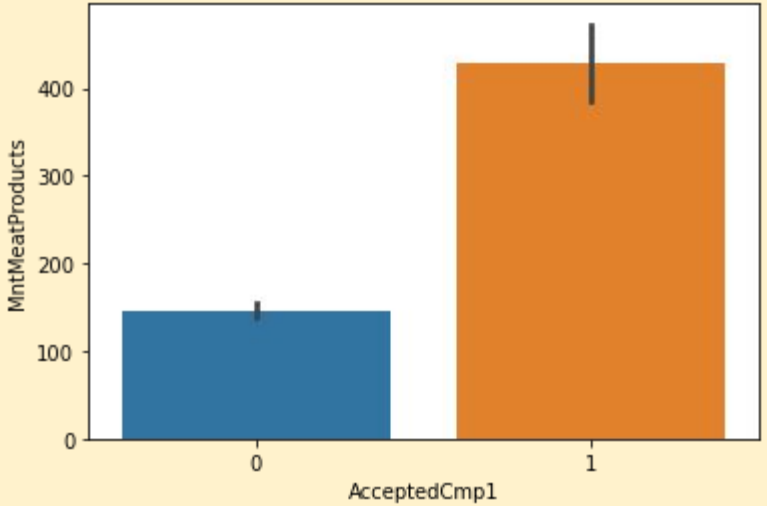

Income
80000
70000
60000
50000
40000
30000
20000
10000
0
0
1
AcceptedCmp1

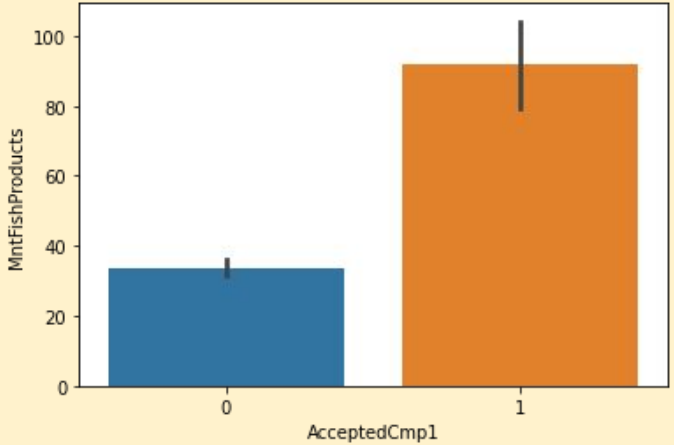

Age
50
40
30
20
10
0
0
1
AcceptedCmp1

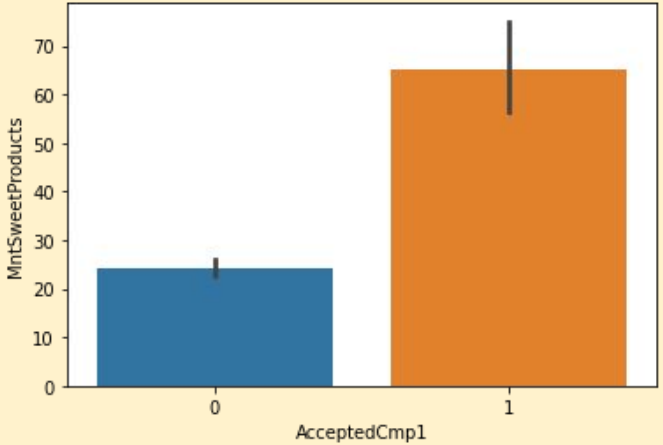

MntWines
800
700
600
500
400
300
200
100
0
0
1
AcceptedCmp1

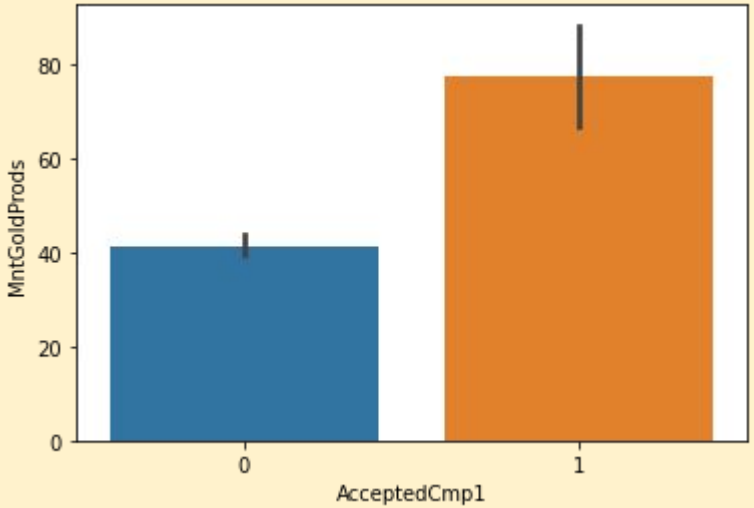

Recency
50
40
30
20
10
0
0
1
AcceptedCmp1

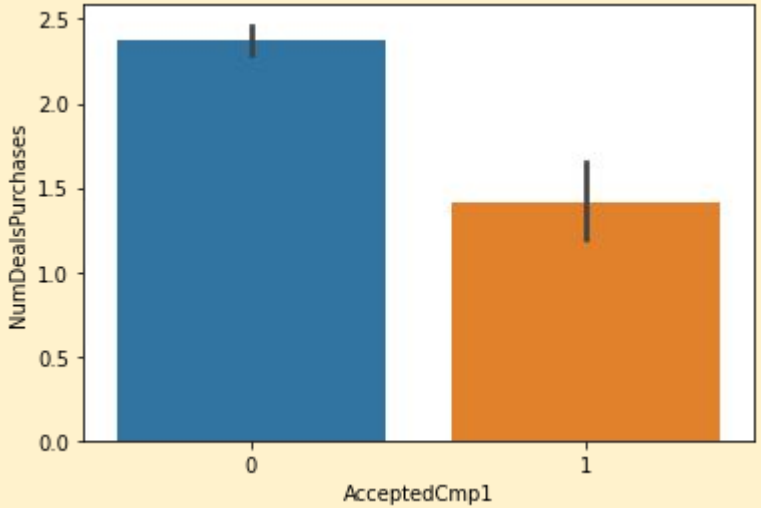

MntFruits
60
50
40
30
20
10
0
0
1
AcceptedCmp1

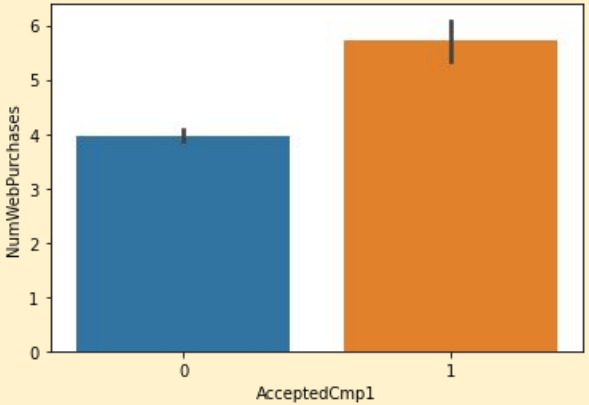

MntMeatProducts
400
300
200
100
0
0
1
AcceptedCmp1

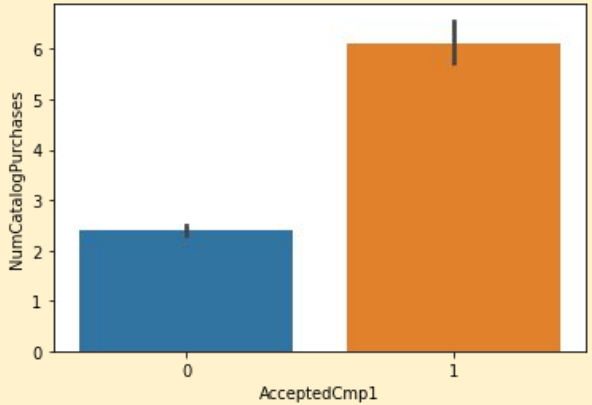

MntFishProducts
100
80
60
40
20
0
0
1
AcceptedCmp1

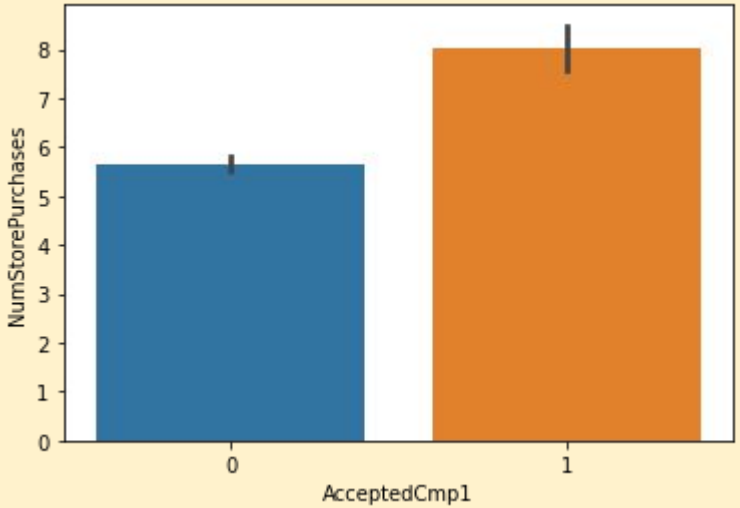

MntSweetProducts
70
60
50
40
30
20
10
0
0
1
AcceptedCmp1

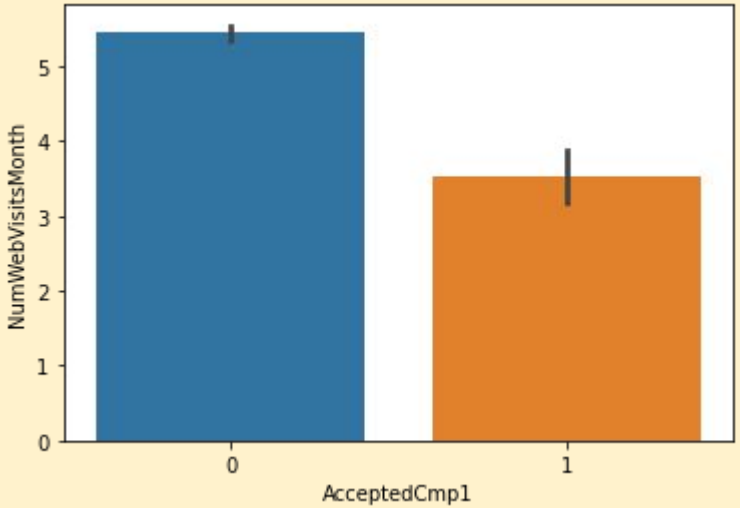

MntGoldProds
0
20
40
60
80
0
1
AcceptedCmp1


NumDealsPurchases
0.0
0.5
1.0
1.5
2.0
2.5
0
1
AcceptedCmp1


NumWebPurchases
0
1
2
3
4
5
6
0
1
AcceptedCmp1


NumCatalogPurchases
0
1
2
3
4
5
6
0
1
AcceptedCmp1


NumStorePurchases
0
1
2
3
4
5
6
7
8
0
1
AcceptedCmp1


NumWebVisitsMonth
0
1
2
3
4
5
0
1
AcceptedCmp1

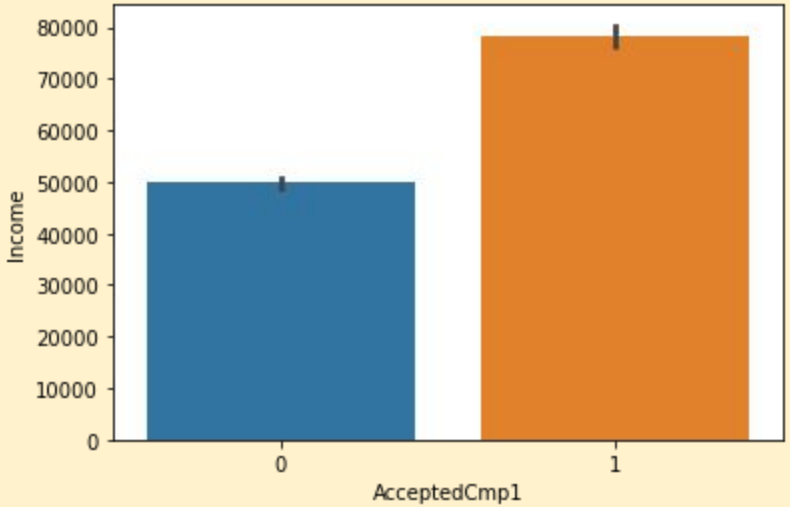
Bivariate analysis using hypothesis testing

Visualization is not enough! Sometimes the relationship between two variables is influenced by another hidden variable(s). To confirm the relationship, we use statistical hypothesis testing with the Z test.

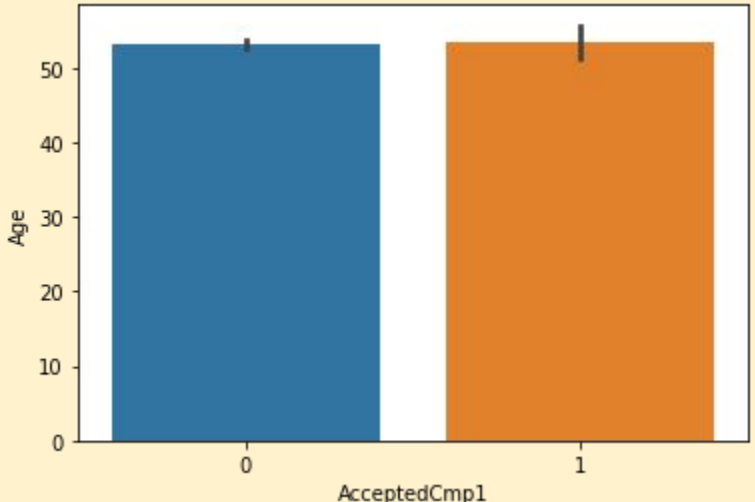
Here, the null hypothesis is- The difference of average between the two groups(for eg- income group who responded to the campaign and income group who didn’t respond to the campaign) is zero.

The alternate hypothesis is- The difference is significant.

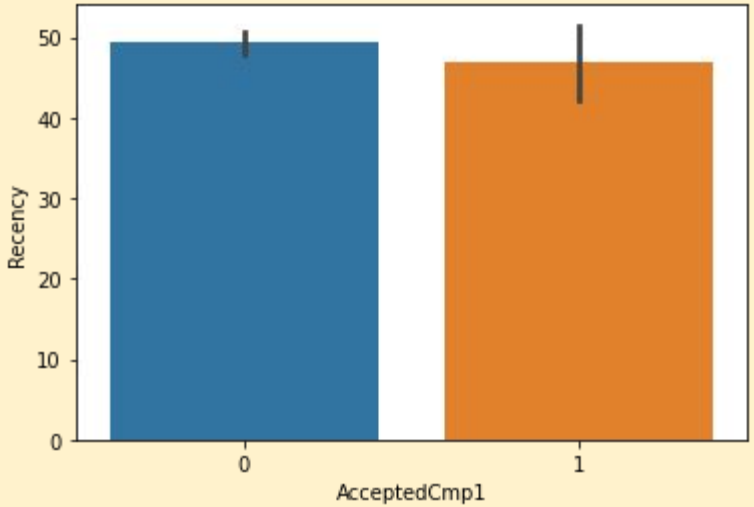
For 95% significance, we will choose the features whose z score>1.96 OR z-score<-1.96(implying p-value<0.05). Even after choosing from them, we will reduce the number of features for the model(s) to improve their predictive power.



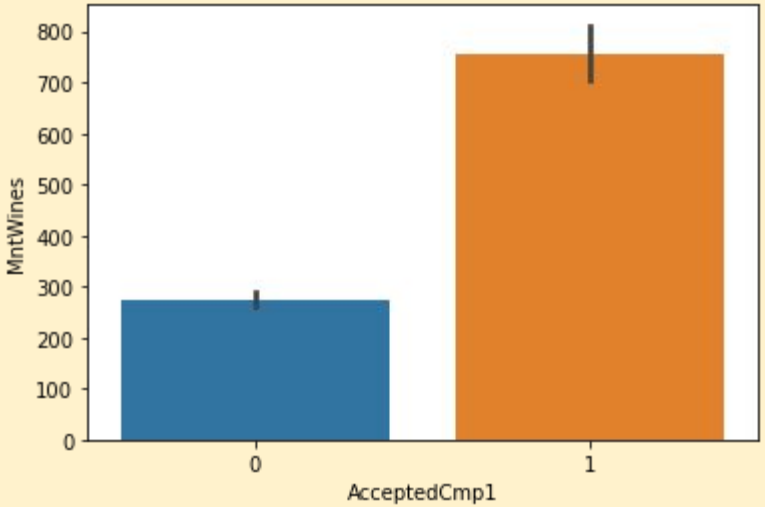
Z score = 26.39877661434138



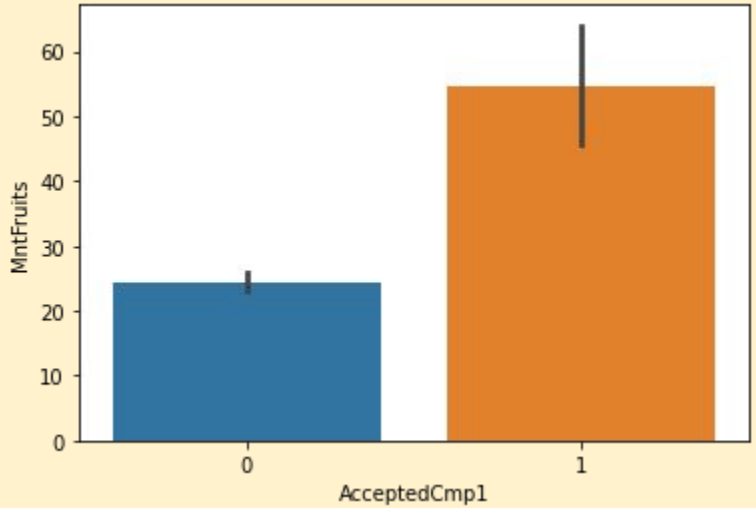
Z score = 0.1635561091405724



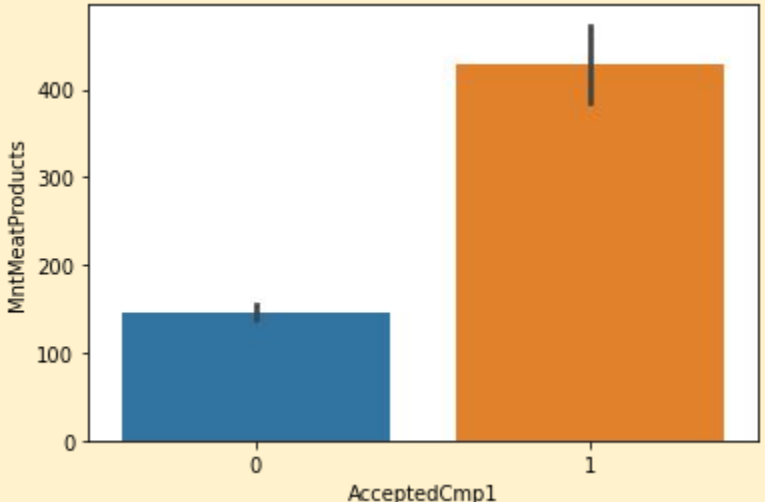
Z score =1.0236788760355076



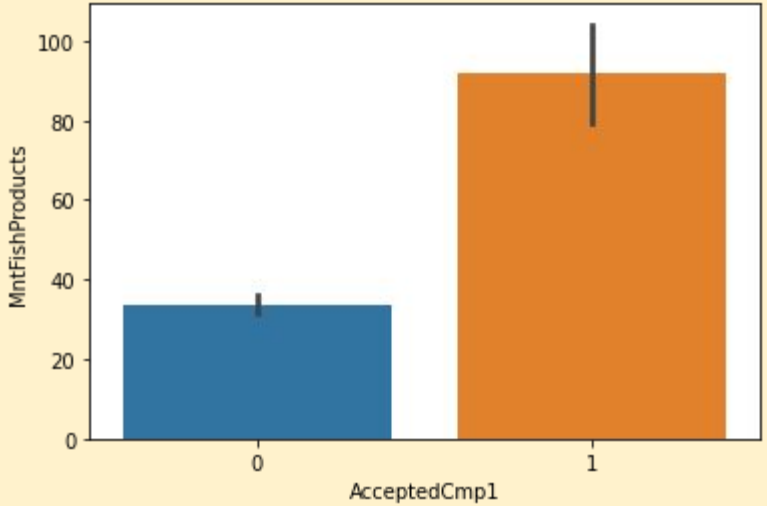
Z score =16.6105544305937



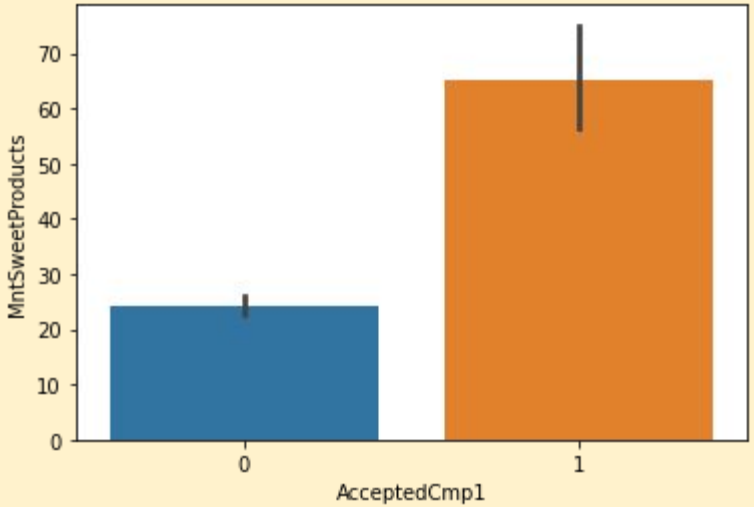
Z score = 6.439549296080812



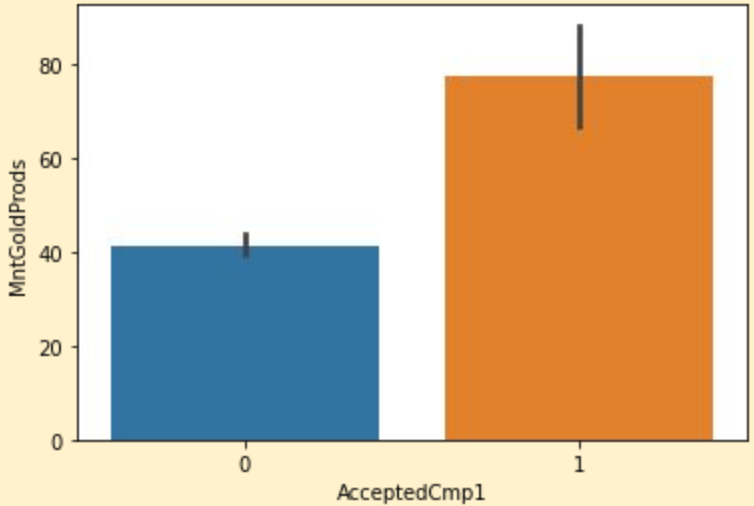
Z score =12.47005085368447



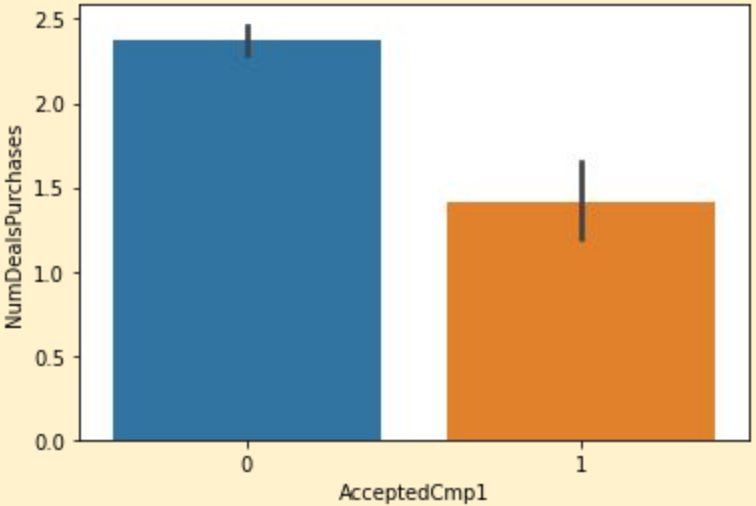
Z score =9.239671900500861



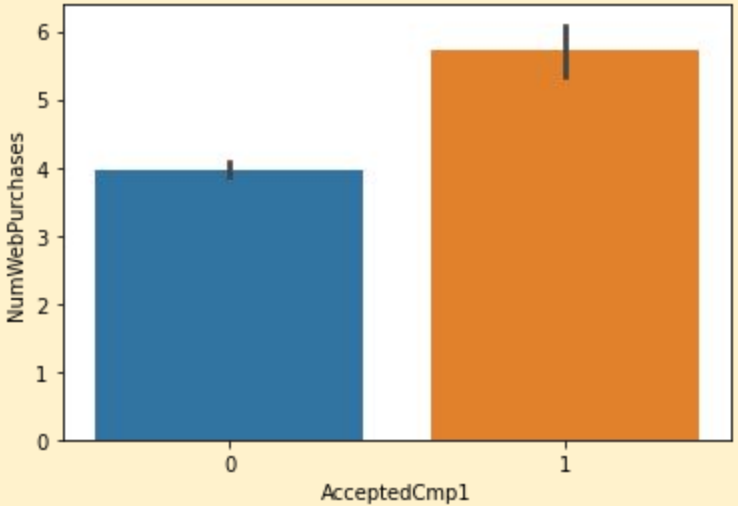
Z score =8.613626999807142



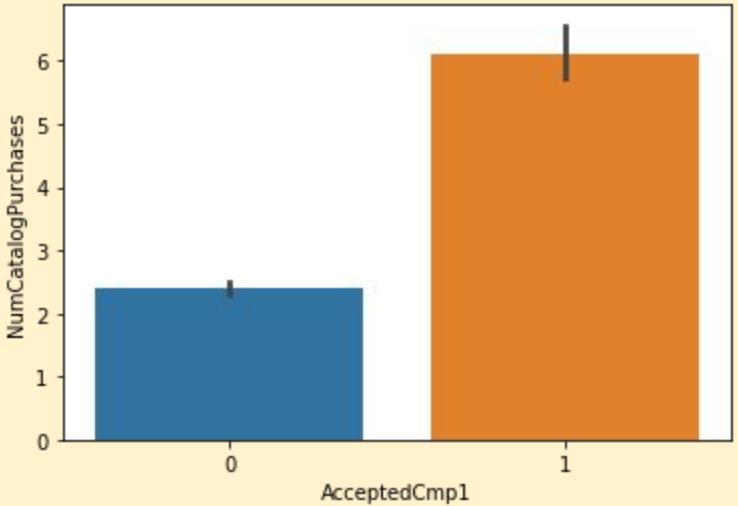
Z score =6.443338308243719



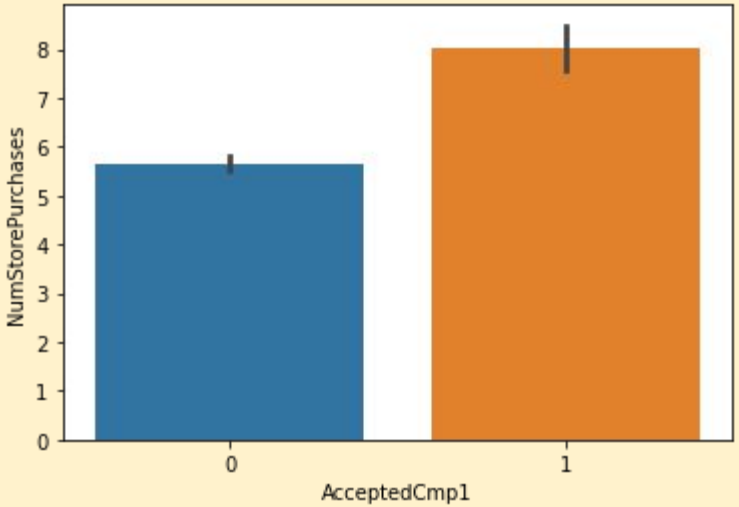
Z score
=-7.802102004939067



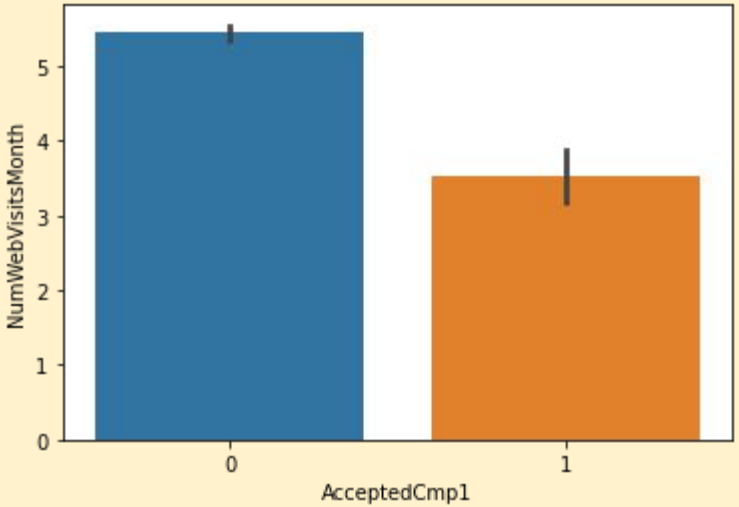
Z score
=9.020796919605909



Z score
=16.85259431730384



Z score
=10.054476335338265



Z score
=-10.80507967575157

Scaling

We scale the continuous features so that the model doesn’t get biased towards the features that have values larger in scale. For eg- Income and MntWines(Amount spent on Wine by the customer in the last 2 years)

Encoding of categorical variables

We use label encoding to encode the two important categorical variables in this dataset-

- Education
- Marital_Status

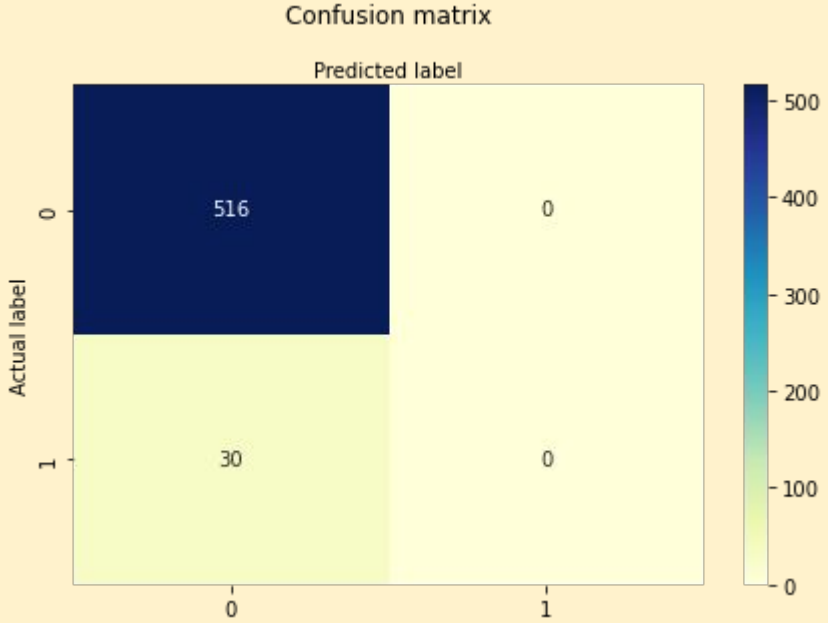
Model fitting

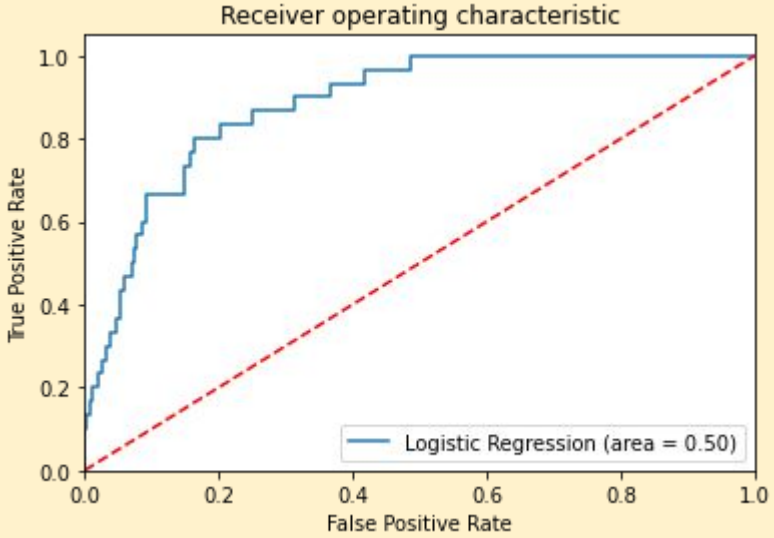
We split the training and test dataset in the ratio 75:25.
We use the following Machine learning algorithms for classification-

- Logistic Regression
- Decision Tree
- KNN
- SVM

Model Evaluation

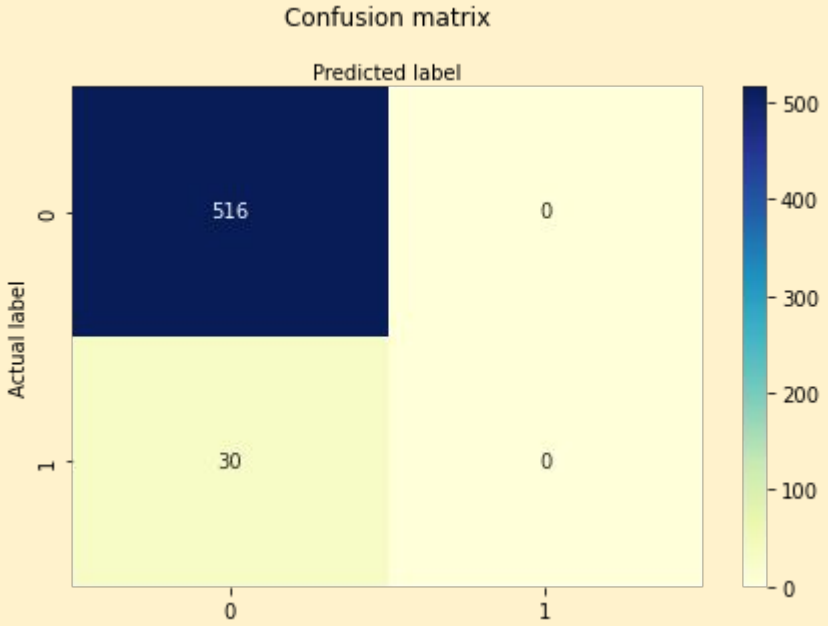
Logistic Regression





Model evaluation

Decision Tree



Evaluation metric

Here, we use accuracy as a metric to decide which model is giving us the best results-

Model	Accuracy
Logistic Regression	0.945054945054945
Decision Tree	0.9194139194139194
KNN	0.9432234432234432
SVM	0.945054945054945

We can see that SVM and Logistic Regression are giving the best results so they will be our preferred models for the predicting the campaign responses for the remaining 4 campaigns using the same methods.

Thank you!