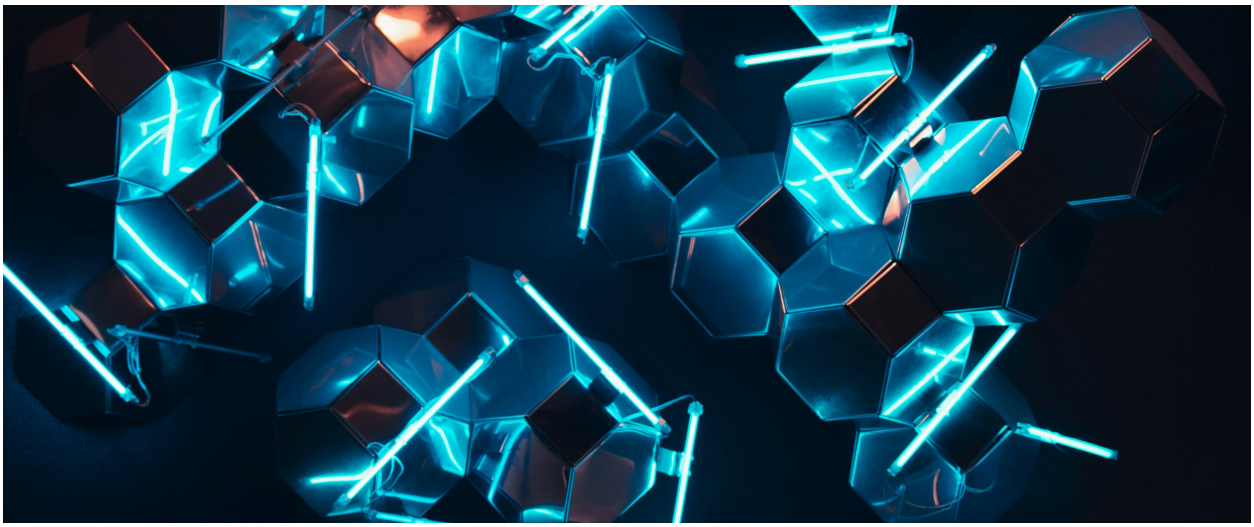


ML Project Report

OVERVIEW

A food company wants to produce the highest profit for the next direct marketing campaign,scheduled for the next month.
A pilot campaign involving 2.240 customers was carried out, customers who bought the offer were properly labeled. The total cost of the sample campaign was 6.720MU and the revenue generated by the customers who accepted the offer was 3.674MU. Globally the campaign had a profit of -3.046MU and the success rate of the campaign was 15%.

Feature	Description
AcceptedCmp1	1 if costumer accepted the offer in the 1st campaign, 0 otherwise
AcceptedCmp2	1 if costumer accepted the offer in the 2nd campaign, 0 otherwise
AcceptedCmp3	1 if costumer accepted the offer in the 3rd campaign, 0 otherwise
AcceptedCmp4	1 if costumer accepted the offer in the 4th campaign, 0 otherwise
AcceptedCmp5	1 if costumer accepted the offer in the 5th campaign, 0 otherwise
Response (target)	1 if costumer accepted the offer in the last campaign, 0 otherwise
Complain	1 if costumer complained in the last 2 years
DtCustomer	date of customer's enrollment 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 in the last 2 years
MntSweetProducts	amount spent on sweet products in the last 2 years
MntWines	amount spent on wines 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

Table 1: Meta-data table

OBJECTIVES

The objective is of the team is to develop a model that predicts customer behavior and to apply it to the rest of the customer base.

Moreover, other than maximizing the profit of the campaign, the CMO is interested in understanding to study the characteristic features of those customers who are willing to buy the gadget. The steps are:

- Data Exploration;

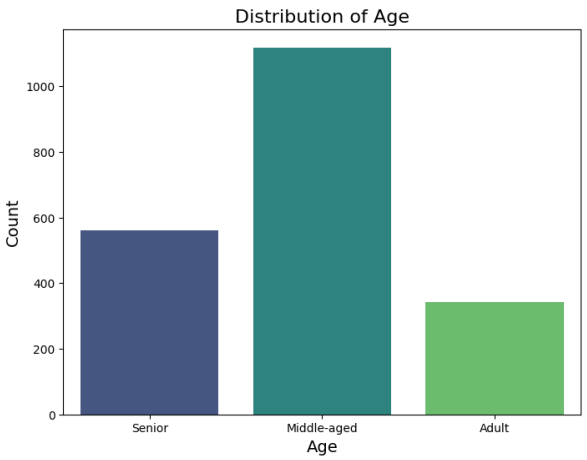
- Segmentation;
- Classification Model;

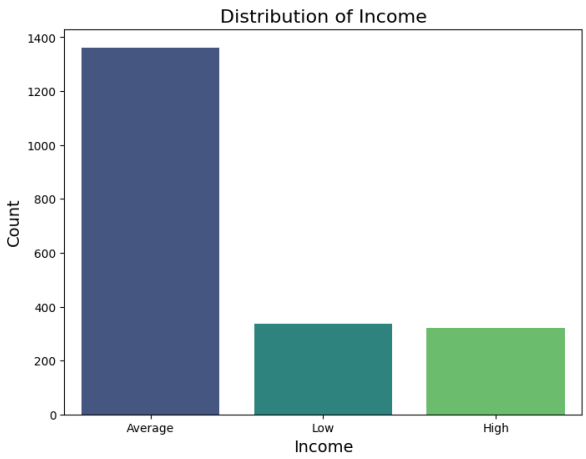
DATA EXPLORATION

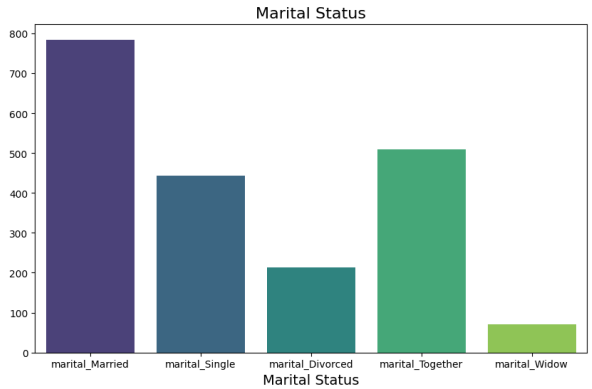
The majority of the customers are married and highly educated, there is no significant correlation between those two categorical information. Also the Majority of customers are middle-aged and have an average income. The Age and Income bins are defined as followed:

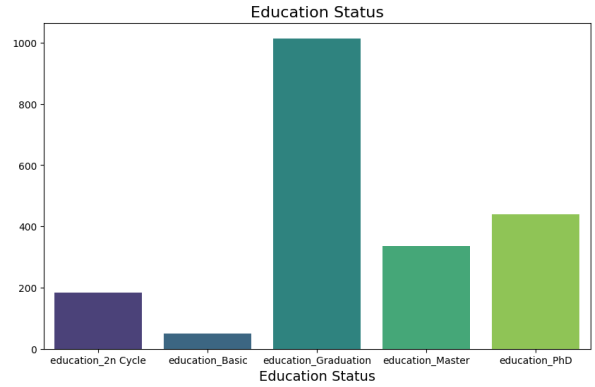
- Low 0 - 30k
- Average 30k - 75k
- High 75k - Max

- Adult 20 - 40 years
- Middle-aged 40 - 60 years
- Senior 60 - Max years



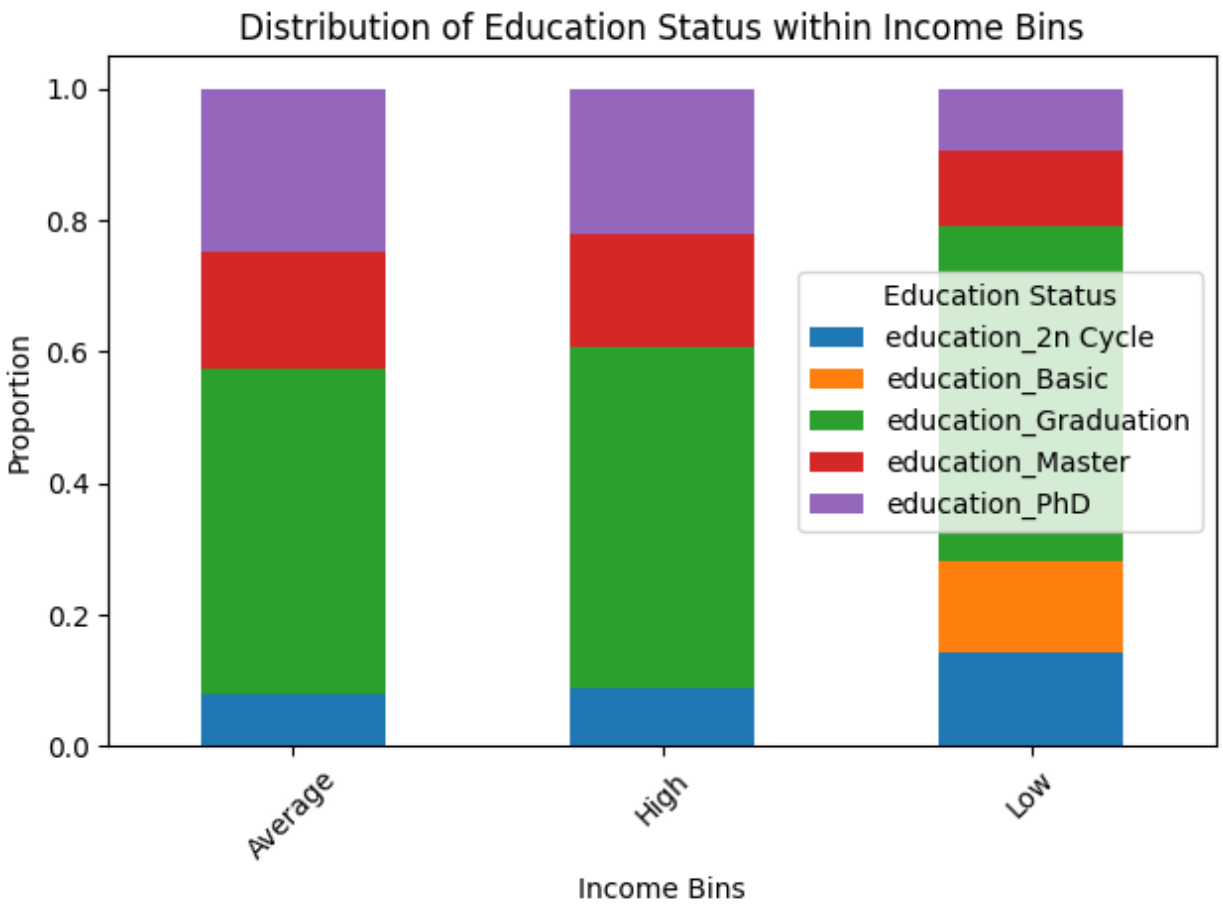




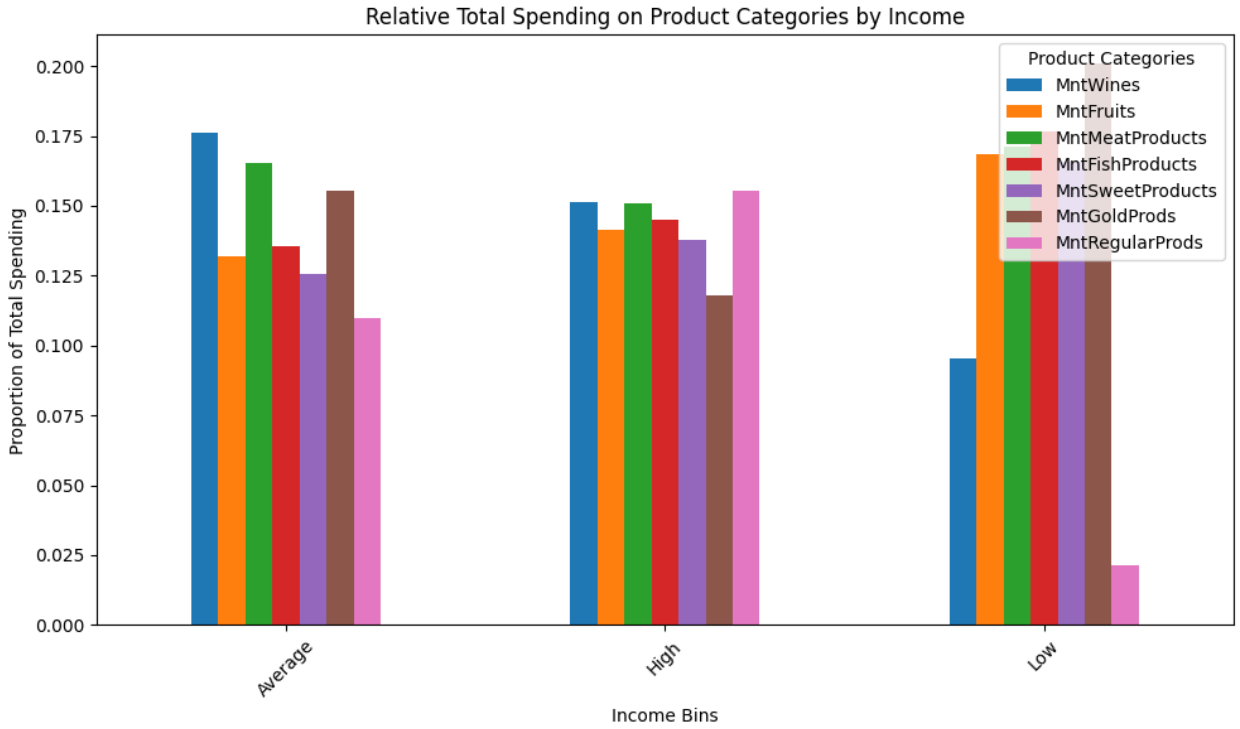


CUSTOMER SEGMENTATION

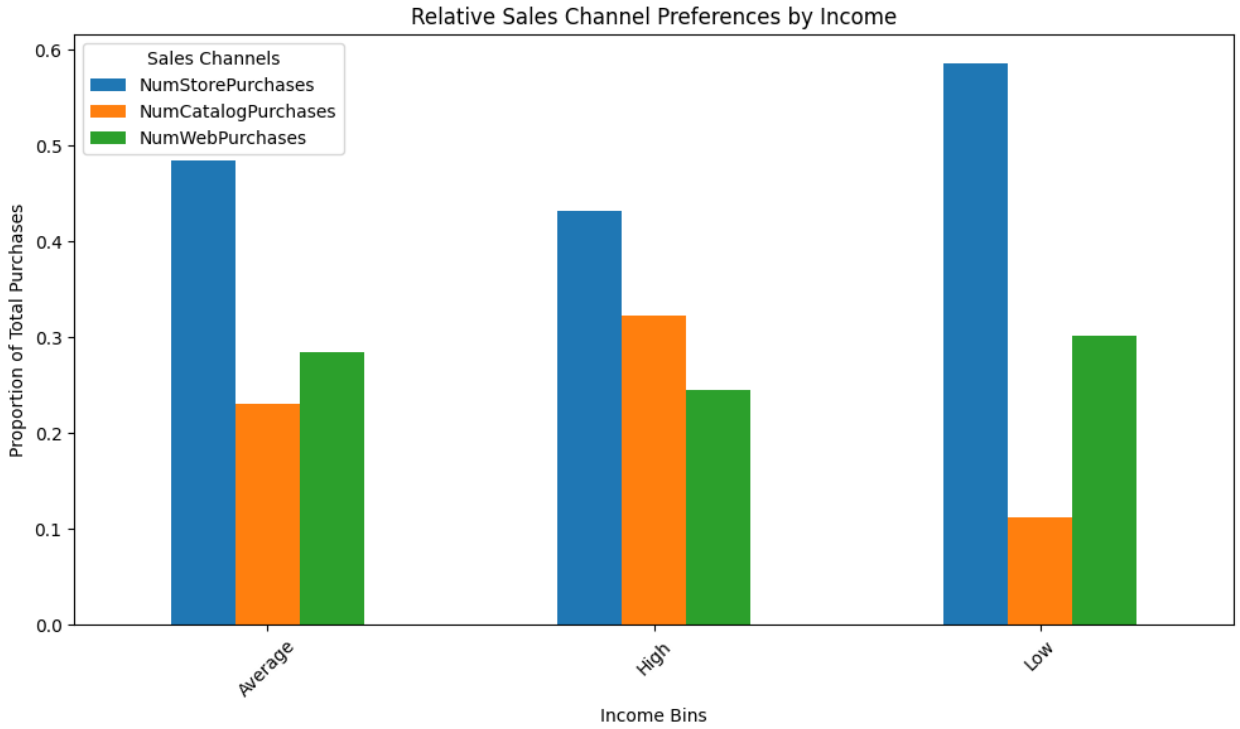
Customer segmentation for customer income related to their education status shows that all of the customers with basic education have a low income. It also show that many low earners have high education and that there are fewer customers with a phd in the high income cluster than in the average income cluster.



Here we can see the proportion of total spending on each product category relative to the total spending across all product categories for each income bin. Interestingly the low income spend much more on gold products than on regular products. In contrast the high income bin spends more on regular products than on gold products



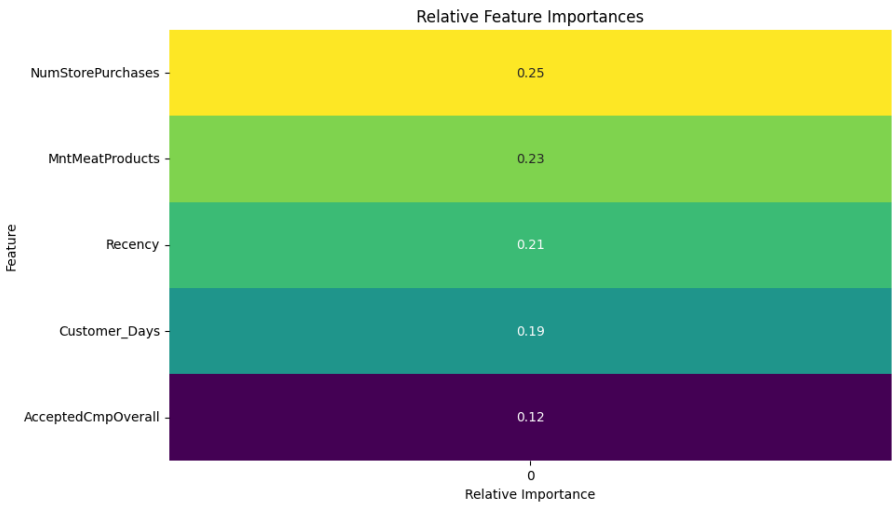
Here we see the proportion of purchases from each sales channel relative to the total purchases across all sales channels for each income bin. We see that Store Pruchases are the most popular sales channel for all income clusters but the higher the income class the more popular seems to get the catalog purchases.



CLASSIFICATION MODEL RESULTS

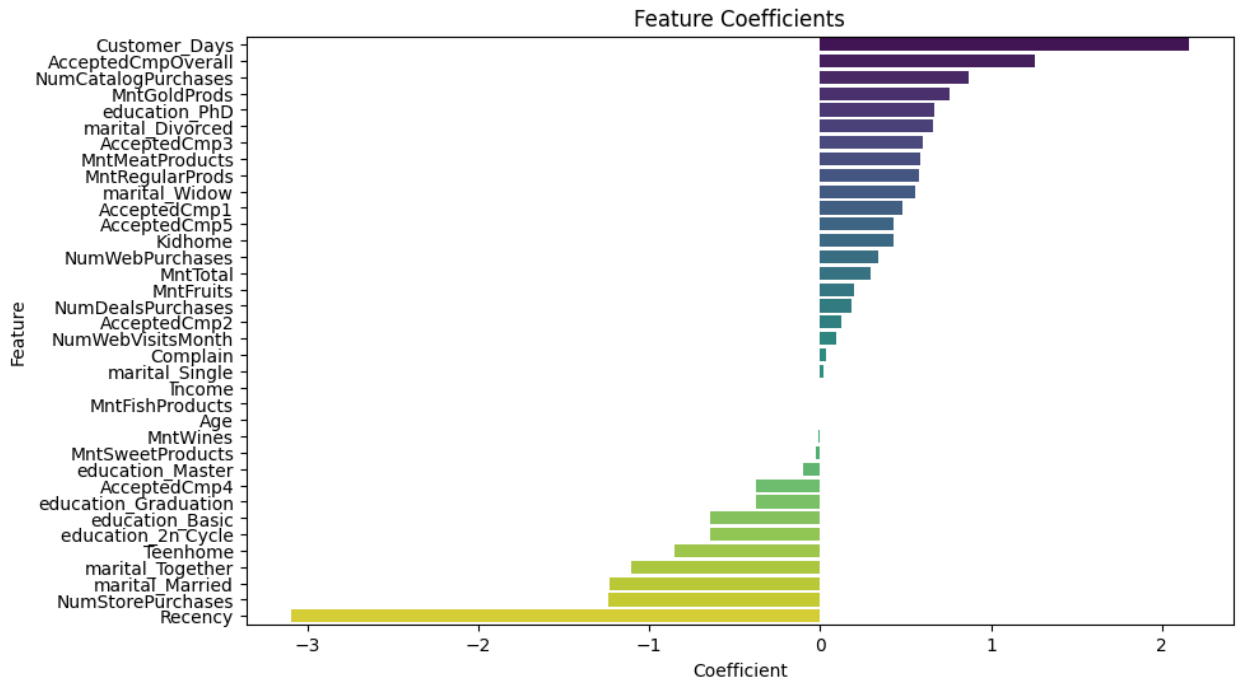
For the campaign I developed a predictive model that classifies if the customer will accept the offers or not. I can classify with 89% of accuracy. (rounded value)

The Model was trained with the shown features, showing their relative importance.

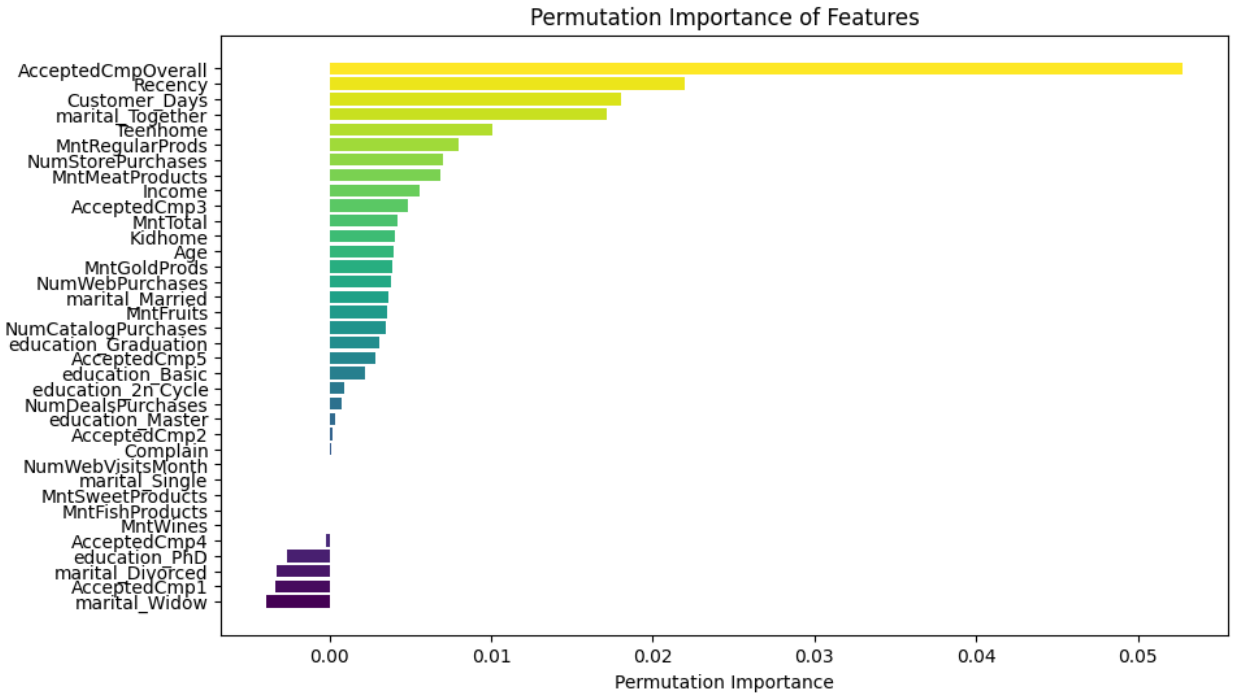


However when I trained the model without feature selection methods we can see different feature importances.

- **Positive coefficient**: An increase in the predictor variable leads to an increase in the target variable.
- **Negative coefficient**: An increase in the predictor variable leads to a decrease in the target variable.

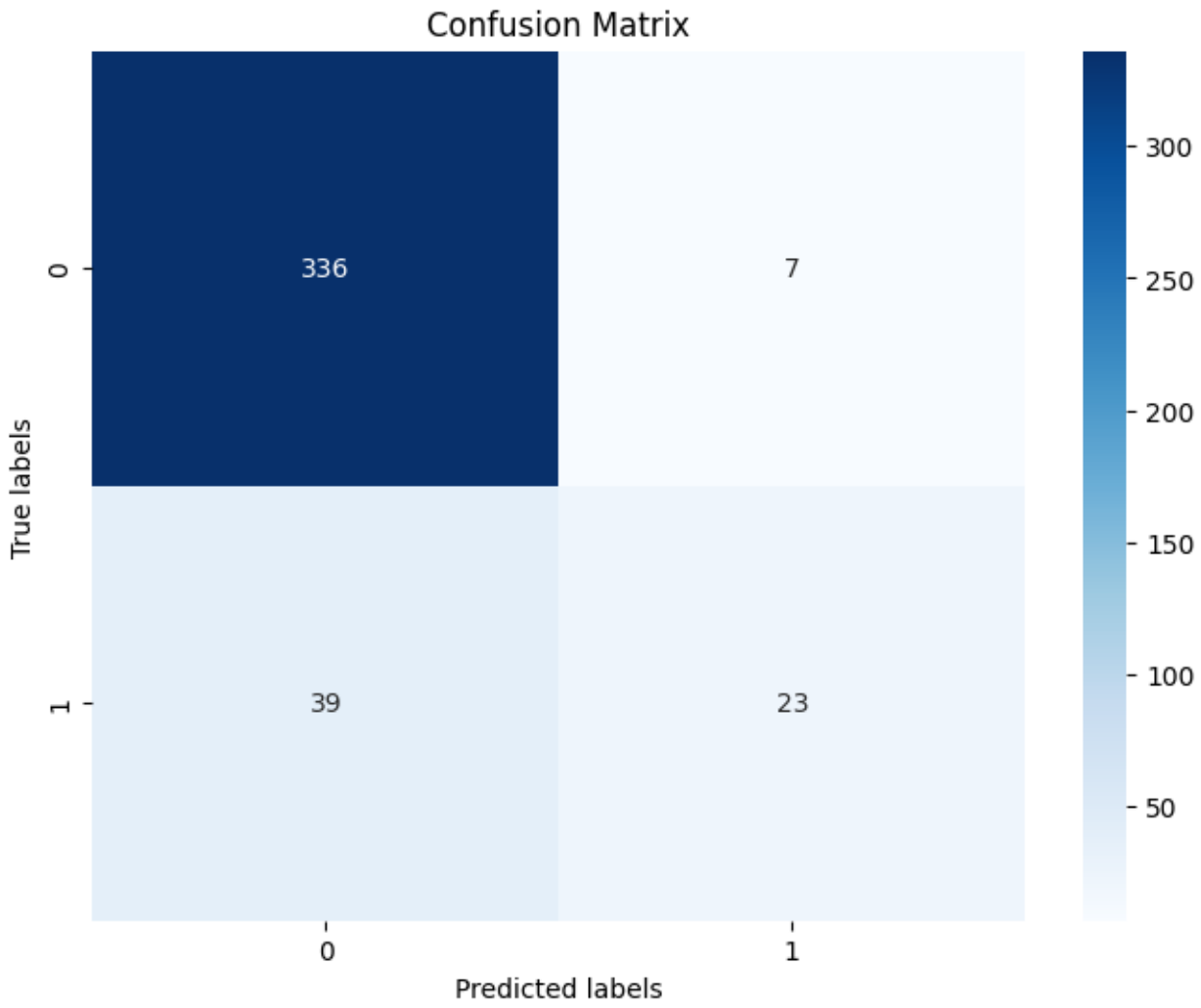


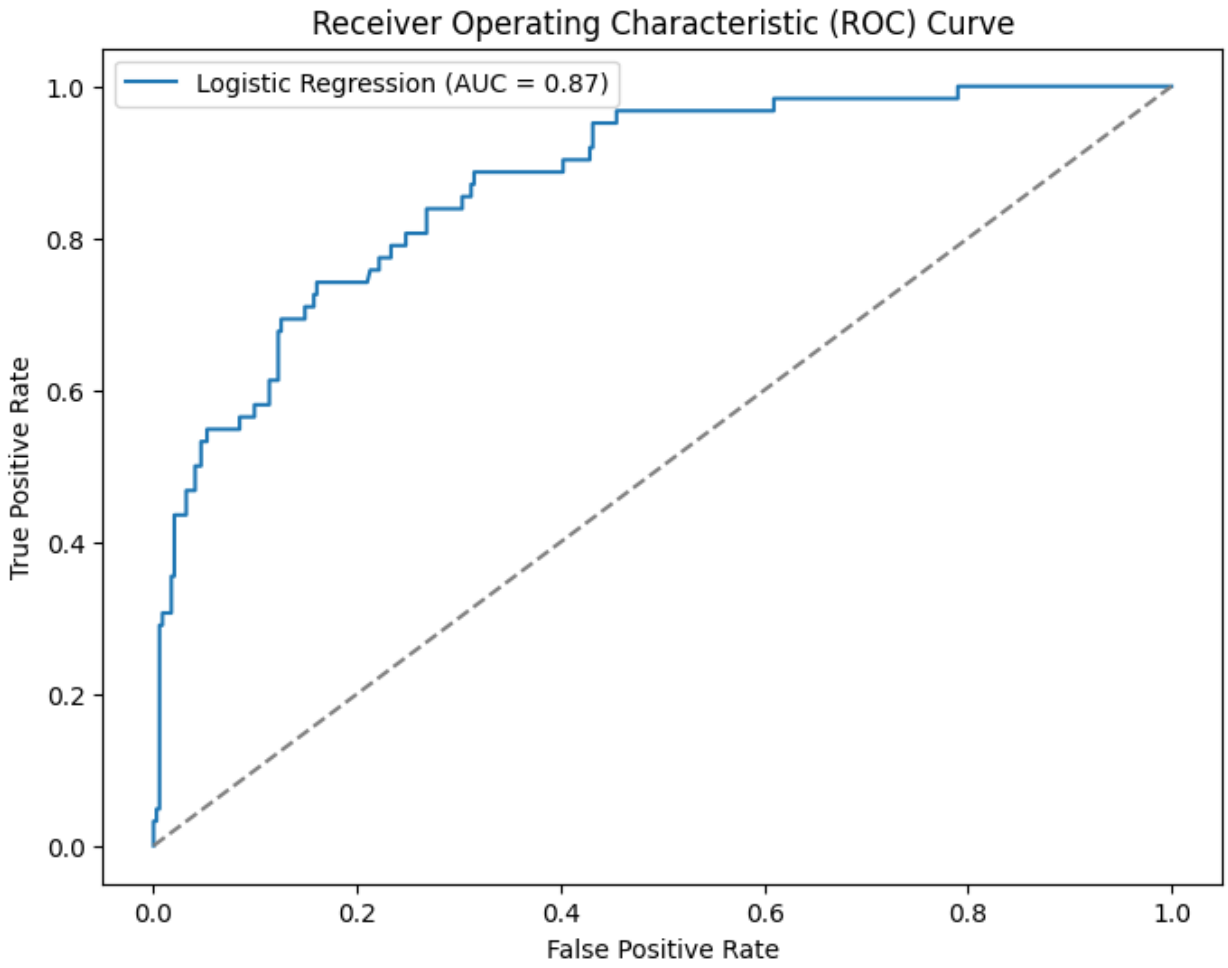
Aside from coeficient based feature importance we can see below the permutation importance of features since it is more interpretable beacuse it captures both linear and nonlinear relationships between features and the target variable and handles correlated features more effectively than coefficient-based methods.



The Logistic Regression Model was chosen, since it performed better than other models. Below we can see the most important metrics

Metrics	Values
Accuracy	0.888
Mean cross-validated score	0.882
Precision	0.766
Recall	0.370
F1 Score	0.5
ROC AUC Score	0.871


Confusion Matrix
336
7
39
23
0
1
True labels
0
1
Predicted labels
300
250
200
150
100
50


Receiver Operating Characteristic (ROC) Curve
Logistic Regression (AUC = 0.87)
True Positive Rate
False Positive Rate
0.0
0.2
0.4
0.6
0.8
1.0