Marketing Campaign Prediction and Customer Segmentation Project

Student: Changping Chen (002109149)

chen.changp@northeastern.edu

Market Campaign Analysis

iFood is the lead food delivery app in Brazil, present in over a thousand cities. Keeping a high customer engagement is the key for growing and consolidating the company's position as the market leader.Globally, the company had solid revenues and a healthy bottom line in the past 3 years, but the profit growth prospects for the next 3 years are not promising. For this reason, several strategic initiatives are being considered to reverse this situation. One is to improve the performance of marketing activities, with a special focus on marketing campaigns.

The objective of the team is to build a predictive model that will produce the highest profit for the next direct marketing campaign, scheduled for the next month. The new campaign, sixth, aims at selling a new gadget to the Customer Database. To build the model, a pilot campaign involving 2.240 customers was carried out. The customers were selected at random and contacted by phone regarding the acquisition of the gadget. During the following months, customers who bought the offer were properly labeled. The total cost of the sample campaign was 6.720 units and the revenue generated by the customers who accepted the offer was 3.674 units. Globally the campaign had a profit of -3.046 units. The success rate of the campaign was 15%. The objective of the team is to develop a model that predicts customer behavior and to apply it to the rest of the customer base. Hopefully the model will allow the company to cherry pick the customers that are most likely to purchase the offer while leaving out the non-respondents, making the next campaign highly profitable. 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.

Data Collection

The data source is masked iFood customer and order data, including customers' demographic features and if they accepted the offer. Beyond that, industry benchmarks like the average revenue generated from a customer, and the average cost for sample market campaigns of a customer, and the detailed marketing campaign channels are also considered.

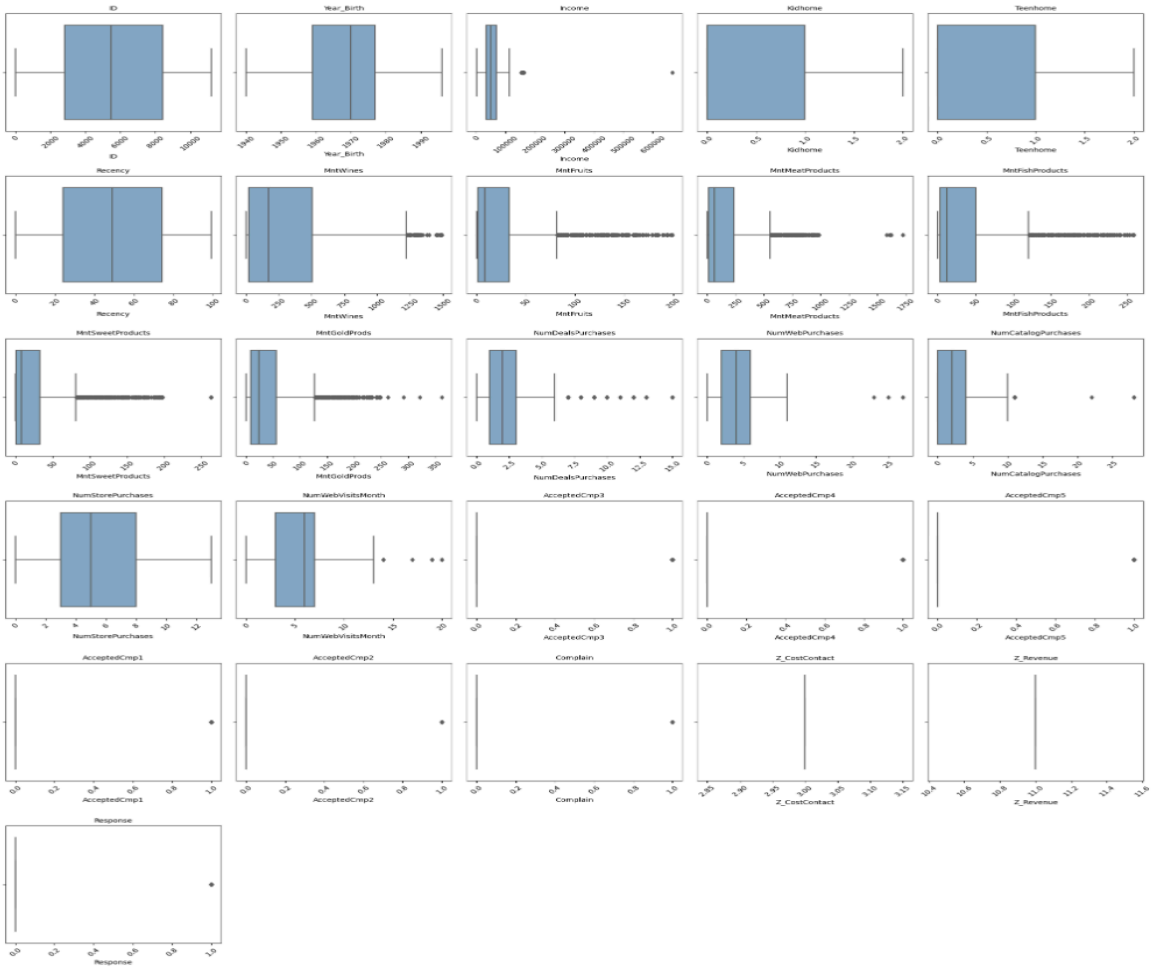
Data Cleaning and Exploration

Step 1: I look at the information of the whole dataset, it has a total 2240 records with 29 features. The data type of features are int64, float64, and object. The features include customers demographic features and the response of the previous fifth market campaigns and the target is the last market campaign response.

Step 2: Then I check the missing and duplicate values. Fortunately, there are only 24 records of income missing values, and no duplicate values. For these missing values, I impute them with

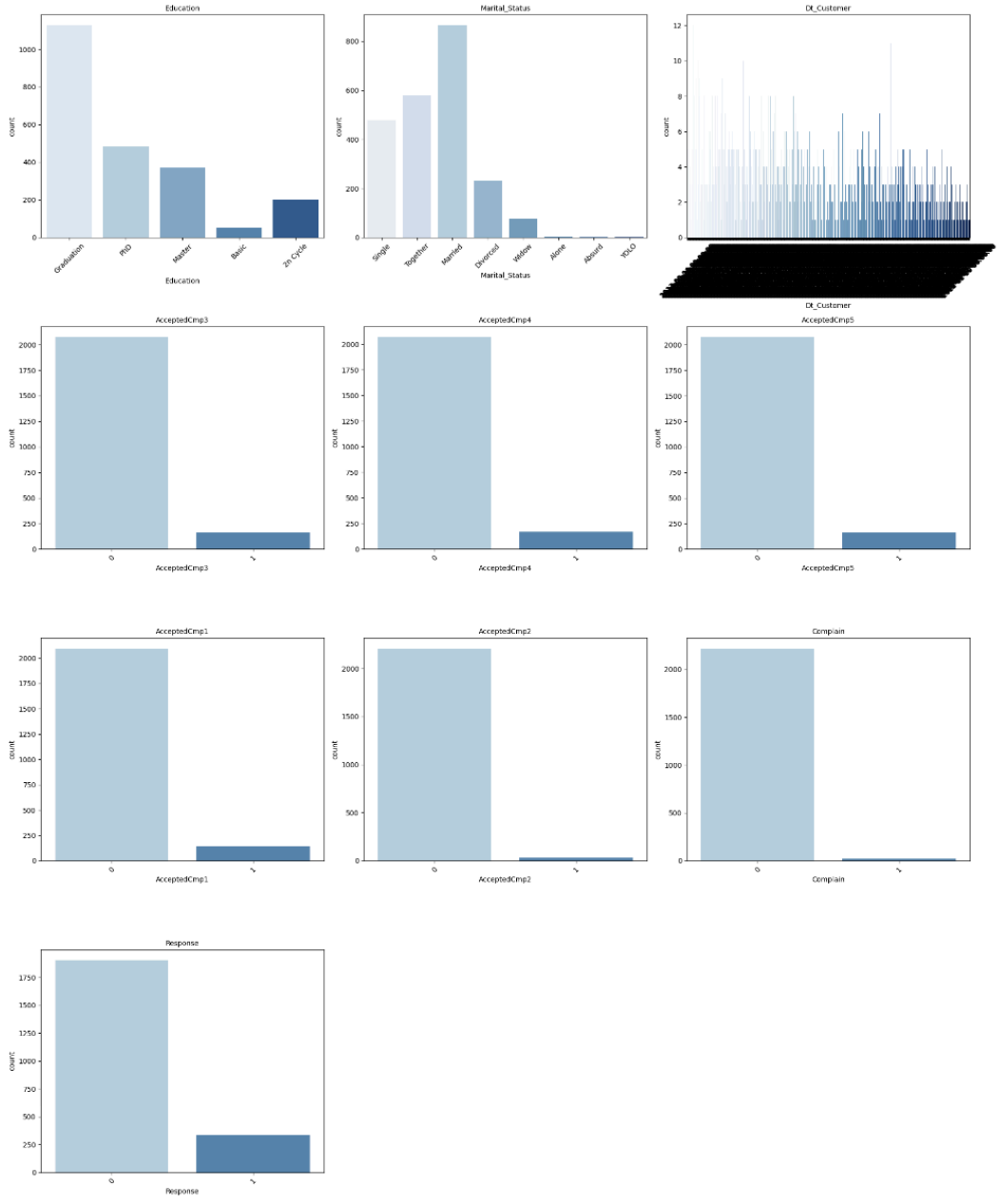
mode value. And then I check the description of the dataset to check the count, mean, standard deviation and also percentile of values. Based on the description, there are some unreasonable values in the date of birth part, for example, some people's age is over 130 based on their date of birth. After checking, I found only 3 customers' dates of birth are out of scope, so we can exclude these unreasonable values directly.

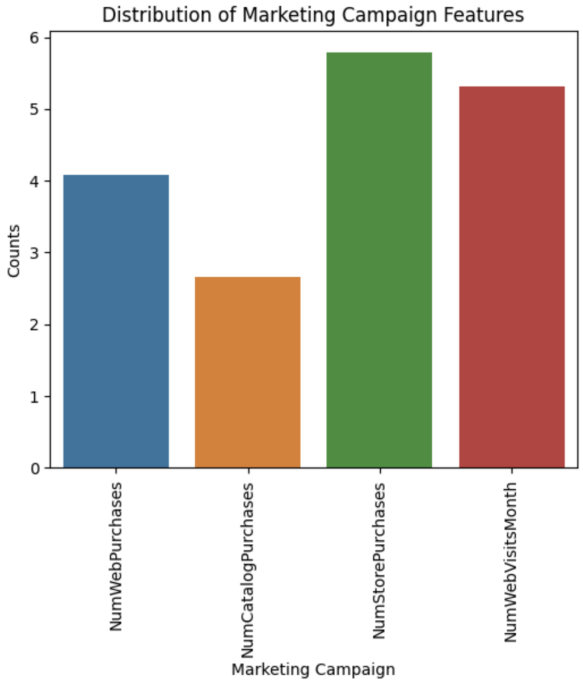
Step 3: Since outliers can impact models performance, I detect the numerical columns outliers by using boxplot. There are some features that have outliers, but not all of them need to be addressed. The amount spent on the products is important information representing the preference of different customers, so we keep them. Therefore, I have only income feature outliers that need to be addressed. I calculated the 25% quantile, 75% quartile and IQR about the Income feature and checked outliers who are lower than 1.5 IQR of quantile 1 and who are 1.5 IQR larger of quantile 3. Since there are only 8 outliers, we can exclude them directly.

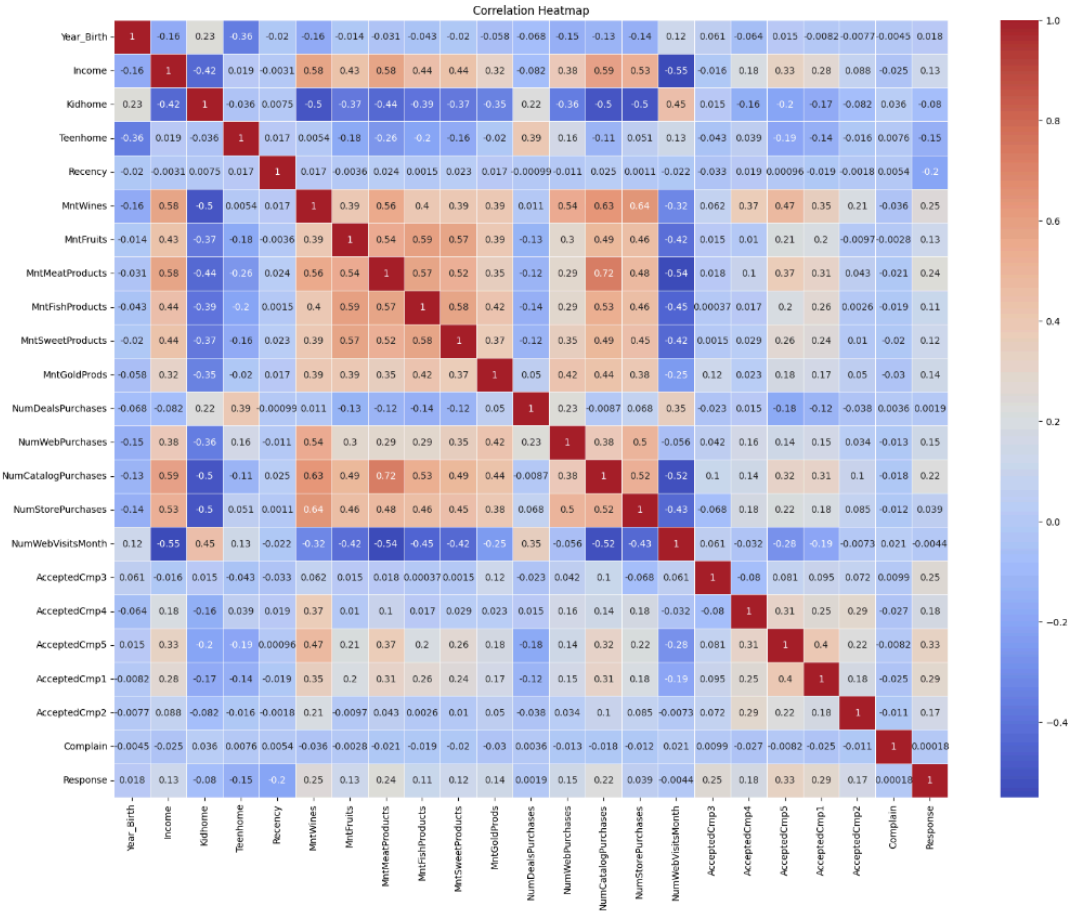


Step 4: After I exclude the outliers, I also need to check the distribution of the categorical features. In education, the majority of customers have a graduation degree, and in Marital status,

main status are married, together and single. From the previous fifth marketing campaign, I can find that the class 1(offer accepted) and class 0 (offer unaccepted) are highly imbalanced, so before I build the model, I should take care of the imbalance dataset. Also, I checked the correlation among numerical features. From the heatmap, I can see catalog purchase and meat product have 0.72 correlation, I can investigate more later about how catalog purchase impacts meat product sales. Other than that, not too many correlations can be detected.




Distribution of Marketing Campaign Features
Counts
0
1
2
3
4
5
6
NumWebPurchases
NumCatalogPurchases
NumStorePurchases
NumWebVisitsMonth
Marketing Campaign


Correlation Heatmap
Year_Birth
Income
Kidhome
Teenhome
Recency
MntWines
MntFruits
MntMeatProducts
MntFishProducts
MntSweetProducts
MntGoldProds
NumDealsPurchases
NumWebPurchases
NumCatalogPurchases
NumStorePurchases
NumWebVisitsMonth
AcceptedCmp3
AcceptedCmp4
AcceptedCmp5
AcceptedCmp1
AcceptedCmp2
Complain
Response
1.0
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4

Feature Engineering

Step 1: I convert data of birth feature into new feature Age, which will be more relevant to our prediction. And we drop the irrelevant ID and Dt Customer features.

Step 2: Based on the data exploration above, I encode and combine some features so that our model can extract the information from our dataset efficiently. I checked the values of different Marital status and I found out that YOLO, Absurd and Alone are only a few customers, I can combine them with Single status customers to reduce the dimensions. Also, since the Marital status is not ordered, I can use 'get_dummies' to encode this feature. Then I check the values of different Education levels. The 2n Cycle education is similar to a Master's degree, so I can combine these two together. Since Education level has its order, I map different education levels with 0,1,2,3. In other words, Basic = 0, Graduation = 1, Master = 2, PhD = 3.

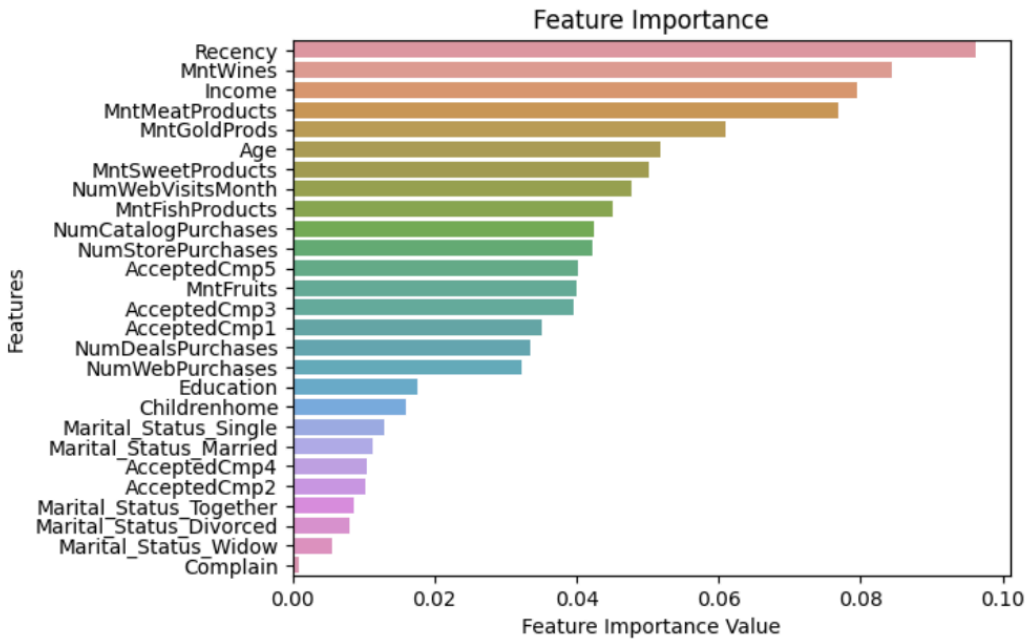
```
Marital_Status
Married     864
Together    579
Single      479
Divorced    231
Widow        77
Alone         3
Absurd        2
YOLO          2
Name: count, dtype: int64
```

```
Education
Graduation    1127
Master         571
PhD            485
Basic           54
Name: count, dtype: int64
```

I also have Teenhome and Kidhome, they are basically the same meaning that how many children in the home, so I combine them together as a new feature Childrenhome.

```
Childrenhome
1    1126
0     637
2     421
3      53
Name: count, dtype: int64
```

Step 3: I use a RandomforestClassifier model to simply fit with our data and check the features importances.



Based on the feature importances graph, we can drop the least important feature which is Complain. And Since Marital status and Children home are less important, we can combine them as a new feature as Family Status to increase the information of the features. Also, the previous fifth AcceptedCmp can be combined as a Total accepted Cmp to increase the information of the features. Eventually, we have 17 features left.

Model Building

Step 1: For better model performance, I use a standard scaler to standardize all the demographic features like income and the amount they purchased, etc. But I left all the binary columns and then I concat standardized columns with the binary columns as a new dataset.

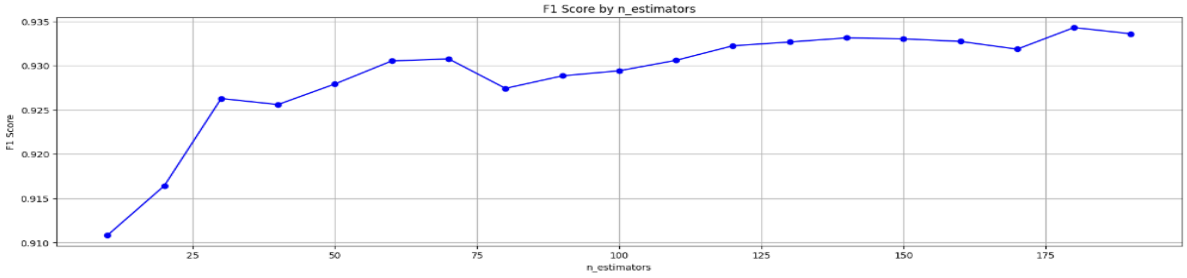
Step 2: As I mentioned before, the dataset is highly imbalanced, so I apply SMOTE for oversampling to make 2 classes balance.

Step 3: Since it is a classification problem, I picked Logistic Regression, Gaussian Naive Bayes, Random Forest Classifier, and XGBoost Classifier models. I built a pipeline for running these models in an efficient way and do 5 fold cross validation to check their performance before I refine them. Also I checked Accuracy, Precision, Recall and F1 score.
Since our campaign has a significant cost associated with targeting a customer (e.g., calling customers individually) and the company want to minimize the risk of targeting non-responsive customers (false positives), so I prioritize precision.

	Models	Accuracy	Precision	Recall	f1_score
0	LogisticRegression	77.008929	38.345865	70.833333	49.756098
1	GaussianNB	68.750000	28.205128	61.111111	38.596491
2	RandomForestClassifier	86.160714	56.250000	62.500000	59.210526
3	XGBClassifier	86.607143	58.823529	55.555556	57.142857

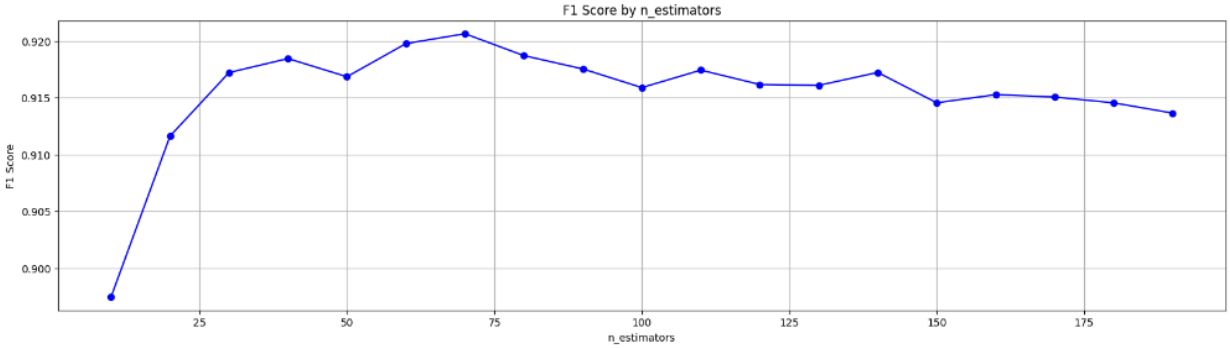
Step 4: After building the models, I start tuning the models. Different models have different hyperparameters, so I used different methods respectively.
For Logistic regression, I chose the Inverse of regularization strength, penalty methods and solver as my tuning parameters, and then I utilized GridSearch to find the best combination for my logistic regression model. Then I got Best parameters for the Logistic Regression model: C: 0.01, Penalty: L2, Solver: liblinear;
For RandomForestClassfier, I first draw a learning curve to get the best n_estimators, and then I use GridSearch to find the rest of the parameters. Finally i got: n_estimators: 180, max_depth: None, min_samples_leaf: 1, min_samples_split: 2;

```
Best average cross-validation F1 score: 0.9342939276717278
Best n_estimators: 180
```



For XGBoost, as what I have done in RandomForestClassfier, I draw a learning curve to get the best n_estimators, and then I use GridSearch to find the rest of the parameters. Finally i got: n_estimators: 70, colsample_bytree: 0.8, gamma: 0.1, learning_rate: 0.1, max_depth: 7, min_child_weight: 1, subsample: 0.8.

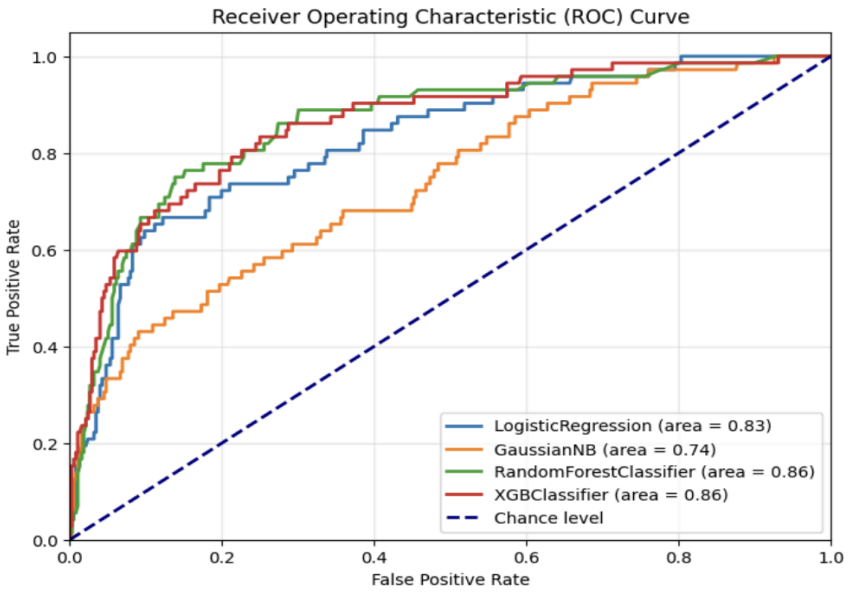
```
Best average cross-validation F1 score: 0.9206394122981356
Best n_estimators: 70
```



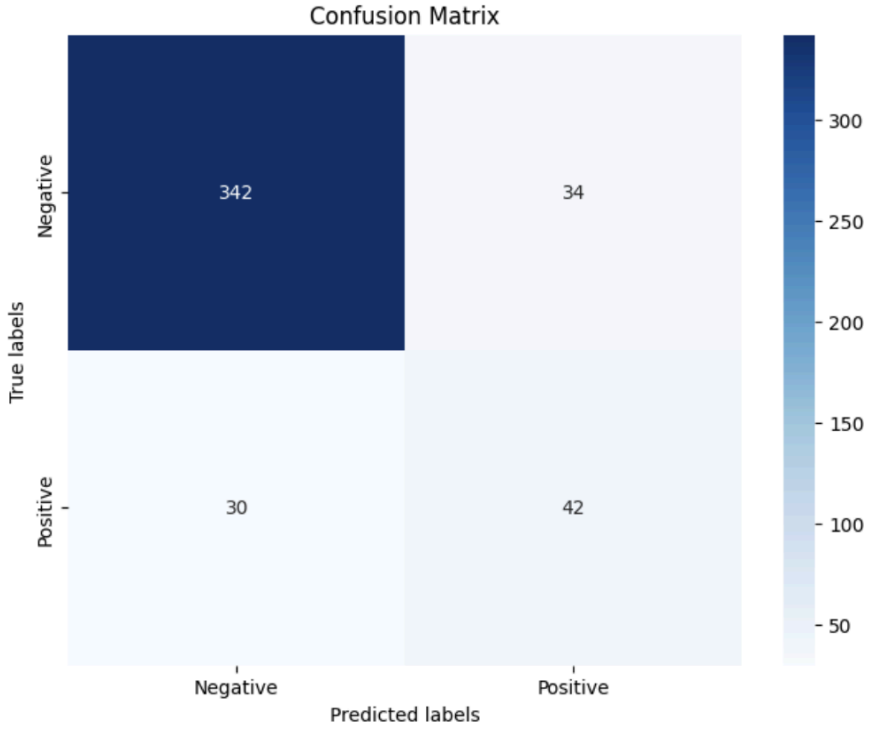
Finally, after tunning the models, we have the improved model performance as below:

	Models	Accuracy	Precision	Recall	f1_score
0	LogisticRegression	74.107143	35.333333	73.611111	47.747748
1	GaussianNB	68.750000	28.205128	61.111111	38.596491
2	RandomForestClassifier	86.607143	57.894737	61.111111	59.459459
3	XGBClassifier	86.607143	58.108108	59.722222	58.904110

Step 5: I check the model's ROC curve to visualize and finally compare all the models and choose the best one.

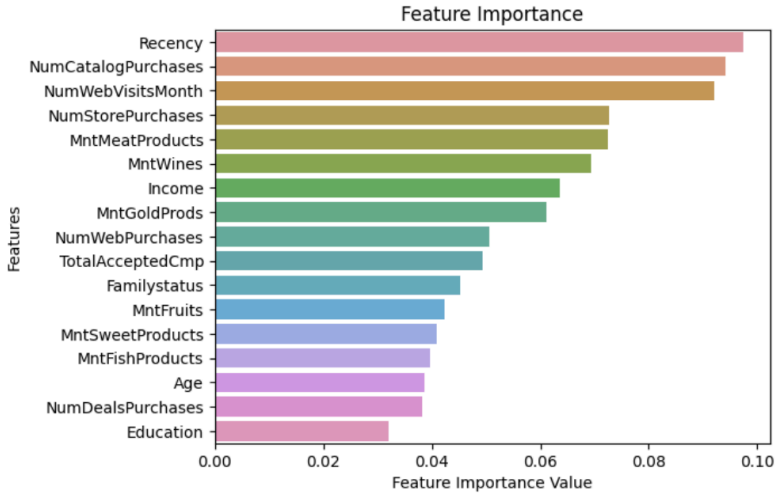


Step 6: Finally, based on the objective of the company, we choose RandomForestClassifier who has the relatively higher Precision and F1 score as our prediction models. And check the confusion matrix of it.

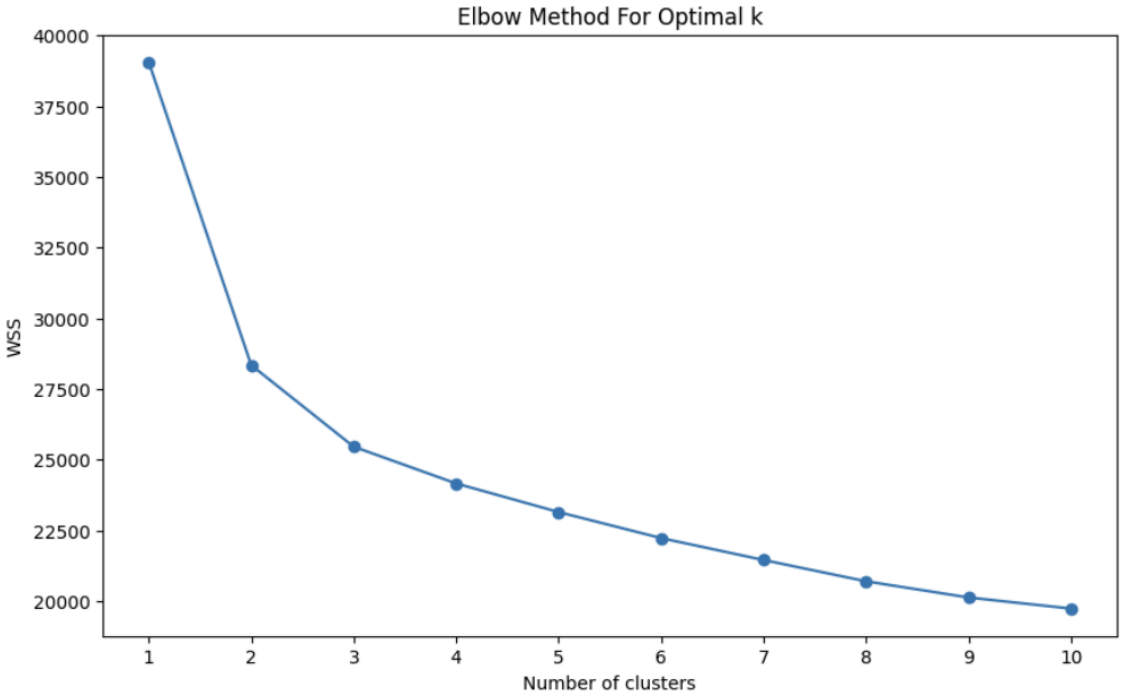


Customer persona

Based on the new model I check the feature importance again.



Based on the above feature, I applied KNN algorithms to study the persona of customers who have the highest response rate. When I built the KNN model, I draw elbow graph to find the best k cluster values.



Based on the above elbow point, K = 3 can be a optimal K value

When K is 3, the model has better performance, so we divide customers as 3 clusters and check customers characteristics.

Cluster	Education	Income	Recency	MntWines	MntFruits	MntMeatProducts	MntFishProducts	MntSweetProducts	MntGoldProds
0	1.604631	35030.819861	49.253002	83.546312	7.469125	39.484563	11.650943	8.088336	23.798456
1	1.730841	70411.804673	48.967290	544.499065	46.770093	305.921495	65.743925	47.776636	65.978505
2	1.000000	666666.000000	23.000000	9.000000	14.000000	18.000000	8.000000	1.000000	12.000000

alsPurchases	NumWebPurchases	NumCatalogPurchases	NumStorePurchases	NumWebVisitsMonth	Response	Age	TotalAcceptedCmp	Familystatus
2.413379	2.791595	0.885935	3.698113	6.619211	0.111492	48.098628	0.116638	2.188679
2.230841	5.500000	4.600000	8.081308	3.901869	0.190654	52.285047	0.495327	1.691589
4.000000	3.000000	1.000000	3.000000	6.000000	0.000000	42.000000	0.000000	2.000000

Key Insights and Recommendations

Insights 1: who are the target customers?

The customer persona with the highest response rate appears to be a well-educated and affluent individual, averaging an income of around 70,411 units. They are likely in their early fifties, with a propensity to spend generously on wines and meats products, indicating a taste for the finer things in life. Their purchasing patterns are diversified, with a strong preference for web and catalog shopping. This cluster has an outstanding 19% response rate to marketing campaigns, suggesting they are engaged and receptive to targeted marketing efforts. They are also likely to have a family, indicated by a family status average of around 2 people.

Insights 2: How much impact does the model have?

	Number	Response %	Cost	Revenue	Profit
All customers	2208	331(15%)	6624	3641	-2983
Target Customers	2208	662 (30%)	6624	7282	**658**

While the total cost remained unchanged at 6,624 units, targeting led to a 30% response rate, a notable increase from the 15% rate seen with the general customer approach. This strategic targeting doubled the revenue to 7,282 units, turning a substantial loss of 2,983 units into a profit of 658 units.

Goal (Number of product sold)	Customer needed	Cost	Revenue	Profit
100	667	2001	1100	-901
100	**334**	987	1100	**113**

For the same sales goal of 100 products, targeting half as many customers (334 versus 667) leads to a significant reduction in costs (from 2001 units to 987 units) and a positive profit of 113 units, as opposed to a loss of 901 units. This indicates a more strategic, focused approach can drastically improve profitability, highlighting the importance of understanding and segmenting the target market to optimize marketing expenditures.

Insights 3: Marketing Campaign strategies recommendations

According to the distribution of marketing channels and the importance of the channels, strategies should focus on catalog purchase and in store purchase, refining catalog marketing and in store purchase experience to entice repeat and new purchases, and bolstering web presence to increase and capitalize on customer visits. An integrated approach that aligns targeted online

advertising with compelling catalog content and an improved in-store experience can create a synergistic effect, driving both online and offline conversions.