-

ABSTRACT

Today, many stores invest a lot of money in advertising their products, in order to increase the clientele and thus increase their profits. Our main goal is classifying customers into two groups – will accept the campaign or will not.

During the project, we used the "Customer Personality Analysis" database. After analyzing the data and classifying the customers according to their potential for accepting the campaign, we think that we have a good recommendation for the store manager.

In the process of choosing the classification model, we examined three different alternatives by using the ROC curve and confusion matrix. Finally, the "Random Forest" algorithm was chosen.

We have created a model that predicts 80% of whether a customer will agree to the campaign.

We believe that with more long-term data we can build better customer segments and improve the model. The improvement shall be focusing on avoiding "FN" values and try to reach the perfect "recall" grade..

1 Introduction

Customer Personality Analysis is a detailed analysis of a company's ideal customers. It helps a business to better understand its customers and makes it easier for them to modify products according to the specific needs, behaviors, and concerns of different types of customers. Customers show different personalities and behavior.
Therefore, while producing quality analysis of the data, will contribute to spending less money on campaigns, and to understanding which customers will likely spend more in the store. For example, instead of creating a general campaign, the store will do it for specific personalities.
The dataset for this project is provided by Dr. Omar Romero-Hernandez.

2 Material and Methods

2.1 Data

The purpose of our project is to predict whether consumers will agree to buy in the store through the campaign that the store will advertise.
Our target column is 'Aceepted_Cmp' which shows whether a customer has agreed to campaigns in the past.
The classification model under supervised learning we built, based on the 'Personality Customer Analysis' data, includes 2240 lines and 29 columns. The data columns describe the customer in terms of year of birth, income, number of children, marital status, education, seniority, and number of purchases in the store. (Correlation tests and histograms are attached in the 'Appendix').

2.2 Data Cleaning

Issue	Manipulate	Reason
Income	Deleting all income > 150000 (7 rows).	Deleting outliers.
Missing cells	Delete 24 rows. (24/2240)	Ignore missing cells.

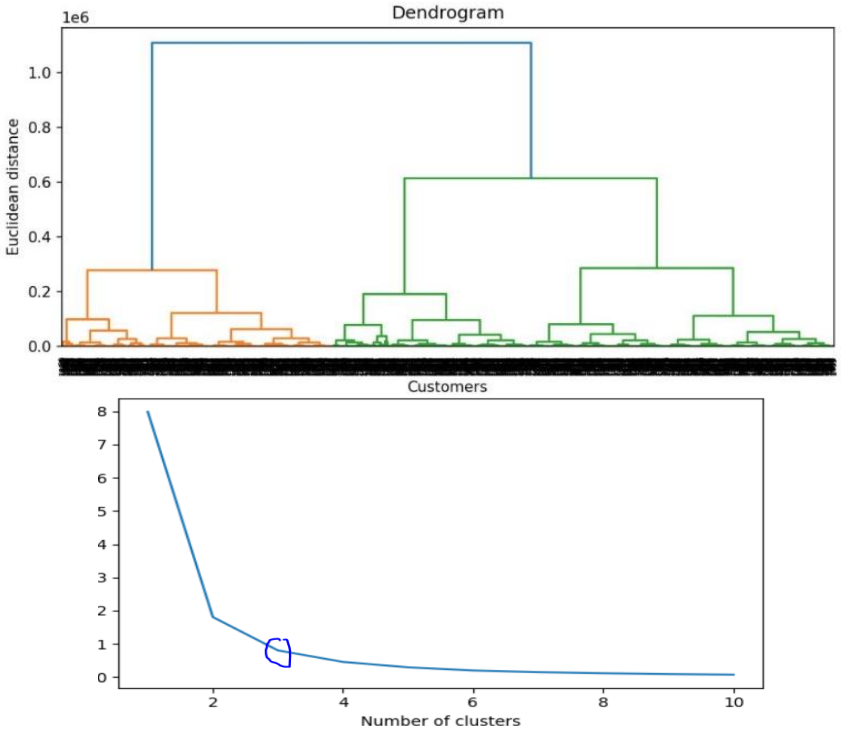
2.3 Data preparation:

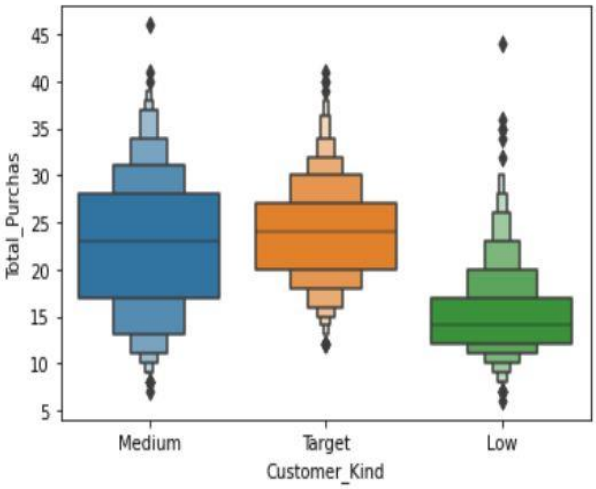
In order to better understand the data for our study, we have made certain columns categorical and combined columns into one column as follows:

a. Marital status - Is the client in a relationship or not.
b. Education - Is the client an academic or not.
c. Customer seniority - the time the customer registered for the store.

d. The amount of money the customer bought - summarizing all the products into a total payment.
e. A number of transactions purchased - the sum of all transactions made in various services on the site.
f. "Accepted_Cmp" - Whether a customer Accepted a campaign or not.

We also examined potential customers, so we utilized the method of hierarchical clustering(for sanity check we used 'Kmeans').The clustering made on the following columns:
income, seniority, and the total amount of money bought by the customer. We found that the optimal separation is to three kinds of customers. We converted the columns to one that calls 'Customer_Kind'. As we can see below, the clustering gave a good indication for understanding the segmentation of the customers who will accept the campaign.

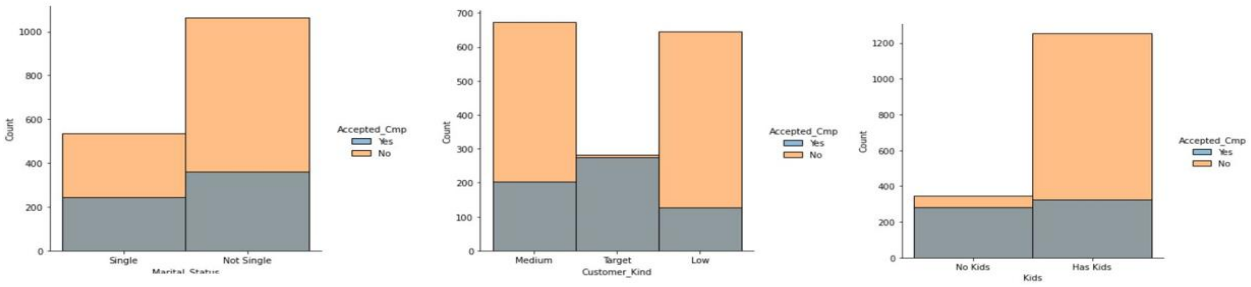




2.4 Preparations for algorithms

1. We converted all the columns into a categorical data type.
2. For finding the columns that match our target column ("Accepted_Cmp"), we performed correlation tests (Chi-Square) and found that the most suitable columns for predicting are:
   a. Customer Kind: P-value = 0.003.
   b. Marital Status: P-value = 0.01.
   c. Kids : P-value = 1.35×10^{-9}

3. The data split for 70% training and 30% test.
4. From the features we got the indication for prediction by the following plots:



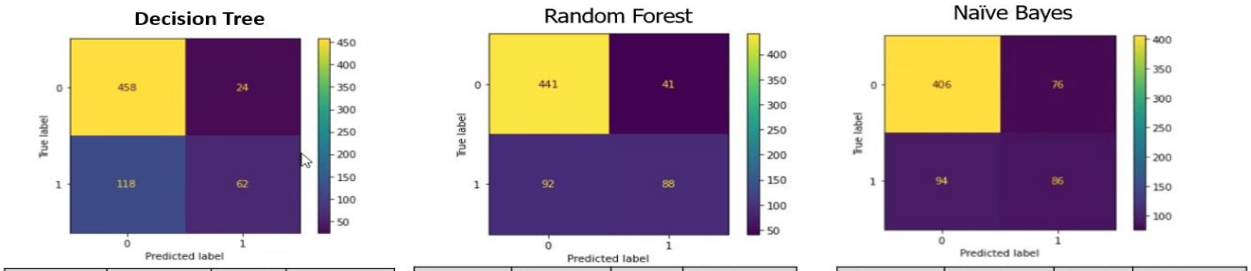
3 Numerical Analysis and Results

3.1 Analysis of the results

While predicting the results for 662 customers, we performed three algorithms:

1. Decision Tree -The columns were inserted into a 'one hot' matrix, and we chose that the depth of the tree would be 3. The "Marital Status" column has the highest 'Gini' grade. The results: Correct – 520, Incorrect – 142.
2. Naïve Bayes – The columns data type that inserted were categorial (Guesiaan). The results: Correct – 492, Incorrect – 170.
3. Random Forest. The results: Correct – 529, Incorrect – 133.

We calculated the confusion matrix for each algorithm.



Accuracy	Precision	Recall	F_1 Score
0.8	0.72	0.34	**0.46**

Accuracy	Precision	Recall	F_1 Score
0.8	0.68	0.49	**0.57**

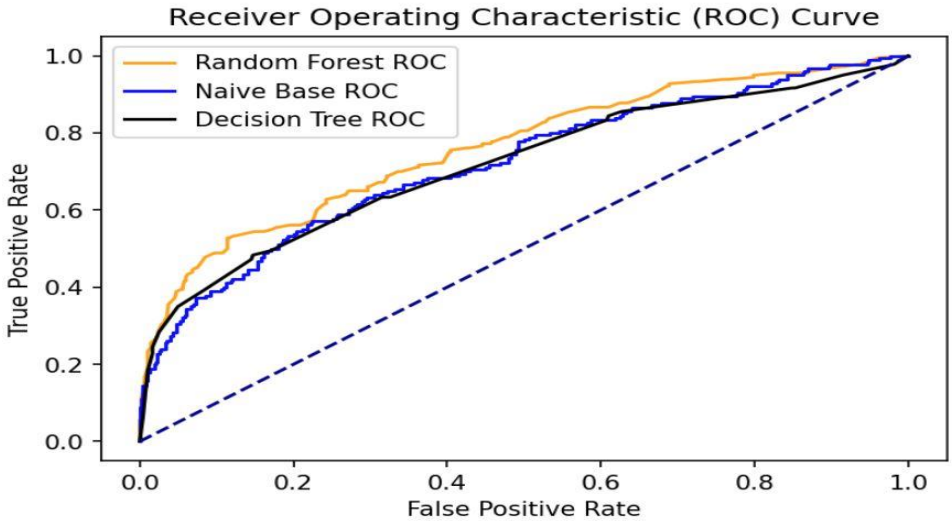
Accuracy	Precision	Recall	F_1 Score
0.75	0.53	0.48	**0.50**

3.2 Discussion

After analyzing the results, the algorithm with the best performance is Random Forest. According to our business goal, we want to predict in the best way the customers that will accept the campaign out of the actual number. Our target is to get a high score in the "Recall" grade, specifically get the "FN" value to a minimum.

Another way of spending less money on campaigns is to decrease the number of customers that were labeled as 'Yes' but actually, they are 'No'(FP).

We can notice in the ROC curve that "Random Forest" has the minimum distance for the perfect classifier.



4 Conclusions

In this project, our statement was helping the store to adjust the advertising campaigns for customers kinds, in order to get the optimal result from it.

For example, if the store wants to set a new meat product campaign, the model will analyze the customers that have a high percentage of buying the new product.

Our model gives the store manager a prediction according to 3 characteristics:

1. The customer's potential to buy - is characterized by his income, seniority, and his total payment. If the customer will classify as a 'target' customer, probably will accept the campaign.
2. The marital status of the customer.
3. Customer without kids has a bigger percentage to accept the campaign.

Our recommendation for further work is to collect long-term data regarding the customer characteristics and his treatments. Customer segmentation is the key for the prediction of whether the client will accept the campaign. Furthermore, the focus shall be on the 'Recall' grade and try to reach the perfect score of '1' for spending less money on consumers that won't get engaged with the advertisements and finding all the potential consumers that will respond.

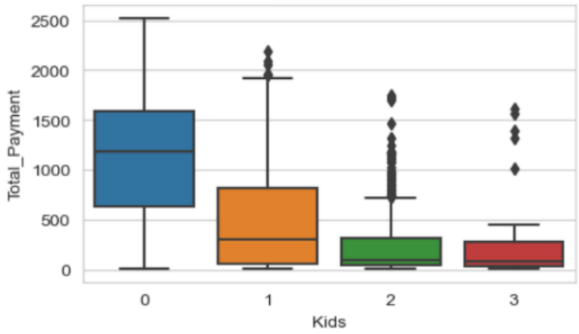
5 References

[1] https://www.kaggle.com/imakash3011/customer-personality-analysis

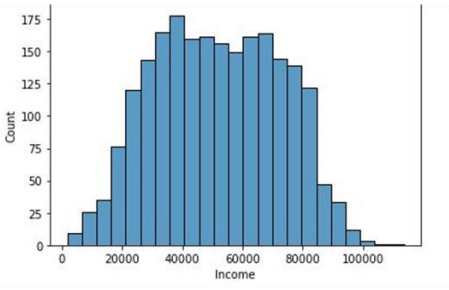
6 Appendix

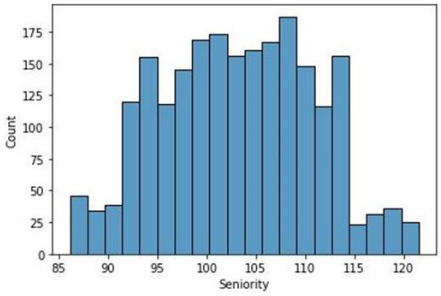
6.1 Correlation between numeric columns:

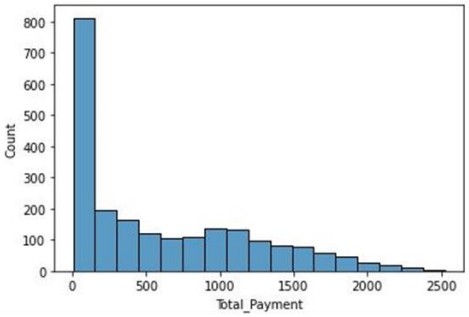
Kids coulmn

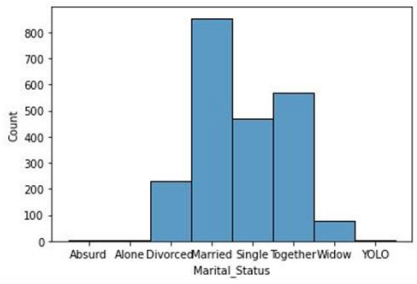


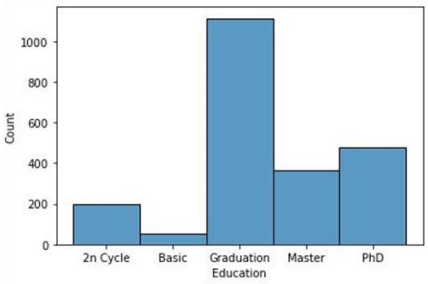
6.2 Histograms :











6.3 Data Pre- Processing

Issue	Manipulate	Reason
Marital_Status	Converting to category type	Saving data storage.
Education		
Dt_customer	Converting to date type	Saving data storage.
Creating 'Age' column	Creating a new column. Age = 2021-Year_birth	Knowing customer current age.
Creating 'Total_Payment'	Summaries all products payments columns.	To get the aim, we want to look at the general payment that customers spend.
Income	deleting all income > 150000 (7 rows).	Deleting outliers.
Missing cells	Delete 24 rows. (24/2240)	Ignore missing cells.
Creating 'Kids'	Creating a new column. Kids = Teenhome+kidHome	Review overall kids at home.
Creating 'Total_camp'	Summaries all promotion columns.	Knowing if any campaign succeeds on a specific customer.
Drop columns	kidhome, teenhome, dt_customer, products columns, promotion columns, Response.	Using these columns data for creating new columns.
Creating 'Seniority'	Creating a new column. Seniority = today()-dt_customer	Knowing customer seniority.

6.4 Raw Data

Column	Description	Data Type
Year Birth	Customer's birth year	Int64
Education	Customer's education level	Object
Marital_Status	Customer's marital status	Object
Income	Customer's yearly household income	Float64
Kidhome	Number of children in customer's household	Int64
Teenhome	Number of teenagers in customer's household	Int64
Dt_Customer	Date of customer's enrollment with the company	Object
Recency	Number of days since customer's last purchase	Int64
Products (6 columns)	Amount spent on each product in last 2 years	Int64
Promotion (6 columns)	if customer accepted the offer in the 1st campaign, 0 otherwise	Int64