**Abstract of Customer Sentimental Analysis**

|  |  |
| --- | --- |
| **Name** | **Student ID** |
| Jagmit Singh | 0774287 |
| Parminder Kaur Chana | 0766562 |
| Srijan Rajput | 0774278 |
| Supreet Kaur | 0760707 |

**Supervised By**: Dr. Savita Seharawat

**Submission Date:** 13th Oct 2021

**Abstract**

Nowadays advertisements have become more interesting, and customer oriented rather those irritating ads which were of no interest for all. Companies are targeting the consumers with customized advertisements. Ads of the products which are of the customers’ interest. This is not just saving an Organizations’ Marketing cost but also helping to target end users with effective Ads. So, this is what has interested us to select this dataset, where we will try to make the consumer segment based on some selective attributes and then target them with respective Ad-campaigns which should turnout as an effective marketing strategy to enhance the sales or for promotion of a new product. This dataset has entities like customers’ demographic data, data related to their shopping pattern, data related to the campaigns run by Organization and data on type of methods opted for shopping by end users. So, considering all these different variables, we will form the customer segments and then based on the most relevant attributes customers will be send the promotional offers for products of their choice only and not some random offers.

**Keywords**: Decision Trees classification, KMeans,Logistic Regression, Clustering, Target Marketing etc.

**Research Questions:**

* Use data to identify different customer segments and evaluate their responses to different marketing initiatives.

Target Variable: Response

* Detect the spending behaviour of customer for different products.

Target Variable: MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProduct, MntGoldProds

* Find out key points from the analysis that could help the marketing campaign

**Tools to be used:**

NumPy and Pandas Libraries for Data analysis, Jupiter Notebook, Seaborn and Matplotlib for Visualization purposes.

**Target**: Need to perform clustering on the data to summarize the customer segments.

**Context**:

The purpose of analysis would be to work on target marketing and in order to achieve that we’ll start by performing clustering for customer segmentation. Clustering is a useful technique for dividing the data points in several groups in such a way that each data point in the same group holds similar properties as compared to data points in the other groups.

**Questions to be asked:**

**1.What do you already know about the topic?**

We have the brief knowledge about the Customer Personality analysis where customized way of advertisements has been used to do marketing for their product by the organization. Specially for the products that have more customer demand. Besides, we are familiar with the utilization of target marketing e.g., when we make some google searches, we could see the advertisements block on website we open such as YouTube related to the searches we have made**.**

**2.What do you have to say critically about what is already known?**

The classification and regression techniques those were used in the earlier implementations helps me to decide to proceed with the next implementations. Like which model is best fit for my data. Do I need to use logistic Regression or Linear Regression? Which optimized value of K-means is best suitable for predicting the model? How clustering can be performed so that I can figure out which variable has more impact on my response variable.

**3.Has Anyone else ever done anything exactly the same?**

We found some similar analysis done on Kaggle.References are mentioned in the footnotes.In those analysis  the objective is to consider which product is most likely to be bought, by considering their requirements, daily needs, and their purchasing pattern. On which products the company need to focus to do promotion rather than performing market for each product that not only waste company time but also their money. Basically, focus is put on four attributes that are  people, promotion, product, and place.

The target is to perform the clustering by using the three clusters where he had categorized the customer segments into three parts those are highly active customers, moderately active customers, and least active customers in terms of Education, age, Martial\_Status, Income, Kids, Expenses, Age, and day engaged.

**Data source:**

<https://www.kaggle.com/imakash3011/customerpersonalityanalysis?select=marketing_campaign.csv>

**Descriptive Statistics:**

**Data details:**

Data have **2240** observations and **29** attributes, which mainly includes the customers’ attributes, amount spend on products, campaign techniques.

**Attributes**

**People**

ID: Customer's unique identifier - [ID]

Year\_Birth: Birth year of customer - [Year]

Education: Education level of customer ['Graduation', 'PhD', 'Master', 'Basic', '2n Cycle']

Marital\_Status: Marital status of customer ['Single', 'Together', 'Married', 'Divorced', 'Widow', 'Alone’, ‘Absurd', 'YOLO']

Income: Yearly income of customer [1732 to 666666]

kidAtHome: Number of children in customer's household [0-2]

teenAtHome: Number of teenagers in customer's household. [0-2]

dt\_Customer: Date of customer's enrollment with the company [Date]

Recency: Number of days since customer's last purchase [0-99]

Complain: 1 if customer complained in the last 2 years, 0 otherwise

**Products**

amtWines: Amount spent on wine in last 2 years [0-1493 in USD]

amtFruits: Amount spent on fruits in last 2 years [0-199 in USD]

amtMeatProducts: Amount spent on meat in last 2 years [0-1725 in USD]

amtFishProducts: Amount spent on fish in last 2 years [0-259 in USD]

amtSweetProducts: Amount spent on sweets in last 2 years [0-263 in USD]

amtGoldProds: Amount spent on gold in last 2 years [0-362 in USD]

**Promotion**

noDealsPurchases: Number of purchases made with a discount [0-15%]

AcceptedCmp1: 1 if customer accepted the offer in the 1st campaign, 0 otherwise [0,1]

AcceptedCmp2: 1 if customer accepted the offer in the 2nd campaign, 0 otherwise [0,1]

AcceptedCmp3: 1 if customer accepted the offer in the 3rd campaign, 0 otherwise [0,1]

AcceptedCmp4: 1 if customer accepted the offer in the 4th campaign, 0 otherwise [0,1]

AcceptedCmp5: 1 if customer accepted the offer in the 5th campaign, 0 otherwise [0,1]

Response: 1 if customer accepted the offer in the last campaign, 0 otherwise [0,1]

**Place**

noWebPurchases: Number of purchases made through the company’s web site [0-15]

noCatalogPurchases: Number of purchases made using a catalogue [0-28]

noStorePurchases: Number of purchases made directly in stores [0-13]

noWebVisitsMonth: Number of visits to company’s web site in the last month [0-20]

The data has 3 Categorical attributes i.e. Education, Marital\_Status and Dt\_Customer and rest all the variable are numerical.

**Data Structure:** By default, the panda’s data frame load nonnumeric attributes as an object type. For the categorical we need to convert them programmatically to categorical to optimize the memory and increase the usability. Once this is done(highlighted in red in the screenshot below).We have calculated the statistical parameters (mean, median, min, max, mode and skewness) for all numerical variables.

**Table

Description automatically generated with medium confidence**

**Missing Values:** From the response shown in above screenshot we can observe that there are some missing values in each column (highlighted in yellow) and income tops the list for missing values. Pandas automatically detect and fill the blank values with NA values. Using the is null method, we can confirm that both the missing value and “NA” are recognized as missing values. In the above screenshot we can observe that NA values are matching with the null values (count of null values are placed next to yellow highlighted column).

**Cleaning Values:** Next step is to analyze if there is any pattern for the missing values especially income.

1. Begin with the Dt\_customer we found that data has 8 missing values which constitute to 0.35% of the overall data. Since it is difficult to fill in the registration date for the customers as the data has no sequence for date and the percentage of missing values are very less, we will drop the records with missing values. The total number of records after deletion are 2232.
2. For handling the missing income values which constitute to 1.3% of the total data we are simply replacing that with **median** value which is **51381.5**
3. To handle the missing values for rest of the columns we will use knn transform. Since knn works for numerical data only we convert the categories columns i.e. Education and Marital\_Status to numerical by using the corresponding label codes for categorical data. This formatted data is passed to KNN for KNN imputation with k=5. And then rollback the label code to original categorical data.

**Outliers:** We have outliers on mostly sales related attributes.

amtWines, amtFruits, amtMeatProducts, amtFishProducts, amtSweetProduct, amtGoldProds. Boxplots below gives the idea of outliers present in different attributes.

Boxplots below gives the idea of outliers present in different attributes.

A picture containing chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated with low confidence

Chart, histogram

Description automatically generated

Chart

Description automatically generated with low confidence

Chart

Description automatically generated with medium confidence

Chart

Description automatically generated

Chart, histogram

Description automatically generated

Chart, histogram

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Chart

Description automatically generated

Frequency Table

Chart, bar chart

Description automatically generated

Chart, bar chart

Description automatically generated

**Correlation:** gives the relation b/w two variables and shows how the change in one variable impact other. Its values vary b/w -1 to 1. No strong correlation has been observed b/w any of the variables.

**Table

Description automatically generated**

**Correlation Plot:**

**Chart, bar chart

Description automatically generated**

**Heat Map:**

**A picture containing text, scoreboard

Description automatically generated**

**­­**

KNN classification: We have drawn plot with K values against error rate to find the optimal value of the K for which the model gives the maximum accuracy and value has been observed to be K=10. Accuracy has been calculated around ~ 85% for KNN model.

Chart, scatter chart

Description automatically generated

Logistic Regression model: basically, used where binary target variable is involved. We have made calculations for optimum threshold value using the ROC curves and value has been calculated to .1 for which minimum false positive has been observed. The accuracy of the model for this threshold value has been calculated to 87%.

Chart, scatter chart

Description automatically generated

Decision Trees:

The algorithm calculates the information gain corresponding to each column and the for the attribute for which the information gain is maximum we use that as the root node. Timeline

Description automatically generated

Timeline

Description automatically generated

Clustering: Clustering is a technique to make subset of groups such that data points that contains similar properties are contains in the same group. In our dataset, data points are divided into five clusters that we have computed using the KMeans Clustering technique, to compute most appropriate value for k we have used elbow method from there we have computed that our dataset will be divided into 5 clusters.

Chart, line chart

Description automatically generated

References:

* https://www.kaggle.com/imakash3011/customer-personality-analysis

[https://www.kaggle.com/imakash3011/customer-analysis-eda-report-clustering](https://www.kaggle.com/imakash3011/customer-analysis-eda-report-clustering" \t "_blank)