Project Proposal

Course: Data Mining.

Lecturer: Dr. Omer Tzuk.

Ido Stern

Niv Ben Avraham

Stav Atias

"Customer Personality Analysis"

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

Project Aim Predicting if a customer will accept a campaign for buying in the store, according to his type.

Meta Data Number of rows: 2240. Number of columns: 29. Dtypes: int64 (25), float64(1), and object(3). Missing cells - 24, duplicate rows - 0.

Columns that we used on the project:

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 campaign, 0 otherwise	Int64

Problem Statement

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. 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.

EDA

Data pre-processing and cleaning:

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.

Issue	Manipulate	Reason
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.

- **Our target column is "Total_camp" (Int type).**

Graphs

Figure2

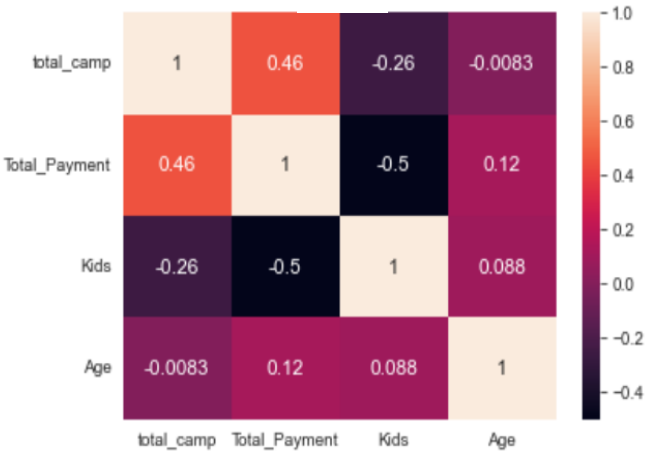

Figure1

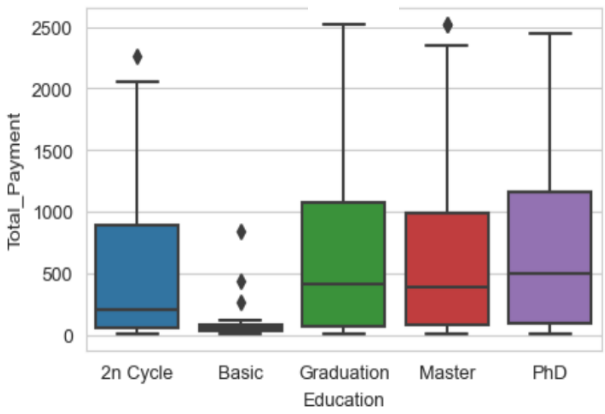

Figure 3

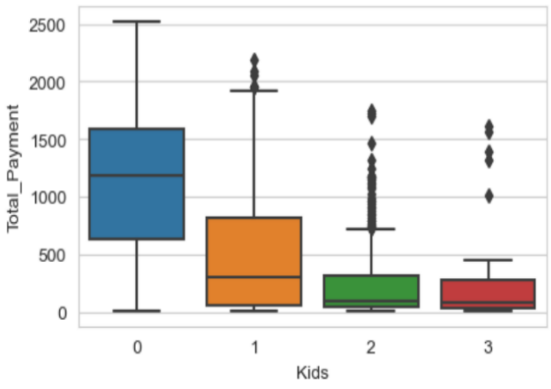

Figure 1: after analyzing the numeric columns, we made a correlation test. The results are:

1. Positive correlation between 'Total_Payment' and 'total_camp' (0.46).

2. Negative correlation between 'Kids' and 'Total_Payment' (-0.5).

Figure 2: Educated people are spending more money.

Figure 3: The fewer kids you have, the less money you spend.

Algorithm Proposal

After analyzing the data and following the research question, we chose to predict using a classification that belongs to supervised learning. The algorithm we will use is the 'decision tree', to predict whether a customer will accept the campaign.