Adapting to Customer Trends

CSE 4095-003

Kaitlyn Ha, Loc Hoang, Allen Choun

Abstract

Customer trends are ever changing no matter when and where a business or company is located, and knowing when they will and how to accommodate these trends can be the decision on whether a business will succeed or fail. Thanks to data science techniques we have a number of tools at our disposal to interpret and predict these trends. With the Customer Personality Analysis dataset, 3 methods were used to analyze the data and what possible future trends may look like: Linear Regression, K-Means Clustering, and Time Series Analysis. The purpose of the dataset is to provide the business the opportunity to identify weak points and improve them. After analyzing this dataset, we came to various conclusions that could assist this business in improving profitability such as focusing on higher end products, providing customers with member exclusive benefits and mirroring more similar, yet successful companies.

Introduction

The Customer Personality is a dataset from Kaggle that is meant to offer insights into the diverse behaviors, preferences, and attributes of customers across various demographics. By leveraging this dataset to understand customer attributes and preferences, businesses are able to optimize their operations to better meet customer needs, ultimately leading to increased revenue streams and improved customer satisfaction.

Background

Data can be found in everything nowadays, with companies actively using and selling the public's information they're able to use this data to track and see how their customers may act. As mentioned before in lecture, Target was able to market to specific customers who bought key items and applied to certain customer groups. Millions of people go shopping every year, with each of those years, people change, they have children, move to a new home, change their likes

and dislikes, it's important that when selling to customers for the business to understand when to adjust their business plan when necessary.

Companies regularly post annual reports which can include statistics on their memberships if they're wholesale clubs, sales growth, customer demographics, performance, etc. and all of this information can be found easily accessible online. In this report the variables used throughout this project will include the following: ID(Customer's unique identifier), Year_Birth, Education, Marital_Status, Income, Kidhome, Teenhome, Dt_Customer(Date the customer enrolled with the company), Recency(Days since last purchase), Complain(1 if there was a complaint within the last 2 years, 0 otherwise), MntWines, MntFruits, MntMeatProducts, MntFishProducts, MntSweetProducts, MntGoldProds, NumDealsPurchases, AcceptedCmp1, AcceptedCmp2, AcceptedCmp3, AcceptedCmp4, AcceptedCmp5, Response, NumWebPurchases, Num Catalog Purchases, NumStorePurchases, and NumWebVisitsMonth. Characteristics beginning with "Mnt" described the amount a customer had spent on a specific type of product, characteristics beginning with "Accepted" described 1 for if the customer accepted a Nth campaign and 0 otherwise, and characteristics beginning with "Num" described the amount of purchases through a specific method.

This project aims to use the dataset, Customer Personality Analysis, to find trends and related factors in order to provide a new motion for the company. Through the three methods mentioned earlier, each method will take into account 2240 observations along with their relevant attributes to interpret, predict, and assess the next possible steps in order to maximize the company.

Methodology

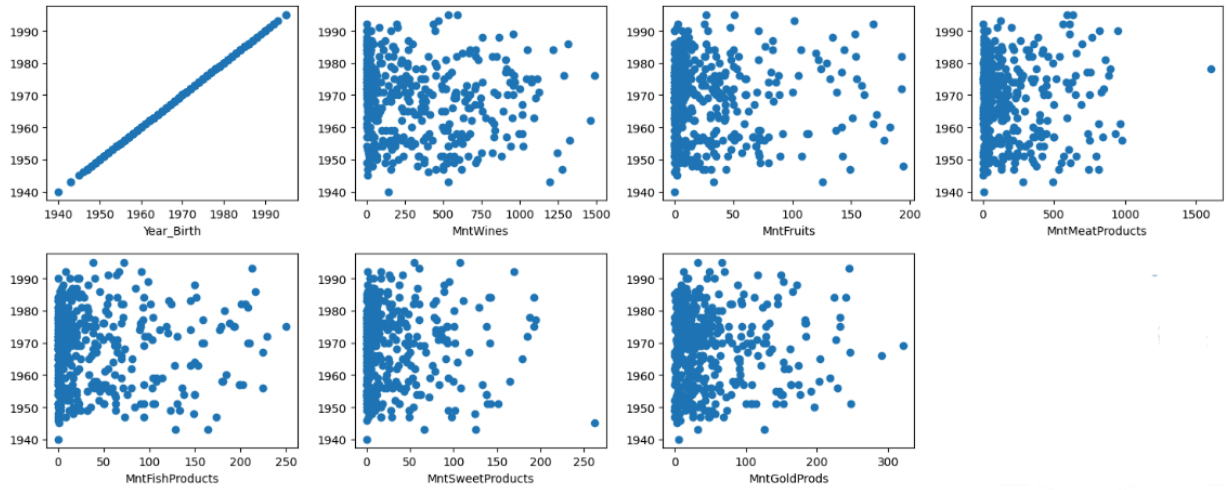
Linear Regression

Initially, for each observation the attributes mainly focused were the following:

- Year_Birth
- MntWines
- MntFruits
- MntMeatProducts
- MntFishProducts

- MntSweetProducts
- MntGoldProds

At first when working with the data, Year_Birth was used as the target while the amount spent on each product was used as predictors in order to observe any type of relationship between the variables, however, the results when plotting between Year_Birth and all the amounts spent were not favorable.

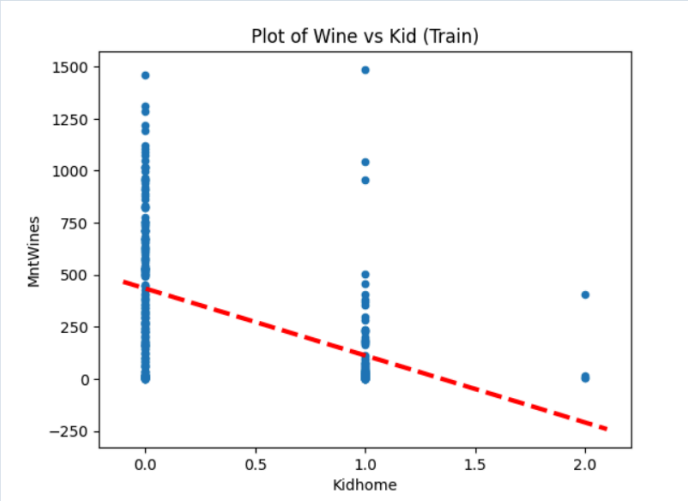


Through this realization and discussion, Year_Birth was translated into Age to better visualize the customer demographic. When this was done, it was observed that there were many outliers within the data set, ages being over 100, resulting in the reconsideration of the data range. The data was then adjusted to limit to the ages of 94 and below. After more experimentation, it was determined that the initial target and predictor variables were not suitable for testing. It was then concluded after observing the attributes, it was noticed the highest selling or most spent on product was the wine products. Noticing this, rather than compare one target to the products of the company, a single product was chosen as a target and based on that, the following predictors were selected: Income, Education, Recency, and Kidhome. The attributes described the individual customer's yearly income, their education level, days since last purchase, and the amount of children in their household. More specifically, after viewing a correlation between the variables, it can be seen that there was a highly negative correlation between MntWines and Kidhome.

	Income float64	Recency float64	Kidhome float64	intercept float64	MntWines float64
Inc...	1	-0.01994712713	-0.2359876904	nan	0.3254588782
Rec...	-0.01994712713	1	0.0741358768	nan	0.06445427482
Kid...	-0.2359876904	0.0741358768	1	nan	-0.4946474935
inte...	nan	nan	nan	nan	nan
Mn...	0.3254588782	0.06445427482	-0.4946474935	nan	1

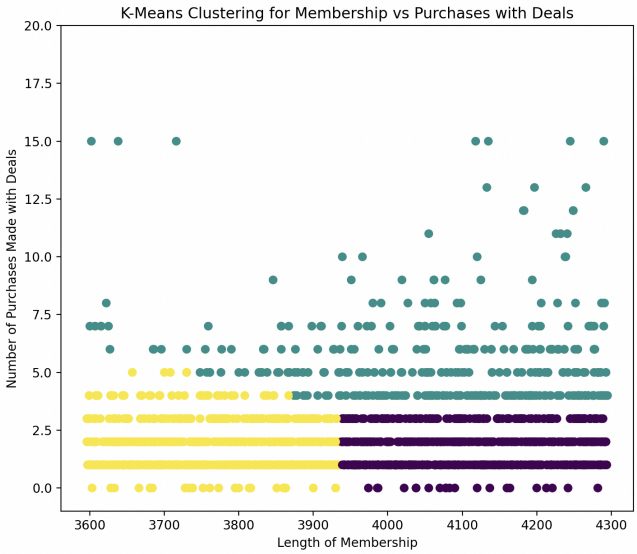
	coef float64	std err float64	t float64	P>|t| float64
inte...	434.5134	20.547	21.147	0
Kid...	-321.7364	31.213	-10.308	0

Although there is a very large mean squared error, this was most likely due to the orientation between MntWines and Kidhome as you can not have decimals of children and the dispersion of Kidhome was either 0, 1, and 2.



K-Means Clustering

To analyze the dataset is to utilize attributes to create new possibilities. Using K-means clustering to compare length of membership to number of discounts used can help divide customers into groups based on similarity. The business is then able to utilize the characteristics of the most loyal customers to change the marketing of their products. This graph is produced with the MATLAB, Pandas and Scikit-learn libraries in Python.



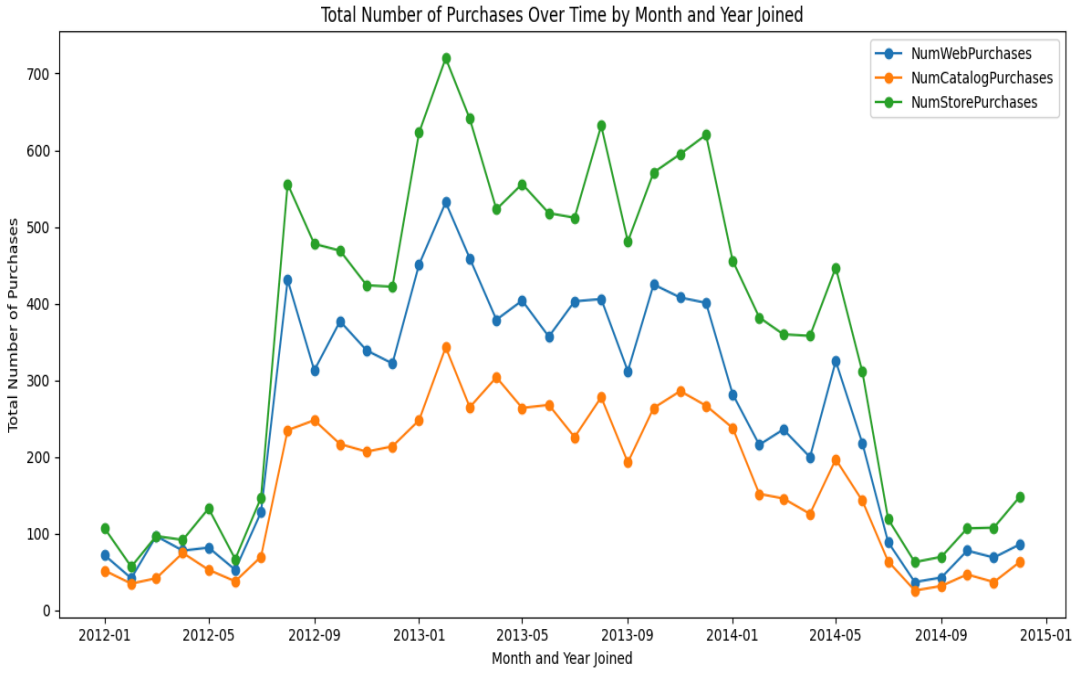
By taking the attribute Dt_Customer, which is the date of the member's enrollment with the company, and finding the number of days since, the days since the enrollment can be calculated using the datetime library. Since the dataset only records up to a certain date, the ratio heavily outnumbers the ratio that was most common 10 years ago. The x-axis shows days from lowest to highest, which means the dates must be interpreted as 2014 on the left side of the graph and 2012 on the right. 2012 seems to have the highest consistent number of discounts being used by members. Looking at the clusters, there is a group that varies in length of membership, but has the highest number of discounts used. This customer group has shown loyalty due to their long standing membership and utilization of the business's promotions.

Using K = 3 will produce 3 clusters of somewhat similar points. This is the ideal number of clusters in this context because it clearly separates the members based on use of promotions. Using K = 2 splits the group into two clusters horizontally, which will only explain what the axis already does, that one group has been members for longer. The topmost cluster separates the customers for further analysis. The business is able to take advantage of these specific records and identify other similar attributes such as age, income and amount spent on products to devise possible business routes.

Time Series Analysis

Analyzing data through the measure of time is a very effective method when it comes to making data-driven decisions. Effectively, we can narrow down the facts, metrics, and data to guide our business decisions with our goals and to further improve our initiative of ultimately creating more revenue.

There are three primary ways to make a purchase within the business. Customers can make a purchase either through the website / online store (NumWebPurchases), catalog via phone call or mailing forum (NumCatalogPurchases), or in store (NumStorePurchases). Utilizing the attribute Dt_Customer, we are effectively able to evaluate the date of the member's enrollment with the company and compare the time they've been a member to the usage rates of which methods of purchasing they have used.



Upon analysis, we have created a graph fixed to the company's last two years. We grouped each customer into their respective date categories (Month and Year Joined) since their enrollment to the company. Afterwards, we took the summation of each method of purchase and plotted it respectively to the group of customers. We can see the clear trends of each purchasing method. Ironically, each method of purchase peaked in 2013; however, towards the end of 2014, we can see a clear decline in overall number of purchases similar to how they have started.

Results

Through linear regression, it was concluded that depending on the amount of children within a customer's household, affected the amount of purchases on wine products. As a result, it was determined that the company should focus advertisements towards wine products on households with fewer dependents, as regardless of age, with or without, the less children present within a household, the more likely a customer was to purchase wine products.

After assessing the results of K-means clustering, it was concluded that most customers utilized available promotions on various products. To assist the business in profit, a possible solution is providing membership owners with exclusive deals and promotions. This could be in the form of a credit card, point system or app rewards which may increase customer spending and loyalty.

By analyzing past methods via time series analysis that have proven successful in generating revenue, the business can make informed choices about where to allocate resources and focus their efforts in the future. This approach can help them capitalize on what has worked well in the past and potentially replicate that success in the present or future endeavors.

Reflection

While using the data to make computations, it was noticed that much of the models were overfitted, with over 2000 rows, it would have been much more beneficial to have minimized the data as to not be overwhelmed with more observations than needed.

It was also noticed there were many outliers, with the Year_Birth attribute going as far back as the late 1800s, many of the customers enrolled with the company were most likely no longer present. Through this, the dataset was accommodated by being scaled to the reasonable age limit of 94 and below.

As a group, we wished we would have used a more complex and defined dataset. A dataset like this was very broad and limiting. For instance, this dataset only provided purchases of items that were in the last two years (2012-2014) and the business, itself, was extremely vague. Everything was up for interpretation - we interpreted that this business was a wholesale club similar to a Costco due to its items it was selling. We decided that the ultimate goal of a

business is to thrive through the maximization of revenue. However, companies deal with datasets like this all the time so the challenge in performing an analysis was definitely rewarding.

References

[Customer Personality Analysis Dataset from Kaggle]