**Increase Marketing Efficacy: Using Python to Identify Potential Customers and Maximize Campaign Profits**

IDS 400

Final Project Report

Arifa Baber

Jing Wen Chui

Anh Pham

Francisco Sanchez

Joseline Tanujaya

**Introduction**

Marketing campaign refers to a planned series of activities to promote service, product or brand to achieve a larger business goal. The importance of having a marketing campaign in a business is to provide a direction for companies to reach customers in different market segments with customer marketing strategies. Customer marketing differs from marketing strategy as marketing strategy analyzes sales activity through different angles and aims to overcome competitions from competitors in the industry while utilizing the findings to boost sales and exposure. By having a successful marketing campaign for the business, not only can it increase the sales, but it can also bring a significant impact to the business in the long run.

By understanding the purpose of marketing campaigns, the objective of this project is to increase the marketing efficacy by using Python to identify potential customers in the retail food sector. We managed to find a data set from Kaggle that contains some important features that will enable us to deeply analyze and find out what is the most approachable campaign to different audience groups. The nature of this Marketing Campaign data set enables a relatively high usability; it contains a series of product categories and customer demographics, including marital status, age group, education level, type of marketing campaign, and the number of purchases through different platforms.

**Project Overview**

For this project, we are using a dataset from Kaggle which can be found in this [link](https://www.kaggle.com/rodsaldanha/arketing-campaign). This dataset contains 2,240 observations, which are records of customer’s responses from marketing campaigns done by a company. Initially, the dataset contains 29 columns, these explanations are obtained from the original data source:

1. ID - The customer’s ID
2. Year\_Birth - the customer’s year of birth
3. AcceptedCmp1 - 1 if customer accepted the offer in the 1st campaign, 0 otherwise
4. AcceptedCmp2 - 1 if customer accepted the offer in the 2nd campaign, 0 otherwise
5. AcceptedCmp3 - 1 if customer accepted the offer in the 3rd campaign, 0 otherwise
6. AcceptedCmp4 - 1 if customer accepted the offer in the 4th campaign, 0 otherwise
7. AcceptedCmp5 - 1 if customer accepted the offer in the 5th campaign, 0 otherwise
8. Response (target) - 1 if customer accepted the offer in the last campaign, 0 otherwise
9. Complain - 1 if customer complained in the last 2 years
10. DtCustomer - date of customer’s enrolment with the company
11. Education - customer’s level of education
12. Marital\_Status - customer’s marital status
13. Kidhome - number of small children in customer’s household
14. Teenhome - number of teenagers in customer’s household
15. Income - customer’s yearly household income
16. MntFishProducts - amount spent on fish products in the last 2 years
17. MntMeatProducts - amount spent on meat products in the last 2 years
18. MntFruits - amount spent on fruits products in the last 2 years
19. MntSweetProducts - amount spent on sweet products in the last 2 years
20. MntWines - amount spent on wine products in the last 2 years
21. MntGoldProds - amount spent on gold products in the last 2 years
22. NumDealsPurchases - number of purchases made with discount
23. NumCatalogPurchases - number of purchases made using catalogue
24. NumStorePurchases - number of purchases made directly in stores
25. NumWebPurchases - number of purchases made through company’s website
26. NumWebVisitsMonth - number of visits to company’s web site in the last month
27. Recency - number of days since the last purchase
28. Z\_CostContact - cost of sending one campaign to the customer
29. Z\_Revenue - revenue if the customer responds to the campaign

**Project Objective**

The objective brings us to identify the business problems that we would like to address and verify through data visualizations and machine learning algorithm to make predictions:

1. Which cohort is the company’s true target market? By identifying this, the company can focus its conversion marketing efforts (marketing campaigns with an objective to recruit new customers) to a specific target market.
   1. Exploratory data analysis can tell us at which age a customer is more likely to be converted, what their marital status is, and how educated they are.
   2. How do the customers behave in relation to their purchases? By solving this, we attempt to understand how the customers conduct their purchases to make future strategic decisions.
   3. To achieve this, we will observe interactions between variables and their visualizations. Finding the relationship between the customer’s website visits and their online purchases is one of the correlations we will look into.
2. Which customer will respond to the next campaign? By answering this question, we can target the people who are more likely to respond to the campaign, therefore focusing the marketing efforts and budgets on a predicted cohort only. This serves to maximize campaign profitability.
   1. To do this, we will attempt to build machine learning algorithms to predict the “Response” variable.

**Business Value**

This project has three business values:

First of all, it allows marketers to identify the true target customers under different segments. It can be seen that there are a lot of factors that can directly and indirectly affect the decision making of customers. In this project, we plan to look at the important common sides to narrow down those factors such as age, marital status and education level. Through finding the relationship between those factors and customer decision making, marketers can improve their marketing strategy to target the right customer segments as well as the effectiveness of recruiting new customers.

Secondly, it helps marketers to understand more about the relationship between customer behavior in purchasing different products through different methods (website, catalog, store, and deal purchases). This project also helps marketers to learn about whether there is a positive correlation between the number of website visits from customers and real time data that recorded their actual purchase through the website after surfing. By these observations, marketers can make more effective future strategies in marketing.

Lastly, to be able to recognize which customers would respond to the 6th campaign, marketers are able to feed and train data using different machine learning algorithms to build the most effective algorithm in order to predict the future “Response” variable based on the dataset from Kaggle. By predicting the outcome on whether customers are going to respond on the 6th campaign, it helps marketers to propose reasonable cost distribution in investing in different marketing methods and maximize campaign profitability.

**Project Analysis**

In conducting our analysis, the following steps we took are:

1. Data preparation and cleaning: In this step, we made sure there is no NULL or NA value that would interfere with our analysis, and derived new variables from the existing data that will be more informative.
2. Exploratory data analysis: In this section, we will observe the existing variables in terms of their summary statistics and distribution to understand the nature of the individual variables. Through exploration, we can also better understand the true conditions of the business we are studying.
3. Visualize data to observe interactions between variables: In this section, we seek to observe the relationship between independent variables to deepen our understanding of the data in order to be able to interpret model results.
4. Predict the customers who are most likely to respond to the next campaign: This is the core focus of this project, where we aim to increase marketing cost efficiency and therefore increase profits.

**Data Preparation**

To standardize the findings and analysis that will be made in the following sections, having a clean and straightforward data set is important to increase productivity in general. By having a clean data set, it allows us to deliver high quality information and make sure the deliverables are clear for clients or users when they review the analysis during the decision making process. The master data is cleaned through the following updates:

Update/ Remove Column

1. ID column is being dropped as it does not contribute findings to the analysis

2. Age column is created by calculating the current year minus the year of birth of the customers to group the customers by age for future analysis

3. Spending column is created by adding all products categories including ‘MntWines’, ‘MntFruits’, ‘MntMeatProducts’, ‘MntFishProducts’, ‘MntSweetProducts’, ‘MntGoldProds’

4. Has\_Child column is created to identify if customer has a child

1. Categorize customer by how many campaigns they respond to at the 6th Campaign:
   1. 'Repeat' if the customer has made a response within the previous 5 campaigns
   2. 'Just Converted' if the customer has not made a response within the previous 5 campaigns, but made a response to the 6th
   3. ‘'Not Converted' if the customer has not made a response within the previous 5 campaigns, and does not respond to the 6th
   4. 'Not Responding' if the customer responded to the previous campaigns but not on the 6th campaign
2. Derive the age at which a person becomes a customer
3. Clean Marital\_Status column: ‘Alone’, ‘YOLO’, ‘Absurd’ are converted to ‘Single’.
4. Derive household size based on Marital\_Status, Kidhome and Teenhome

Missing Data

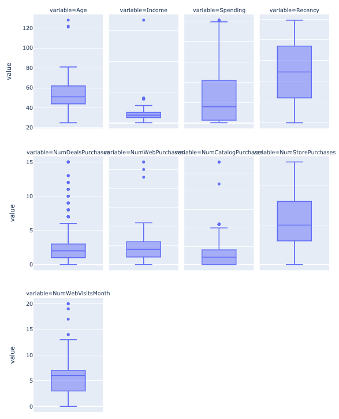
Other than the simple updates and data combination to have a better idea of the particular variables, it is also important to identify the missing values in this real-world data set, where missing values can occur due to the process of data entry errors, data collections, or customers not feeling comfortable to disclose their personal details. Hence, this could be the case where income is confidential information, and all of the missing value comes from the Income variable. Based on the findings, the number of missing Income data is as follows along with its proportion against the data set:



Understand that missing data shall not be zero as it would affect the overall accuracy and distribution for our future analysis, K Nearest Neighbors algorithm is used to identify the predicted missing income data with n\_neighbours of 5.

Remove Outliers

It is important to identify potential outliers and possibly replace them as they indicate bad data, which may have been inputted incorrectly or may have occurred due to random variation. Boxplot is one of the methods to visualize the distributions of all data. With that, we chose numerical variables that have the potential to be useful in our future analysis, proceeding with the boxplot plotting:



This is derived through the method of interquartile range method by calculating the interquartile range:

* Q1 = first quartile
* Q3 = third quartile

where:

IQR = Q3 - Q1

With the condition of having any value that is lower or greater than 1.5 times below or above the third or the first quartile of the IQR will be defined as an acceptable outlier. Other than that, any value that is lower or greater than 3 times below or above the third or the first quartile of the IQR will be defined as an not acceptable outlier as it is considered as an extreme data point when compared to the rest of the data points within the range.

With this consideration, all variables’ outliers except Income are not removed as they are significant to the data set to indicate they are in different bins, whereas income has an extreme outlier with only one data point of $666,000 per year.

Code Assignment

For easier referencing in machine learning later on, some variables are being assigned to codes by adding columns to label the types of the variables with integers. For example, Marital Status consists of eight types of marital status which includes ‘Single', 'Together', 'Married', 'Divorced', 'Widow', 'Alone', 'Absurd', 'YOLO', and we are categorizing them into 4 codes:

|  |  |
| --- | --- |
| **Marital\_Status\_Code** | **Marital\_Status** |
| 1 | Alone, Single, Absurd, YOLO, Widow |
| 2 | Divorced |
| 3 | Married |
| 4 | Together |

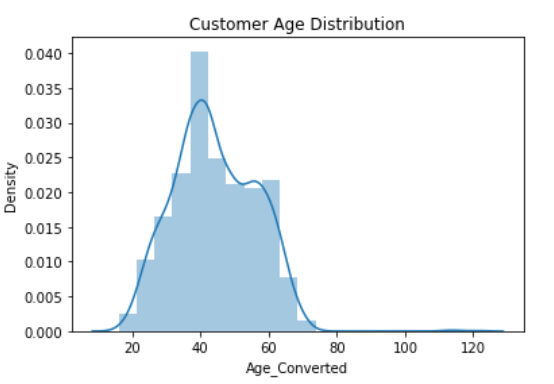
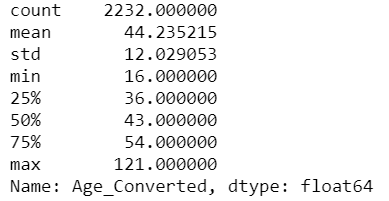
The same procedure applies to Education variables:

|  |  |
| --- | --- |
| **Education\_Code** | **Education** |
| 1 | 2nd Cycle |
| 2 | Graduation |
| 3 | PhD |
| 4 | Master |
| 5 | Basic |

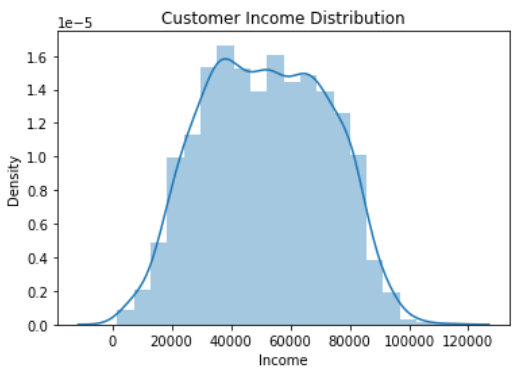
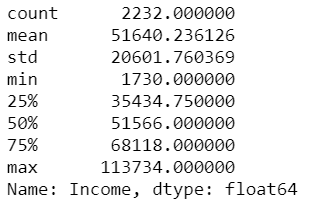
**Exploratory Data Analysis**

Demographic Variables

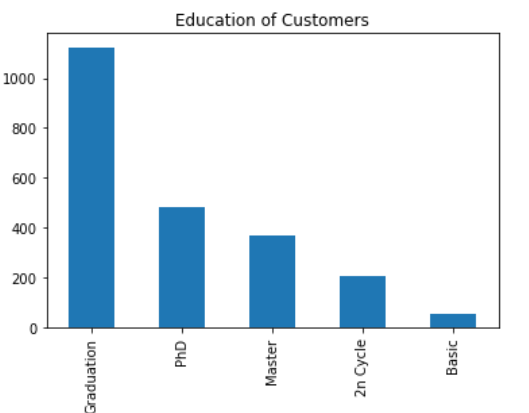
We will first explore the demographic variables independently.

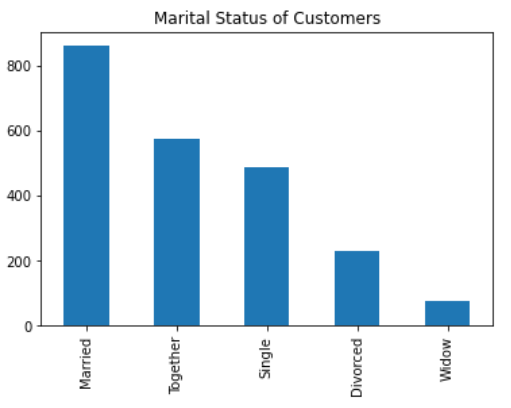
This analysis is showing that our customers age follows a rough Gaussian/Normal distribution with mean around 40-50 years old.

This analysis is showing that our customer's income follows a rough Gaussian/Normal distribution with significant high-side outliers.

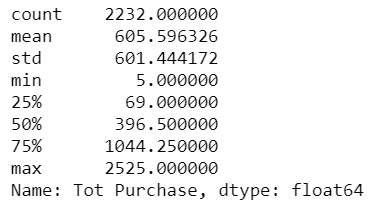
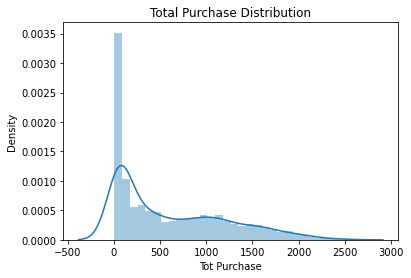


Our customers mostly go to college. We will see if the difference in education levels mean something to their purchasing power and behavior later in the analysis.

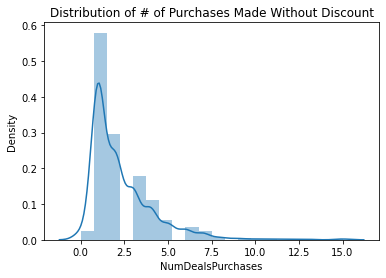
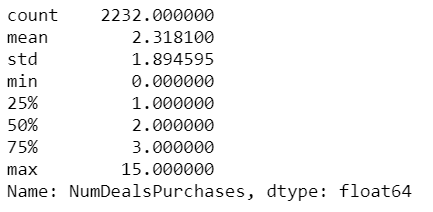


Our customers are mostly married or living with a partner. We will see if the differences in Marital\_Status mean something to their purchasing behavior later in the analysis.

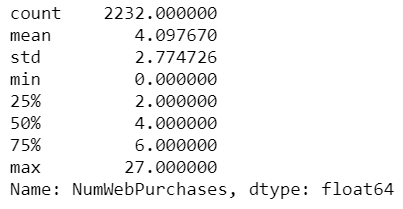
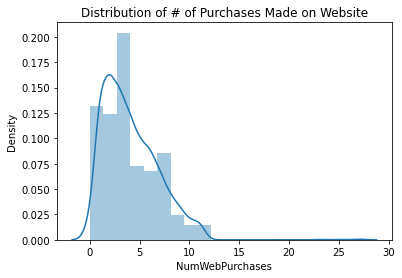
Purchasing behavior variables



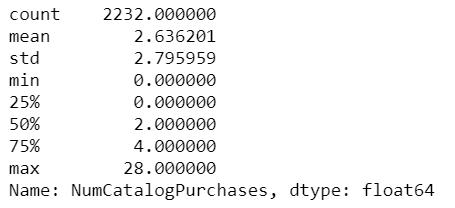
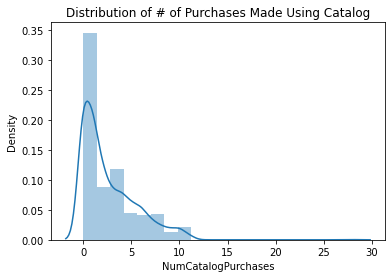
The distribution is severely right tailed. Half of the customers have accumulated purchases amounting to less than $396. However some extremes are also observable.

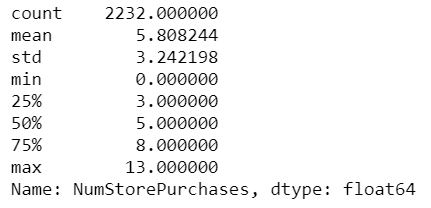
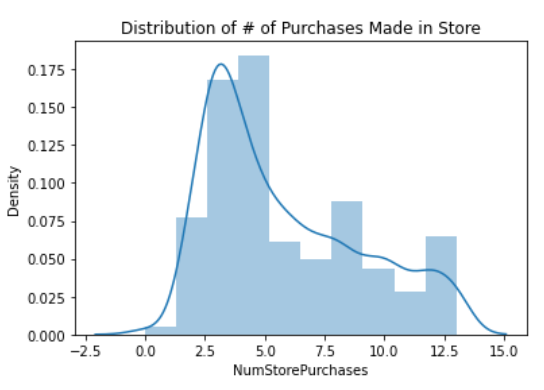
The distribution is severely right tailed. Most customers shop with discounts.



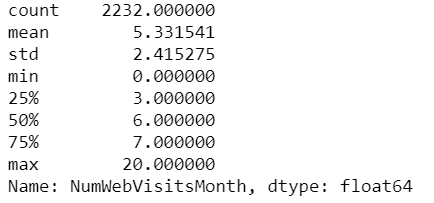
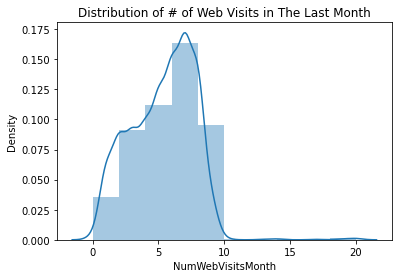
Most of the customers only have a small number of purchases made through the website.



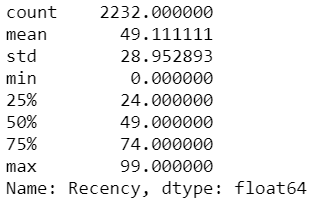
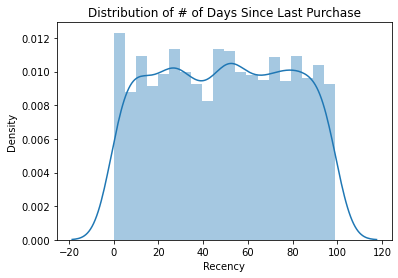
The distribution is skewed to the right. Most customers shop a small number of times using the catalog.



This analysis is showing that the number of store purchases made directly in store follows a skewed right tail distribution.

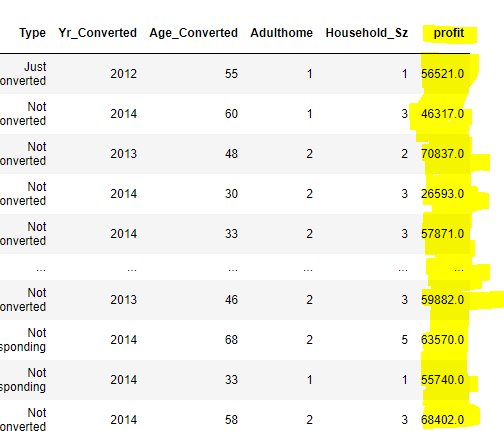


This analysis is showing that the number of website visits in the last month by a customer follows a skewed left tail distribution.

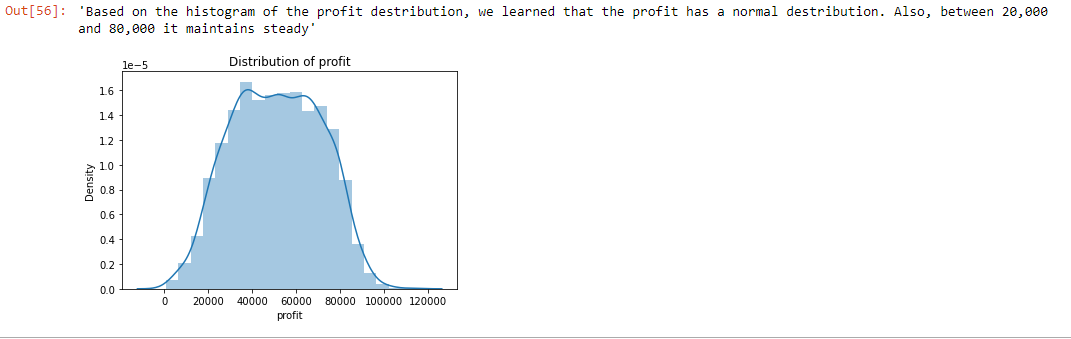


This analysis is showing that the number of days since the last purchase is consistent throughout 100 days. This indicates that there is no apparent point of time when a customer is more likely to stop being a customer to anticipate, nor there is a cycle to observe.

As we analyze the variables from the dataset, we realized that we can create a profit variable by calculating customers’ income menus total purchase expense. The goal is to learn how much our customers are willing to spend and suggest improvements by targeting different demographic groups, such as education level and marital status.



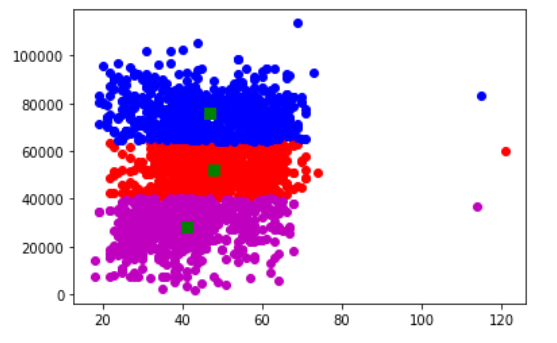
When analyzing the individual variables, we desired to extract data and turn it into relevant information. Based on the analysis, we identified that there was a need to create a profitable variable. Thus, we discovered a normal distribution among the customers. We learned that there is a consistency between 30,000 to 80,000 and that there are fewer instances making 20,000 or less.



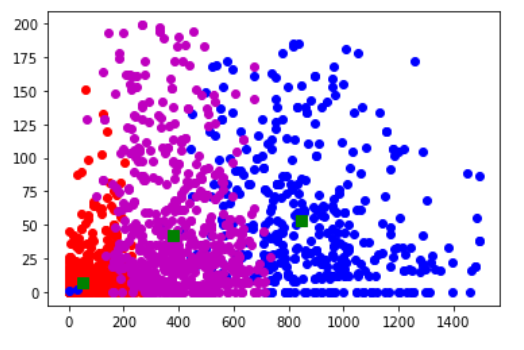
K-Means Clustering

We attempted to do K-means clustering based on different variables.

Based on demographics (variables included are: 'Age', 'Income', 'Spending', 'Kidhome', 'Teenhome')

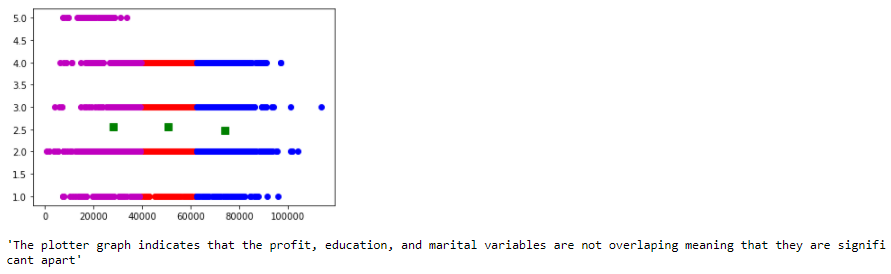


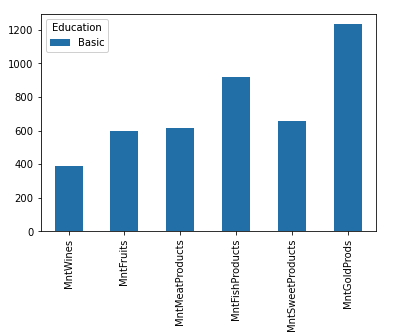
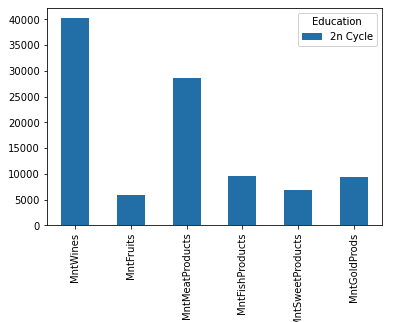
Based on purchase history, the variables included are: 'MntWines', 'MntFruits', ‘MntMeatProducts’ ‘MntFishProducts’, ‘MntSweetProducts’, ‘MntGoldProds’, 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', and 'NumWebVisitsMonth'.

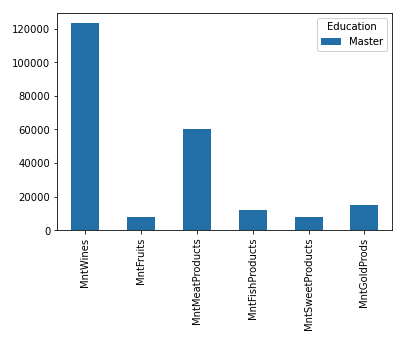
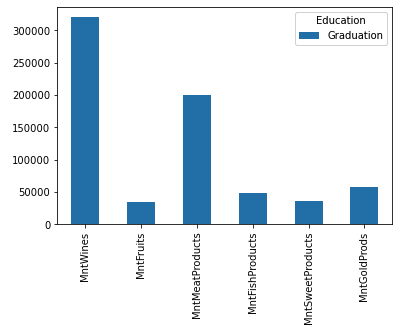
**

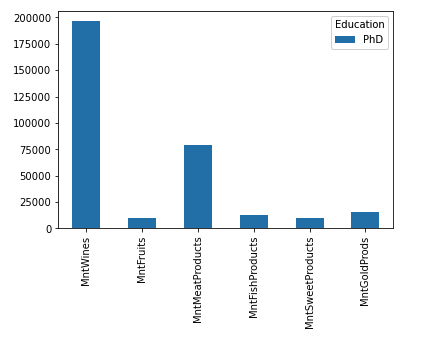
Clustering based on purchase history is evidently less clear than clustering based on demographics. This is indicated by the heavily overlapping regions on different clusters on the clustering based on purchase history. The clustering based on demographics presents a clearer segmentation.

Once we created the ‘profit’ variable, we wanted to continue exploring the data. However, to move forward, we tried to cluster and learn more about the data. Why is it essential to cluster data? Clusters support innovation and growth. They start in many ways, but all offer knowledge sharing, partnership, infrastructure, a skill pool, and career opportunities. For big businesses – being present in a relevant cluster is a great way to sense the direction of innovation and to find new partners. Therefore, we decided to analyze the data through K-means clustering. Thus, we wanted to cluster the profit, education, marital status. The goal is to see if these variables are significant .Based on the graph above, we learned that the variables profit, education, and marital status are not overlapping, meaning that they are significant apart. Therefore, we concluded that it is okay to create a deep analysis of these variables.

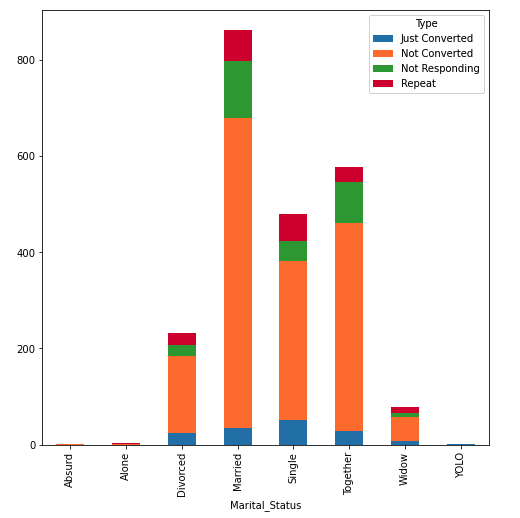
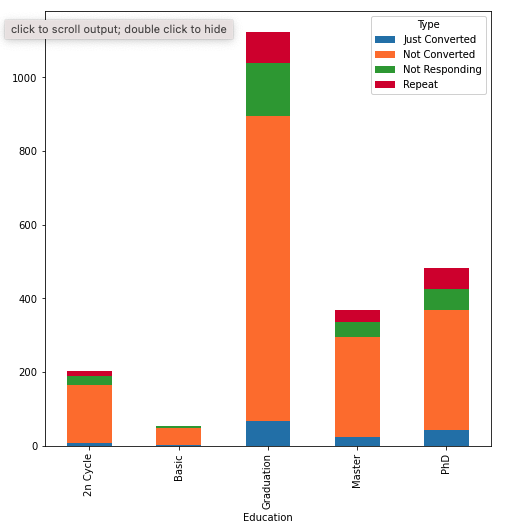








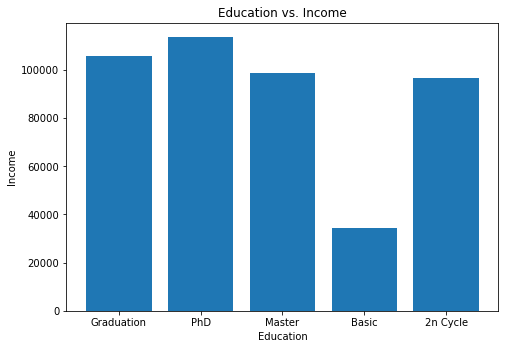
The 5 graphs above are the total number of sales for each product and for each level of education. The products included are: “MntWines”, “MntFruits”, “MntMeatProducts”, “MntFishProducts”, “MntSweetProducts”, and “MntGoldProds”. Here we can see that the sale of “MntWines” is the highest in higher education such as Phd, Master, Graduation and 2n Cycle while the sale of “MntGoldProds” is significantly higher in Basic level of education compared to others.



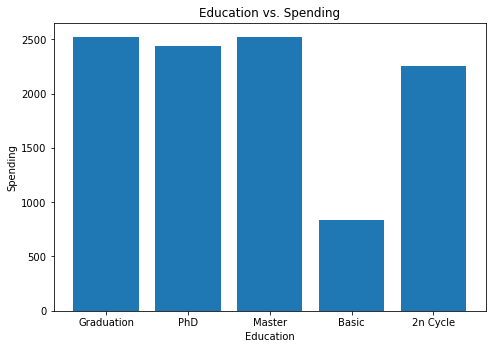
Next, we also explore the type of customer: “Just Converted”, “Not Converted”, “Not Responding” and “Repeat” in each type of education and marital status.

Education Variables

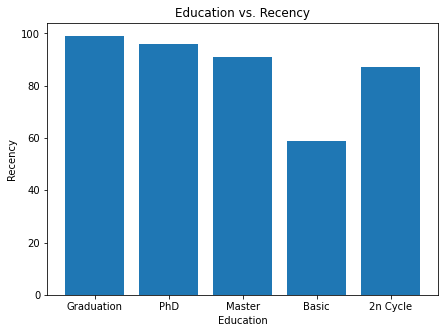
After evaluating various graphs and variables thus far, it was determined that Education is an important factor in deciding whether certain customers will respond to the 6th campaign or not. After careful consideration, we decided to compare Education to important demographic variables such as Income, Spending, and Recency to see the effects of Education.



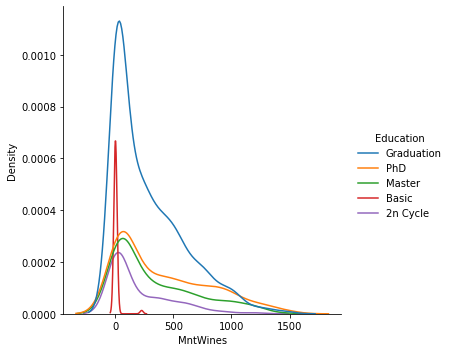
In the histogram labeled Education vs. Income, it can be seen that customers who have a degree in higher education earn more income annually. Customers with a PhD earn the highest income, followed by Graduation, Master Education, 2n Cycle, and Basic Education being the category with the lowest income.



Continuing with our comparison analysis of Education and significant variables, in the histogram presented above labeled Education vs. Spending, we decided to compare the educational level with the amount they spend by using the Spending variable. In the Education vs. Spending chart, it can be seen that customers with a higher education degree are spending more compared to those with a Basic Education. It is shown that customers with a Graduation, PhD, and Master Education are spending almost the same amount of money at a significantly higher rate compared to customers who have a Basic and 2n Cycle degree.



In the third comparison and analysis of the Education variable and other significant demographic variables, in the histogram labeled Education vs. Recency, we are comparing customer education level and recency. In marketing terms, recency is defined as the time elapsed sense a customer bought a product/service. Thus, in this Education vs. Recency chart, it can be visualized that customers with a degree in higher education have a higher Recency rate, where customers with a Graduation education level have the highest Recency rate of 100, followed by PhD, Master Education, 2n Cycle, and Basic Education level with the lowest Recency rate.

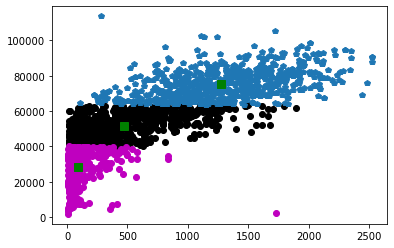


In our earlier stages of data exploration and collection, we identified MntWines to be the most successful product sold out of the six products that were mentioned in the data set. Thus, continuing with the significance of the Education variable, we decided to compare Education with MntWines to find out which education level is spending the most on the MntWines product.

After careful consideration, we decided to use a density plot to observe the distribution of MntWines bought and education level. When looking at the density plot, it can be seen that customers with a Graduation education level have the highest concentration of MntWines bought, followed by Basic, PhD, Master Education, and lastly, 2n Cycle.

K-mean Clustering

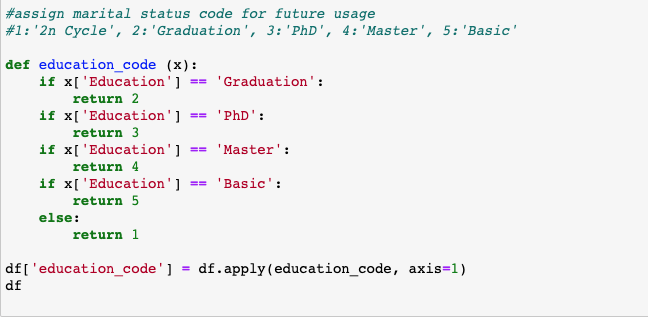
After doing a deep exploration of the Education variable and how it compares with other significant variables in our data set, such as Income, Spending, Recency, and MntWines, a k-means clustering algorithm using k= 3 was produced to see the significance of these variables when clustered together.

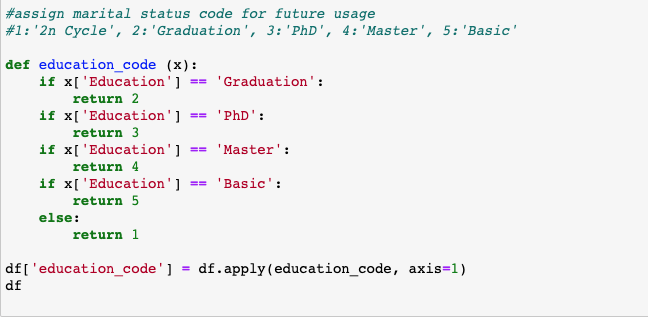


**Textual Data Analysis**

Text Data Cleaning and Preprocessing Data

To be able to explore and apply the variables in the dataset to train the machine learning model, we had to preprocess some textual data. To clean and standardize all the data, we converted “Education” and “Marital\_Status” columns’ data to “education\_code” and “marital\_status\_code” respectively. For this step, we only use this preprocessing data to make the EDA easier for us. Later on in the machine learning part, we actually use a more effective way by using get\_dummies() function to create dummy variables to train in the dataset.

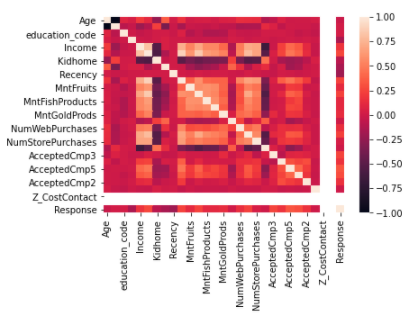




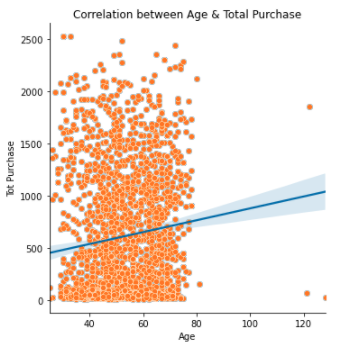
**Heatmap**

Examining the correlation between variables

To be able to check whether the variables in the dataset are related to each other, we ran the corr() function in our code. We took the top 15 variables and plotted them against each other on the heatmap. To create a heatmap, we used the heatmap function from Seaborn. From the heatmap we attached below, it can be seen that we were able to observe a decent correlation between the variables and the “Response” column. Normally, correlation strength is measured from -1.00 to +1.00 and when the value is closer to -1.00 or +1.00, it implies the stronger relationship between the two variables. When doing the EDA, we discovered that the dataset is quite unbalanced that it has more “No” response than “Yes” (1898 versus 334). Looking at the heatmap, we can not see the significant correlation between those variables due to the unbalanced in the “Response” column. In order to discover more about the correlation, we proceed to plot correlation and do the Chi-Squared test.



**Correlation Plot**

****

While the heatmap does show the correlation between some variables, it does not provide full visualization of the correlation of the important variables that we explore in the previous section. Hence, we made use of the correlation plot to identify if there is a strong, weak, or even no relationship between Age, Total Purchase and Income variables.

The correlation plot shows a greater positive relationship between Income and Total Purchase. In comparison to that, the correlation between Age and Total Purchase have less strong positive relationship and the distribution of data does not equally distributed as compared to the income and total purchase plot.

**Chi-Squared Test**

It is noted that the Chi-Squared test is one of the important tools to determine if there is an association between categorical variables. The purpose of doing Chi-Squared Test is to identify if there is a co relationship between the Education and Conversion rate. By tabulating the two variables, we then can have a rough visualization of the data distribution across two variables before the performance of statistical hypothesis tests.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Type** | **Just Converted** | **Not Converted** | **Not Responding** | **Repeat** |
| 2n Cycle | 8 | 157 | 24 | 14 |
| Basic | 1 | 47 | 5 | 1 |
| Graduation | 68 | 826 | 146 | 884 |
| Master | 25 | 269 | 43 | 32 |
| PhD | 44 | 324 | 57 | 57 |

Having a statistical way to make conclusions from data allows us to be more confident on the performance lift knowing that it might be due to random chance alone. Hence the following are our hypotheses:

Ho: There is no relationship between Education and Conversion Rate (Type)

Ha: There is a relationship between Education and Conversion Rate (Type)

We formed the analysis plan with the following threshold and we came to the conclusion of having p value of 0.0002375:

|  |  |
| --- | --- |
|  | **Value** |
| Degree of Freedom | 1 |
| Chi Square | 13.508644118 |
| p-Value | 0.0002374670 |
| Alpha | 0.05 |
| Critical Value | 3.8414588206 |

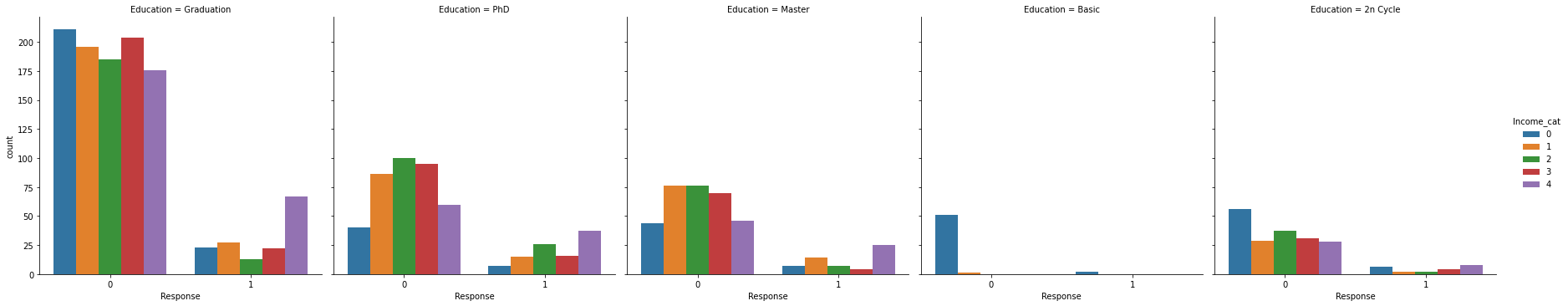
As the p-value is less than the alpha value, we reject Ho for Ha and conclude that there is a relationship between Education and Conversion Rate (Type).

**Machine Learning Metrics**

There are 5 preceding marketing campaigns, to which a customer may or may not respond (by making a purchase). Our task is to predict whether or not a customer will respond to the next (6th) campaign. The customers’ responses are stored in binary variables: 1 for yes an 0 for no in the column ‘Response’.

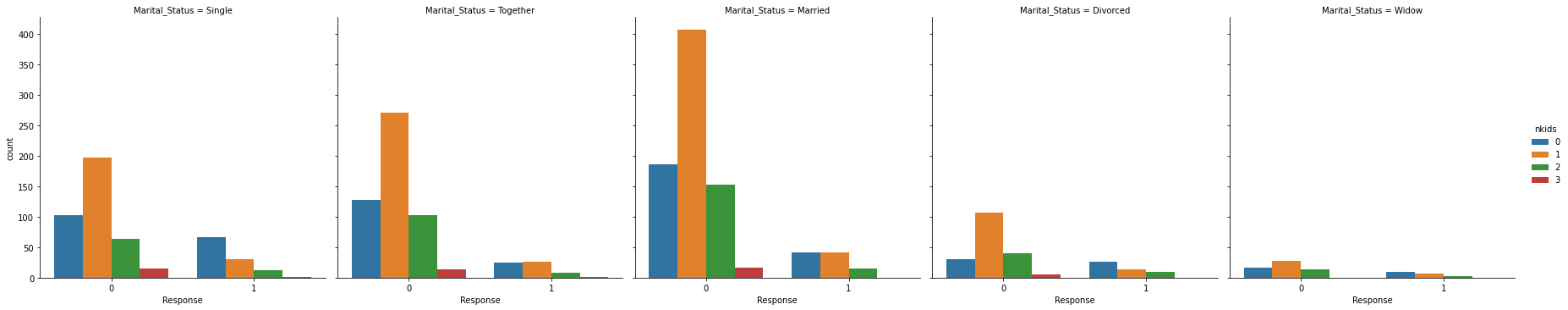
The goal of building machine learning algorithms is to effectively predict responses to the 6th campaign. Therefore, we can target people who are more likely to respond, hence saving marketing cost and maximizing revenue.

Alluding to our analysis, we visualize the data once more in relation to the response variable. We first visualize the interaction between the variable ‘Education’ with ‘Income’ towards ‘Response’. For the purpose of this analysis, we change the ‘Income’ variable to a categorical variables of 5 bins based on the quartiles, ranging from 0 (the lowest income) to 4 (the highest income)



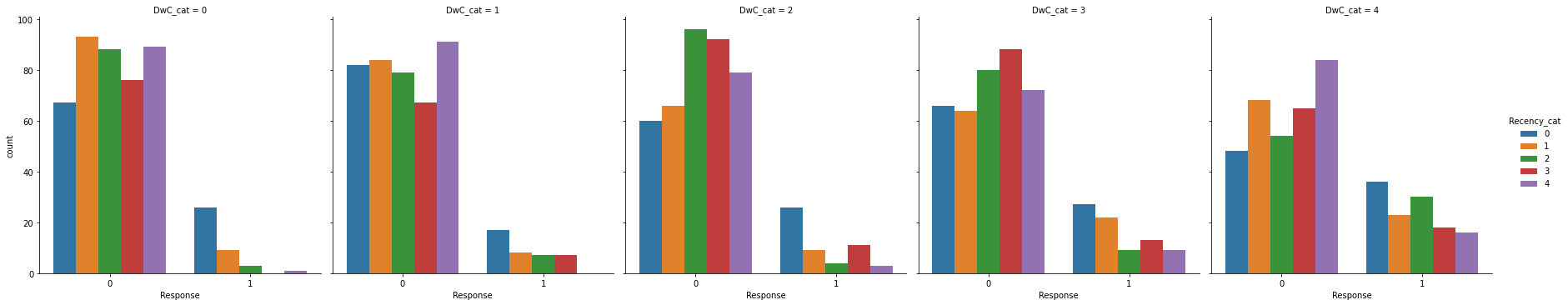
The highest rejection percentage is apparent on the education level = "Graduation". However for this education level, income category 4 is more likely to respond to the campaign. Income category 4 with education level of Graduation, PhD and Master are more likely to respond to the campaign.

We then visualize the interaction between the variable ‘Marital\_Status’ with ‘nkids’ (number of kids in the house) towards ‘Response’. For the purpose of this analysis, we derive ‘nkids’ from ‘Kidhome’ and ‘Teenhome’.



Across all Marital\_Status, people with 0 kids are more likely to respond. Highest rejection rate is apparent from people with 1 kid, across all Marital\_Status.

Finally, we visualize the interaction between the variable ‘DwC’ (stands for Days with Company) that accounts for how many days a customer has been a customer, with ‘Recency’, that depicts how many days it has been since the customer’s last purchase. We derive the DwC variable from the customer’s join date and 31 December 2014, which is the end of the year when the dataset was retrieved.



The longer a customer has been a customer for the company, the more likely they are to respond to the 6th campaign. Across all stages of the customer account's life within the company, customers that purchased recently are more likely to respond to the 6th campaign.

With these observations, we will move forward with creating the machine learning algorithms.

In making our models, we decided to keep the following independent variables:

'Age', 'education\_code', ‘Marital\_Status', 'Income', 'Spending', 'Kidhome', 'Teenhome', 'Recency', 'MntWines', 'MntFruits', 'MntMeatProducts', 'MntFishProducts', 'MntSweetProducts', 'MntGoldProds', 'NumDealsPurchases', 'NumWebPurchases', 'NumCatalogPurchases', 'NumStorePurchases', 'NumWebVisitsMonth', 'AcceptedCmp3', 'AcceptedCmp4', 'AcceptedCmp5', 'AcceptedCmp1', 'AcceptedCmp2', 'Complain', 'Days\_with\_company'.

The dependent variable is ‘Response’, which is a binary variable of 0 (no response) and 1 (response), indicating a customer’s response to the 6th campaign.

In the next steps, we followed through with this processes:

1. Make sure categorical variables are properly coded
2. Create dummy variables where necessary (where categorical variables are not ordinal)
   1. In this case, the Marital\_Status is converted to dummy variables.
3. Separate to training and testing sets
   1. We set aside 20% of the observations as a validation set.
4. Create k-folds validation
   1. K-folds cross validation results in the mean of 0.8521185110790285 and standard deviation of 0.02053905698962178.
5. Build machine learning algorithms. We made the following models:
   1. Logistic regression
   2. Linear Discriminant Analysis
   3. K Nearest Neighbors
   4. Decision Tree
   5. Gaussian Naive Bayes
   6. Support Vector Machine

Evaluate Models

This model is expected to achieve a certain business objective, that is to maximize profit. We will start by examining the profits in the no-model scenario.

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Per customer ($)** | **Customers** | **Total ($)** |
| **Cost** | (3) | 447 | (1341) |
| **Revenue** | 11 | 61 | 671 |
| **Profit/Loss** | | | (670) |

Under the no-model scenario, the company sends the campaign to every customer in the list, incurring a cost of $1,341 for the 6th campaign. Out of the 447 recipients, only 61 people responded, generating a revenue of $671. The total loss of the campaign is therefore $670.

With the new models, we will recalculate the profit. We will explain the process using the example of the Logistic Regression model:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Pred | |  |
|  | 0 | 1 | Tot |
|  | 0 | 369 | 17 | 386 |
| 1 | 50 | 11 | 61 |
|  | Tot | 419 | 28 | 447 |

Accuracy = 85% (correctly classified)

Precision = 39% (correct positives)

We are particularly interested in the accuracy and precision of our models, for the reason that we care about how many responses are predicted correctly out of the predictions of 1s.

In the Logistic Regression Model scenario, we will send out the campaign to 28 people who are predicted to respond to the 6th campaign. Out of the 28, 11 people responded. The profit calculation of this scenario is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Per customer ($)** | **Customers** | **Total ($)** |
| **Cost** | (3) | 27 | (84) |
| **Revenue** | 11 | 11 | 121 |
| **Profit/Loss** | | | 37 |

The profit jumped from a loss of $670 to a profit of $37. The model scenario is evidently better. Now we do this calculation for all models, and we come to the following tabulation:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | | | **No Model** | | | **Model** | | |
| **Model** | **Accuracy** | **Precision** | **Cost ($)** | **Rev ($)** | **Profit/ Loss ($)** | **Cost ($)** | **Rev ($)** | **Profit/ Loss ($)** |
| LR | 0.850112 | 0.392857 | 1341 | 671 | -670 | 84 | 121 | 37 |
| LDA | 0.89038 | 0.65 | 1341 | 671 | -670 | 120 | 286 | 166 |
| KNN | 0.834452 | 0.314286 | 1341 | 671 | -670 | 105 | 121 | 16 |
| DT | 0.856823 | 0.477612 | 1341 | 671 | -670 | 201 | 352 | 151 |
| Naïve Bayes | 0.756152 | 0.273585 | 1341 | 671 | -670 | 318 | 319 | 1 |
| SVM | 0.863535 | N/A | 1341 | 671 | -670 | 0 | 0 | 0 |

Through the table above we can see that the Linear Discriminant Analysis is the best at classifying the positives correctly (65% precision). Consequently, it generated the most profit of $166. The second best model is the Decision Tree, which boasts a 48% precision and generates $151 in profit.

**Conclusion**

We went through some steps in order to answer our business questions as per described in the earlier parts of this report, which can be summarized as follows:

1. Who are our customers and what is their purchasing behavior?
2. Who are more likely to respond to our marketing campaigns?

The two questions are going to be answered through data visualization and machine learning algorithms respectively. In answering these questions, we aim to guide future business decision making and maximize profits.

Our analysis began with data cleaning and preparation, which is crucial in setting the foundation for a firm and reliable analysis. The main analysis can be compartmentalized into 2 parts: 1) data visualization, which aims to explain the current conditions of the business, and 2) predictive modelling, which aims to help future campaigns.

Through our analysis, we have achieved our business objectives. We understood the true profile of the buyers and their purchasing behaviors through visualization, and predicted which customer is more likely to respond to a marketing campaign through machine learning.

**Recommendations**

Through the insights from this analysis we can derive the following suggestions:

1. To recruit new customers, we can direct our marketing efforts to the customer demographics that is evident from our data visualization (aged 40-50 years old, at least college educated, married/living together with a partner, earns around $38,000-$68,000).
2. Create different promotion schemes to target different purchasing behaviors:
   1. Give a reminder to come back to the store soon after the customer shops, because people who bought recently are more likely to respond to a campaign.
   2. Send clearance products only to people who tend to buy discounted products, because there are people who do not buy discounted products.
   3. Create a special catalog for people who do not have kids yet, because they are more likely to respond to a campaign.
   4. Give customers who shop in store coupons for shopping online, because our store purchases are strong but our online purchases are still growing.
3. Use our Linear Discriminant Analysis predictive model in future marketing campaigns, and send the campaigns to people who are predicted to respond to the campaign by the model .