Victoria Hall

DSC550-T301 Data Mining

Dr. Brett Werner

March 4, 2022

Final Project Submission: Data Mining with Customer Personality Analysis

Notebook Link - <https://github.com/vihall95/Data_Mining_CustomerPersonalityAnalysis>

Introduction:

A Customer Personality Analysis is an analysis of the ideal customer for a company. Customers behave in a number of ways and that behavior is common among different segments of customer groups. Demographic details including age, education status, marital status, income, and the presence of children can be used to identify these segments. In order for a business to increase customer engagement and plan resources according to what provides the most value for the lowest amount of money spent, customer segmentation can be used to focus advertising campaigns. The data set from Kaggle contains demographic information as well as whether or not customers accepted different add campaigns. The goal of the project is to use the features from this data to accurately predict what customers should be targeted for future advertising campaigns.

Most businesses settings have to balance costs with benefits and a Customer Personality Analysis can minimize costs by containing the scope of the project to those who are mostly likely to respond. For example, if a business had access to 10,000 customers in which they could advertise, and the cost of the advertisement was five dollars per advertisement, if they chose to advertise to all 10,000 customers, that would cost $50,000. In targeting all of the customers, there is potential for loss costs because some customers might not respond to the ad campaign or, they would have bought the product regardless of the campaign. However, using a combination of machine learning algorithms, the company could make predictions about the best customer segment to send the advertisements too. They could then limit the scope of who receives the campaign treatments to those who are mostly likely to engage because of that campaign.

These costs reductions would be the forefront of a pitch to group of stakeholders to gain buy in for the project. The targeted approach of this tactic significantly reduces the costs and promises more success due to the segmentation of customer groups. We can also identify the features of the different campaigns that caused the different segments to engage in.

The dataset for this project comes from Kaggle and was uploaded for the purpose of analyzing segmentation. The intention of the analysis is to help a business modify its products based on customer segments.

Milestone one allowed for an initial review of the dataset and the aspects of it that need further review for cleaning and modeling in the future. After loading in the dataset as data frame, the shape, size and data types were reviewed. After reviewing the frame, there were several columns that contained demographic information about each of the cases. Demographic information could inform the segmentation of customers, so I created a set of histograms to further inspect these potential features of our model. The results of those histograms can be seen below.

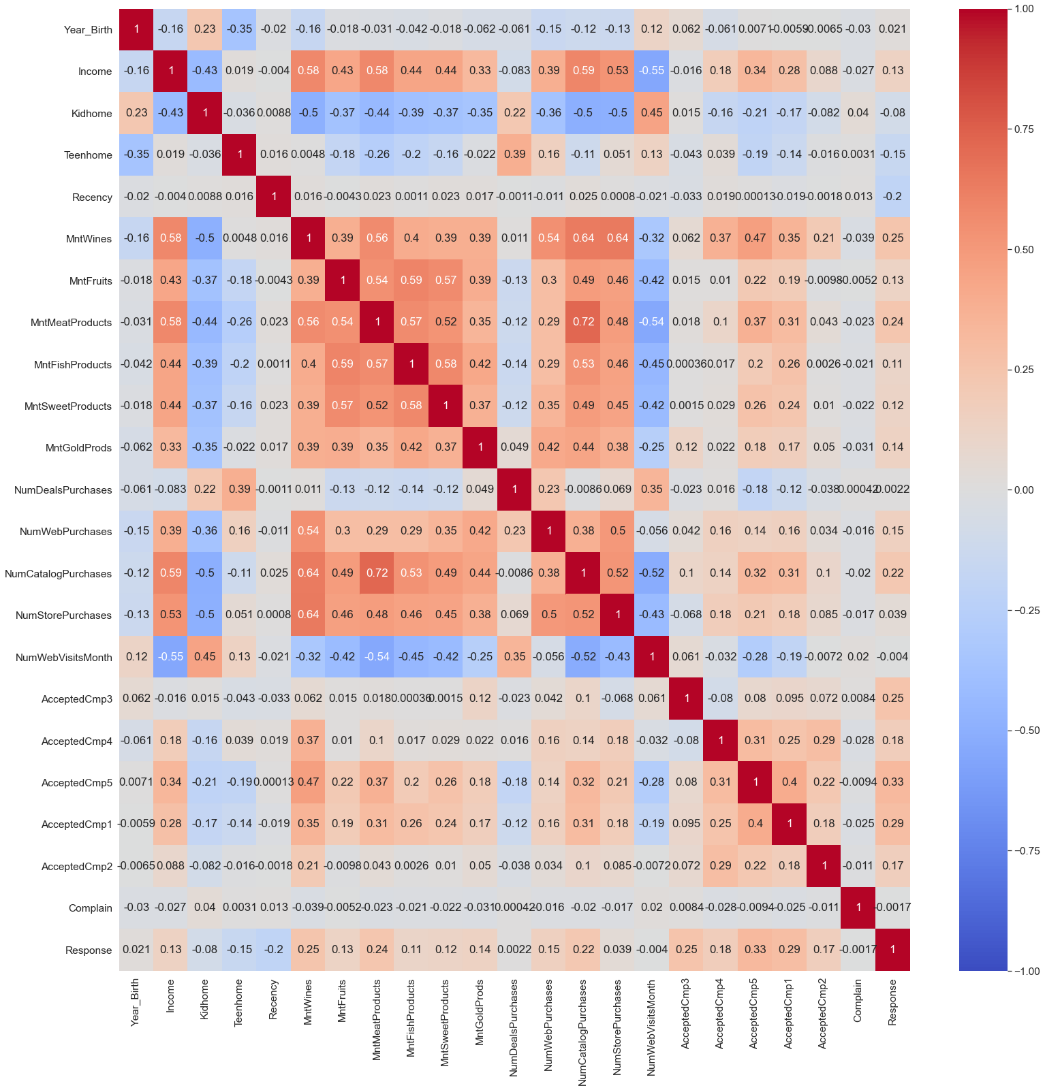
Logo

Description automatically generated with medium confidence

Histograms are typically used to summarize the frequency distribution of the data. However, they also present a unique opportunity to see what responses our in our frame and what is useful. Birth year, income, number of kids in home, and number of teens in the home are relatively straight forward. Most individuals were born between 1950 and 2000, though there may be some outliers in terms of being older. Income shows that we again have an outlier at the high end of income though most of the respondents were similar. We can also see that the majority of the respondents lived in household without kids or with only one kid or teenager. The education and marital status histograms show the distribution of responses as well, but it also highlights some issues with the data. The education labels are not clear and there appears to be joke answers to marital status question like ‘YOLO’ or absurd. This shows elements that will need to be cleaned before any models are trained.

Next, I reviewed the amount of money that each case spent on different items. I used boxplots so that the outliers were clearly identifiable, and we could see the main quartile ranges of the data. In each instance, we see large tails towards high amounts spent of the items. For example, we can see most customers spent between 0 and 500 dollars on wine. However, there are outliers that spent way more, up to 1,400 dollars. By using these boxplots, I was able to identify the presence of outliers that A picture containing text, silhouette

Description automatically generatedwill need to be dealt with later in the cleansing and model building stages.

The final aspect of the data I reviewed in milestone one was whether there was correlation between the features of the data. I used a heatmap which I annotated with the correlation value and colored so that more depth of color, the greater the correlation there was between the variables. Interestingly, we saw the highest levels of correlation among the amounts spent on different products. For example, spending on meat products was correlated with spending on fruits. It also showed elements that may not be important to the model construction because they had little correlation with any other features. Fortunately, no features, appeared to be over-correlated. The image below shows that correlation heat map.

Milestone two provided the opportunity to cleanse the data and prepare it for model building and execution. The first thing I did was to check for missing values. The only column with missing values was income and there were 24 missing values. The heatmap showed that income does correlate some of the other features, so I dropped the cases with missing income.

After removing missing values, I dealt with the data types in dataset. I converted so date information to datetime objects and transformed the date of enrollment into a categorical feature of the group that identified three groups based on seniority of enrollment. This helps support the segmentation of the analysis by creating another group-based feature. After this I renamed some columns so that it was easier to tell what information they provided. I also grouped responses to marital status and education into clear to understand groups. I also modified the data so that kids and teenagers in household were combined.

After these cleansing steps, I dealt with outliers. Milestone one showed that there was outliers in the income column. I decided that because the purpose of this task was to identify similar customer groups that the best option would be to remove outliers. I did not want a couple of cases to have undue influence on the model. In order to address outliers, I looked at the mean income and the value of one standard deviation. I ended up using Z Score to find values that were more than three standard deviations from the mean. I removed those cases with a Z Score greater than three.

The final step of milestone two was dealing with categorical variables. I created dummy variables for those categorial features that had a Boolean response of 1 or 0 to indicate the presence or absence of that categorical feature. Once, I created those features my dataset was prepared for modeling.

Before I started building and evaluating models in milestone 3, the first step I took was to split the data into training and test sets so that we could prevent over fitting and test the model on unseen data. The Boolean responses to whether the customer accepted the campaigns were our targets and the rest of the data were features. I did this for each of the six advertising campaigns so that each model would be tailored to that specific type of campaign.

I began model evaluation by building first a standard K Means clustering model. I chose this model because the purpose of this project was to identify similar customers and group them into segments. Clustering takes the observations and groups them into a number of clusters where each observation is grouped with to the cluster with the closest mean.

To do this I created a pipeline that scaled the data and then used it to fit the model. I set the initial number of clusters to three to begin with just as guess. After I fit each of the six models, I looked at the results of each model to see if it could accurately predict whether the customer responded to the campaign. I looked at the accuracy as well as the F1 score. The F1 score is found by taking the harmonic mean of precision and recall. Because this is a classification model, I choose these to evaluate the model performance. The performance of all six models was relatively similar. The most accurate model was about 56% accurate and the best F1 score was .47. This model only performed slightly better that random guessing.

Before I moved onto a new model to evaluate, I decided to use the elbow method to determine the optimal amount of clusters for each of the models. The elbow method uses the sum of the squared distances between the data points and their assigned clusters. The point in which the line bends is the optimal number of clusters. All six models performed best with the three clusters originally chosen.

The next classification model I test was a logistic regression model. Logistic regression models as mostly used on binary targets and the campaign responses are binary. Similar to the clustering model, I created a pipeline that scaled the features and then fit the model and looked at accuracy and F1 scores. In general, the logistic regression model performed better than the clustering model. Again, each of the models had similar results. The F1 scores were all above 40 and the accuracy was a little better but still not much better than random guessing. The best of the six models was 56 percent accurate.

After this, I decided to use a logistic regression model as the predictor of the campaign response. In order to improve this model’s performance, I wanted to dig further into the segmentation of the customers. Our correlation heatmap was showing there was some correlation among our features, but nothing in particular stood out. Because I was unable to glean any exact idea of what features were important, I decided to approach the data as an unsupervised learning task first. I used K Means clustering to find latent groupings that I was unable to identify. After identifying those clusters, I created labels for those newly identified groups and used them as a feature of each case. After creating the features, I turned these into dummy variables to feed into the logistic regression model. Then, similar to the other models, I created a pipe that scaled the data and then fit logistic regression model that was then used to predict responses. Each model for the six campaigns performed significantly better and in most cases doubling the accuracy and the F1 score. The most accurate model was for the second campaign at 92% accuracy and an F1 score of 92.

Conclusion

The model and analysis show that using clustering and then predicting led to highly accurate predictions of campaign responses. It shows that unsupervised learning tasks can help support other models’ accuracy. The model is probably not quite ready to be deployed. The dataset was relatively small with about 2000 cases. This doesn’t necessarily mean the models are inaccurate, but we could represent the population better if our models were trained with larger data sets. The complication of using this clustering technique is that we don’t necessarily know what features lead to the clusters that were created. So, it is hard to say something like people with 3 kids respond to this particular ad campaign. More work needs to be done to dissect these identifications. But overall, the model could be used in a business situation. With new customer information, we could feed that into the model and figure out what clusters those customers belong too and then use our business knowledge about the details of the campaign to better inform which customer segments to target.

References

Brendel, C. (2021, December 5). *Cluster-then-predict for classification tasks*. Medium. Retrieved March 4, 2022, from https://towardsdatascience.com/cluster-then-predict-for-classification-tasks-142fdfdc87d6

Patel, A. (2021, August 22). *Customer personality analysis*. Kaggle. Retrieved March 4, 2022, from https://www.kaggle.com/imakash3011/customer-personality-analysis

Vatsal. (2021, February 24). *K-means explained*. Medium. Retrieved March 4, 2022, from https://towardsdatascience.com/k-means-explained-10349949bd10

Widyadwatmaja, A. (2021, November 23). *Customer personality analysis segmentation (clustering)*. Medium. Retrieved March 4, 2022, from https://medium.com/@andhikaw.789/customer-personality-analysis-segmentation-clustering-1b68a62a61a2