Duke

E-Commerce Cosmetics Store Marketing Analysis

Final Project | EGRMGMT 590

Team 5
Ayush Jithin
Kyle Morris
Aakash Vaghani
Kayako Yamakoshi

Duke

Table of Contents

Duke

Context

In the past 20 years, e-commerce has been exploding in usage and value. As of 2020, global e-commerce sales reached roughly $4 billion USD. Due to the COVID-19 pandemic, online shopping has become increasingly important and an opportunity exists for companies to further improve their online sales revenue.

As students interested in marketing analytics, this cosmetics store ecommerce dataset provides us the chance to analyze customer behavior, identify target customer segments, and develop a model to predict customer purchases. In order to narrow the scope of our project, we chose to use data from a shop which specialized in cosmetics, which will reduce the number of categories of products we will have to account for.

Data Overview

The dataset chosen for this project, taken from an online cosmetics store and whose source can be seen in the references section, consists of user activity observations, which include viewing, adding to cart, removing from cart, and purchasing. Each unique observation has a user id, time, product id, event id, and session id. This combination of variables allows us to identify each user, identify the times when they are active, what kind of product they are interested in, and what products they eventually purchase.

event_time <chr>	event_type <chr>	product_id <dbl>	category_id <dbl>	category_code <chr>	brand <chr>	price <dbl>	user_id <dbl>	user_session <chr>
2019-10-01 00:00:00 UTC	cart	5773203	1.487580e+18	*NA*	runail	2.62	463240011	26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:03 UTC	cart	5773353	1.487580e+18	*NA*	runail	2.62	463240011	26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:07 UTC	cart	5881589	2.151191e+18	*NA*	lovely	13.48	429681830	49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:07 UTC	cart	5723490	1.487580e+18	*NA*	runail	2.62	463240011	26dd6e6e-4dac-4778-8d2c-92e149dab885
2019-10-01 00:00:15 UTC	cart	5881449	1.487580e+18	*NA*	lovely	0.56	429681830	49e8d843-adf3-428b-a2c3-fe8bc6a307c9
2019-10-01 00:00:16 UTC	cart	5857269	1.487580e+18	*NA*	runail	2.62	430174032	73dea1e7-664e-43f4-8b30-d32b9d5af04f

Fig. 1: Columns available in the cosmetics ecommerce data set.

This dataset consists of 8,782,890 unique observations and 9 columns, encompassing a time span of 5 months from October 2019 to February 2020.

Limitations

The most immediately apparent limitation is the sheer size of the dataset. With over 5 million unique observations, the data will need to be analyzed and processed in smaller pieces to match the computational power of our computers.

Another limitation is the variables themselves. For our uses, many of the variables do not provide any insight since we do not have a way to determine which products match up with each product

id and which category matches with each category id. For these reasons, we did not use these variables in our models. In addition to this, the *category_code* column consisted of 98% NA values, rendering it worthless. Therefore, we removed *category_code* from our data frame. The majority of observations in *brand* also had NA values, and therefore brand was not used in our models.

Exploratory Data Analysis

The Exploratory Data Analysis allows the team to develop familiarity with the data and derive some key insights. The first step was to understand user activity as a function of time. This analysis would allow us to determine when user activity is highest and subsequently make marketing decisions.

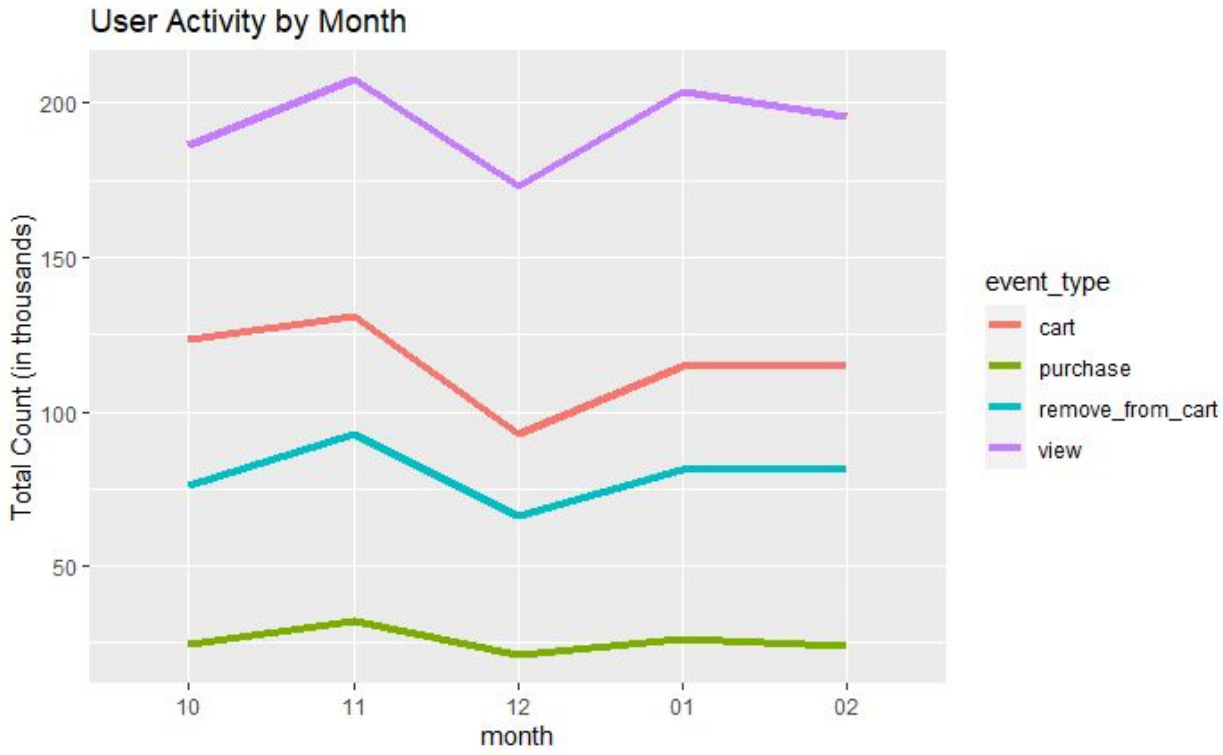

Fig. 2: User activity levels over a five month period.

As shown in Figure 2, we observed that the cart activity was highest in November. Considering the large size of our complete dataset and the potential to derive more important insights from the month of November, the team decided to work on the data from November for subsequent analyses.

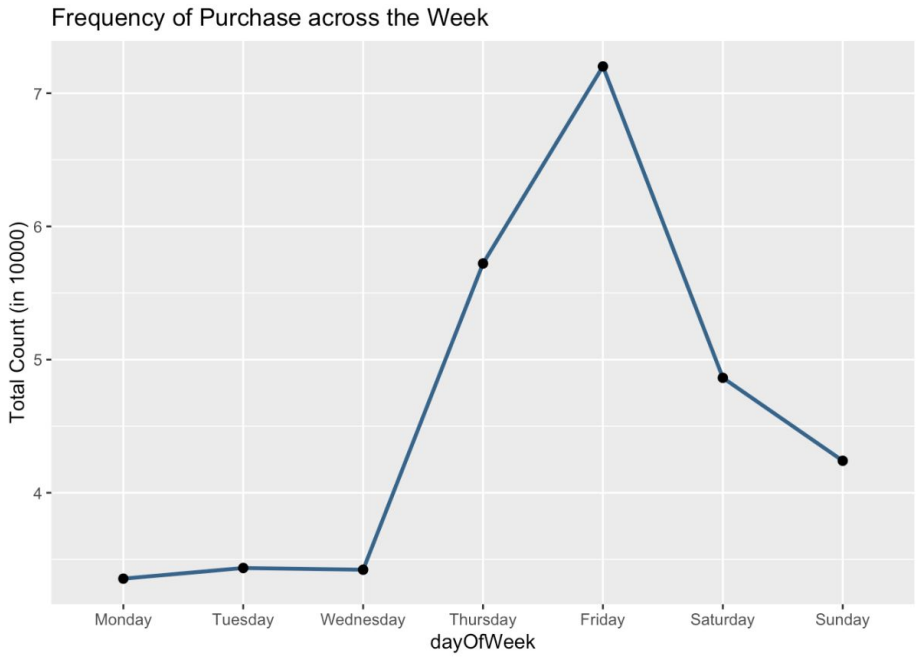

Fig. 3: Most purchases are made on Saturday.

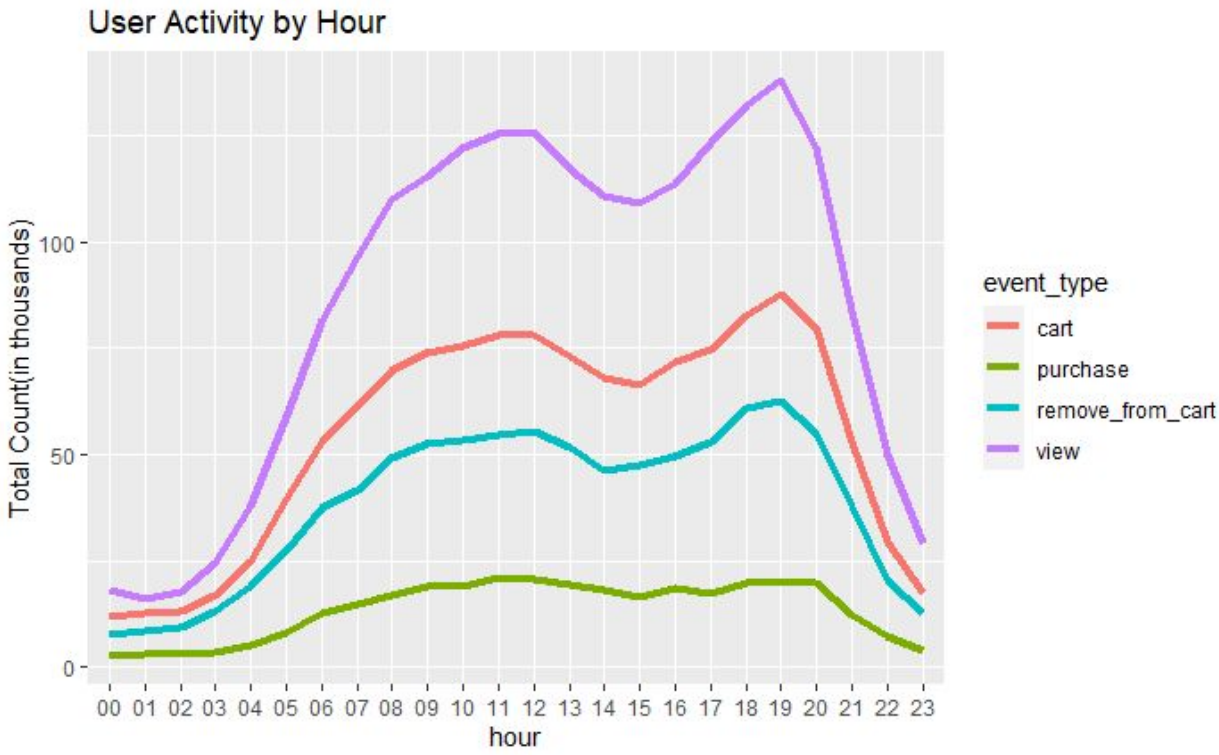

Fig. 4: User activity peaks at noon and 7pm.

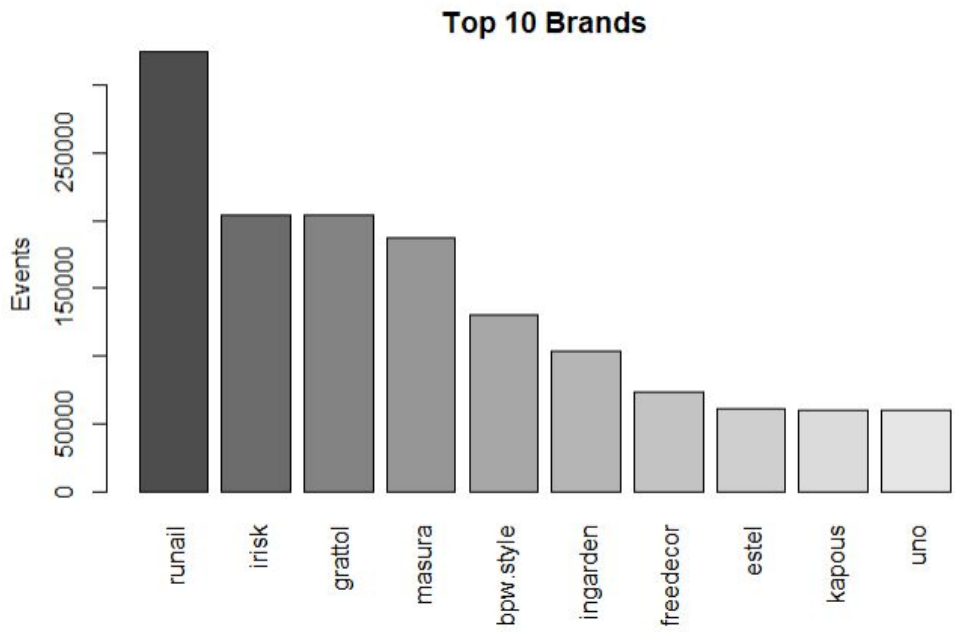

Fig. 5: Top 10 brands which generated the most activity on the site.

Another key insight derived from our EDA was the list of top 10 brands based on user activity. These brands represent those products that were viewed and purchased the most times, making them ideal choices to focus on for online marketing promotions and advertising.

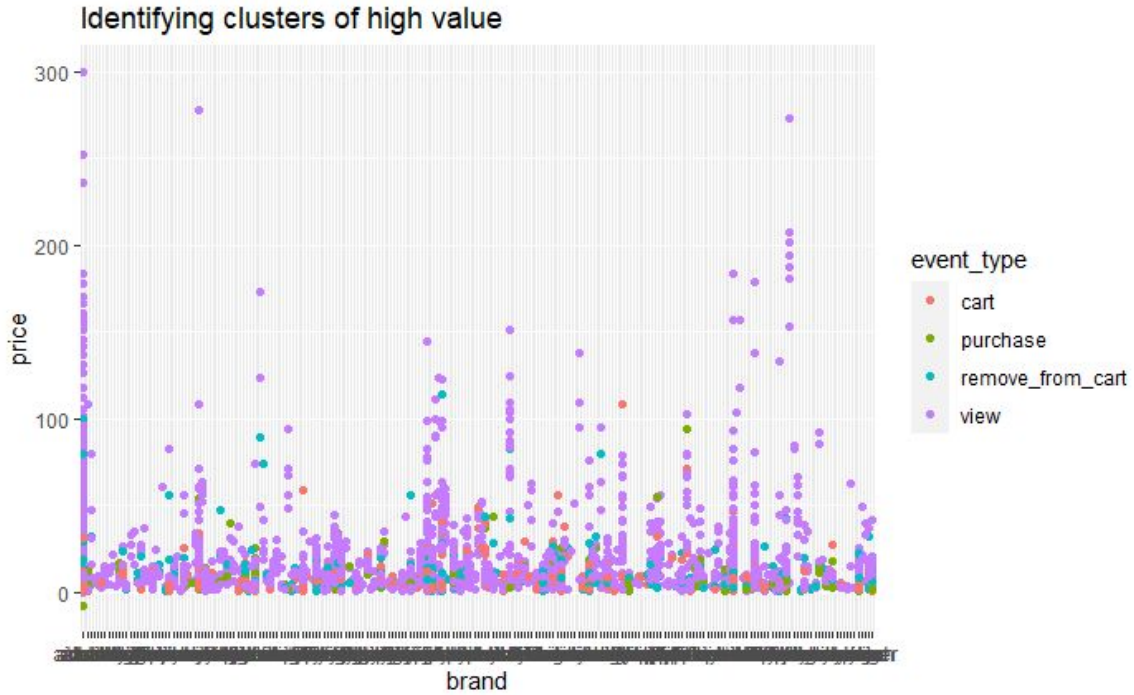

Fig.6: Plotting user activity of products against their prices to identify clusters of high value purchases

Another interesting visualization (Fig.6) that the team developed during the exploratory analysis was a plot of points denoting user activity of different products grouped under their brands. The green points signify purchases and this visualization can be used to identify clusters of high value purchases and more specifically, high value products that are purchased frequently.

Data Preprocessing

Data Cleaning

Based on the data summary, 98% of *category_code* and 43% of *brand* were NAs, and would not provide significant insights to the model. Furthermore, 7,662 observations had prices less than or equal to zero in the dataset, which accounted for only 0.2% of the dataset. We removed these insignificant columns and observations.

Next, the columns were transformed. Most original columns in the dataset were categorical, except for *event_time* and *price*. The categorical columns were transformed to a factor type, *event_time* to a POSIX time data type, and *price* to a number.

Feature Engineering

We engineered six features to be used as key predictors in our models. Since Figure 3 and 4 showed considerable impacts of time of the day and day of the week on purchase conversion, we added columns for these two features. Figure 2 also showed the user activity spikes around holiday seasons, therefore we created a feature, *days_till_christmas*. Three other features were created by aggregating the dataset by *user_id*. Those features represent days since last purchase, the number of items purchased, and total spending respectively.

Data Modeling

The e-commerce company needs to make sure whether they are meeting their target consumers' needs and launching effective marketing plans to stay competitive. In other words, having a good understanding of their existing customers and the factors driving their purchases is crucial to measuring the effectiveness of their marketing strategy. We have decided to take two approaches to gain those insights; clustering and classification. Clustering will be used to segment the existing customers and to identify profitable target segments. Classification will be used to identify the factors which have a significant impact on customer's purchase decision during the final stage of the marketing funnel, which is the transition from cart to purchase.

Clustering Model

The clustering model is built based on the RFM methodology, a segmentation method that focuses on recency, frequency, and monetary value of customer behaviors. We chose the RFM segmentation approach as it is simple and efficient, and provides high-level information on customers even within large databases. Our goal is to identify the segment that has a high potential to increase lifetime value.

We extracted 263,760 observations where *event_type* is equal to purchase to focus on the existing consumers that have made a purchase in the past. The predictors for this clustering model were days since last purchase, the number of items purchased, and total spending per customer. Given the nature of the dataset, it was challenging to identify whether multiple items were purchased at the same time and to measure how frequently a customer comes back to the website to purchase products. We made an assumption that the number of items purchased and how many times a customer comes back have a correlation to some degree, and decided to use the number of items as a proxy for frequency.

Data Standardization

As Figure 7 shows, the distribution of recency was almost normal. On the other hand, the distributions of frequency and monetary value were very skewed to the right and included many outliers.

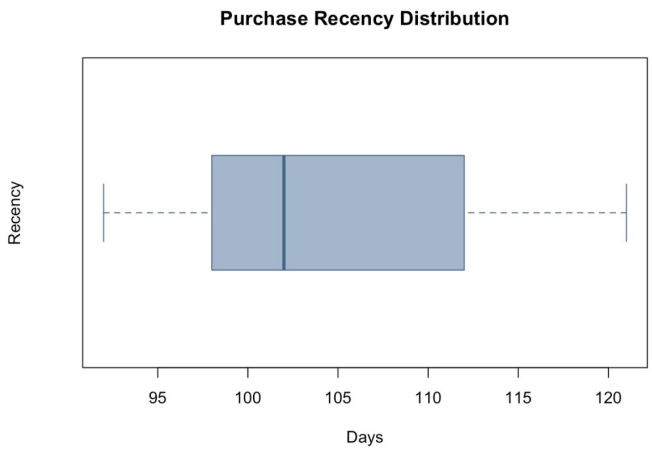

Fig. 7 : Distribution of days since last purchase.

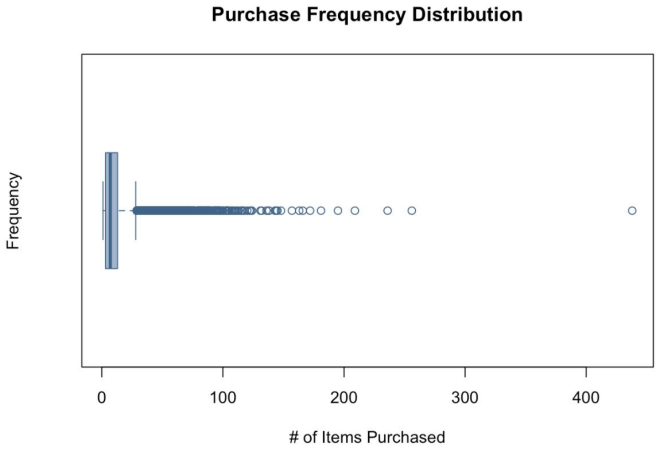

Fig. 8: Distribution of items purchased.

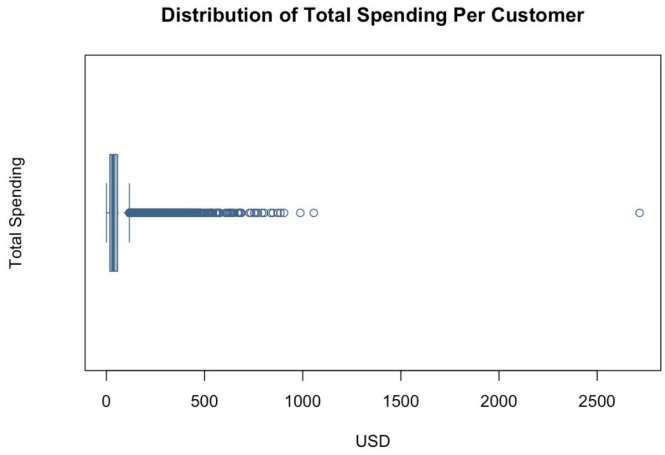

Fig. 9: distribution of total spending per customer.

The outliers would have significant impacts on the model and decrease its accuracy. Therefore, we set a threshold of $200 USD for monetary value and 25 times for frequency, and removed observations that were above the threshold. There were 28,220 observations left after filtering the data. This dataset was then standardized to put every predictor on the same scale. The distribution of each predictor after standardization is shown below in Figure 10.

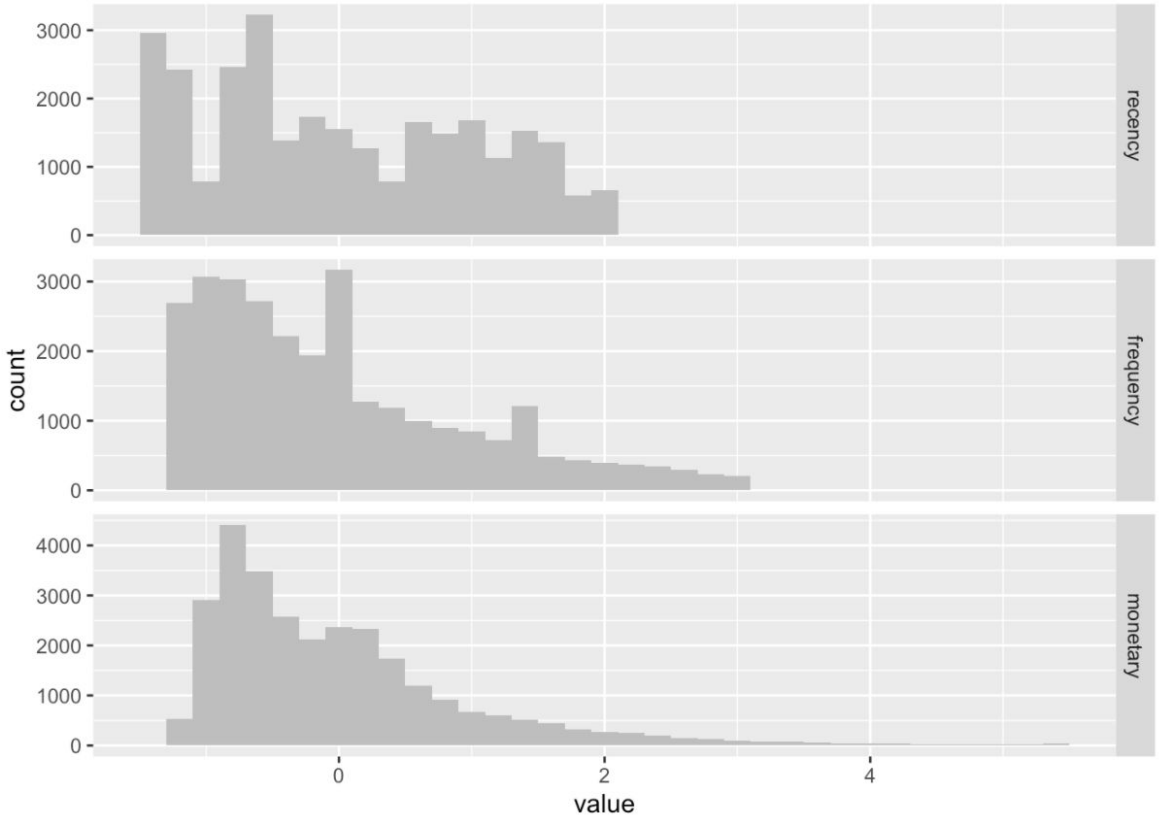

Fig. 10: Distributions of standardized recency, frequency, and monetary value.

Model Fitting

We implemented our clustering model using a k-means algorithm. K-means works relatively quickly with *O(n)* compared to other algorithms. After comparing the performance of k-means models with varying *k*, we decided to use *k*=3, following the elbow chart below.

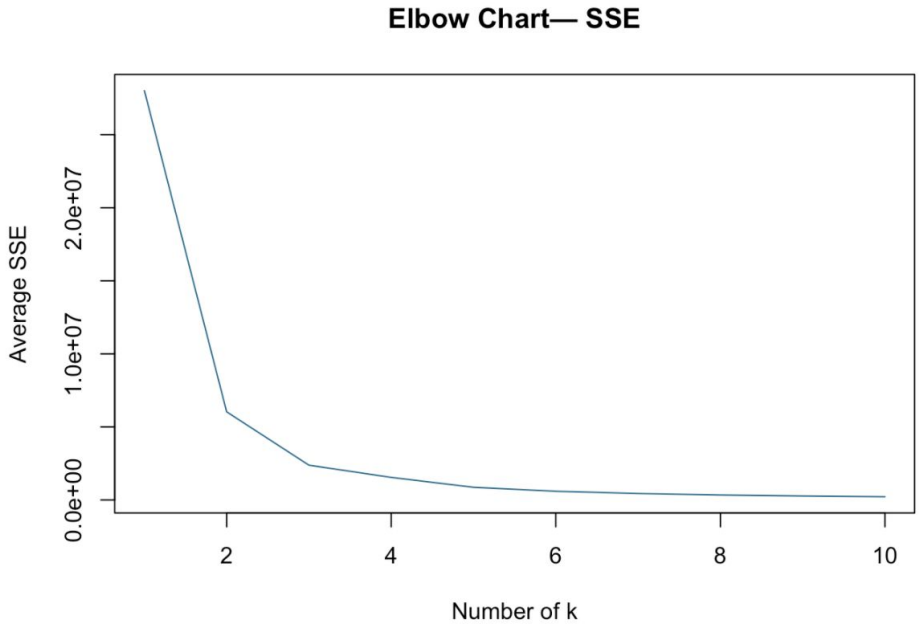

Fig. 11: Elbow chart shows k=3 is optimal.

The k-means model clearly separated data points into three different segments.

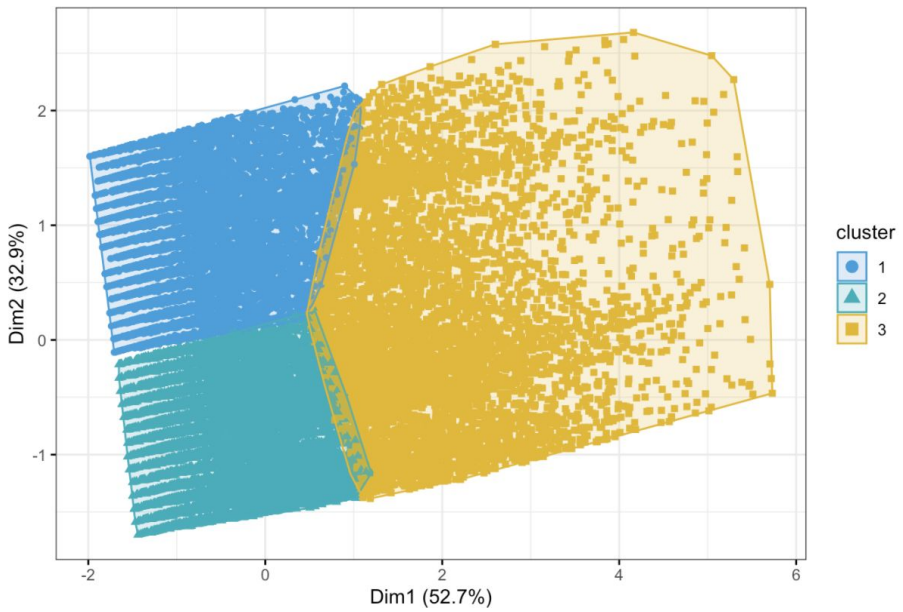

Fig. 12: Existing customers are clustered into three segments.

Analysis & Insights

Using data from the RFM model, Median Number of Items Purchased was plotted against Median Days Since Last Purchase in order to visualize the 3 segments our model identified. The results can be seen in Figure 13 below.

Fig. 13: Visualization of 3 distinct customer segments.

We then developed unique names to represent the purchasing behavior of each segment. The "One and Done's", or segment 1, are customers that purchased only a few items, and haven't made another purchase from the store in over 1 month. For this reason, we have chosen to not prioritize our marketing efforts towards them. "Necessity Shoppers", or segment 2, purchase the most recently, but in relatively low amounts. This indicates that they are only buying what they need at that point in time. This segment might be incentivized to purchase more with coupons. The third and final segment identified is the "Variety Shoppers", or segment 3. These shoppers make purchases relatively often and in large amounts, with a median number of items purchased of 15 (2.5X that of "Necessity Shoppers"). For these reasons, the majority of marketing efforts should be directed at this customer segment.

To gain further insight into the purchasing behavior of each segment, we identified the number of unique brands each customer in each segment purchased over a one month period, seen in Figures 14, 15, and 16 below.

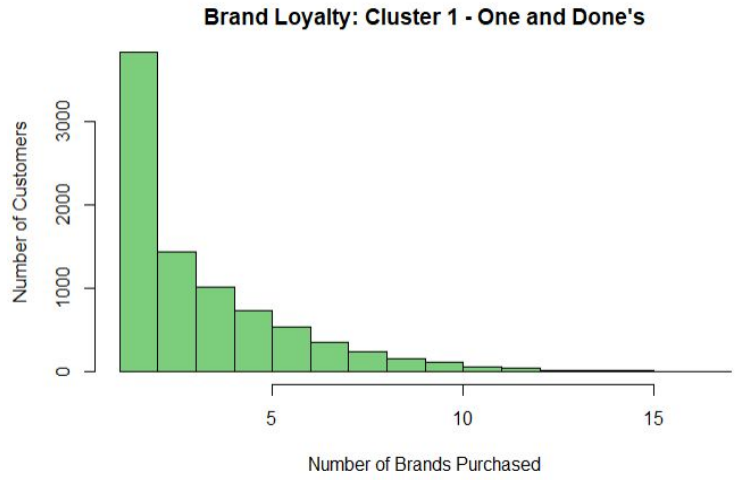

Fig. 14: Histogram of the number of unique brands purchased by One and Done's.

Fig. 15: Histogram of the number of unique brands purchased by Necessity Shoppers.

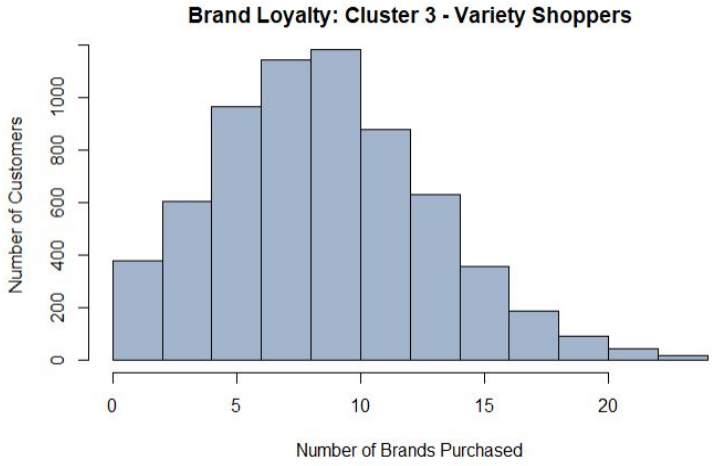

Fig. 16: Histogram of the number of unique brands purchased by Variety Shoppers.

From these three visualizations, we can clearly see that Variety Shoppers on average purchase a much larger variety of different brands. This implies that they are more willing than Necessity Shoppers and One and Done's to try new brands. Because of this, Variety Shoppers should be targeted with promotions for brands and products new to the store, as well as ones they have not purchased before. To entice brand-loyal Necessity Shoppers to purchase more, the store could offer coupons for some of their most purchased brands as well as promote products from those brands.

Classification Models

The dataset focuses on user activity and allows us to predict purchase outcomes based on certain predictors. We implemented two classification models to determine if a product that was added to cart would be purchased or removed from the cart. The end goal was to develop a model to predict customer purchases and identify which factors play the largest roles in leading to a purchase.

Naive Bayes

The first model that we implemented was the Naive Bayes Classification model. While this is often a good way to start classification, the main limitation of the model is that it assumes that all predictors are independent. The training and testing data were split in a 70:30 ratio. The resulting confusion matrix of the model is shown in Figure 17.

		Predicted	
		Purchased	Removed from Cart
Actual	Purchased	5.80%	11.75%
	Removed from Cart	20.05%	62.40%

Fig. 17: Confusion matrix for the Naive Bayes classification model.

The resulting accuracy of the model was about 68%. Naive Bayes is computationally efficient and was able to run quickly, even on our large dataset. This makes it a good performance baseline to compare with our other models.

Random Forest

The second model that we implemented was the Random Forest (RF) classification model. RF is a classification technique that does not fit a single model to the data, but instead uses a collection of models to apply the best fitting model to classify the available data. Similar to Naive Bayes, we

split the data into testing and training datasets. We used the training data to develop a model and then run it on the testing data. The model classified the testing data as shown in Figure 18.

		Predicted	
		Purchased	Removed from Cart
Actual	Purchased	0.71%	0.03%
	Removed from Cart	23.72%	75.55%

Fig. 18: Confusion Matrix from Random Forest

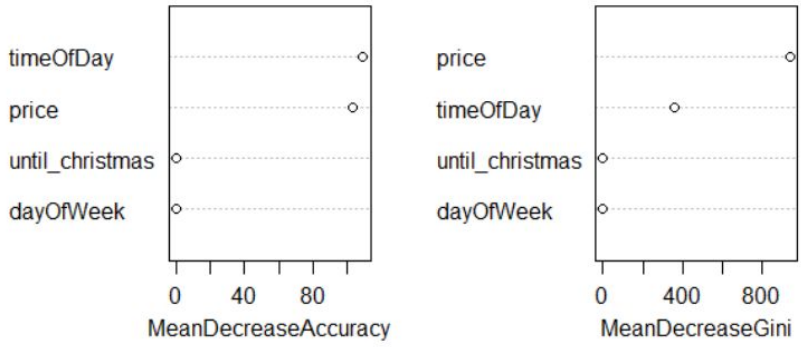

Fig. 19: Importance of Predictors using Random Forest

The accuracy of the RF model was 76%, which is slightly higher than the Naive Bayes classification model. Thus, we were able to improve prediction accuracy. In this manner, RF is beneficial in telling us whether a customer will purchase the product or remove it from their cart. Depending on customer activity, the e-commerce website can make a prediction and tailor marketing messages accordingly.

Recommendations

Our clustering analysis provided us with three major segments — *"Variety Shoppers", "Necessity Shoppers", and "One and Done's"*. The classification models then showed price and time of the day are important predictors to know if the customer will purchase an item or remove it from their cart.

Based on these learnings, we have devised a few recommendations for this particular e-commerce website. Firstly, assuming that this e-commerce website has limited resources and

wishes to focus marketing efforts, they should target *"Variety Shoppers" and "Necessity Shoppers"* as they are likely to visit the site more often and bring in recurring revenue.

From Figure 14-16, we see that *"Variety Shoppers"* purchase an average of eight different brands as compared to the other customer segments where the average is one. If the e-commerce website plans to introduce new brands on their platform, they should market it to the *"Variety Shoppers"* segment first as this segment is more likely to buy other brands.

As for *"Necessity Shoppers"*, it would be important to focus on factors such as 'price' and 'time of day' as they impact purchase behavior. We can also incentivize the users checkout experience by offering coupons and having discounted hours during the day, to increase customer activity.

Lastly, we developed two classification models that were able to predict whether a customer will purchase or remove from cart. We can use these models to analyze user behavior, and if we predict that they will purchase a product, we can then push personalized marketing messages that can help nudge them to complete their purchase on the e-commerce website.

Future Work

Our goal through this project was to improve purchase conversion rates for this e-commerce website by using data analysis techniques. We were successful in doing so by building models that would help the store in understanding customers and the factors that drive purchases. However, if we had more time and resources to work on this project, the following are some things we would have worked on.

The Kaggle dataset that we used has cart activity for the months of October 2019 through February 2020. In our analysis, we only considered November as we saw the highest activity during that month. However, in order to improve our model we would consider all the months. This would allow us to build new and improved models that take into account the varying levels of cart activity throughout those five months.

Given computational power constraints, we were only able to test a few different values of mtry for our random forest model. In order to find the optimal random forest model, we would like to tune parameters, such as the number of trees and mtry. We would also be open to trying new models, other than k-means and random forest, to find one which has the best results with our data.

As we did our research to improve our existing models, we learned about xGBoost which stands for eXtreme Gradient Boost which is typically used after building a Naive Bayes model. The xGBoost model could further improve the predictions, thus giving us higher levels of accuracy for our model.

As we did not have enough information about brand data, we were unable to use it in any of our models. But if we use the entire dataset of the five months from Kaggle, we could possibly have enough data to use *brand* as a predictor for our models too. This would enable the e-commerce website to make recommendations based on brands previously purchased.

Lastly, using the brand data, we could also develop clusters that show association between different brands. This will help us recommend brands similar to the ones the users purchased. This would be similar to Amazon's "User's also bought" section, and it would help in building a lasting relationship with the user.

These are just some of the ideas we have for the future. However, as we conduct exploratory data analysis on the larger data, we expect to uncover other trends and may need to feature engineer some other predictors.

References

Dataset: Michael Kechinov,
https://www.kaggle.com/mkechinov/ecommerce-events-history-in-cosmetics-shop