MGTA 634 Marketing Analytics

MSc Management Analytics

Group Case Study Report

July 27, 2023

Sarah Mansoor, Mengxi Dai, Yiyang Zhang, Hai Ninh Vu

Problem Identification

The problem Shine Cosmo-Cosmetics Inc. is facing is whether to operate with regular marketing practices or use data-driven marketing through business analytics for the year 2018. This involves making a strategic decision that will significantly impact the company's marketing approach and competitiveness in the market. The problem lies in deciding whether to use these data-driven results to focus on customers or on merchandise. Either way, the decision to adopt data-driven marketing will have a substantial impact on Shine's future growth and profitability. With sales growth slowing down and more competition in the cosmetic sector, choosing the right approach is vital for the company to maintain its market share, bring in new customers, and keep existing ones. There is an approaching deadline for finalizing the 2018 marketing plan and as it is currently late October 2017, Shine needs to decide. Delaying the decision may lead to missed opportunities during critical marketing periods and reduce profits.

The hypothesis in this case study revolves around the decision to focus on customers or merchandise to increase sales for Shine Cosmo-Cosmetics Inc. The company is facing a slowdown in sales growth, which is due to increasing competition and changing customer behavior. To address this challenge, there are two proposed approaches. The first approach focuses on the young demographic cohort (aged 21 to 34), Shine's best customer segments, and provides them with coupons that can be applied to any items. The second approach promotes specific items or item bundles heavily to attract new customers who have not previously purchased from Shine. This hypothesis seeks to determine which strategy is more effective in driving sales growth, customer retention, and overall profitability.

Data Preparation

Data cleaning is an important step in the data analysis procedure to ensure that the data is accurate, consistent, and ready for further analysis. The first step was to import the Pandas library. The dataset containing the transaction information was stored in an Excel file which was imported using Pandas' read_excel() function. A new column for transaction IDs was added to uniquely identify each transaction. This made it easier to track and analyze individual transactions. The dataset merged into a single DataFrame based on the SKU column using the SKU_list. The discount column was in a non-numeric format, so it was converted into numeric format. By subtracting the discount amount from the original price, an actual price column was

created, representing the original price before applying the discount. To analyze the profitability of each transaction, a profit column was computed and created by subtracting the cost of goods sold from the actual price. Lastly, a total profit column was generated to represent the total profit earned from each transaction. With the necessary columns added and the data cleaned to ensure consistency, we could begin with exploring the data.

Exploratory Data Analysis

After cleaning the data, we explored the dataset. Firstly, we examined its shape, which consisted of 4827 rows, and found 4679 unique customer IDs, implying that 148 customers made multiple purchases. We calculated various metrics, including average price, average discount, total quantity, total return, total profit, and the date of the last order. Once these metrics were calculated we used data visualization to explore the data further. The histogram of total profit implied that most customers produced profits around $200, with the majority falling between $100 and $300 (see Appendix A.1). The histogram of average price revealed that customers typically spent between $55 and $75 on a single product, with an average of about three products per purchase (see Appendix A.2). The customer tenure histogram revealed that most customers had only made a single purchase, supporting the number of unique customer IDs (see Appendix A.3). The age histogram revealed a continual decline from 21 years old to 30 years old, with more customers in their early 20s compared to their late 20s (see Appendix A.4). This is fitting for the demographics expected for a cosmetics company. The boxplot of tenure versus total profit revealed that though average profit increased with tenure, customers with a tenure of 1 had a more diverse range of values for profit (see Appendix A.5). Lastly, the histogram of transaction count by gender showed a suggestively higher number of transactions by female customers, with women having around 4000 transactions compared to men under 1000 transactions (see Appendix A.6). These exploratory analyses reveal important trends and patterns within the data, which provides valuable insights to inform strategies and decisions for Shine Cosmo-Cosmetics Inc.

Data Modelling

CLV

In lifetime value (CLV) analysis, the churn rate for the industry is assumed to be 50% per year, and the average customer lifespan is estimated at 2 years, based on the median tenure from the dataset. A discount rate of 10% is used to calculate the present value of future cash

flows. The CLV formula takes into account various factors such as profit from sales, return costs, order processing costs, and customer acquisition costs. The cohort data is grouped and annualized to make meaningful comparisons.

Upon analyzing the results, it is evident that customers belonging to cohort 5, with ages ranging from 32 to 34, have the highest CLV of 469 (see Appendix A.9). This suggests that targeting this particular age group for new customer acquisition campaigns could be more financially beneficial for the business. The CLV value is an estimate of the total net profit the company can expect to generate from a customer throughout their relationship with the business. By considering the acquisition cost and the likelihood of customer churn, this analysis allows the company to make informed decisions about resource allocation and marketing strategies.

RFM

The RFM (Recency, Frequency, Monetary) model is used to segment and understand customer behavior based on three key metrics: recency of their last purchase, frequency of purchases, and total monetary value spent. The RFM model analyzes transactional data of customers and assigning scores to each customer. Recency calculates the time since a customer's last purchase. Frequency calculates the number of transactions made by a customer within a specified period. Monetary calculates the total amount spent by a customer over a given period. By combining these three, the RFM model produces a wide-ranging customer segmentation that groups customers into different categories based on their recency, frequency, and monetary scores. In our RFM analysis, we first created an RFM data frame and set the number of bins to 2. Next, we computed the recency score by sorting the data frame based on 'day_from_last_order' in ascending order. We then created a list to store the recency scores and assigned these scores to each observation. Adding the 'recency_score' column to the data frame allowed us to capture the recency aspect of customer behavior. Moving on to the frequency score, we sorted the data frame based on the 'day_from_last_order' column in descending order to identify the most frequent customers. A list was created to store the frequency scores, and each observation was assigned an appropriate score. The 'frequency_score' column was subsequently added to the data frame, providing insights into the frequency of customer transactions. In the final step, we computed the monetary score by sorting the data frame based on the 'total_spend' column in descending order, highlighting customers with higher spending patterns. A list was created to store the monetary scores, and

each observation was assigned a corresponding score based on their total spending behavior. By adding the 'monetary_score' column to the data frame, we gained a comprehensive view of customers' monetary value to the business. From this procedure of using the RFM analysis, we were able to segment customers based on their recency, frequency, and monetary characteristics. This process provided valuable insights for targeted marketing and customer relationship management strategies.

MBA

Market Basket Analysis (MBA) is applied to two datasets: customer data and order data. The goal of MBA is to identify interesting associations between items that are frequently bought together in transactions. This analysis can be beneficial for retailers and marketers to understand customer behavior and optimize their product offerings and promotions.
In the customer data MBA, we first select relevant columns related to Stock Keeping Units (SKU) from the "df_customer" dataframe, which represent the different products customers have purchased. Next, we convert the SKU columns into boolean values, indicating whether a customer has bought each SKU or not. We then apply the Apriori algorithm to find frequent itemsets, which are combinations of SKUs that tend to occur together in customer transactions. A minimum support threshold of 0.5 is used to consider only the significant itemsets. Subsequently, we generate association rules based on the frequent itemsets using the "lift" metric. Lift measures the strength of association between two items in a rule, indicating how much the presence of one item affects the likelihood of the other item being purchased. The rules are sorted based on their lift values, and we extract the top 10 association rules for further analysis (see Appendix A.7).

Similarly, in the order data MBA, we create an order dataframe where each row represents a unique order, and the columns are SKU dummy variables indicating the presence or absence of each SKU in an order. Applying the Apriori algorithm to this dataset, we identify frequent itemsets with a minimum support threshold of 0.01. Association rules are then generated using the "lift" metric, and we filter out rules with a lift value below 1 to focus on significant associations. The top 10 association rules are extracted to gain insights into item affinities within customer orders.

Interpreting the results, the association rules provide valuable information on item dependencies and co-occurrence patterns. Strong associations between certain items, as indicated by high lift values, suggest that customers tend to purchase these items together. Retailers can leverage this knowledge to create targeted marketing strategies, such as bundling

related products to encourage cross-selling. Additionally, understanding item affinities within orders can help optimize inventory management and supply chain operations, ensuring that frequently co-purchased items are readily available to customers.
In conclusion, Market Basket Analysis is a powerful technique for uncovering hidden patterns in customer transaction data, enabling businesses to make data-driven decisions to enhance sales and customer satisfaction.

Recommender System

A recommender system is a vital tool in the digital marketplace, providing personalized suggestions to users based on their behavior and preferences. In our specific context, we are seeking to build a recommender system for Shine Cosmo-Cosmtics Inc. and provide personalized product recommendations to each customer. This is a non-trivial task due to the vast number of SKUs, each representing a unique product. The key challenge lies in accurately predicting which products a customer is most likely to purchase based on their past buying behavior as well as the behavior of similar customers. To address this, we reshaped the dataset and let it has “customer_id”, “SKU”, and “quantity” columns, then we performed 4 recommender system algorithms, which are:

1. User-Based Collaborative Filtering (UBCF)
2. Item-Based Collaborative Filtering (IBCF)
3. SVD++
4. CoClustering

In conclusion, building a recommender system for Shine Cosmo-Cosmetics Inc. presents a challenge due to the vast number of SKUs and the need to provide product recommendations to each customer. By reshaping the dataset to include "customer_id," "SKU," and "quantity" columns, we prepared the data for analysis. We then implemented four different recommender system algorithms, including User-Based Collaborative Filtering (UBCF), Item-Based Collaborative Filtering (IBCF), SVD++, and CoClustering, to predict which products a customer is most likely to purchase based on their buying behavior and the behavior of similar customers. We evaluated the performance and accuracy of these algorithms to determine the most effective approach for personalized product recommendations. This method will enable Shine Cosmo-Cosmetics Inc. to improve customer satisfaction and increase sales.

Clustering

Using customer segmentation techniques is essential for Shine Cosmo-Cosmetics Inc. to know its customer base better and tailor marketing strategies correspondingly. We applied clustering algorithms, k-means, to group customers with similar attributes together. The Elbow method helped determine the optimal number of clusters, and we found 2 clusters would work best for our analysis. After training the K-means model with the optimal number of clusters and predicting cluster labels for each customer, we joined the PCA components with the cluster labels to create a scatter plot for visualization.

Upon thorough analysis of the centroids of two distinct customer clusters, it is discernible that each cluster encapsulates a unique set of characteristics (see Appendix A.8). Predominantly, Cluster 0 is composed of customers who are slightly older, with an average age of 29.8 years, and have maintained a longer affiliation with the company. These customers exhibit a discernable affinity for higher-priced items, a preference that is evident in the superior average price and actual price of their purchases. In addition, this cluster demonstrates an enhanced inclination towards certain products, such as SKU_SCS72046. Intriguingly, Cluster 0 also has a higher proportion of male customers.

On the other hand, Cluster 1 is characterized by customers who are marginally younger, with an average age of 26.8 years, and have a relatively shorter tenure with the company. Customers in this cluster typically opt for less expensive items and display a distinct preference for different products, notably SKU_FCS48021. This cluster is predominantly populated by female customers.

These discernible patterns accentuate the heterogeneity within our customer base, emphasizing the necessity for customized marketing strategies to effectively engage each distinct segment. Further, the integration of clustering analysis as a supplementary technique would augment the efficacy of the final recommendation solution.

Alternative evaluation

Each of these techniques is cross-validated and evaluated on the basis of two metrics: Root Mean Square Error (RMSE) and Mean Absolute Error (MAE). Both metrics are popular for evaluating the accuracy of a prediction, with lower values indicating better performance.

Here are the results:

	RMSE Mean	RMSE Std	MAE Mean	MAE Std
User-Based Collaborative Filtering (UBCF)	0.4067	0.0086	0.2251	0.0065
Item-Based Collaborative Filtering (IBCF)	0.4069	0.0104	0.2241	0.0046
SVD++	0.3846	0.0044	0.2186	0.0018
CoClustering	0.4107	0.0055	0.2347	0.0052

From the results, it is clear that SVD++ performs the best as it has the lowest RMSE and MAE among all the methods. This means that SVD++ has the smallest average prediction error, making it the most accurate method for this specific dataset and problem according to the chosen evaluation metrics.

Implementation plan

To enhance customer engagement and optimize revenue generation, this study proposes a two-pronged approach focusing on both existing and potential customers. These strategies are informed by a thorough analysis of customer purchasing behavior, employing techniques such as Recency, Frequency, Monetary (RFM) scoring, Market Basket Analysis (MBA), and recommender systems.

The first approach targets existing customers, specifically those with RFM scores of 112 and 122, indicative of their profitability but less frequent recent engagement. Personalized incentives, such as coupons or discounts on frequently purchased products, are proposed to stimulate re-engagement and increase transaction frequency. Further, to promote up-selling and cross-selling, customers with relatively lower total expenditure and average purchase price are identified. By leveraging insights from MBA to ascertain frequently co-purchased items and utilizing recommender systems to suggest higher-end alternatives, this strategy aims to elevate average purchase values. This dual-faceted approach presents an opportunity to not only reinforce customer loyalty but also enhance overall customer lifetime value.

The second approach focuses on the acquisition of new customers, particularly within the 32-34 age demographic, identified as having the highest Customer Lifetime Value (CLV). Targeted outreach efforts, guided by the MBA, advocate for the promotion of popular and basic

product bundles, such as SCS72048 and FCS48028, SCS72044 and SCS72043, SCS72046 and FCS48021. The coupling of these targeted offerings with the predictive power of recommender systems allows for a high degree of personalization in attracting potential customers, thereby maximizing the efficacy of customer acquisition initiatives.

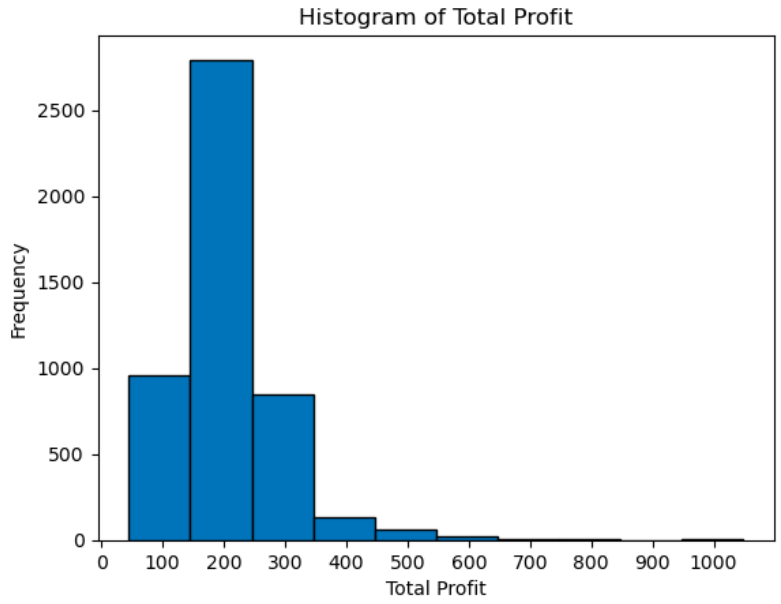
In essence, these strategies, underpinned by rigorous data analysis and personalized customer targeting, aim to cultivate customer loyalty, encourage spending, and attract high-value new customers. Consequently, this approach contributes to sustainable business growth through increased revenue and enhanced customer satisfaction.

Conclusion

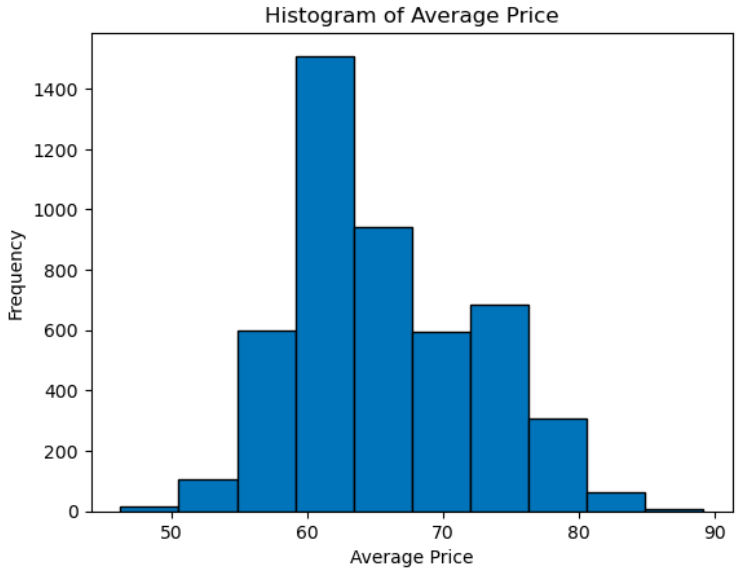
In conclusion, with Shine Cosmo-Cosmetics Inc. facing the critical decision-making deadlines, the company's future growth and profitability depend on the strategic choice, given the slowdown in sales growth and increased competition in the cosmetic sector. By implementing various data analysis techniques, including RFM scoring, Market Basket Analysis, and recommender systems, we got valuable insights into customer behavior and preferences. With the RFM analysis we segmented customers based on their recency, frequency, and monetary characteristics, enabling targeted marketing and customer relationship management strategies. Market Basket Analysis provided information on item affinities within customer orders, guiding product bundling and cross-selling strategies. And lastly the recommender systems offered personalized product recommendations. Evaluating recommender system algorithms showed that SVD++ performed the best, as it achieves the lowest RMSE and MAE metrics. This makes SVD++ the best method for personalizing product recommendations for Shine Cosmo-Cosmetics Inc. Our implementation plan advises a dual-pronged approach, focusing on existing and potential customers. By offering personalized incentives to existing customers and targeting new customers with high potential CLV, Shine can improve customer loyalty, increase transaction frequency, and draw in new customers. This approach optimizes revenue and promotes business growth.

Appendix

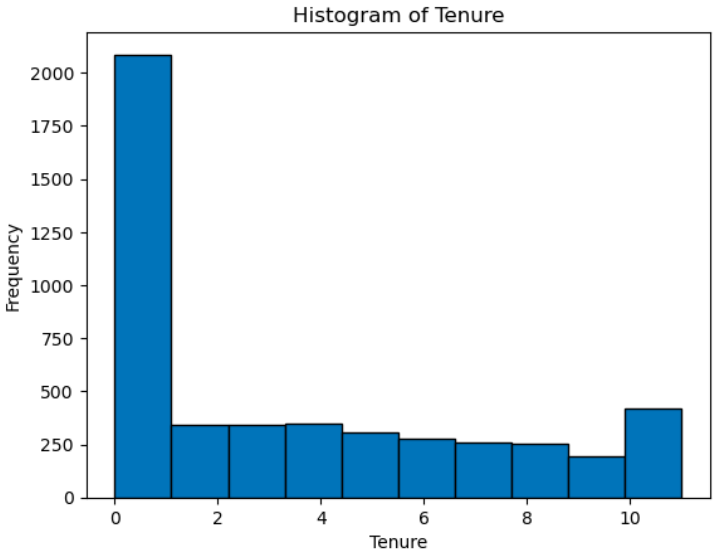
Appendix A.1



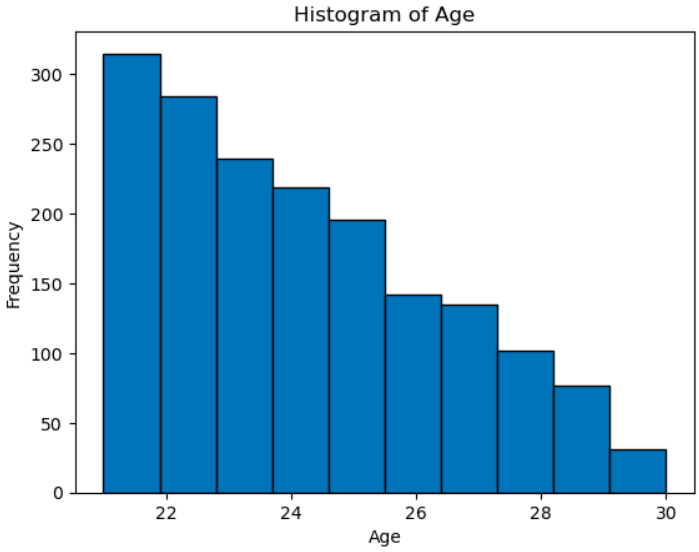
Appendix A.2



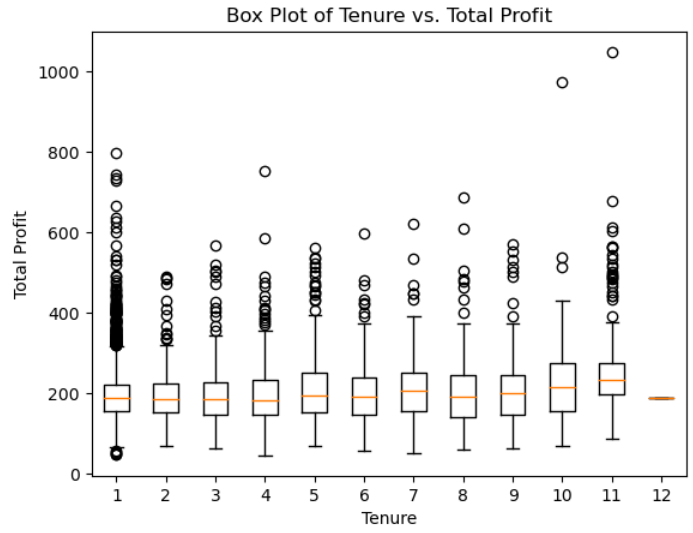
Appendix A.3



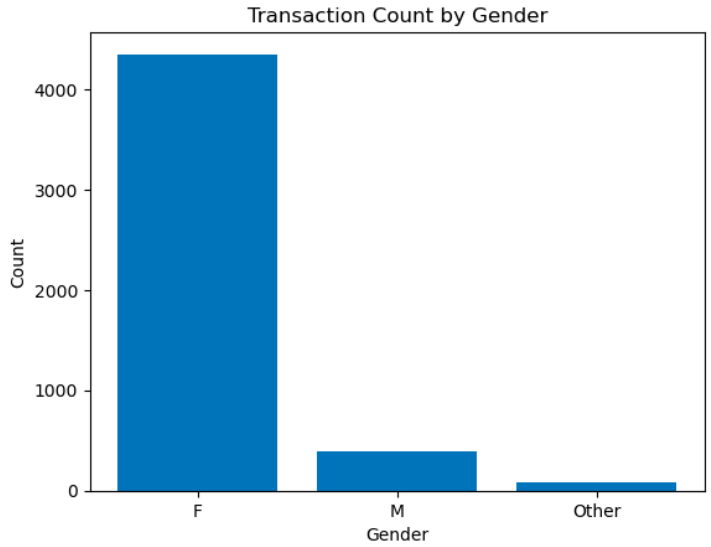
Appendix A.4



Appendix A.5



Appendix A.6



Appendix A.7

	antecedents	consequents	antecedent support	consequent support	support	confidence	lift	leverage	conviction	zhangs_metric
2	(SCS72048)	(FCS48028)	0.089599	0.093459	0.058492	0.652821	6.985086	0.050118	2.611161	0.941165
3	(FCS48028)	(SCS72048)	0.093459	0.089599	0.058492	0.625854	6.985086	0.050118	2.433281	0.945173
6	(SCS72043)	(SCS72044)	0.107513	0.103347	0.070407	0.654870	6.336624	0.059296	2.598013	0.943641
7	(SCS72044)	(SCS72043)	0.103347	0.107513	0.070407	0.681268	6.336624	0.059296	2.800122	0.939256
5	(FCS48031)	(SCS72042)	0.093320	0.089654	0.049021	0.525298	5.859151	0.040654	1.917719	0.914686
4	(SCS72042)	(FCS48031)	0.089654	0.093320	0.049021	0.546778	5.859151	0.040654	2.000521	0.911002
0	(SCS72046)	(FCS48021)	0.100680	0.104597	0.061214	0.608000	5.812809	0.050683	2.284192	0.920658
1	(FCS48021)	(SCS72046)	0.104597	0.100680	0.061214	0.585236	5.812809	0.050683	2.168270	0.924685

Appendix A.8

	Cluster 0	Cluster 1
transaction_id	25076.346	22601.290
price	98.032	57.373
quantity	1.098	1.101
age	29.802	26.805
tenure	5.055	2.945
return	0.058	0.057
Retail price	98.032	57.373
Cost	60.424	33.948
actual_price	89.730	51.804
profit	29.306	17.856
total_profit	30.814	18.901
spend	107.455	63.346
sku_FCS48020	-0.000	0.007
sku_FCS48021	-0.000	0.104
sku_FCS48022	0.125	0.000
sku_FCS48023	0.122	0.000
sku_FCS48024	-0.000	0.046
sku_FCS48025	0.000	0.010
sku_FCS48026	-0.000	0.045
sku_FCS48027	0.125	0.000
sku_FCS48028	0.001	0.093
sku_FCS48029	0.000	0.110
sku_FCS48030	0.001	0.012
sku_FCS48031	0.000	0.093
sku_FCS48032	-0.000	0.047
sku_FCS48033	0.125	0.000
sku_SCS72041	0.000	0.016
sku_SCS72042	-0.000	0.092
sku_SCS72043	-0.000	0.107
sku_SCS72044	-0.000	0.106
sku_SCS72045	0.122	0.000
sku_SCS72046	0.378	0.000
sku_SCS72047	0.000	0.012
sku_SCS72048	0.001	0.092
sku_SCS72049	0.000	0.008
discount_0.0	0.579	0.550
discount_0.1	0.082	0.075
discount_0.15	0.084	0.079
discount_0.2	0.083	0.092
discount_0.25	0.084	0.104
discount_0.3	0.087	0.100
gender_F	0.792	0.903
gender_M	0.194	0.081
gender_Other	0.014	0.017

Appendix A.9

cohort_group	average_price	average_discount	purchased_quantity	product_return	spend	profit	number_order	CLV
1	62.612248	0.098126	22.428571	1.194617	1411.883375	409.263188	16.169772	347.203448
2	62.998078	0.098139	22.784679	1.186335	1446.618944	418.354555	16.525880	358.090344
3	62.274332	0.090217	18.837306	1.036269	1182.893492	351.447275	13.262176	289.090525
4	66.592105	0.082701	18.424870	0.853886	1251.832332	379.820518	13.127461	336.911427
5	73.386582	0.084275	22.974093	1.255959	1686.431399	497.974197	16.918135	468.921300