Identifying High-Value Customers in a Retail Setting: Evaluating Coupon Usage as an Indicator

Erasmus University Rotterdam - Erasmus School of Economics
Msc Data Science & Marketing Analytics

Bruno Hankel

2023-06-05

Abstract

This research aims to investigate whether valuable customers of a retail company can be identified in an early stage of their engagement. Next to this, coupon usage will be explored as an indicator of customer value. First, a customer base of a retail company will be segmented by weighted RFM clustering. Subsequently, a random forest model will be trained to predict the cluster of a household. The model yields a precision of 77,78% when predicting the cluster with the most valuable households, outperforming a multinomial logistic model. No substantial relation of the personal coupon redemption rate and customer value has been found. This study provides a methodological framework for companies to recognize which customers are valuable, and therefore can lead to an enhanced allocation of marketing resources.

Contents

1 Introduction

Determining what customers are most valuable to your company can lead to a more efficient allocation of (marketing) resources. Investing these resources to attract/retain customers that are expected to yield little revenue might not be the optimal decision. Perhaps it would be more efficient to use these resources to invest in customers that are expected to yield the highest revenue, binding them to your company. This is underscored by the 80/20 rule, stating that 80% of the profit is generated by the top 20% of most profitable customers (Duboff, 1992). To be able to target that top 20%, it is critical to recognize valuable customers. The sooner, the better.

One common way to find the most valuable customers, is to segment a customer base. Once this is done, the most valuable customer segment can be recognized. Customer base segmentation can be done multiple ways: segmenting on amount of transactions and turnover (Marcus, 1998), RFM-based segmentation (Chen et al., 2012), weighted RFM (Khajvand et al., 2011) and many more.

When identifying valuable customers, customer-specific characteristics can be an indicator for a valuable customer. The relation between customer value and coupon usage has, to the best of my knowledge, not been studied yet, although a lot of research regarding coupon usage has been done in the past: Leone (1996) examined the effect of face value on coupon redemption, Bawa et al. (1997) studied coupon proneness among customers and Lichtenstein et al. (1990) explored the psychology of value conscious customers. Because of the possible interesting relation between customer value and coupon usage, and the existing gap in academic literature regarding this subject, this research will also focus on uncovering this relation.

1.1 Problem Statement

Identifying valuable customers will be the main goal of this research. As mentioned, the sooner these customers can be recognized, the better. Therefore the main research question is as follows:

Can we identify the most valuable customers of a retail company in an early stage, and is coupon usage an indicator of customer value?

The main research question will be supported by three sub questions:

1. What demographics distinguish high-value customers from other customers?
2. Is it possible to identify valuable retail customers in an early stage of their engagement?
3. What is the relation of the personal coupon redemption rate and CLV for retail customers?

To answer all questions mentioned above, the customer base of a retailer will first be clustered by using K-means based RFM clustering. The optimal amount of clusters will be determined by several methods, whereafter the clusters will be ranked according to their weighted CLV, resulting in one cluster to be the most valuable. After this, a random forest model will be trained to predict what cluster a customer will be assigned to. The model will be tuned to yield the highest predictive performance for predicting the most valuable customers. To evaluate the predictive performance for customers in an early stage of their engagement, the values for amount of transactions (per week), spending (per transaction) and coupon redemption rate will be computed considering only the first 4 weeks of the engagement of the customers present in the test set. The performance will be compared to a simple multinomial logit model's performance. To determine the relation of the personal coupon redemption rate and customer value, partial dependence plots will be examined.

1.2 Academic Relevance

This research adds to the academic knowledge of the collaboration of K-means clustering and random forest, in the context of identifying valuable customers. Specifically, the study investigates the predictive performance of a random forest using the clusters, resulting from K-means, as dependent variable.

Next to this, it fills the gap in existing literature regarding the relation between coupon usage and customer value. To the best of my knowledge, this if the first paper to examine this relation.

1.3 Managerial Relevance

First, this study provides a methodological framework for managers or decision-makers to recognize valuable customers in an early stage of their engagement. This will benefit companies (retailers especially), providing them guidance to implement it themselves.

Second, it provides insights in the relation between coupon usage and customer value. This creates managerial understaing regarding customers and their coupon usage, allowing managers to translate this knowledge into better decision making.

2 Literature Review

2.1 Identifying Customer Segments

Understanding your customer base is a vital part of a companies' long-term success. Customer relationship management (CRM) is mainly focused on this, and translating insights of a customer base into concrete marketing strategies. Correct segmentation is one of the first and most important parts of customer relationship management. CRM can be defined as 'Managerial efforts to manage business interactions with customers by combining business processes and technologies that seek to understand a company's customers' (Kim et al., 2003). A company's marketing and financial performance is positively affected by CRM, which seeks to build lasting relationships with high-value customers (Soliman, 2011). In order to pursue long-term relations with these valuable customers, this group should first be identified.

To identify segments in a customer base, clustering is often the method of choice. There are many variables to segment a customer base on. Marcus (1998) segmented retail customers based on just two, relative simple, variables: average number of purchases and average purchase amount. This lead to a basic segmentation, purely focused on the transactions. This approach suits small businesses particularly, since implementing more advanced clustering methods, like RFM, is too complex and time consuming. Kim et al. (2006) proposes customer segmentation to enhance CRM based on current value, potential value and customer loyalty. The research managed to look past only the transactions and include a new variable: customer loyalty. The study focuses heavily on improving CRM and marketing strategies, an important part of running a modern business.

In a more recent research, Chen et al. (2012) applied RFM-based customer segmentation, specifically suited for small online retailers. RFM stands for recency, frequency and monetary value, and is a model used to cluster customers based on these values. Advantages of RFM are that it is a cost effective method and allows for easy decision making, it is based on individual customers instead of aggregated groups and it is very strong in identifying valuable customers. A disadvantage is the limited variables RFM clusters on, resulting in possible latent heterogeneity in clusters. Subsequently, a possible disadvantage is that it does not provide very meaningful insights regarding consumer behaviour, given the little behavioural information the model considers (Wei et al., 2010).

Khajvand et al. (2011) elaborates on general RFM clustering by calculating CLV ranking based on the RFM clusters. By multiplying the mean RFM values of each cluster with specific weights, weighted CLV was calculated for each cluster. The weights are determined by questionnaires send to experts of a company's sales department, better known as analytical hierarchy process. The weights are industry specific; a business selling luxurious watches might value monetary value higher than frequency, whilst a retailer might value frequency more.

All mentioned segmentation techniques differ in the way they assign value to a customer, and therefore the variables their clustering is based on. They all cluster on customer value, only that value is defined by different variables. In this study CLV will be used as customer value indicator. CLV (Customer Lifetime Value) represents the present value of all future expected cash flows of a specific customer (Pfeifer et al., 2005). For a sophisticated CLV calculation, information regarding marketing costs, retention rates and margins (Chang et al., 2012) has to be considered. Because that information is not available for this research, we will only be able to focus on the revenue generated by each customer.

2.2 Consumer Coupon Usage

Next to identifying valuable customers, this research focusses on the relation between coupon usage and customer value. Coupon redemption is an academically widely explored subject. Yet, whether customers that use coupons more frequent than others are (not) valuable for companies has not been thoroughly investigated in the academic literature. If we want to dive deeper into this, there should first be an understanding about coupons: What drives companies to embark in coupon campaigns and what drives customers to consume them.

Coupon campaigns aim to drive up sales by benefitting customers financially. But this is not all, as observed by Srinivasan et al. (1995): "Coupons not only have redemptive value, but also can have advertisement value to some customers". Berman (2006) investigated the effect of discount campaigns on long term brand loyalty and observed a positive effect of discount campaigns on long term brand loyalty. This especially applies to programs that provide a discount at the checkout, like coupons, rather than a "buy two get one for free" type of campaign. Allender & Richards (2012) however found a significant effect of price promotions that decreases brand loyalty on the long term. According to that study, consumers who are aware of price promotions are more likely to switch brands than those who are not aware. This is an interesting contradiction, indicating different results of the same phenomenon.

Companies often engage in multiple coupon campaigns, and could learn from previous promotions to enhance future ones. To correctly measure and manage the returns from coupon campaigns for retailers, a data-driven approach should be adopted (Venkatesan & Farris, 2012). Keeping track of the results of a coupon campaign can lead to insights on various topics, including personal coupon redemption. Personal coupon redemption should be viewed as how often a customer uses coupons. This could be expressed as a percentage of the transactions discounted by coupons relative to all transactions of that customer, further referred to as personal coupon redemption rate (1):

$$PersonalCouponRedemptionRate = \frac{AmountOfTransactionsWithCouponUsage}{TotalAmountOfTransactions} \tag{1}$$

Coupon redemption in general is dependent on two factors: coupon attractiveness and consumer coupon proneness (Bawa et al., 1997). Coupon attractiveness refers to the degree to which a coupon is perceived as a good deal by a consumer. Four factors that contribute to coupon attractiveness are identified: the value of the coupon, the difficulty of obtaining the coupon, the product category, and the purchase occasion. Consumer coupon proneness refers to the likelihood that a consumer will use coupons in general. Consumer coupon proneness is influenced by demographic factors such as age, income, and education (Bawa et al., 1997).

Elaborating on coupon attractiveness, Leone (1996) examined the effect of coupon face value on coupon redemption, brand sales and brand profitability. An increasing face value of a coupon influences coupon redemption positively, and marginally increases brand sales. It is suggested there is a trade-off between coupon face value and brand profitability. Higher face value may increase short-term sales, but can have a negative effect on long-term brand profitability. Furthermore, coupons for different product categories will result in different redemption rates. Food and health and beauty products yield higher redemption rates compared to other product categories, possibly due to lower price sensitivity for normal goods compared to luxury goods (Reibstein & Traver, 1982).

As mentioned, consumer coupon proneness is a construct that refers to the propensity of an individual to utilize coupons when purchasing goods or services (Bawa et al., 1997). Within the same customer, coupon proneness can vary between product categories, but is positively correlated across product categories. This means a coupon prone customer in product category A is likely to also be a coupon prone customer in product category B (Swaminathan & Bawa, 2005).

A related construct that should be touched upon as well is value consciousness. Next to coupon attractiveness and coupon proneness, value consciousness is a construct that plays an important role in determining the likelihood of coupon redemption. Value conscious consumers are driven by acquisition utility (tangible benefits such as cost savings) and transaction utility, also referred to as experiential benefits, including the psychological satisfaction when obtaining a good deal (Lichtenstein et al., 1990). While coupon proneness is applicable to the use of coupons only, value consciousness reflects a more comprehensive motivation among consumers to pursue value in all stages of the consumption process, extending beyond merely acquiring goods.

3 Methodology

3.1 Data

‘The Complete Journey’ dataset contains transactional information of 2,500 American households who are frequent shoppers at a retailer, and is accessed via Kaggle. The dataset is published by Dunnhumby, a customer data science business located in the United States. Demographic variables are included for 801 households, and only these households will be considered for this research. Each sold product corresponds to a separate row in the dataset. The 801 relevant households purchased over 1,4 million products within a time span of 102 weeks, adding up to a total of 4,5 million USD spent. Each row contains a basket ID, relating to the basket (products bought in the same visit to the shop) the product was in when it was purchased. Over 140,000 baskets are checked out by the relevant households. Product specific information is included, stating what department it belongs to among other things.

For each transaction the coupon discount has been registered, giving insight in the households’ couponing behaviour. Because of this, the personal coupon redemption rate can be calculated for each household. Next to this, the three departments where each household recorded the highest sales revenue are included in the dataset. *First department* corresponds to the highest sales revenue, *second department* to the department with the second highest sales revenue and *third department* to the department with the third highest sales revenue.

The goal is creating a model that is able to accurately predict a households’ cluster when the household is still relatively new to the company. This will be achieved by only considering data from the first four weeks of the engagement of customers in the test set. This will result in different values for *Average spending*, *Baskets per week* and *redemption rate basket*.

3.2 Methods

3.2.1 Weighted RFM Clustering

This research seeks to predict a household's cluster, obtained by weighted RFM clustering, by training a random forest classifier model. First, RFM clustering should be performed using historical data. RFM stand for three components to describe a customer: Recency, Frequency and Monetary value (Wei, 2010). Recency traditionally is regarded as the time between the customer's last transaction and the final date present in the dataset. A high value of R would then incline a negative influence on the RFM value, since a customer that has not visited in a while could be considered less valuable than a customer that has visited lately. To ensure all RFM components have a positive effect on the RFM value, Recency is represented as the interval between a customer's first visit and his/her last visit, as proposed by Belhadj (2021). A large value of Recency now implies a lengthy engagement of a customer with the firm, which is regarded as positive. Frequency is described as the total baskets a household has checked out, and Monetary value is described as the total amount of a household's spending.

	Dimension	Value
1	Recency	Interval between first and last visit
2	Frequency	Total baskets
3	Monetary Value	Total spent

Table 1: RFM Values

Before the clustering commences, the RFM data should be normalized to ensure equal weight for all variables (Wei, 2010). Min-max normalization is used as a normalization method (2). The method applies a linear transformation to the data, resulting in all values being in a range of 0-1. For variable R, the smallest value will be transformed to zero and the largest value will be transformed to one.

$$R' = (r - r_{min})/(r_{max} - r_{min}) \tag{2}$$

The normalized RFM data will then be used as input for K-means clustering. K-means clustering is an unsupervised machine-learning technique. The algorithm creates non-overlapping clusters, meaning all observations belong to only one group, and was first introduced by Hartigan and Wong (1979). It aims to minimize intra-cluster distances by minimizing the following formula (3), using Euclidean distance as distance function:

$$F = \sum_{j=1}^{k} \sum_{i=1}^{n} ||x_i^{(j)} - c_j||^2 \tag{3}$$

Where k represents the number of clusters, n the number of observations, x an observation and c_j the centroid for cluster j.

This method requires a small amount of tuning, namely selecting the amount of clusters. There are several ways to determine this. For this research the elbow method and the silhouette coefficient will be reviewed. The elbow method applies to a plot with the total within clusters sum of squares on the y-axis and the number of clusters on the x-axis. A kink in the plot, or an "elbow", indicates little marginal gains for including an extra cluster, and hence shows the optimal amount of clusters (Bholowalia, 2014). The silhouette coefficient is a coefficient that indicates how similar datapoints are within a cluster, compared to other clusters (Yuan & Yang, 2019). A high value indicates homogeneity within clusters and heterogeneity across clusters. Once the optimal amount of clusters has been determined, K-means will be executed. To determine on what demographical variables the most valuable customers are different compared to the complete customers base, Pearson's Chi-Square Test will be executed since all demographic variables are categorical.

The resulting clusters will have average values for Recency, Frequency and Monetary value. To translate this into a weighted CLV calculation, these values will be multiplied with corresponding weights (Liu & Shih, 2005). These weights are usually determined by an analytic hierarchy process. Since that is not possible for this research, weights of a different research will be used. As mentioned, these weights are industry specific. Therefore the weights proposed by Belhadj (2021) will not be applicable since they examine the banking industry. Khajvand et al. (2011) however studied a health & beauty retailer, which resembles the dataset of this study the most, and are therefore copied for this study.

The weights used for Recency, Frequency and Monetary value are 0.105, 0.637 and 0.258 respectively. My personal view is that Frequency might be overvalued in these weights and Monetary value undervalued. Nevertheless these weights will be used. To calculate weighted CLV, the following formula (4) should be executed:

$$W_{CLV} = R * W_R + F * W_F + M * W_M \tag{4}$$

In this formula, w stands for the assigned weight. Once this has been done, the clusters will all have a weighted CLV value. These should be ranked, and a variable $CLV\ Rank$ should be added to the dataset.

Random forest clustering was also considered as a method. Although random forest clustering is an interesting clustering method, K-means allows for better interpretation of the clusters, given the ensemble nature of the random forest algorithm. And, given the low dimensional data the clustering is based on, the possible increase of performance by using random forest clustering does not outweigh the decrease of interpretability compared to K-means. Next to this, I find it interesting to combine multiple machine learning methods in the same study.

3.2.2 Random Forest

CLV ranking of the clusters will be the variable of interest for the second stage of the study, since this is the variable we want to predict. This will be done by training a random forest classifier model. Random forest is an ensemble learning method suited for both classification and regression tasks in machine learning, and was first introduced by Breiman (2001). The algorithm constructs a multitude of decision trees that all predict the dependent variable. In the end, majority vote of all predictions decides on the final prediction of the model. All trees are binary decision trees, splitting the data in two in each node (Breiman, 2001). The implementation of this research will be done in R, using the *randomForest* package.

Bootstrap aggregating (or bagging) is a random sampling technique with replacement, and is used to create training bootstrap samples. Each tree is trained on a separate sample. Next to bagging, the random forest classifier also uses feature selection. Feature selection selects the features (or variables) the trees are exposed to and trained on (Breiman, 2001). This differs per tree, making the trees less correlated and therefore less likely to overfit the data (Pal, 2005). This explains why the random forest is called 'random': each tree is trained on a random bootstrap sample, and is trained on random features.

The goal of the random forest trees is to split the data in the most dissimilar groups possible at each split. The Gini Index is used as a measure to select what variable to split on (Breiman, 2001). For each node, the Gini impurity takes a value of 0 when all observations in the node are of the same class, and is 1 when there are multiple classes present in a node (Pal, 2005). The Gini Index of a variable is the weighted average of the Gini impurity value across all nodes. A Gini Index of 0 means the variable selected has created a split that resulted in nodes with perfect purity. This means the nodes both contain only observations belonging to one class, which means the split has created very different groups. Gini impurity is calculated as follows (5):

$$g(N) = \sum_{i \neq j} P(w_i)P(w_j) \tag{5}$$

In this formula, $g(N)$ is the Gini impurity of node N, and $P(w_i)$ is the proportion of the population of the node with class i.

Since the goal of the trees is to make splits in nodes that create very different groups, so the split with the lowest Gini Index is preferred. This can be translated into variable importance as well: a low Gini Index means the variable is capable of splitting the dataset in two different groups, and therefore should be considered an important variable.

As mentioned, the final solution of the model is the majority vote of all trees for classification analysis. Because of the law of large numbers, random forests are less prone to overfitting than other ensemble methods like boosting or bagging (Pal, 2005).

Two parameters need to be set to produce the random forest: the number of decision trees ($Ntree$) and the number of variables to be considered for feature selection ($Mtry$). Since random forest does not overfit, $Ntree$ can be as large as desired. Most researches start with 500 trees, and then tune for the best accuracy (Lawrence et al., 2006). Since this research aims to accurately predict the most valuable customers, tuning will be done to find the highest accuracy of predicting that cluster. Typically, $Mtry$ is determined by taking the square root of the number of input variables (Gislason et al., 2006), mainly because of increased computation time if choosing larger values. Since the used dataset does not contain as much variables, that is no problem or this research.

Mtry will be found by repeated 10-fold cross validation based on accuracy, considering all possible *Mtry* values. The value with the highest accuracy will be chosen.

A training and test set will be determined, using a 70/30 split. To ensure the *CLV Rankings* are represented equally between both sets, a stratified split is performed to create the sets. As mentioned, all continuous values for the test set are computed by only considering the first four weeks of a customer's engagement, and therefore the results of predicting the test set should be considered as predictions for customers that have been with the company for only four weeks. The performance will be compared with a logit model performing the same prediction. To study and compare the data sensitivity of the models, the analysis will be rerun 5 more times, each time adding four weeks of data to the test set. This results in a maximum of 24 weeks worth of data to be considered for the test set. After 24 weeks, the customer might not be considered "new" anymore, and therefore analyses with more than 24 weeks of data will not be considered.

To determine the effect of the personal coupon redemption rate on the probability of an observation being predicted in a certain class, partial dependence plots will be explored. Partial dependence plots are computed by the following formula (6):

$$f(x) = logp_k(x) - \frac{1}{K}\sum_{j=1}^{K} logp_j(x) \tag{6}$$

Where K is the number of classes, $_k$ is the specified class and p_j is the proportion of votes for class $_j$ (Friedman, 2001). These plots should be interpreted cetirus paribus. A summary of the proposed methodology is shown in Figure 1.

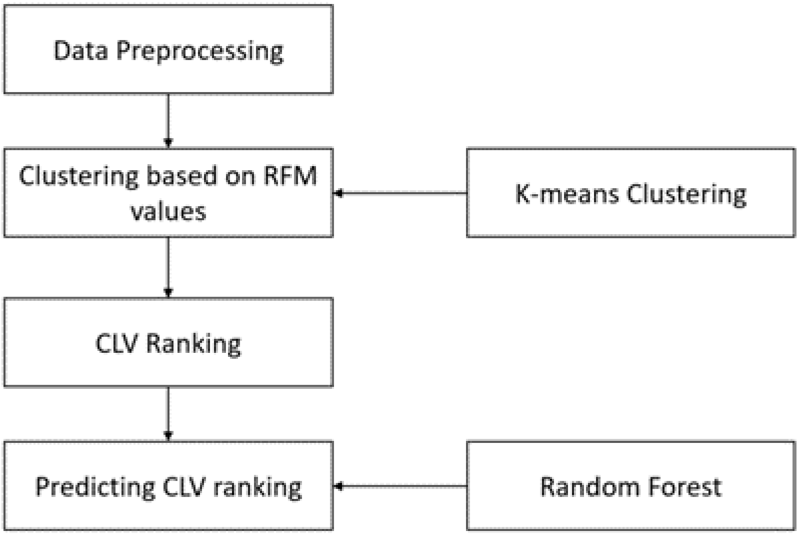

Figure 1: Proposed Methodology

4 Results

4.1 Data Description

Before diving into the results, some descriptive statistics of the data should be introduced. First, all continuous variables are described in Table 2. All variables show relative large standard deviations compared to the mean, indicating the values are widely spread within the range of the variable.

Variable	Min	Mean	Max	SD
Average spending	2.83	36.91	163.43	25.14
Baskets per week	0.26	1.92	14.24	1.23
Redemption rate basket	0.00	0.08	0.69	0.09

Table 2: Descriptives of Continuous Variables

Next to the descriptives shown in Table 2, the correlation between continuous variables should also be considered. Figure 2 visualizes the correlations of all variables with each other. Low correlation between variables is present. The average spending shows a small positive correlation with the redemption rate, and a small negative correlation with the amount of baskets per week, suggesting customers who shop less frequently tend to spend more, on average, per transaction.

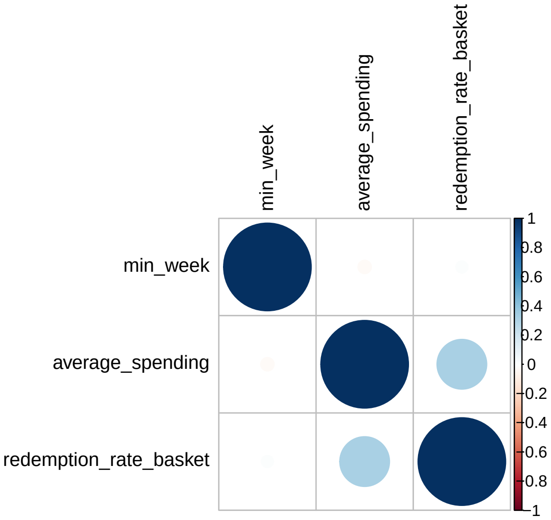

Figure 2: Correlation plot

All categorical variables used for the analysis are listed in Table 3. A frequency table for each categorical variable is given in the Appendix.

	Variable	Class
1	AGE_DESC	Factor w/ 6 levels
2	MARITAL_STATUS_CODE	Factor w/ 3 levels
3	INCOME_DESC	Factor w/ 12 levels
4	HOMEOWNER_DESC	Factor w/ 5 levels
5	HH_COMP_DESC	Factor w/ 3 levels
6	HOUSEHOLD_SIZE_DESC	Factor w/ 5 levels
7	KID_CATEGORY_DESC	Factor w/ 4 levels
8	first_department	Factor w/ 5 levels
9	second_department	Factor w/ 10 levels
10	third_department	Factor w/ 14 levels
11	CLV_ranking	Factor w/ 3 levels

Table 3: List of Categorical Variables

4.2 Cluster Specification

To determine the amount of clusters for K-means clustering, the elbow method and silhouette coefficient will be evaluated. The elbow plot, shown on the left in Figure 3, indicates three clusters is the optimal amount. There is a clear elbow present at $k = 3$, whereafter the marginal gain seems to remain constant. The right plot in Figure 3 visualizes the average silhouette width for a specified amount of clusters. The highest silhouette coefficient, indicating homogeneity within clusters and heterogeneity across clusters, is found at $k = 3$. Because of the results of both methods, the amount of clusters is set at three.

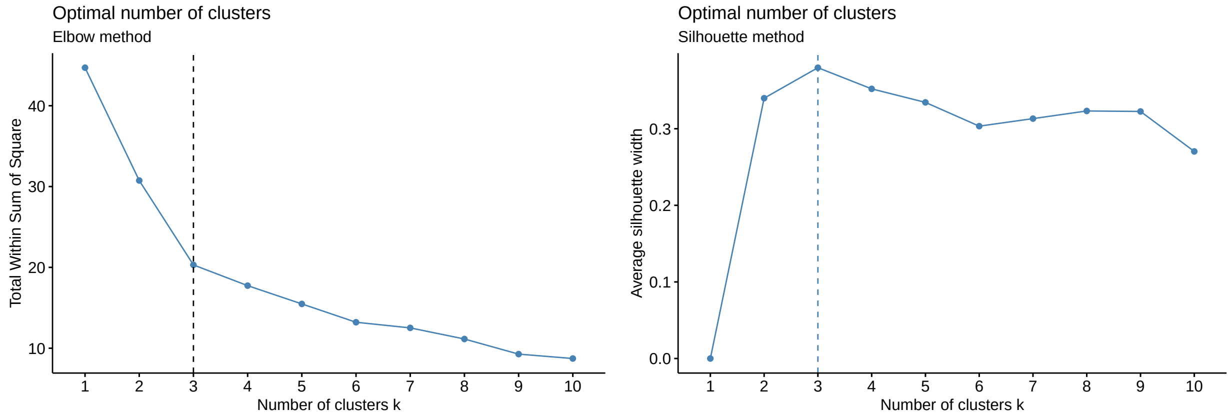

Figure 3: K-means Tuning

The result of K-means clustering with three clusters is shown in Table 4. The mentioned RFM values are averages of each cluster. In terms of *Recency*, cluster two contains the highest average value, yet the percentual differences between the highest and other values in this variable are not as large as in other variables. Cluster one shows the highest values for *Frequency* and *Monetary value* by quite some margin. Cluster one and two appear to have high similarity in *Frequency* and *Monetary value*. This is also portrayed in the weighted CLV of each cluster. Cluster two and three have comparable values, and cluster one has the largest value.

The clusters are ranked by W_{CLV} from highest to lowest in *CLV Rank*. *Average Spending* endorses the *CLV Rank* of each cluster. Clusters two and three again show similar values, and customers in cluster one spend on average more than two times as much as customers in other clusters.

	Recency	Frequency	Monetary	Wclv	CLV_Rank	Size	Average_Spending
Cluster 1	0.73	0.27	0.41	0.35	1	141	11841.16
Cluster 2	0.87	0.10	0.14	0.19	2	294	4586.93
Cluster 3	0.60	0.09	0.12	0.15	3	366	4042.50

Table 4: Cluster Characteristics

According to the RFM weights used in this research, Frequency and Monetary value are the two most important features of the RFM analysis. Therefore it is interesting to visually observe the distribution of the three clusters in a plot of Frequency and Monetary. This plot is shown in Figure 4 and again implies clusters with *CLV Rank* two and three show similar characteristics. It is clear the cluster with *CLV Rank* one shows the largest values for both Frequency and Monetary value.

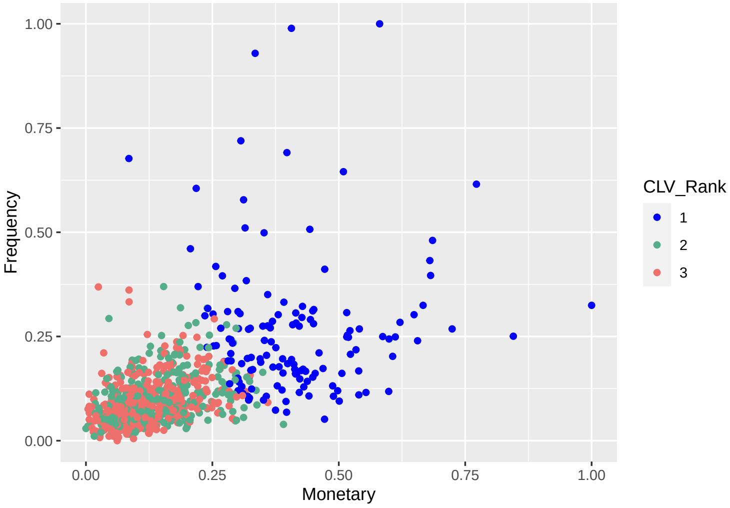

Figure 4: Cluster plot

To determine on what demographic characteristics the most valuable customers differentiate from other customers, Pearson's Chi-Squared Test is executed for the demographic variables in the dataset. The customers with the highest *CLV Rank* are compared with the total customer base. The results are showed in Table 5. A significant p-value (<0,05) indicates a significant different distribution of the variable between the two groups.

	Variable	p-value
1	AGE_DESC	0,06
2	MARITAL_STATUS_CODE	0,09
3	INCOME_DESC	0,00***
4	HOMEOWNER_DESC	0,07
5	HH_COMP_DESC	0,03*
6	HOUSEHOLD_SIZE_DESC	0,08
7	KID_CATEGORY_DESC	0,05

Table 5: Chi-Squared Results

Table 5 shows a significant difference in the distribution between the two groups in income and household composition on a 95% confidence interval. To further examine this difference, the proportions of the levels of both variables are calculated. Table 6 clearly shows the higher incomes being overrepresented for the most valuable customers, and low incomes being underrepresented. The last column shows the percentual difference of the proportion of a level in the most valuable customer segment compared to the total customer base.

	CLV_1	Total	Difference (%)
200-249K	0.00	0.62	-100.00
15-24K	3.55	9.24	-61.62
35-49K	11.35	21.47	-47.15
25-34K	7.09	9.61	-26.22
100-124K	3.55	4.24	-16.46
75-99K	12.77	11.99	6.52
50-74K	26.24	23.97	9.47
Under 15K	9.93	7.62	30.38
125-149K	8.51	4.74	79.40
150-174K	7.80	3.75	108.30
175-199K	3.55	1.37	158.22
250K+	5.67	1.37	313.15

Table 6: Difference in Income

Table 7 shows that in the most valuable cluster, households consisting of two adults and children are overrepresented, whilst single adults are underrepresented.

	CLV_1	Total	Difference (%)
Single Adult	24.11	31.84	-24.26
2 Adults No Kids	36.88	39.70	-7.11
2 Adults Kids	39.01	28.46	37.04

Table 7: Difference in Household Composition

4.3 Predicting Clusters

A logit model has been trained to predict *CLV Rank* with repeated 10-fold cross-validation. The results for predicting the test set, only considering the first four weeks, are shown in Table 8. The model has a general accuracy of 52,48%, and predicts the highest CLV cluster with a precision of 63,79%.

	1	2	3	Precision
1	37	11	10	0.64
2	4	31	41	0.41
3	2	47	59	0.55

Table 8: Logit Confusion Matrix

The random forest model first was tuned for the optimal value of *Mtry* by repeated 10-fold cross validation. *Mtry* has a minimum value of one, and the maximum value is the amount of independent variables, which is 14 in this case. All possible values were considered, and an *Mtry* of 14 was found to result in the highest model accuracy. After this, the optimal *Ntree* was determined by running a model with 500 trees, and selecting the amount of trees with the lowest error rate for the highest CLV cluster. The result of this is shown in Figure 5.

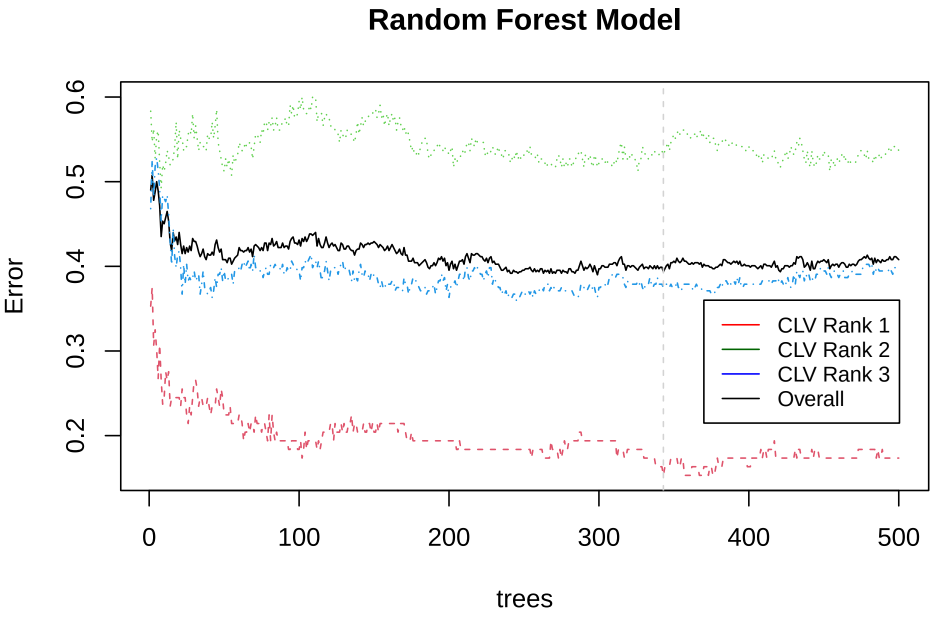

Figure 5: Random Forest Model

We are looking for the value of *Ntree* that results in the lowest error for predicting the most valuable customers. An *Ntree* of 343 showed the lowest error rate for predicting the highest CLV cluster, and is therefore chosen as the optimal amount of trees.

The random forest model yields a total accuracy of 53,72%. Although this may not seem particularly accurate, we are interested in the performance of predicting the most valuable clusters. On this specific part the model yields a precision of 77,78%, meaning 77,78% of households being predicted as *CLV Rank* one actually belong to *CLV Rank* one. The model yields a recall of 77,78%, meaning 77,78%% of all households belonging to *CLV Rank* one are also predicted as *CLV Rank* one. The final predictions of the model are shown in the confusion matrix in Table 9. As mentioned before, the clusters with *CLV Rank* two and three are very similar. The model also finds it difficult to make a distinction between the two, which results in low accuracy for predicting those clusters. An observation with CLV cluster two is often predicted as cluster three (47 times), and vice versa (47 times). This results in the low total model accuracy.

	1	2	3	Precision
1	35	5	5	0.78
2	5	37	47	0.42
3	3	47	58	0.54

Table 9: Random Forest Confusion Matrix

To study and compare the data sensitivity of both models, the process was rerun 5 more times, each time adding 4 weeks worth of data to the test set. Figure 6 shows the trajectory of both models' precision for predicting the most valuable cluster. The random forest model outperforms the logit model at any given amount of data. It should be noted that the difference in performance is the largest when the amount of data is the smallest. This can especially be seen by the steep increase of performance of the random forest model in week four until eight, compared to the more gradual increase of performance of the logit model. After eight weeks however, the random forest model does not really gain much performance, whereas the logit model gently keeps improving.

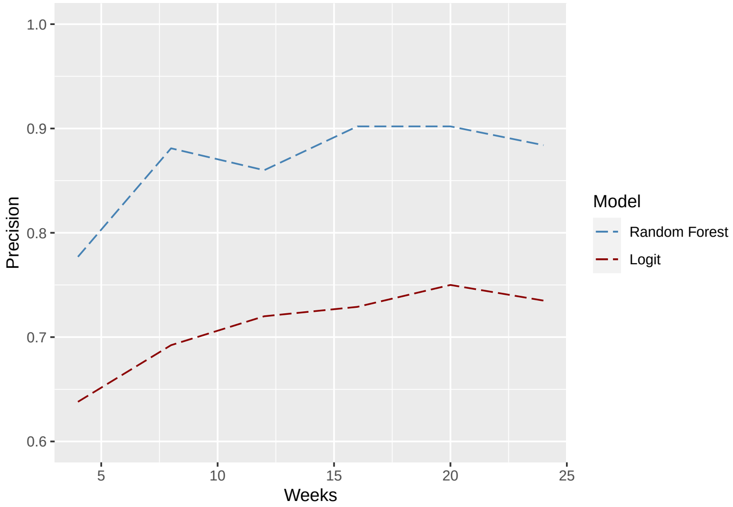

Figure 6: Sensitivity Analysis

To fully understand the random forest model and how the predictions are made, variable importance should be examined. Figure 7 shows the importance of all variables for the trained model, measured in mean decrease of the Gini coefficient. Mean decrease of the Gini coefficient shows the average decrease of the Gini coefficient if the variable is included. A high mean decrease in Gini coefficient indicates that the variable is able to split the sample in relative pure nodes, hence decreasing the Gini coefficient. This implies that the variable is important for the model. Looking at Figure 7, the average amount of baskets checked out per week is the most important variable, followed by the average spending per basket. It makes sense these two variables are the most important, since they have a direct link to revenue, and therefore customer value. The third most important variable is the personal coupon redemption rate, closely followed by a household's income.

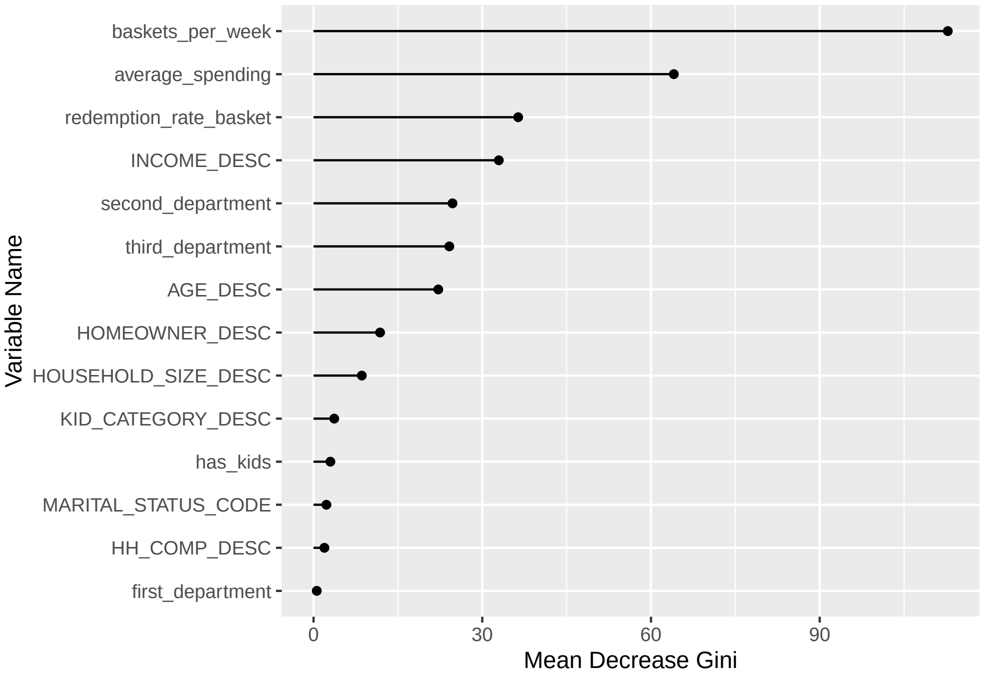

Figure 7: Variable Importance

4.4 Effect of Coupon Redemption Rate

Next to studying whether valuable customers can be recognized, this research also studies the relation between a household's personal coupon redemption rate and *CLV Ranking*. To examine this relation, partial dependence plots are examined. The distribution of *redemption rate basket*, or personal coupon redemption rate, is shown in Figure 8. The mean of *redemption rate basket* for *CLV Ranking* one, two and three is 8.87%, 8.29% and 7.91% respectively. There seems to be little variance across cluster, although customers with CLV Rank one seem to be more frequent coupon users on average.

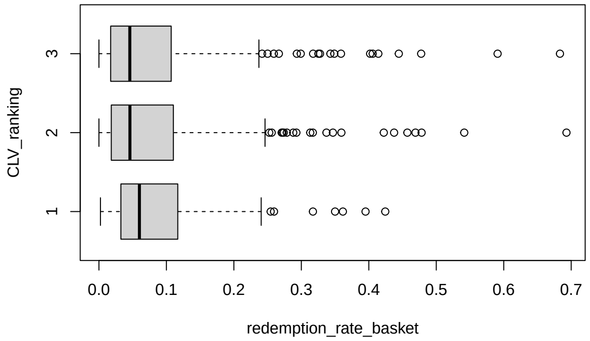

Figure 8: Redemption Rate per Cluster

Starting with the least valuable cluster, Figure 9 should be examined. Interpretation of these plots is somewhat challenging, and formula 6 should be consulted Due to the logarithmic nature of formula 6, small proportions of p_k lead to negative values for $f(x)$, and large proportions of p_k (in combination with small proportions for other classes) lead to positive values for $f(x)$.

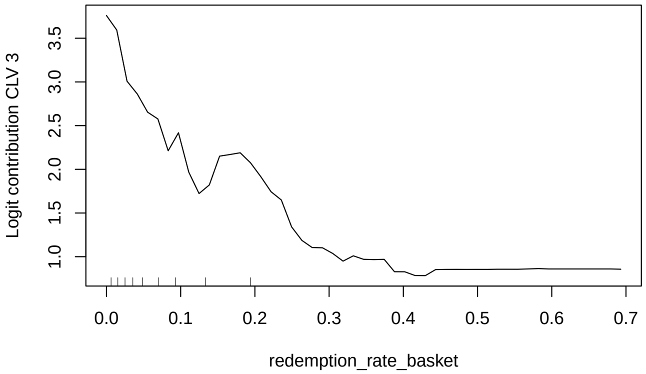

Figure 9: Partial Dependence Plot CLV 3

We can interpret the plot as follows: the line in Figure 9 shows that, as *redemption rate basket* increases, the proportion of the votes in the random forest model for *CLV Rank* three decreases, because the second part of formula 6 remains about constant within the same *CLV Rank*. To quantify this, the probability of a customer to be ranked as *CLV Rank* three is 99,99999% when the redemption rate is 0 (and $f(x)$ is 3,5), whereas the probability is 99,00% when the redemption rate is larger than 0,4, keeping everything else constant. We therefore conclude a negative, but very small effect of the personal coupon redemption rate on the probability of a customer being ranked as the lowest valuable cluster.

Moving on the *CLV Rank* two, a similar pattern as in Figure 9 is present. This makes sense, because the clusters have shown similarity in their characteristics, as shown in Table 4. Figure 10 shows a negative effect of the personal redemption rate on the probability of a household being predicted as CLV Rank two. It indicates that, keeping all other things constant, the probability of a household to be predicted as CLV rank two decreases as the personal coupon redemption rate increases. To again quantify, the probability of a customer to be ranked as *CLV Rank* two is about 99,99999% when the redemption rate is 0 (and the y-value of the partial dependence plot is 3,5), whereas the probability is 90,00% when the redemption rate is larger than 0,4.

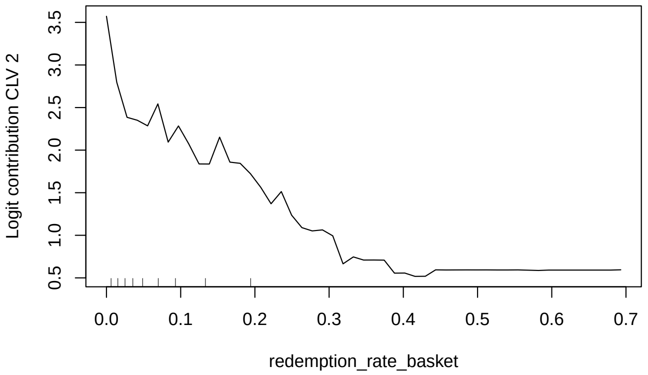

Figure 10: Partial Dependence Plot CLV 2

Finally, Figure 11 shows the partial dependence plot of the personal redemption rate on CLV Rank 1, the most valuable customers. Here we see a different trajectory than the previous plots. Figure 11 shows that the probability of a household to be predicted as *CLV Rank* one increases as the personal redemption rate increases. To quantify this, the increase of -7 to -2 on the y-axis indicates an increase of the probability to be ranked as *CLV Rank* one of 10^5. This might seem like a lot, but the base level of -7 indicates a probability of 10^{-7}, which is a very small probability. For the maximum value of -1,5, the probability to be ranked *CLV Rank* one is 0,03, keeping everything else constant. Hence, the effect is considered positive, but marginal.

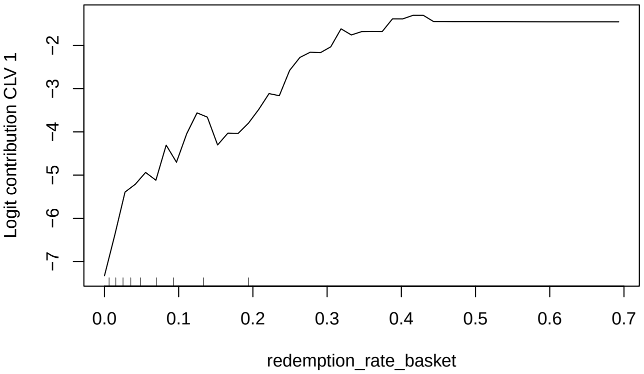

Figure 11: Partial Dependence Plot CLV 1

5 Discussion

The results of K-means based weighted RFM-clustering illustrated one cluster of retail customers being more valuable than the other two. On average, households in this cluster spend more than double the amount of the other clusters. According to the Chi-Square tests, households of the highest *CLV Rank* differentiate from the other households in two demographic characteristics: income and household composition. Therefore, the first sub question "What demographics distinguish high-value customers from other customers?" is answered as follows: the most valuable households have a higher income compared to other households, and more frequently consist of two adults with children.

After that, a random forest model was tuned and trained to yield the highest possible accuracy for predicting *CLV Rank* one, the cluster with the most valuable customers. The model delivered an accuracy of 77,78% when predicting the highest *CLV Rank* of the test set, which consisted of only the first 4 weeks of data for the customers, to resemble new clients. The same prediction was done by a logit model, which yielded a precision of 63,79% when predicting the highest *CLV Rank*. To study the data sensitivity of both models, a sensitivity analysis was performed. Four more weeks of data was considered when calculating the continuous variables of the test set, and the analysis was run again. This was done five times, considering a maximum of 24 weeks for the test set. The results are shown in Figure 6. The random forest model outperforms the logit model for any amount of weeks considered, and especially excels when there is four to eight weeks of data available, compared to the logit model. Taking this all into consideration, the second sub question "Is it possible to identify valuable retail customers in an early stage of their engagement?" is answered as follows: This research concludes a model can be trained to successfully identify valuable retail customers in the early stages of their engagement with the company. A random forest model outperforms a logit model, and shows particularly higher precision when the customer is with the company for four to eight weeks. After that period the precision does still marginally increase, but the difference with the logit model decreases.

To answer the third and last sub question "What is the relation of the personal coupon redemption rate on CLV for retail customers?", partial dependence plots derived from the random forest model were examined and interpreted. Figure 11 shows an upward trajectory, indicating a higher probability to be classified as a high value customer as the redemption rate increases. This effect is positive, but very marginal. For the two lower value clusters, the effect was negative, and also marginal. Because of the marginality, this research can not determine a substantial effect of the personal coupon redemption rate on CLV for retail customers.

This research has provided a way for retailers to identify valuable customers in the early stages of their engagement. Because companies now recognize which customers are worth to invest marketing resources in (and which customers are not) in an early stage, enhanced marketing budget allocation, and in turn a higher profit margin can be expected.

An interesting fusion of machine learning techniques is used for this paper. The integration of the results of K-means clustering in a random forest model yields accurate results when predicting whether a customer is valuable or not, contributing to the academic knowledge of the collaboration of these techniques. Next to this, the relation of the coupon redemption rate and customer value was explored. Although no substantial results were found, this is (to the best of my knowledge) the first study to examine this relation.

Despite the fact that this research found interesting results, it also has some drawbacks. This research is based on revenue, due to lack of information about costs and margin. A focus on profit instead of revenue might be more meaningful, and could lead to different results. Especially regarding the effect of coupon usage on customer value, since more coupon usage could lead to lower profit margins. Next to this, the final dataset contains 801 households. Although that is sufficient, an increased sample size could perhaps enhance the statistical power and generalizability of this research. Furthermore, the two most important variables of the random forest model (according to variable importance) to predict *CLV Ranking* are derived from values that are used to determine the *CLV Ranking*, namely the total of sales (Monetary in RFM) and the total amount of baskets (Frequency). Although the test set of the model only considers four weeks of data, and therefore the values in the test set are different than they would be if the complete dataset would be considered, this could be seen as a pitfall of the model. Lastly, because customers are grouped, information about individual customers is lost.

It would be interesting for further research to focus on profit instead of revenue, and compare results. Next to this, further exploration of the effect of coupon usage on customer value should be done, since the effect has not been determined yet. This could be done when the data allows to compute a correct value for CLV. Then, a model can be trained to predict CLV, and can be investigated to find the relation between CLV and the personal coupon redemption rate. Also, repeating this research using a dataset where some customers are actually new would be interesting to do, since the used dataset in this research only consists of regular customers from over the years.

6 References

Allender, W. J., & Richards, T. J. (2012). Brand Loyalty and Price Promotion Strategies: An Empirical Analysis. Journal of Retailing, 88(3), 323–342. https://doi.org/10.1016/j.jretai.2012.01.001

Bawa, K., Srinivasan, S. S., & Srivastava, R. (1997). Coupon Attractiveness and Coupon Proneness: A Framework for Modeling Coupon Redemption. Journal of Marketing Research, 34(4), 517. https://doi.org/10.2307/3151968

Belhadj, T. (2021). Customer Value Analysis Using Weighted RFM model: Empirical Case Study. https://www.asjp.cerist.dz/en/downArticle/196/7/3/170867

Berman, B. (2006). Developing an Effective Customer Loyalty Program. California Management Review, 49(1), 123–148. https://doi.org/10.2307/41166374

Bholowalia, P. (2014). EBK-Means: A Clustering Technique based on Elbow Method and K-Means in WSN. International Journal of Computer Applications - IJCA. https://www.ijcaonline.org/archives/volume105/number9/18405-9674

Breiman, L. (2001). Random Forests. Machine Learning 45, 5–32, https://doi.org/10.1023/A:1010933404324

Chang, W. S., Chang, C., & Li, Q. (2012). Customer Lifetime Value: A Review. Social Behavior and Personality, 40(7), 1057–1064. https://doi.org/10.2224/sbp.2012.40.7.1057

Chen, D., Sain, S. & Guo, K. (2012). Data mining for the online retail industry: A case study of RFM model-based customer segmentation using data mining. J Database Mark Cust Strategy Manag 19, 197–208 (2012). https://doi.org/10.1057/dbm.2012.17

Duboff, R. S. (1992). Marketing to Maximize Profitability. Journal of Business Strategy, 13(6), 10–13. https://doi.org/10.1108/eb039521

Friedman, J. (2001). Greedy Function Approximation: A Gradient Boosting Machine. The Annals of Statistics, 29(5). https://www.jstor.org/stable/2699986

Gislason, P., Benediktsson, J. A., & Sveinsson, J. R. (2006). Random Forests for land cover classification. Pattern Recognition Letters, 27(4), 294–300. https://doi.org/10.1016/j.patrec.2005.08.011

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. Applied Statistics, 28(1), 100. https://doi.org/10.2307/2346830

Khajvand, M., Zolfaghar, K., Ashoori, S., & Alizadeh, S. (2011). Estimating customer lifetime value based on RFM analysis of customer purchase behavior: Case study. Procedia Computer Science, 3, 57–63. https://doi.org/10.1016/j.procs.2010.12.011

Kim, J., Suh, E., & Hwang, H. (2003). A model for evaluating the effectiveness of CRM using the balanced scorecard. Journal of Interactive Marketing, 17(2), 5–19. https://doi.org/10.1002/dir.10051

Kim, S. W., Jung, T. H., Suh, E., & Hwang, H. S. (2006). Customer segmentation and strategy development based on customer lifetime value: A case study. Expert Systems With Applications, 31(1), 101–107. https://doi.org/10.1016/j.eswa.2005.09.004

Lawrence, R. L., Wood, S., & Sheley, R. L. (2006). Mapping invasive plants using hyperspectral imagery and Breiman Cutler classifications (randomForest). Remote Sensing of Environment, 100(3), 356–362. https://doi.org/10.1016/j.rse.2005.10.014

Leone, R. (1996). Coupon face value: Its impact on coupon redemptions, brand sales, and brand profitability,. Journal of Retailing, 72(3), 273–289. https://doi.org/10.1016/s0022-4359(96)90030-5

Lichtenstein, D. R., Netemeyer, R. G., & Burton, S. (1990). Distinguishing Coupon Proneness from Value Consciousness: An Acquisition-Transaction Utility Theory Perspective. Journal of Marketing, 54(3), 54–67. https://doi.org/10.1177/002224299005400305

Liu, D., & Shih, Y. (2005). Integrating AHP and data mining for product recommendation based on customer lifetime value. Information & Management, 42(3), 387–400. https://doi.org/10.1016/j.im.2004.01.008

Marcus, C. (1998). “A practical yet meaningful approach to customer segmentation”, Journal of Consumer Marketing, Vol. 15 No. 5, pp. 494-504. https://doi.org/10.1108/07363769810235974

Pal, M. (2005). Random forest classifier for remote sensing classification. International Journal of Remote Sensing, 26(1), 217–222. https://doi.org/10.1080/01431160412331269698

Pfeifer, Phillip E., Mark E. Haskins and Robert M. Conroy (2005). Customer Lifetime Value, Customer Profitability, and the Treatment of Acquisition Spending. Journal of Managerial Issues, 17(1), 11–25. https://www.jstor.org/stable/40604472

Reibstein, D. J., & Traver, P. A. (1982). Factors Affecting Coupon Redemption Rates. Journal of Marketing, 46(4), 102–113. https://doi.org/10.1177/002224298204600411

Soliman, H. S. (2011). Customer relationship management and its relationship to the marketing performance. International journal of business and social science, 2(10). https://science.donntu.edu.ua/ius/duzik/library/1.pdf

Srinivasan, S. S., Leone, R., & Mulhern, F. J. (1995). The Advertising Exposure Effect of Free Standing Inserts. Journal of Advertising, 24(1), 29–40. https://doi.org/10.1080/00913367.1995.10673466

Swaminathan, S., & Bawa, K. (2005). Category-specific coupon proneness: The impact of individual characteristics and category-specific variables. Journal of Retailing, 81(3), 205–214. https://doi.org/10.1016/j.jretai.2005.07.004

Venkatesan, R., & Farris, P. (2012). Measuring and Managing Returns from Retailer-Customized Coupon Campaigns. Journal of Marketing, 76(1), 76–94. https://doi.org/10.1509/jm.10.0162

Wei, J. T., Lin, S. Y., & Wu, H. H. (2010). A review of the application of RFM model. African Journal of Business Management, 4(19), 4199-4206. https://academicjournals.org/article/article1380555001_Wei%20et%20al.pdf

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. J, 2(2), 226–235. https://doi.org/10.3390/j2020016

7 Appendix

	Level	Frequency
1	19-24	0.06
2	25-34	0.18
3	35-44	0.24
4	45-54	0.36
5	55-64	0.07
6	65+	0.09

Table 10: Frequency Table of Age

	Level	Frequency
1	Married	0.57
2	Single	0.43
3	U	0.00

Table 11: Frequency Table of Marital Status

	Level	Frequency
1	100-124K	0.04
2	125-149K	0.05
3	15-24K	0.09
4	150-174K	0.04
5	175-199K	0.01
6	200-249K	0.01
7	25-34K	0.10
8	250K+	0.01
9	35-49K	0.21
10	50-74K	0.24
11	75-99K	0.12
12	Under 15K	0.08

Table 12: Frequency Table of Income

	Level	Frequency
1	Homeowner	0.63
2	Probable Owner	0.01
3	Probable Renter	0.01
4	Renter	0.05
5	Unknown	0.29

Table 13: Frequency Table of Homeowner Description

	Level	Frequency
1	2 Adults Kids	0.28
2	2 Adults No Kids	0.40
3	Single Adult	0.32

Table 14: Frequency Table of Houshold Composition

	Level	Frequency
1	1	0.32
2	2	0.40
3	3	0.14
4	4	0.07
5	5+	0.08

Table 15: Frequency Table of Houshold Size

	Level	Frequency
1	0	0.70
2	1	0.14
3	2	0.07
4	3+	0.09

Table 16: Frequency Table of Kids Category

	Level	Frequency
1	DRUG GM	0.01
2	GROCERY	0.98
3	KIOSK-GAS	0.00
4	MISC SALES TRAN	0.00
5	NUTRITION	0.00

Table 17: Frequency Table of First Department

	Level	Frequency
1	DELI	0.02
2	DRUG GM	0.45
3	GROCERY	0.02
4	KIOSK-GAS	0.19
5	MEAT	0.13
6	MEAT-PCKGD	0.04
7	MISC SALES TRAN	0.02
8	NUTRITION	0.01
9	PASTRY	0.00
10	PRODUCE	0.13

Table 18: Frequency Table of Second Department

	Level	Frequency
1	COSMETICS	0.00
2	DELI	0.05
3	DRUG GM	0.23
4	KIOSK-GAS	0.15
5	MEAT	0.17
6	MEAT-PCKGD	0.15
7	MISC SALES TRAN	0.01
8	MISC. TRANS.	0.00
9	NUTRITION	0.01
10	PASTRY	0.00
11	PRODUCE	0.21
12	SALAD BAR	0.00
13	SEAFOOD	0.00
14	SPIRITS	0.00

Table 19: Frequency Table of Third Department

	Level	Frequency
1	1	0.18
2	2	0.37
3	3	0.46

Table 20: Frequency Table of CLV Ranking