A STUDY OF CASHIER BEHAVIOR BASED ON RISK CLASSIFICATION USING CLUSTERING ALGORITHMS

Archana Satyamurthy, Fiona Patricia Moss, Haojun Zhu and Thao Phung

May 2, 2017

University of Wyoming

Department of Computer Science

COSC-45705010-04. Data Mining-2017-Amateur Data Miners

UNIVERSITY of
WYOMING

A STUDY OF CASHIER BEHAVIOR BASED ON RISK CLASSIFICATION USING CLUSTERING ALGORITHMS

Archana Satyamurthy, Fiona Patricia Moss, Haojun Zhu and Thao Phung

University of Wyoming
Department of Computer Science
1000 E University Ave
Laramie, WY 82071
`cosc.uwyo.edu`

May 2, 2017

1 Abstract

Our project focuses on identifying and classifying risky cashier behavior based on the Total Risk Factor. To classify the cashier behavior at Safeway, we needed to cluster the data. For this, we explored four clustering algorithms for classifying the cashier behavior: K-Means, Hierarchical, Agglomerative (Ward) and Affinity Propagation. The data was first transformed, normalized, and clustered. We followed two approaches to cluster the data. The first approach was to apply the four clustering algorithms to obtain the clusters for all the ten regions and observe which clustering algorithms provided meaningful clusters. The accuracy score of each algorithm was obtained with the help of confusion matrix. Based on the accuracy obtained, we observed that Hierarchical and Agglomerative clustering provided better clusters than K-Means and Affinity Propagation. In order to further clarify the performance of K-Means and Affinity Propagation algorithms, the second approach was applied to cluster the data. This method involved choosing applying K-Means to all the combinations of three columns or features and applying Affinity Propagation to all the combinations of two, three and four columns/features. Based on the results we obtained from both methods, we concluded that Hierarchical and Agglomerative clustering algorithms are suitable for classifying the data provided to us. We have worked only on Safeway data and the scope of our project is limited to Safeway.

2 Introduction

In the contemporary world, the retail industry faces challenges or risk at every stage. One such risk is shrinkage of inventories or risk of cashier behavior. These risk factors if not managed efficiently and are not remedied, will affect the profitability of the company and may even lead to its closure because profit margins are quite thin now-a-days. This is why these risk factors need to be identified, analyzed and preventive measures should be taken to minimize the risk. By doing so, the retail industry would be able to provide cheaper and better service to customers and thereby increase its profit.

Identifying risky cashier behavior could help Safeway to take preventive measures such as providing more training and taking other related measures like changing the risk management policy or operation system, etc.. This will help the cashiers to improve their performance and be less prone to mistakes. It would also help Safeway to compare risk factors to determine which risk factors were significant in calculating risk.

Safeway provided us with the Cashier Risk Report[2] which included various risk factors that contributed to the total risk factor for each cashier in ten regions in the United States. The main objective of our project is identifying and classifying the risky cashier behavior based on the Total Risk Factor. In order to achieve this, we have used four clustering algorithms: K-Means Clustering[7], Hierarchical Clustering[8], Agglomerative (Ward) Clustering[9] and Affinity Propagation[6].

3 Methodology

The methodology of our project involved the following four major steps:

3.1 Sanitizing the dataset

The first step for our project was to sanitize the data in order to make it easily readable. For our project, we were provided with an excel sheet "University of Wyoming Risk Reports.xlsx" report[1]. Cleaning the data involved removing unnecessary rows and dis-joining some of the merged columns. The resultant dataset was in a readable format.

3.2 Transforming/Normalizing the dataset

The second step was to transform the data by normalizing it. This step was crucial to obtain accurate results. Each column or risk factors in the "University of Wyoming Risk Reports" had different ranges. In order to have a uniform range for all the columns, we scaled the data. Transformation of the data played a vital role in the methodology because significant improvement in accuracy scores were observed.

We plotted the values of the Total Risk Factor and realized that the distribution is mostly normal but skewed to the left. Since we wanted to divide the Total Risk factor into five groups ranging from lowest to highest risk and classify cashier behavior based on these groups, we grouped all the values of the Total Risk Factor on the basis of this distribution for K-Means Clustering, Hierarchical Clustering and Agglomerative Clustering.

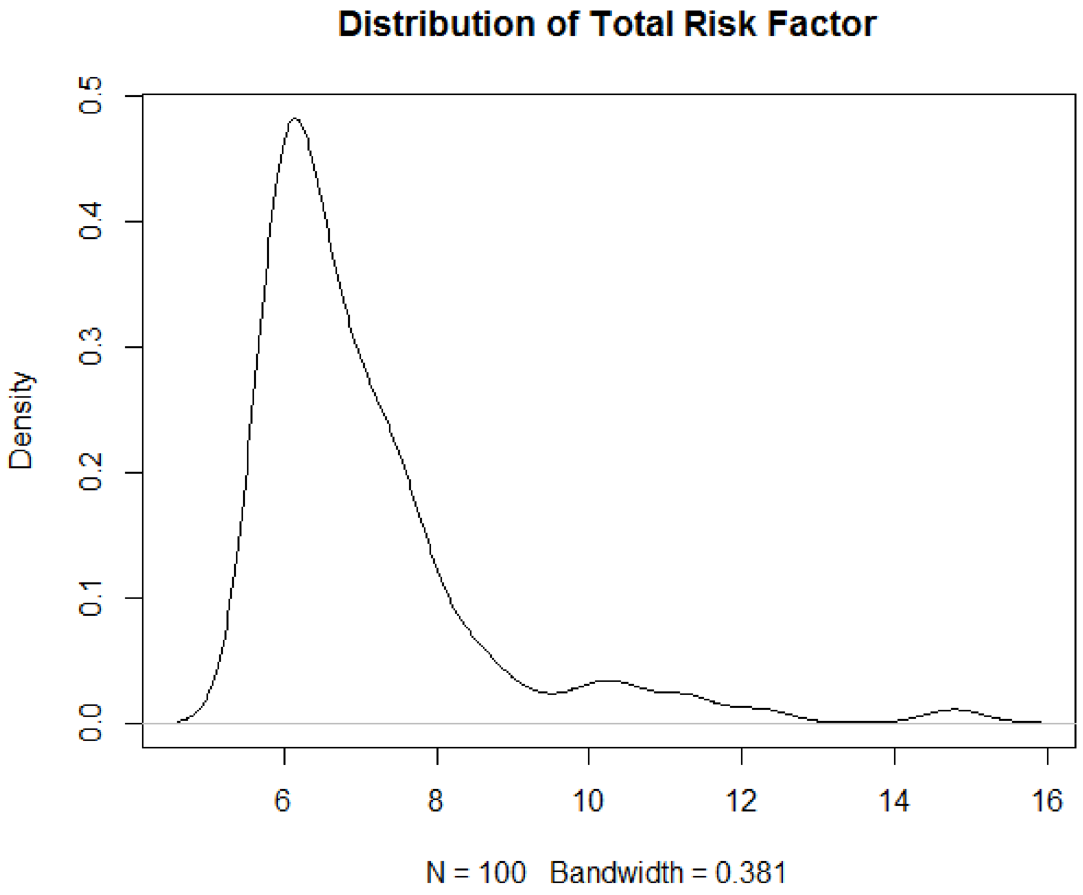


Figure 1: Distribution of the Total Risk Factor on the Original Scale

We used Box-Cox transformation[5] particularly for Affinity Propagation to somewhat normalize the distribution of the total risk factor score. We selected $\lambda = -4$ according to the Box-Cox plot to transform the total risk factor score Y:

$$Y' = \frac{Y^{\lambda} - 1}{\lambda}$$

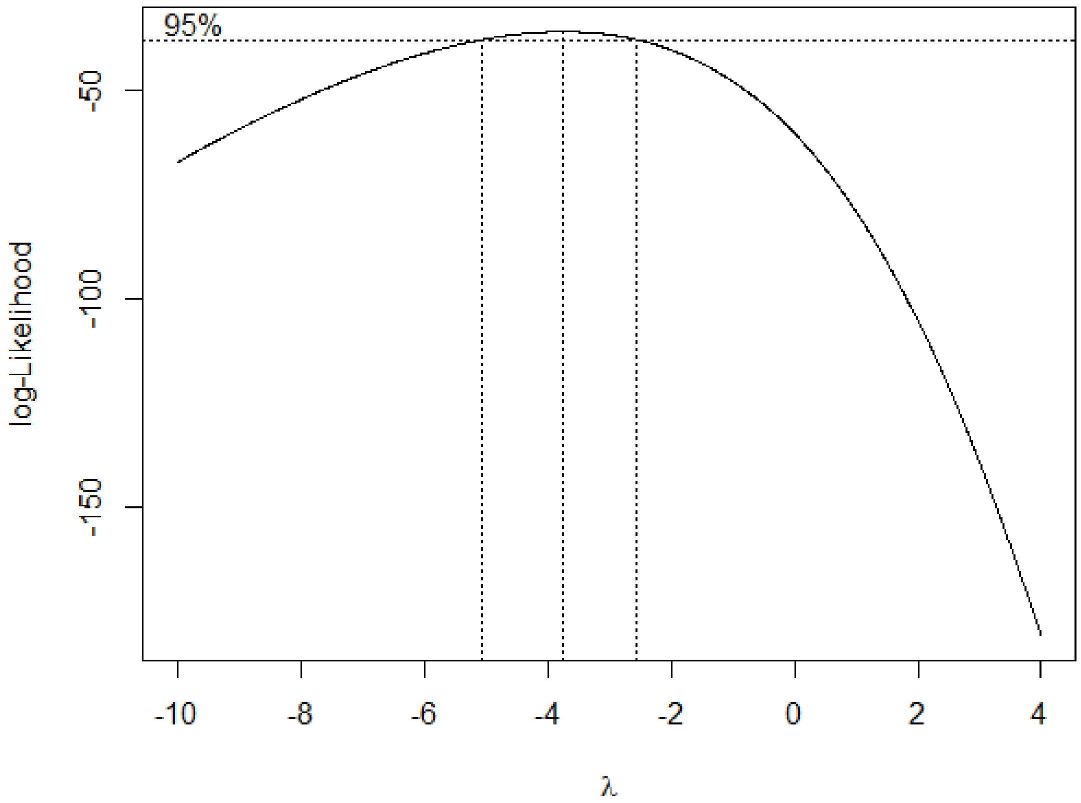


Figure 2: Profile Likelihood Plot for the Parameter of Box-Cox Transformation

After the Box-Cox transformation, the distribution was no longer highly skewed to the left. In fact, it appeared to be a mixture Gaussian distribution. For the purpose of this project, we could separate the total risk factor scores into several more equally distributed groups.

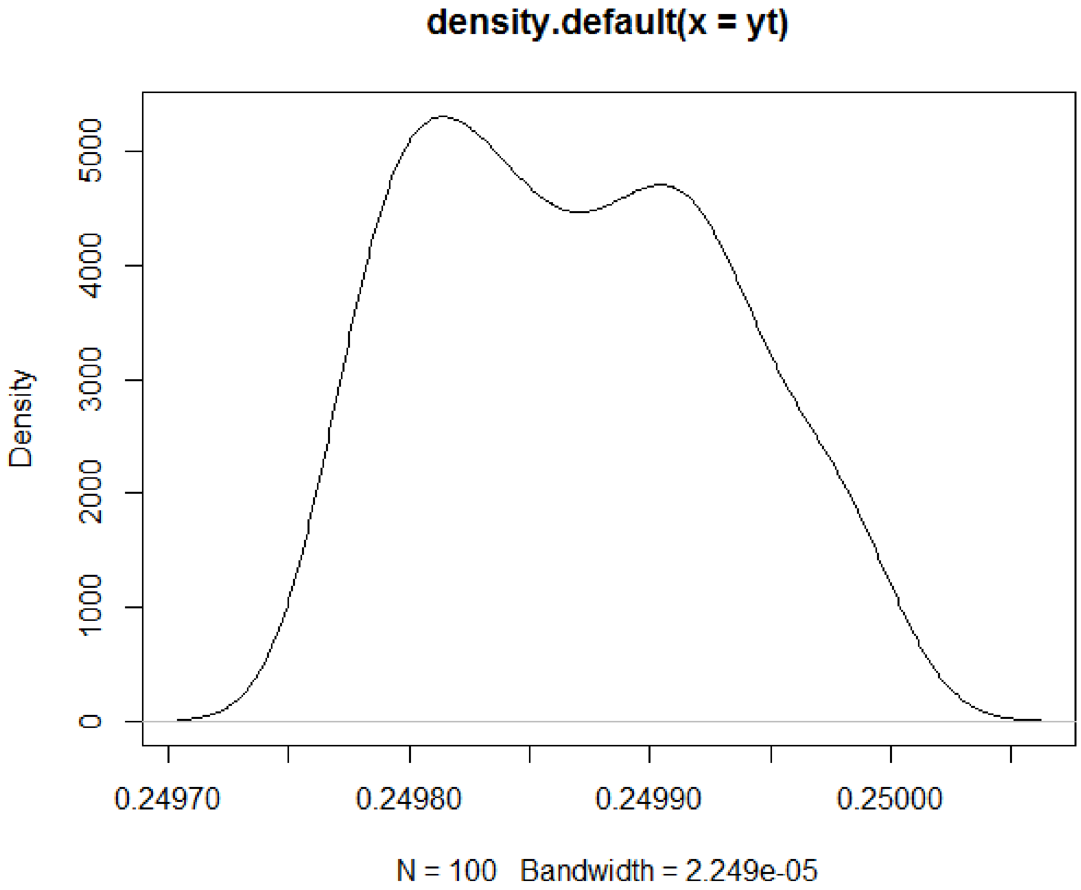


Figure 3: Distribution of the Total Risk Factor after Box-Cox Transformation

3.3 Clustering the data using various clustering algorithms

The next step was to cluster the data and was divided into the following two approaches:

3.3.1 Applying all the clustering algorithms to all the regions

The first approach was to directly apply the four clustering algorithms: K-Means, Hierarchical, Agglomerative (Ward) and Affinity Propagation to all the regions in our dataset. The cluster packages of R programming language[3] were used to perform this step. We then tested all the four algorithms on the dataset and clustered the data based on the Total Risk factor. The accuracy of each algorithm was calculated by developing the confusion matrix[4] which determined how accurately the algorithms were clustering the data. The accuracy score of each algorithm was obtained by comparing its clusters to the actual clusters obtained by observing the distribution of the total risk factor. To get better accuracy scores, we also considered clusters that were incorrectly clustered by one cluster i.e. which were off by 1 cluster.

3.3.2 Applying combinations of factors to K-Means and Affinity Propagation

The results of the first approach of all the algorithms were compared and it became evident that K-Means and Affinity Propagation did not have high accuracy scores and meaningful clusters. Hence for further clarification, we used the second approach only for K-Means and Affinity Propagation to conclude how many factors or columns were required to obtain a good accuracy value for K-Means and Affinity Propagation. We also intended to find the most important factors contributing to risk and consider them with this method of choosing factors. For this approach, all the combinations of two, three and four columns in the dataset were obtained and considered as factors. K-Means and Affinity Propagation algorithms were applied to each combination of three factors and each combination of two,three and four factors respectively, and the results were evaluated to know if any combination produced meaningful clusters. Due to constraints of computing resources, this approach was applied for only 'Denver' region. Just like the first approach, we considered clusters which were off by one even in this approach to get better accuracy scores.

3.4 Finding the most applicable algorithm to cluster our dataset

After getting the results from the above approaches, we determined the most suitable algorithm to cluster our dataset based on the accuracy score. We considered clusters which were off by one even in this approach to get better accuracy scores in both the approaches.

4 Evaluation

The first approach was applied to all regions. The observations for each algorithm are as follows:

1. **K-Means and Affinity Propagation:** K-Means clustering algorithm provided clusters that were not meaningful. The accuracy values were not consistent. Also, similar observations were seen for Affinity Propagation.

2. **Hierarchical Clustering:** Initially, the aim of our project was cluster the data using K-Means. Due to inconsistent values obtained from K-Means algorithm, We shifted our focus to hierarchical clustering. The accuracy of hierarchical clustering was much better than K-Means.

3. **Agglomerative Clustering:** Agglomerative algorithm was the next choice after hierarchical clustering. This algorithm also had better accuracy than K-Means.

The observations for the second approach are as follows:

K-Means and Affinity Propagation: We implemented K-Means on all combinations of three columns and Affinity Propagation on all combinations of two, three and four columns separately for 'Denver' region. The accuracy of both algorithms was still considerably low.

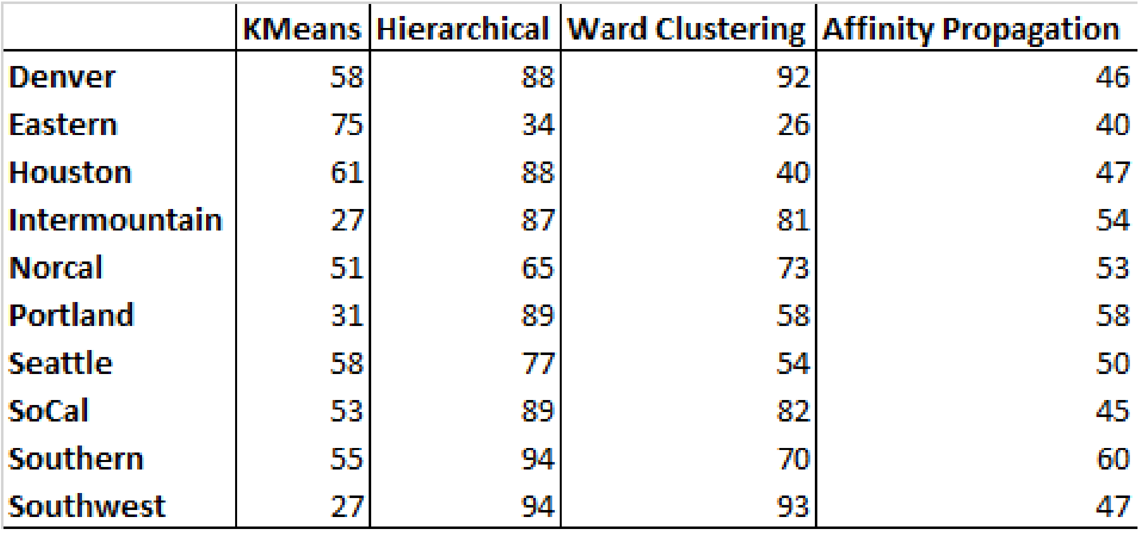
5 Results

5.1 Overview

The results show that Hierarchical and Agglomerative clustering algorithms are suitable for clustering our data. These two algorithms use the distance measure between pairs of observations to cluster the data and that's why these algorithms could be relevant to cluster our data.

K-Means algorithm is not appropriate for our data. This is due to the distribution of the data. K-Means requires clusters with equal radii in the data. However, the data that we were given did not have any equal radii clusters, which is the reason for the inconsistent results obtained from K-Means algorithm.

Affinity Propagation works by identifying the exemplars and cluster the data based on these exemplars. Identifying these good exemplars in our data could be difficult for this algorithm. Hence this could be a reason why the accuracy was less than Hierarchical and Agglomerative Clustering.



	KMeans	Hierarchical	Ward Clustering	Affinity Propagation
Denver	58	88	92	46
Eastern	75	34	26	40
Houston	61	88	40	47
Intermountain	27	87	81	54
Norcal	51	65	73	53
Portland	31	89	58	58
Seattle	58	77	54	50
SoCal	53	89	82	45
Southern	55	94	70	60
Southwest	27	94	93	47

Figure 4: Prediction Accuracy % of Different Algorithms Using All 28 Features

We observed that the prediction accuracy for K-Means clustering varied greatly. This was possibly due to the fact that the starting position of the clustering centroids were chosen randomly.

5.2 Affinity Propagation

We implemented Affinity Propagation clustering algorithm on all features and various subsets of the features.

First, we implemented Affinity Propagation on all 28 features and the 378 combinations of 2 features. For the latter, Affinity Propagation clustering algorithm was implemented on each of the pairwise combinations of 2 features. The results are visualized in the figure below. We color-coded the data points so as to obtain a clear visualization of the clustering results.

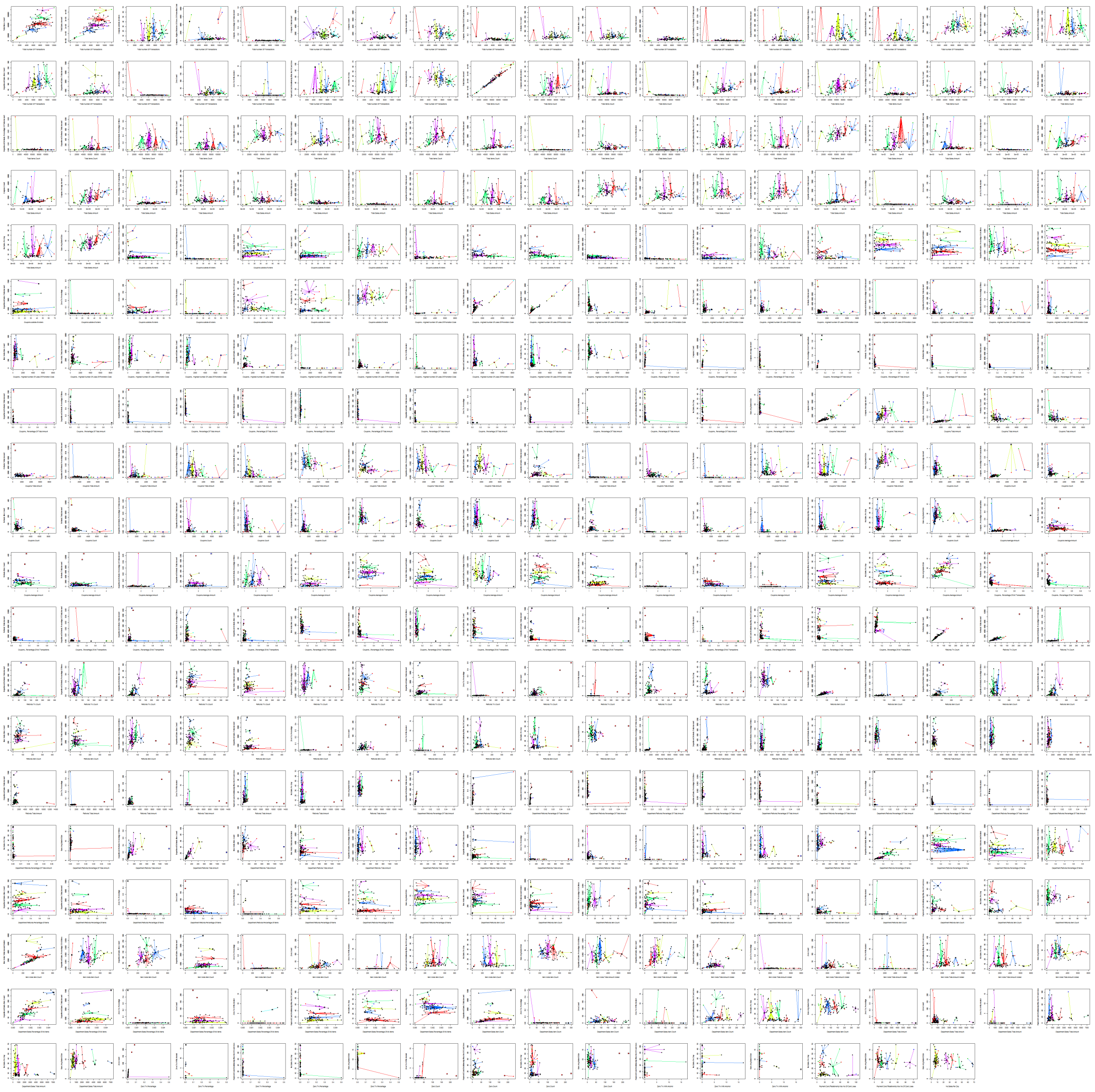
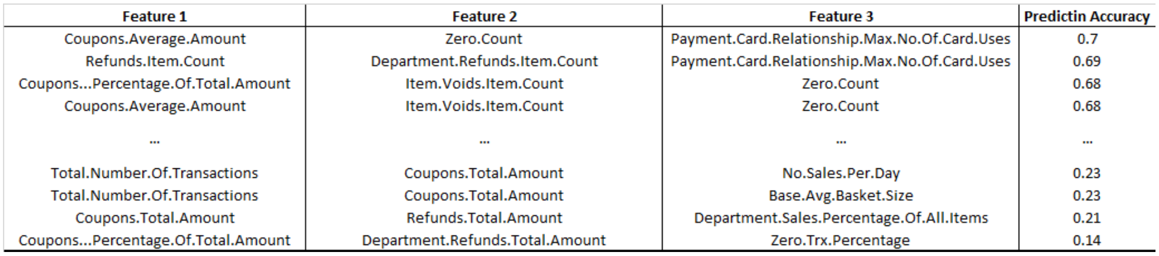


Figure 5: Scatterplot Matrix of Affinity Propagation Clustering Results on Combinations of 2 Features

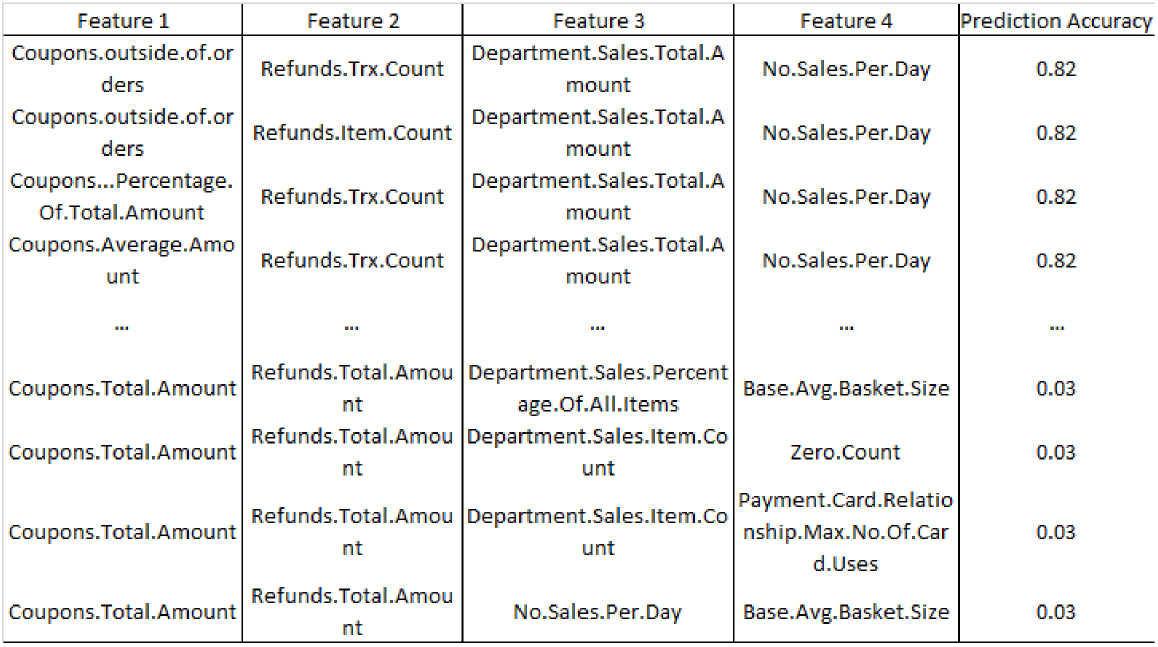
When using a subset of 3 features, we have 3,276 different combinations ($C^3_{28} = 3276$). The clustering prediction accuracy varies from as high as 70% to as low as 14%.



Feature 1	Feature 2	Feature 3	Predictin Accuracy
Coupons.Average.Amount	Zero.Count	Payment.Card.Relationship.Max.No.Of.Card.Uses	0.7
Refunds.Item.Count	Department.Refunds.Item.Count	Payment.Card.Relationship.Max.No.Of.Card.Uses	0.69
Coupons...Percentage.Of.Total.Amount	Item.Voids.Item.Count	Zero.Count	0.68
Coupons.Average.Amount	Item.Voids.Item.Count	Zero.Count	0.68
...	...	...	...
Total.Number.Of.Transactions	Coupons.Total.Amount	No.Sales.Per.Day	0.23
Total.Number.Of.Transactions	Coupons.Total.Amount	Base.Avg.Basket.Size	0.23
Coupons.Total.Amount	Refunds.Total.Amount	Department.Sales.Percentage.Of.All.Items	0.21
Coupons...Percentage.Of.Total.Amount	Department.Refunds.Total.Amount	Zero.Trx.Percentage	0.14

Figure 6: Prediction Accuracy of Affinity Propagation Clustering using Combinations of 3 Features

Next, we also explored the prediction accuracy using a subset of 4 features, with $C_{28}^{5} = 98280$ different combinations. We did not try different subset sizes, as the demand on computing resources was exhaustive. For example, $C_{28}^{10} = 13123110$, and $C_{28}^{14} = 40116600$.

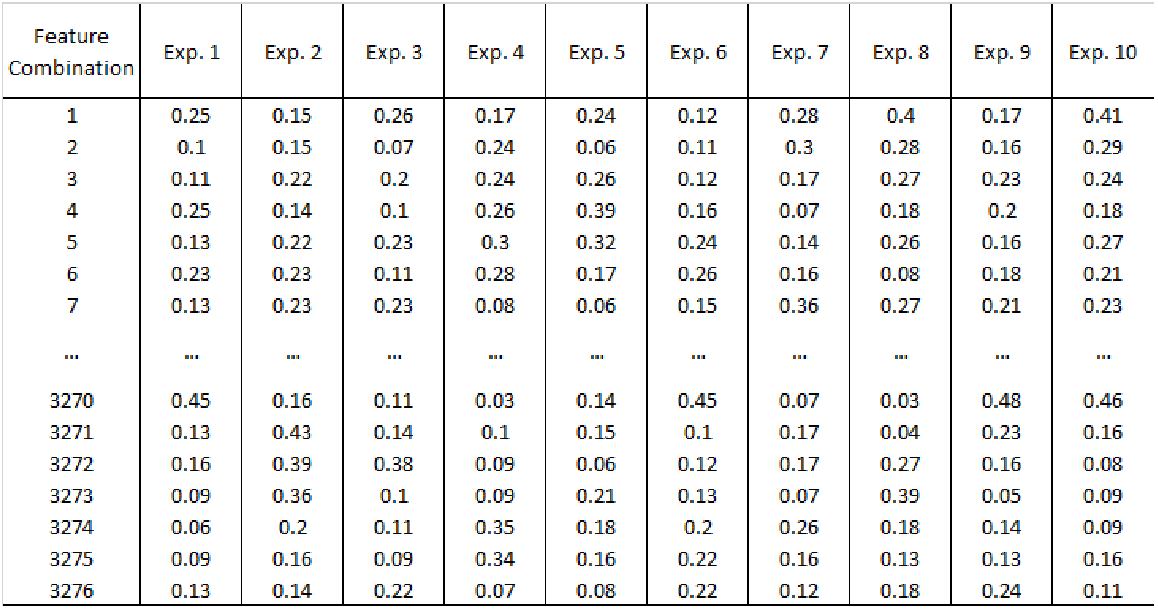


Feature 1	Feature 2	Feature 3	Feature 4	Prediction Accuracy
Coupons.outside.of.orders	Refunds.Trx.Count	Department.Sales.Total.Amount	No.Sales.Per.Day	0.82
Coupons.outside.of.orders	Refunds.Item.Count	Department.Sales.Total.Amount	No.Sales.Per.Day	0.82
Coupons...Percentage.Of.Total.Amount	Refunds.Trx.Count	Department.Sales.Total.Amount	No.Sales.Per.Day	0.82
Coupons.Average.Amount	Refunds.Trx.Count	Department.Sales.Total.Amount	No.Sales.Per.Day	0.82
...	...	...	...	...
Coupons.Total.Amount	Refunds.Total.Amount	Department.Sales.Percentage.Of.All.Items	Base.Avg.Basket.Size	0.03
Coupons.Total.Amount	Refunds.Total.Amount	Department.Sales.Item.Count	Zero.Count	0.03
Coupons.Total.Amount	Refunds.Total.Amount	Department.Sales.Item.Count	Payment.Card.Relationship.Max.No.Of.Card.Uses	0.03
Coupons.Total.Amount	Refunds.Total.Amount	No.Sales.Per.Day	Base.Avg.Basket.Size	0.03

Figure 7: Prediction Accuracy of Affinity Propagation Clustering using Combinations of 4 Features

5.3 K-Means Clustering

We have shown below the instability of K-Means clustering algorithm. For the Denver region, we selected different combination of 3 features ($C_{28}^{3} = 3276$). For each of the different combination of 3 features, we implemented K-Means clustering algorithms 10 times (experiment 1 to experiment 10). We observed that for the same combination of features (same row), the prediction accuracy varied greatly.



Feature Combination	Exp. 1	Exp. 2	Exp. 3	Exp. 4	Exp. 5	Exp. 6	Exp. 7	Exp. 8	Exp. 9	Exp. 10
1	0.25	0.15	0.26	0.17	0.24	0.12	0.28	0.4	0.17	0.41
2	0.1	0.15	0.07	0.24	0.06	0.11	0.3	0.28	0.16	0.29
3	0.11	0.22	0.2	0.24	0.26	0.12	0.17	0.27	0.23	0.24
4	0.25	0.14	0.1	0.26	0.39	0.16	0.07	0.18	0.2	0.18
5	0.13	0.22	0.23	0.3	0.32	0.24	0.14	0.26	0.16	0.27
6	0.23	0.23	0.11	0.28	0.17	0.26	0.16	0.08	0.18	0.21
7	0.13	0.23	0.23	0.08	0.06	0.15	0.36	0.27	0.21	0.23
...	...	...	...	...	...	...	...	...	...	...
3270	0.45	0.16	0.11	0.03	0.14	0.45	0.07	0.03	0.48	0.46
3271	0.13	0.43	0.14	0.1	0.15	0.1	0.17	0.04	0.23	0.16
3272	0.16	0.39	0.38	0.09	0.06	0.12	0.17	0.27	0.16	0.08
3273	0.09	0.36	0.1	0.09	0.21	0.13	0.07	0.39	0.05	0.09
3274	0.06	0.2	0.11	0.35	0.18	0.2	0.26	0.18	0.14	0.09
3275	0.09	0.16	0.09	0.34	0.16	0.22	0.16	0.13	0.13	0.16
3276	0.13	0.14	0.22	0.07	0.08	0.22	0.12	0.18	0.24	0.11

Figure 8: K-Means Clustering on Combinations of 3 Features Experiments

An interesting observation from the first approach is that for the 'Eastern' region, K-Means had better accuracy than Hierarchical and Agglomerative Clustering. We believe that this could be due to distribution of the data for the Eastern region containing equal radii clusters.

6 Conclusions and Future Work

Our project explored four clustering algorithms and implemented them on the Safeway dataset. The evaluation of the results obtained reveal that the accuracy of K-Means and Affinity Propagation is less than Hierarchical and Agglomerative Clustering and that Hierarchical and Agglomerative clustering algorithms are reasonable choices to cluster the dataset. Both the algorithms use distance measure to obtain the clusters and we believe that this factor could be the reason for a good accuracy.

Future Work could include understanding the parameters that help in obtaining the clusters, i.e., what columns/features of our data could be responsible to achieve good clusters.

We could also explore different combinations of subset of all features, as we previously discussed and demonstrated for K-Means Clustering and Affinity Propagation Clustering algorithms in the previous section, to find the best subset of features that best separate the cashier risk behaviors into 5 groups. However, this approach poses a very high demand on the computing resources. Considering the limited time-frame we have for this course project, we did not perform this exhaustive search problem. Based on the different combinations we have tried, we were able to identify a subset of 3 features that achieved a 70% prediction accuracy.

References

[1] Safeway. University of Wyoming Risk Reports. Feb. 2011. Raw data. N.p.

[2] Safeway-Albertson-Cashier Risk Report Documentation. Rep. N.p.: Sysrepublic, 2011. Print.

[3] "Clustering in R" http://www.statmethods.net/advstats/cluster.html

[4] "Confusion Matrix" http://www.dataschool.io/simple-guide-to-confusion-matrix-terminology

[5] "Box-Cox Transformation" https://stat.ethz.ch/R-manual/R-devel/library/MASS/html/boxcox.html

[6] "Affinity Propagation in R" https://cran.r-project.org/web/packages/apcluster/vignettes/apclu

[7] "K-Means Clustering in R" https://www.r-bloggers.com/k-means-clustering-in-r/

[8] "Hierarchical Clustering in R" https://www.r-bloggers.com/hierarchical-clustering-in-r-2/

[9] "Agglomerative (Ward) Clustering in R" https://stat.ethz.ch/R-manual/R-devel/library/cluster/html/agnes.html