Machine Learning Based Market Models in Business Operation

Annie Li

Portola High School, Irvine, Ca 92618, USA

Keywords: Machine Learning, Market Models, Business Operation, Data Visualization

Abstract: Shopping Centers Are Not Only Places for People to Engage in Commercial and Social Activities But Are Also Important Business and Entrepreneur Sites. for These Areas to Maintain High Customer Satisfaction and Maximal Profits, It is Crucial to Operate These Centers Efficiently and with Optimal Performance. However, Making Good Decisions to Improve a Business is a Complex Ordeal. This Project Utilizes Machine Learning to Design and Propose Market Models as a Method to Establish Business Strategies Suitable for Shopping Center Operation and Management. Three Market Models Comprised of Clustering Model, Principal Component Analysis (Pca), and Association Model Are Proposed in This Paper. Clustering Model is Used to Categorize the Different Stores into Multiple Groups. This Allows Us to Provide Specific Business Service for Each Group of Stores. Pca is Utilized to Organize Our Data into a Cleaner and Easier Format. Finally, the Association Model is Used to Find the Relationships between Different Stores and Store Categories, Which Provide Us with Information Regarding Popular Stores, Where Stores Should Be Located, and How to Provide Promotions from the Viewpoint of Business Operations.

1. Introduction

With the Countless Amenities Provided by Modern Shopping Malls, Buying Various Items and Visiting Shopping Centers Are Common Activities for Consumers Across the Nation. According to the Motley Fool Survey, Women Spend an Average of 399 Hours and 46 Minutes Shopping a Year, from Scouring the Mall for the Perfect Pair of Jeans to Loading Up the Cart with Groceries [1]. This Demonstrates That Shopping Centers Play a Big Role in People's Everyday Lives.

Shopping Centers, Defined by the International Council of Shopping Centers, Are "Groups of Retail and Other Commercial Establishments That Are Planned, Developed, Owned, and Managed as a Single Property" [2]. to Maintain These Areas as Attractive Commercial Playgrounds for Customers, It is Necessary for These Centers to Have More Refined Management on Stores, Allowing Them to Achieve Higher Customer Satisfaction and Larger Profits. However, as Markets/Businesses Continue to Grow, It Becomes More Complicated for Stores to Operate Efficiently and with Optimal Performance. Businesses Are Constantly Faced with Economic and Social Issues: How to Produce Enough Revenue, How to Enhance Their Relationships with Other Stores and Customers, Etc. Despite These Pressing Challenges, Many Managers of These Shopping Centers Continue to Follow Traditional Strategies, Such as Focusing on Managing Stores on an Individual Basis and Putting Stores in Random Areas. as a Result of These Poor Tactics, Business Efficiency, Growth, and Profitability Can Be Significantly Reduced.

Thanks to the Advancement of Modern Information Technology, Such as Artificial Intelligence (Ai) and Machine Learning, Various Business and Market-Related Methodologies Have Become Capable of Finding Ways to Measure the Performance of Shopping Centers. Ai is an Area of Computer Science That Emphasizes the Creation of Intelligent Machines That Work and React Like Humans. It Utilizes Machine Learning Algorithms Ultimately [3]. Machine Learning Can Be Used to Find Relationships within Data, Allowing It to Predict Future Trends. Generally, Machine Learning Encompasses Supervised and Unsupervised Learning. If a Model is Generated from a Set of Labeled Data, the Learning is Called *Supervised* [4]. for Example, the Classification Model and Regression Model Can Be Categorized as Supervised Learning. by Contrast, If a Model is Generated from a Set of Unlabeled Data, the Learning is Called *Unsupervised* [4], and the Program Must Analyze the Data for Rules and Patterns. for Example, the Clustering Model and Association

DOI: 10.25236/ISMHI.2019.132

Model Can Be Categorized as Unsupervised Learning.

Motivated by the Challenges to Establish Business Strategies Suitable for Shopping Center Operations and Activities, Three Important Market Models Are Designed and Proposed in This Paper. the Contributions Are Three-Fold:

●*Clustering Model* is Used to Categorize Stores into Various Groups, So That Different Stores Can Receive Different Strategies. This Model Allows for the Implementation of More Accurate Business Strategies Based on the Characteristics of the Stores.

●*Pca Model* is Used to Reduce the Dimension of the Dataset, So That We Can Decrease the Scale of Data for Analysis. the Data of Less Dimension Can Be Presented Clearly Using Data Visualization.

●*Association Model* is Used to Find the Association between Stores and Store Categories, So That the Stores Can Be Reasonably Arranged Based on Their Relationships, Allowing for the Achievement of Maximum Efficiency.

Furthermore, Applications That Utilize the Proposed Market Models Can Depict the Relationships between Various Stores and Their Customers and Competitors, Providing Useful Information to Shopping Center Management. Such Topics Include What the Most Popular Store Categories, Which Stores Should Be Located Near Each Other Based on Their Associations, When to Provide Promotions, and How to Advertise Stores and Track Customer Values.

The rest of this paper is organized as follows: Section 2 describes related works. Section 3 presents notations, source data, and problem formulation. Based on the transaction dataset, Section 4 provides our methodology for data analysis. In Section 5, the software tools used are introduced and the extensive experimental results are shown to validate our methodology. Section 6 concludes this paper and discusses possible future research directions.

2. Related Work

Many businesses today utilize AI and machine learning to make better decisions, increase business profitability and efficiency, establish better customer relationships, and instill better management methods.

In [5], Bilen eta al proposed a method to solve the business location estimator issue by Regression Model. In [6], benchmark suites are used to explore and evaluate various machine learning techniques that will allow for economic model calibration and harmonization. However, the infinite action space, simultaneous decision making, and imperfect information pose a computational challenge. Also, this study mainly focuses on video games, rather than business operation. In [7], methods of machine learning are considered to optimize the business process of sales management in retail trade. The research utilizes unsupervised machine learning algorithms, such as clustering, and implements dimensionality reduction methods. Dhandayudam et al. proposed a clustering algorithm to divide the customers into K groups in terms of Recency (R), Frequency (F) and Monetary (M) [8]. However, the optimal value of K is less discussed.

Although the research papers mentioned above utilize machine learning models to solve the specific issues, the results of the study cannot be demonstrated clearly, shown graphically for business strategies. In fact, with a plethora of data visualization methods, business owners can view useful scenarios and track metrics-crucial measurements of a business's growth and success.

In [9-11], many impressive data visualization examples are provided to show the power of data visualization. In [12], data visualization can help organization assess the direction of business. It enables to qualify answers to key business questions. For example, when you try to answer that how you can reduce operational costs, you can qualify your strategies by backing it up with data.

3. Preliminaries

3.1 Notations and Descriptions

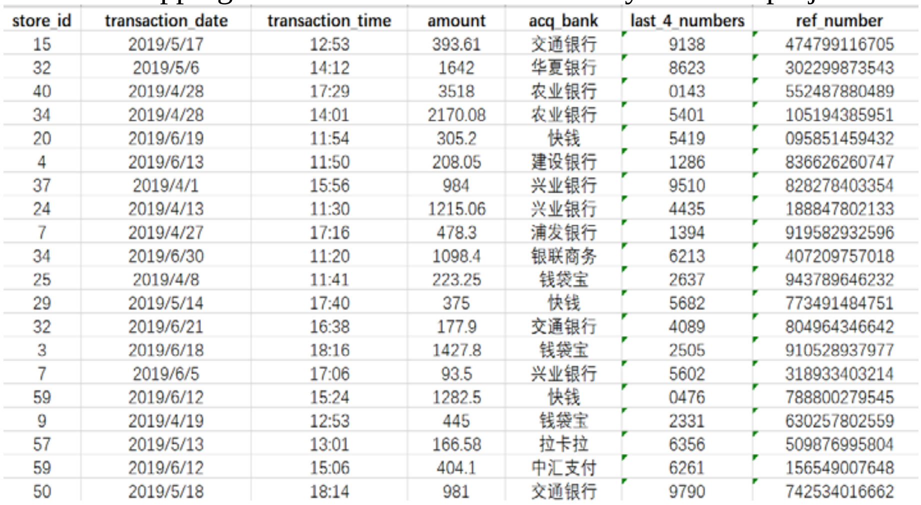
The notations throughout the paper are listed in Table 1.

Table 1 Notations

Symbol/Abbreviation	Description
ETL	Extraction-Transformation-Loading. It is used for data pre-processing or data cleaning before data mining.
KPI	Key Performance Identifier. It is a measurable value that demonstrates how effectively a company is achieving key business objectives. In the field of business operation, examples of KPIs include total revenue, per customer transaction, etc.
T	Transaction dataset: Each entry of *T* is a transaction record that happened in a store within the shopping center. The cardinality of *T* can be represented as $\|T\|$.
Clustering Model	It is an unsupervised model in machine learning used for classification. In detail, given a set of data points, each having a set of attributes, clusters will be found such that 1) data points in one cluster are similar to one another; 2) data points in separate clusters are less similar to one another.
PCA model	Principal Component Analysis. It is an unsupervised model in machine learning that functions in dimensionality reduction.
Association Rule Model	It is an unsupervised model in machine learning for association. In detail, given a set of records including multiple items, rules that demonstrate the occurrence of an item based on the occurrences of other items in the records will be found, e.g., an implication expression of the form X → Y, where X and Y are itemsets.

3.2 Data Description

Transaction data from UnionPay of China collected within 6 months (from 2019/1/1 to 2019/6/30) at a certain shopping mall was used as data for analysis in our project.



store_id	transaction_date	transaction_time	amount	acq_bank	last_4_numbers	ref_number
15	2019/5/17	12:53	393.61	交通银行	9138	474799116705
32	2019/5/6	14:12	1642	华夏银行	8623	302299873543
40	2019/4/28	17:29	3518	农业银行	0143	552487880489
34	2019/4/28	14:01	2170.08	农业银行	5401	105194385951
20	2019/6/19	11:54	305.2	快钱	5419	095851459432
4	2019/6/13	11:50	208.05	建设银行	1286	836626260747
37	2019/4/1	15:56	984	兴业银行	9510	828278403354
24	2019/4/13	11:30	1215.06	兴业银行	4435	188847802133
7	2019/4/27	17:16	478.3	浦发银行	1394	919582932596
34	2019/6/30	11:20	1098.4	银联商务	6213	407209757018
25	2019/4/8	11:41	223.25	钱袋宝	2637	943789646232
29	2019/5/14	17:40	375	快钱	5682	773491484751
32	2019/6/21	16:38	177.9	交通银行	4089	804964346642
3	2019/6/18	18:16	1427.8	钱袋宝	2505	910528937977
7	2019/6/5	17:06	93.5	兴业银行	5602	318933403214
59	2019/6/12	15:24	1282.5	快钱	0476	788800279545
9	2019/4/19	12:53	445	钱袋宝	2331	630257802559
57	2019/5/13	13:01	166.58	拉卡拉	6356	509876995804
59	2019/6/12	15:06	404.1	中汇支付	6261	156549007648
50	2019/5/18	18:14	981	交通银行	9790	742534016662

Fig.1 Transaction Dataset

As shown in Fig.1, each entry of the dataset consists of seven attributes: store ID, transaction date, transaction time, amount, name of account bank, last four digits of credit card, and reference number of credit card. For example, the first entry in the transaction record shows that a customer, who holds a credit card of last four digits 9138, paid $393.61 for some items in store 15 on 2019/5/17 at 12:53.

To obtain the most recent results for data analysis, the most recent three months (from 2019/4/1 to 2019/6/30) of data is extracted from T by data ETL.

3.3 Problem Formulation

For business operation, the aim is to efficiently manage the stores that are located within the shopping mall.

PROBLEM 1: The first problem is to divide the stores into several groups based on T. This will later allow us to provide special strategies and services to different stores. The clustering model is utilized to accomplish this task. Formally, this problem can be expressed as follows: given the transaction dataset T, we derive all possible KPI metrics for each store, and correspondingly divide all stores into K clusters where the value of K should be optimal.

PROBLEM 2: The second problem is to find the relationship among various stores based on T. Here the association rule model is adopted. Formally, this problem can be expressed as follows: given the transaction dataset T, we derive the sequence of stores that each customer has purchased, and correspondingly obtain strong association rules.

4. Methodology

4.1 Overview

As shown in Fig.2, based on the collected source data, data ETL is used to find KPIs for each store. Based on the derived KPIs, the stores will be divided into different clusters by applying the clustering models. In addition, the association rule model is applied to find the relationships among different stores and store categories. Finally, data visualization is provided to demonstrate the results of data mining.

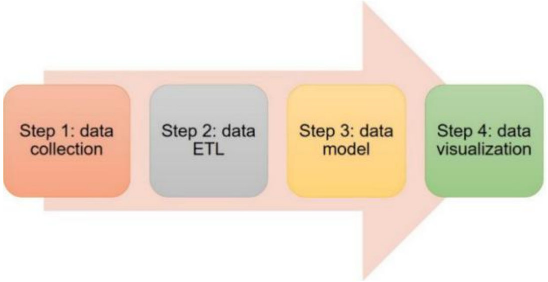


Fig.2 Data Flow Diagram

4.2 Store Clustering

4.2.1 Data Etl

In transaction dataset *T*, store_id will identify the store's name and the category it belongs to through a reference table. Transaction_date can be used to help compute some measurement metrics, such as number of transactions that occurred on weekdays and weekends. Furthermore, summing all amounts in *T* happened in store *i* can be used to derive the total amount for store *i*. Furthermore, amount can be used to find other KPIs, such as per customer transaction, which is also one of the most important KPIs in business operation. In addition, last_4_numbers and ref_number help identify the unique transaction record.

Based on the above analysis, given the transaction dataset *T* and store *i*, four important KPI metrics are calculated as follows:

Total amount $TM(i)$. For each entry *e* of *T*, we have $TM(i)=\sum e.\text{amount}$ where $e.\text{store\_id}=i$. In addition, the number of transactions related to store *i* is denoted as $N(i)$.

Number of transactions on weekdays $NTDAYS(i)$. Given transaction dataset *T*, $NTDAYS(i)$ is calculated by counting the number of transactions of store *i* happened between Monday to Friday.

Number of transactions on weekends $NTENDS(i)$. Given transaction dataset *T*, we have $NTENDS(i)= N(i)\text{-}NTDAYS(i)$.

Per customer transaction $PCT(i)$. Given transaction dataset *T*, we have $PCT(i)=TM(i)/N(i)$.

In addition, notice that the first metric can be derived from other metrics, i.e., $TM(i)=PCT(i)*(NTDAYS(i)+NTENDS(i))$, we removed the metric $TM(i)$, thus reducing the number of metrics from four to three without the loss of information.

4.2.2 Data Clustering

After obtaining three KPI attributes for each store, the main steps for data clustering are shown as follows.

```
Step 1: Select K stores in shopping center (represent initial centroids).
Step 2: Repeat.
Step 3:  Form K clusters by assigning all stores to the closest centroid by Euclidean Distance.
Step 4:  Recompute centroid of each cluster until the centroids do not change.
```

Here the Euclidean distance mentioned in step 3 is defined to be the distance between two points X (x_1, x_2…, x_m) and Y (y_1, y_2…,y_m) with m-dimension: Dist([x_1,x_2,…,x_m],[y_1,y_2,…,y_m]) =

$$\sqrt{\sum_{i=1}^{m}(x_i - y_i)^2} .$$

Note that the value of K should be determined before the clustering algorithm runs. This means that we need to find the optimal value of K. To solve this problem, another metric called Sum of Squares Error (SSE) used to measure the efficiency of the clustering algorithm is introduced: Given a set X of n points in a d-dimensional space and an integer K, group the points into K clusters C= {C$_1$,C$_2$,…,C$_K$}, we have $SSE = Cost(C) = \sum_{i=1}^{K}\sum_{x\in C_i}(x - c_i)^2$ where c_i is the centroid of the points in cluster C_i

One method to find the optimal number of clusters is the elbow method. The idea of the elbow method is to run K-means clustering on the dataset for a range of values of K (say, K from 1 to 10 in the examples), and for each value of K calculate SSE. Obviously, the value of SSE will be proportionally reduced with the increase of K. Therefore, we will find the optimal value of K such that as the increase of K, the decrease ratio of SSE is maximal, e.g., K=arg MAX g(K) where g(K)=[SSE(K-1)-SSE(K)]/SSE(K-1) and K is an integer number.

4.3 Dimension Reduction

The PCA algorithm is used for dimensional reduction. The steps are as follows:

Step 1: Standardization

Standardization transforms data to a comparable range, creating less biased results.

Step 2: Covariance matrix computation

This steps helps find relationships between attributes/variables in the data. We compute the covariance matrix because the data variables may contain redundant information. The sign of covariance is significant, as it can represent relationships between variables. For example, if the sign of covariance is positive, then two variables increase or decrease together. If the sign is negative, then one variable increases or decreases, while the other decreases or increases (inversely related).

Step 3: Compute eigenvectors and eigenvalues of covariance matrix

The principal components of the data are found by computing eigenvectors and eigenvalues from the covariance matrix.

Step 4: Feature vector

We choose whether to keep all the principal components or leave out ones of lesser eigenvalue (lower significance).

Step 5: Reorient data

Data is reoriented using the formula shown:

$$FinalDataSet = FeatureVector^T * StandardizedOriginalDataSet^T$$

After PCA, the dimensionality of store metrics can be reduced, while keeping most of the original data.

4.4 Store Association

Some terms used in this part are provided as follows.

Itemset. A collection of one or more items in a transaction. That is, an itemset is called a k-itemset, if it contains k items, e.g., {milk} is a 1-itemset and {milk, bread} is a 2-itemset. In our project, the itemset is defined to be a collection of one or more items.

Support. Given the transaction dataset T and a rule A→B, support is defined as the probability of these two items happening at the same time, e.g., support(A→B)=Pro(AB)/|T|.

Confidence. It is defined as the probability of an item occurring, given another item has already occurred, e.g., conf(A→B) = $Prob$(AB)/$Prob$(A).

Frequent Itemset. Given a threshold of support $minsup$, an itemset I is called a frequent Itemset if the support of *I* is greater than or equal to *minsup*.

Strong Association Rule. Given a threshold of support *minconf*, a rule $X \rightarrow Y$ is called a strong association rule if $X \rightarrow Y$ is a frequent itemset and confidence of $X \rightarrow Y$ is greater than or equal to *minconf*.

The main steps in the association algorithm are as follows:

Step 1: Identify frequent itemsets from data

Find all frequent itemsets. The Apriori principle states that the support of an itemset never exceeds the support of its subsets. if an itemset is frequent, then all its subsets must also be frequent. If an itemset is not frequent, then its supersets cannot be frequent. This is known as the anti-monotone property of support. After the execution of step 1, the support of each elements in the frequent itemsets is not less than the threshold *minsup*.

Step 2: determine possible rules from frequent itemsets

Based on the frequent itemsets we have obtained, we enumerate all possible rules and find the strong association rules, such that the confidence of rule is not less than the threshold *minconf*.

5. Results

5.1 Software Introduction

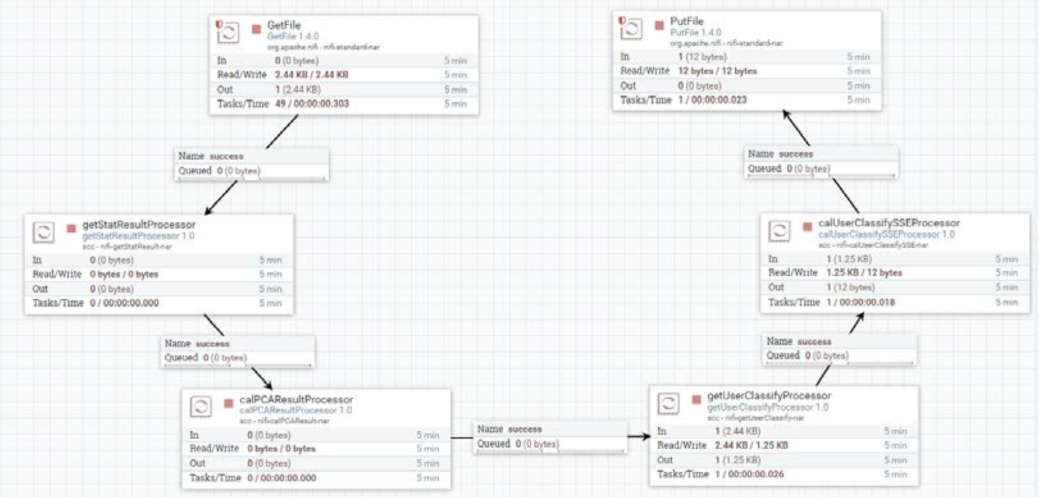
NiFi supports powerful and scalable directed graphs of data routing, transformation, and system mediation logic. Some of the high-level capabilities and objectives of Apache NiFi include web-based user interface, high configuration, data provenance, and designs for extension and security [13].

Power BI is a business analytics service created by Microsoft. It aims to provide interactive visualizations and business intelligence capabilities with an interface simple enough for users to create their own reports and dashboards [14].

In our experiments, Apache Nifi and Power BI are used to organize and compute data, as well as present our results visually. In detail, Nifi is used to organize the process for data ETL. The customized interfaces and components of Nifi give directions to read the data file, implement business logics/procedures, and find specific metrics based on the data attributes from the source file. After ETL procedure, the values of four metrics will be output: store ID, amount per customer transaction, number of transactions on weekdays/weekends. Then, Power BI is utilized to create graphs used for data visualization.

5.2 Clustering Demonstration

The complete data process for clustering is shown in Fig.3.



GetFile
GetFile 1.4.0
org.apache.nifi - nifi-standard-nar
In 0 (0 bytes) 5 min
Read/Write 2.44 KB / 2.44 KB 5 min
Out 1 (2.44 KB) 5 min
Tasks/Time 49 / 00:00:00.303 5 min

Name success
Queued 0 (0 bytes)

getStatResultProcessor
getStatResultProcessor 1.0
scc - nifi-getStatResult-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

calPCAResultProcessor
calPCAResultProcessor 1.0
scc - nifi-calPCAResult-nar
In 0 (0 bytes) 5 min
Read/Write 0 bytes / 0 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 0 / 00:00:00.000 5 min

Name success
Queued 0 (0 bytes)

getUserClassifyProcessor
getUserClassifyProcessor 1.0
scc - nifi-getUserClassify-nar
In 1 (2.44 KB) 5 min
Read/Write 2.44 KB / 1.25 KB 5 min
Out 1 (1.25 KB) 5 min
Tasks/Time 1 / 00:00:00.026 5 min

Name success
Queued 0 (0 bytes)

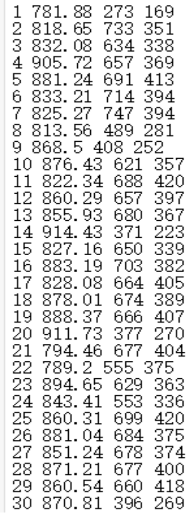
calUserClassifySSEProcessor
calUserClassifySSEProcessor 1.0
scc - nifi-calUserClassifySSE-nar
In 1 (1.25 KB) 5 min
Read/Write 1.25 KB / 12 bytes 5 min
Out 1 (12 bytes) 5 min
Tasks/Time 1 / 00:00:00.018 5 min

Name success
Queued 0 (0 bytes)

PutFile
PutFile 1.4.0
org.apache.nifi - nifi-standard-nar
In 1 (12 bytes) 5 min
Read/Write 12 bytes / 12 bytes 5 min
Out 0 (0 bytes) 5 min
Tasks/Time 1 / 00:00:00.023 5 min

Fig.3 Clustering Flow Diagram in Nifi

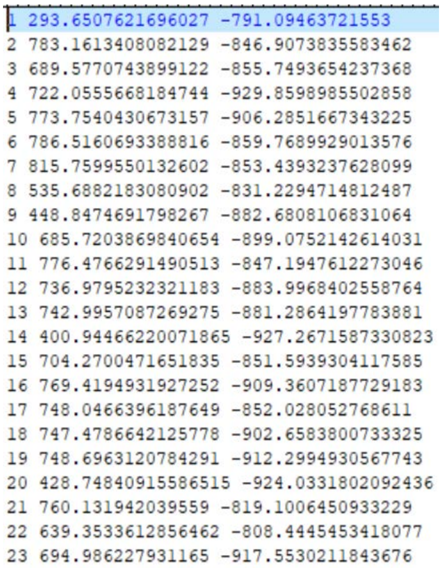
Firstly, a processor called *getStatResultProcessor* is developed in Nifi to derive three KPI metrics for each store. Given the original data, represented in Fig.1 of Section 3.2, the processor organizes the data into multiple columns, depicted in Fig. 4. As shown in this figure, the first column labels each record by its store ID, the second column displays amount per customer transaction, and the third and fourth columns indicate transactions on weekdays and weekends respectively.



```
1 781.88 273 169
2 818.65 733 351
3 832.08 634 338
4 905.72 657 369
5 881.24 691 413
6 833.21 714 394
7 825.27 747 394
8 813.56 489 281
9 868.5 408 252
10 876.43 621 357
11 822.34 688 420
12 860.29 657 397
13 855.93 680 367
14 914.43 371 223
15 827.16 650 339
16 883.19 703 382
17 828.08 664 405
18 878.01 674 389
19 888.37 666 407
20 911.73 377 270
21 794.46 677 404
22 789.2 555 375
23 894.65 629 363
24 843.41 553 336
25 860.31 699 420
26 881.04 684 375
27 851.24 678 374
28 871.21 677 400
29 860.54 660 418
30 870.81 396 269
```

Fig.4 Result of Processor Getstatresultprocessor in Nifi

Secondly, a processor called *getPCAResultProcessor* is developed in Nifi to reduce the dimension of the data. The result of this processor is shown in Fig.5. Comparing Fig.5 with Fig.4, we can observe that the number of dimensions has reduced from four to three.



```
1 293.6507621696027 -791.09463721553
2 783.1613408082129 -846.9073835583462
3 689.5770743899122 -855.7493654237368
4 722.0555668184744 -929.8598985502858
5 773.7540430673157 -906.2851667343225
6 786.5160693388816 -859.7689929013576
7 815.7599550132602 -853.4393237628099
8 535.6882183080902 -831.2294714812487
9 448.8474691798267 -882.6808106831064
10 685.7203869840654 -899.0752142614031
11 776.4766291490513 -847.1947612273046
12 736.9795232321183 -883.9968402558764
13 742.9957087269275 -881.2864197783881
14 400.9446622007186 5 -927.2671587330823
15 704.2700471651835 -851.5939304117585
16 769.4194931927252 -909.3607187729183
17 748.0466396187649 -852.028052768611
18 747.4786642125778 -902.6583800733325
19 748.6963120784291 -912.2994930567743
20 428.7484091558651 5 -924.0331802092436
21 760.131942039559 -819.1006450933229
22 639.3533612856462 -808.4445453418077
23 694.986227931165 -917.5530211843676
```

Fig.5 Result of Processor Getpcaresultprocessor in Nifi

Thirdly, two processors called *getUserClassifyProcessor* and *calUserClassifySSEProcessor* are used to cluster all the stores into certain groups and calculate the corresponding value of SSE, respectively. Given different values of *K*, the value of SSE is shown in Table 2. The relationship between *K* and SSE is shown in Fig.6.

Table 2 K Vs. Sse

Value of K	Value of SSE
2	257473.601
3	181314.594
4	144297.902
5	114148.342

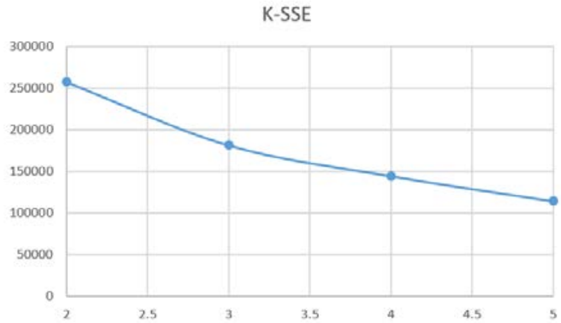


Fig.6 Relationship between K and Sse in Clustering

Based on the data in Table 2 and Fig.6, we can derive that the optimal number of *K* is three due to the fact that the value of SSE for *K*=3 results to a maximal deduction ratio by 29.58%.

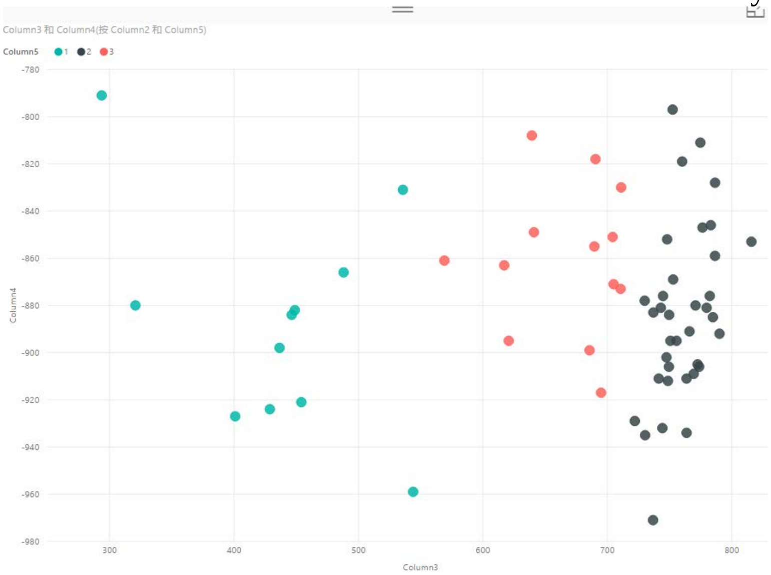
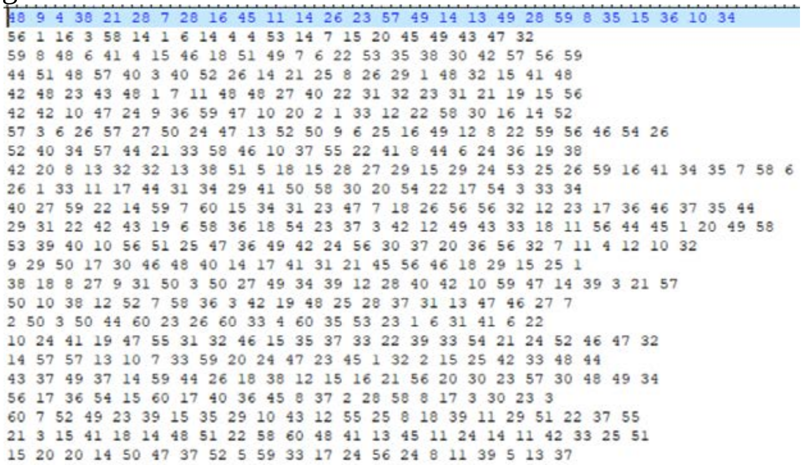


Fig.7 Clustering Result in Power Bi (K=3)

Given the number of clusters, 3, the clustering result demonstrated in Power BI is shown in Fig.7. From this figure, based the reduced attributes of stores, we can observe that all stores can be divided into three groups clearly, which means that the clustering algorithm proposed in this project is effective.

5.3 Association Demonstration

Given the transaction dataset T, the sequence of store id for each customer is derived. Part of data is shown in Fig.8.



```
48 9 4 38 21 28 7 28 16 45 11 14 26 23 57 49 14 13 49 28 59 8 35 15 36 10 34
56 1 16 3 58 14 1 6 14 4 4 53 14 7 15 20 45 49 43 47 32
59 8 48 6 41 4 15 46 18 51 49 7 6 22 53 35 38 30 42 57 56 59
44 51 48 57 40 3 40 52 26 14 21 25 8 26 29 1 48 32 15 41 48
42 48 23 43 48 1 7 11 48 48 27 40 22 31 32 23 31 21 19 15 56
42 42 10 47 24 9 36 59 47 10 20 2 1 33 12 22 58 30 16 14 52
57 3 6 26 57 27 50 24 47 13 52 50 9 6 25 16 49 12 8 22 59 56 46 54 26
52 40 34 57 44 21 33 58 46 10 37 55 22 41 8 44 6 24 36 19 38
42 20 8 13 32 32 13 38 51 5 18 15 28 27 29 15 29 24 53 25 26 59 16 41 34 35 7 58 6
26 1 33 11 17 44 31 34 29 41 50 58 30 20 54 22 17 54 3 33 34
40 27 59 22 14 59 7 60 15 34 31 23 47 7 18 26 56 56 32 12 23 17 36 46 37 35 44
29 31 22 42 43 19 6 58 36 18 54 23 37 3 42 12 49 43 33 18 11 56 44 45 1 20 49 58
53 39 40 10 56 51 25 47 36 49 42 24 56 30 37 20 36 56 32 7 11 4 12 10 32
9 29 50 17 30 46 48 40 14 17 41 31 21 45 56 46 18 29 15 25 1
38 18 8 27 9 31 50 3 50 27 49 34 39 12 28 40 42 10 59 47 14 39 3 21 57
50 10 38 12 52 7 58 36 3 42 19 48 25 28 37 31 13 47 46 27 7
2 50 3 50 44 60 23 26 60 33 4 60 35 53 23 1 6 31 41 6 22
10 24 41 19 47 55 31 32 46 15 35 37 33 22 39 33 54 21 24 52 46 47 32
14 57 57 13 10 7 33 59 20 24 47 23 45 1 32 2 15 25 42 33 48 44
43 37 49 37 14 59 44 26 18 38 12 15 16 21 56 20 30 23 57 30 48 49 34
56 17 36 54 15 60 17 40 36 45 8 37 2 28 58 8 17 3 30 23 3
60 7 52 49 23 39 15 35 29 10 43 12 55 25 8 18 39 11 29 51 22 37 55
21 3 15 41 18 14 48 51 22 58 60 48 41 13 45 11 24 14 11 42 33 25 51
15 20 20 14 50 47 37 52 5 59 33 17 24 56 24 8 11 39 5 13 37
```

Fig.8 Sequence of Store Id of Each Customer

Apriori algorithm is used to find the association rules, including the relationships between different store IDs and the relationships between store categories. The values of confidence for all strong association rules are imported into Power BI. To show the association clearly, Sankey diagram is loaded into Power BI, as shown in Fig.9 and Fig.10.

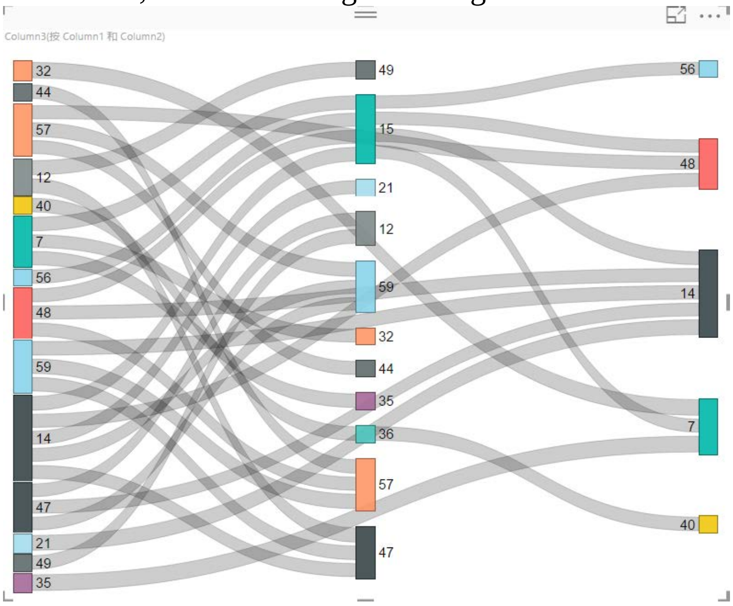


Fig.9 Association between Different Store Id

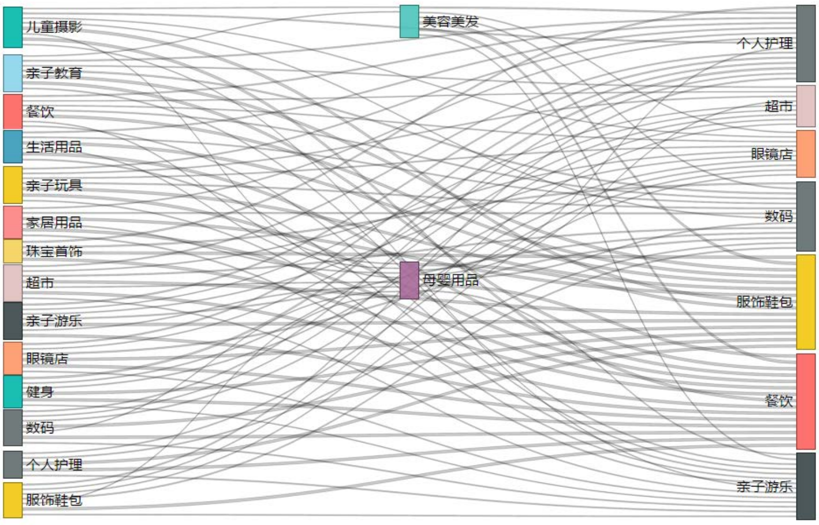


Fig.10 Association between Different Store Category

The Sankey diagram makes it easier to tell which stores are the most popular, which stores are most connected during shopping activities, and the relationships between the stores in different categories. For example, in Fig.9, we can make the conclusion that store 14 and store 15 are highly associated (the height of store 14 and store 15 in Fig.9 is large). Thus, the customer who purchased in any store will make a transaction in store 14 or store 15 with a high possibility. Similarly, in Fig.10, we can arrive at the conclusion that the categories, clothes (yellow bar) and dining (red bar), are two main fields in this shopping mall.

6. Discussion

By applying unsupervised machine learning algorithms to the transaction data, we can find metrics crucial for business operations. We further divide the stores into different groups by KPI metrics (given a set of data points, each having a set of attributes/metrics) and obtain association rules between different stores and store categories. These can be used to determine business strategies for the shopping center and increase the performance and efficiency.

Power BI, as a report tool, makes it much easier to find rules among the different cluster groups, as well as within each individual group. The graphs from Power BI depict the association between various stores labeled with Store IDs and association between store categories. From these models, we can better decide which stores to put near each other based on the associations between different store categories.

Utilizing real time data to help create more accurate and more beneficial business models is a potential goal for the future. In addition, finding a wider variety of metrics could generate more specific results and business models. Perhaps, by finding associations between items in a store, we will be able to offer business advice to individual stores, such as when to do promotions and how to partner with other stores to reach optimal performance.

References

[1] https://www.nydailynews.com/life-style/average-women-spend-399-hours-shopping-year-survey-finds-article-1.116819

[2] https://www.icsc.com/news-and-views/research/shopping-center-definitions.

[3] https://www.techopedia.com/definition/190/artificial-intelligence-ai.

[4] https://machinelearningmastery.com/supervised-and-unsupervised-machine-learning-algorithms/.

[5] T. Bilen, M. E. Ozcevik , Y. Yaslan, and S. F. Oktug.(2018).A Smart City Application: Business Location Estimator using Machine Learning Techniques.20th International Conference on High Performance Computing and Communications, Health,no.12, pp.1314-1321.

[6] M. Prause and J. Weigand.(2018).Market Model Benchmark Suite for Machine Learning Techniques.Computational Intelligence Magazine, Health,no.13(4),pp.14-24.

[7] N. V. Razmochaeva1, D. M. Klionskiy, and V. V.(2018). Chernokulsky.The Investigation of Machine Learning Methods in the Problem of Automation of the Sales Management Business.International Conference on Quality Management, Transport and Information Security, Information Technologies (IT&QM&IS), Health,no.6,PP.376-381.

[8] P. Dhandayudam and I. Krishnamurthi.(2012).An Improved Clustering Algorithm for Customer Segmentation.International Journal of Engineering Science and Technology (IJEST), Health,no. 4(2), pp.695-702.

[9] https://www.maptive.com/17-impressive-data-visualization-examples-need-see/.

[10] https://blog.hubspot.com/marketing/great-data-visualization-examples.

[11] https://www.tapclicks.com/innovative-data-visualization-examples/.

[12] https://www.upwork.com/hiring/for-clients/data-visualization-can-help-make-better-decisions-business/.

[13] http://nifi.apache.org/.

[14] https://en.wikipedia.org/wiki/Power_BI.