E-Commerce Customers Segmentation using RFM Framework

Akhil Chitreddy, Yashwanth Reddy Pothireddy, Aishwarya Kurnutala,

chitreddy.a@northeastern.edu, pothireddy.y@northeastern.edu, kurnutala.a@northeastern.edu

Abstract

Improving sales through marketing is key for every e-commerce organization. It would not be ideal if the organization uses the same marketing strategy for all the customers. To understand customers better, we use clustering as a way to classify customer sand provide tailored marketing. Our approach was to cluster customers using the recency, frequency and monetary(RFM) features which can be done by using RFM analysis. RFM analysis aims to find customer's features which fall under the recency, frequency and monetary parameters. We then use the features to get a rfm score which helps us to classify customers into 11 segments. We use the RFM features along with other features to perform unsupervised machine learning to form clusters. We chose clustering techniques such as K-Means, K-prototype, agglomerative clustering and DBSCAN to draw a comparison to pick the best algorithm. Agglomerative clustering gave the best performance among all with a silhoutte coefficient of 0.789 and calinski harabasz index of 34210.362. The number of clusters generated were four which were later identified using the segments obtained from RFM analysis. Using the four clusters, we could successfully identify similar type of customers through which we could launch a communication campaign for every cluster of customers.

The code for this project can be found at: Github Link

Introduction

With the arrival of the digital age, businesses are setting up their platforms on the internet. The business models set up are not only fast and efficient but are also easily accessible to the customers. As the competition grows, every business focuses on analysing customer behaviour and provide them with personalized results to keep them engaged with the product. As the database gets updated with new customers, the diverse characteristics among them become visible. These characteristics can be utilized to segment customers into various segments which can be helpful for launching tailored marketing programs. Customer segmentation using traditional techniques could be relatively simple and may not provide the desired results. Therefore it becomes crucial to pick the appropriate clustering algorithm to efficiently cluster customers. By this clustering technique, sales can be promoted using tailored marketing programs.

One of the most popular method used to segment customers is RFM (Recency Frequency Monetary) model. Customer features such as psychographic, behavioural, geographic and demographic are the essential aspect for any customer segmentation algorithm. We can segment customer's behaviour using RFM analysis on the transaction data of the customer. Three variables which will be utilized are Recency(recent transactions by the customer), frequency(number of purchases by the customers) and monetary(the purchase value of the customer in a given time period). The customers who are highly valued are the ones who have lowest recency and highest monetary and frequency. The features obtained from RFM analysis can then be used by the unsupervised clustering algorithms.

Analyzing customer's behaviour using clustering techniques is frequently used to identify the key characteristics. Clustering algorithms such as K-Means can be used as it performs optimization on the objective function. Agglomerative clustering, k-prototype and DBSCAN are other algorithms which perform clustering. Our approach was to start with the baseline clustering models and try various other algorithms which use different approaches for clustering. We chose baseline models such as agglomerative clustering, K-Means. We also used algorithms such as K-prototype which is an improvement of the K-Means where it can handle mixed data and DBSCAN which uses the density of the data points to form clusters. Among all the algorithms, agglomerative clustering was chosen as it gave the best performance metrics. Using the clusters generated from agglomerative clustering, we could map the RFM segments from the RFM analysis to the customer IDs. With the mappings, we can notice similar segments fall into the same cluster. Therefore we can take the appropriate action for promoting the sales for the group of customers who fall into the same cluster. For example, in a cluster 1, we can observe customers who are vulnerable to lose interest on the product and in cluster 2, we can observe customers who are loyal to the product. Therefore it becomes easy to launch a communication campaign for a set of customers in every cluster.

Background

It has become essential for organizations to differentiate customers based on their various factors as the demand for online services increased. The segmentation allows to attract more number of customers to engage with the product and makes them remain loyal towards the product. Every customer expects personalised results when they engage with a product which can be effectively implemented using the segmentation of the customers. Businesses frequently divide their clientele into several groups based on statistical information about their gender, revenue, region, age, life stage, sessions etc. Better segmentation are produced by combining data from several sources. Your chances of turning sales into a contract will significantly increase if you can send the appropriate message at the right time to the right people. Advertisers might benefit from receiving a clear methodology while preparing for the future thanks to client division. As a result, the need for consumer segmentation utilizing diverse techniques in many fields is increasing rapidly.

The framework which is leveraged in our project is the RFM framework which can successfully manage customer relationships. One of the key tools for evaluating customer value and profitability is this model. The RFM model, which was first put forth by Arthur Hughes, incorporates three crucial factors for assessing the worth of an e-commerce website to customers: R recent consumption, F frequency of consumption, and M consumption volume . Customer value and these three factors are closely related. Three variables are computed during the time frame given in this article. There must be a one-to-one match between three variables in order to determine the value of the customer's life.

Related Work

Hughes introduced the idea of RFM for the first time in 1994[5]. Several works utilised RFM after it was defined. Customers with higher R, F, and M scores are more likely to make a new purchase, according to a preliminary study [7]. As a result, numerous efforts have been made to divide up the consumer base based on RFM values. A. Dursun and M. Caber suggested a different model to cluster hotel guests. This methodology is used to identify loyal consumers, lost customers, new customers, promising customers, and highly potential customers [3]. The linear programming (LP) method described in [2] combines information from RFM analysis with budgeting information for a particular campaign. The approach can assist direct marketers in deciding whether to maintain or sever ties to a specific RFM consumer segment. Customer behavior analysis is integrated with the idea of time series clustering and forecasting in [1]. Customer's dynamic behavioral patterns are also recorded through modeling consumer behavior using the time series approach. [4] demonstrates the idea of employing density-based algorithms in addition to centroid-based algorithms like k-means for consumer segmentation. [6] suggests a technique for segmenting customers based on the adaptive particle swarm optimization (PSO) algorithm and the modified K-means algorithm. The adaptive learning PSO (ALPSO) technique is suggested to increase optimization accuracy because the present PSO approach is prone to falling into a local extremum.

Project description

The first step in our project is to perform RFM analysis. The RFM model is an analysis tool which help organizations to identify their valuable customers. This model work on the purchasing history of the customer. RFM is an abbreviation that stands for Recency, Frequency, and Monetary. In our analysis, We discovered each customer's Recency, Frequency, and Monetary values in order to better understand them by determining when his/her most recent purchase was, how many times he/she has purchased, and how much he/she has spent with the company. The three important factors are:

Recency value: This represents the amount of time that has passed since a customer's last purchase, including a visit to the website using a mobile app or web browser. Recency is an important indicator of whether a customer is likely to respond to a new marketing offer.

Frequency value: In this case, the number of times a customer made a purchase or otherwise engaged with your brand within a certain time frame is represented by Frequency. Frequency is an effective metric since it shows how deeply loyal the customers are to the brand. Greater frequency indicates a higher degree of customer loyalty.

Monetary value: This shows the amount of money a customer has spent on the brand's products and services over a specific time period. Customers who have spent the most in the past are more likely to spend more in the future, so monetary value is an important statistic.

The quintile technique is used in the RFM scoring procedure to quantify customer behavior. The first quintile with the highest values (the least for recency) is denoted by the number 5. 4 represents the following quintile, and so on. Finally, 555, 554, 553,..., 111, present all of the customers. The best and poorest customer groups are 555 and 111, respectively.

Using RFM analysis data was split into 11 segments as shown in the below figure:

Segment	Activity
Champions	Bought recently, order often and spend the most.
Loyal	Orders regularly. Responsive to promotions.
Potential Loyalists	Recent customers who spent good amounts.
New Customers	Bought most recently.
Promising	Potential loyalist a few months ago. Spends frequently and a good amount. But the last purchase was several weeks ago.
Need attention	Core customers whose last purchase happened more than one month ago.
About to sleep	Made their last purchase a long time ago but in the last 4 weeks either visited the site or opened an email.
Cannot Lose Them	Made the largest orders, and often. But haven't returned for a long time.
At Risk	Similar to 'Cannot Lose Them' but with smaller monetary and frequency value.
Hibernating customers	Customers who made smaller and infrequent purchases before but haven't purchased anything in a long time.
Lost	Made last purchase long time ago and didn't engage at all in the last 4 weeks.

Figure 1: RFM Segments with Activity

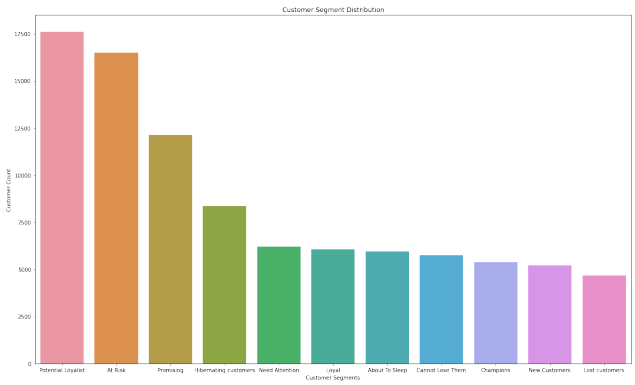

Figure 2: Customer Segments Distribution

Exploratory Data Analysis(EDA)

We performed Analysis to visually explore the data in order to gain useful insights. As there are missing values and Outliers in the data we performed Data cleaning and Outlier Analysis. First we performed Geospatial analysis of the data to gain insights about Brazilian zip codes and its latitude/longitude coordinates. We used the Brazilian zip codes which consists of 5 digits to plot the maps.

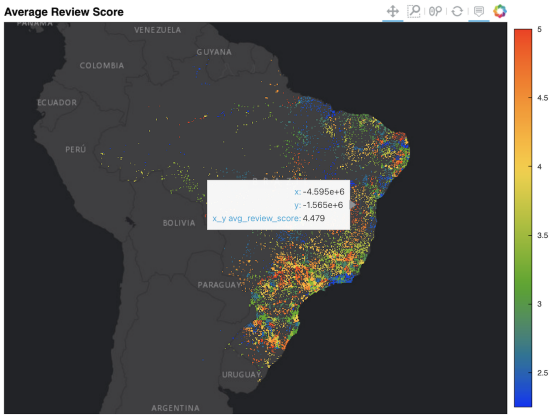

Figure 3: Average review Score

We use the first three digits of the Geolocation zip codes because dealing with five digits may result in relatively few samples. The coordinates are plotted on a map here. Customers appear to be concentrated primarily on Brazil's east coast, notably in the south. The Average Review score is represented in this plot by taking the mean of all review scores and grouping them by the customer's zip code prefix. Customers in the Northeast Region are more likely to rate purchases negatively, whereas customers in the South East Region are more likely to rate purchases positively. And customers from the south region provided the majority of the reviews, while customers from the northwest region provided very few reviews.

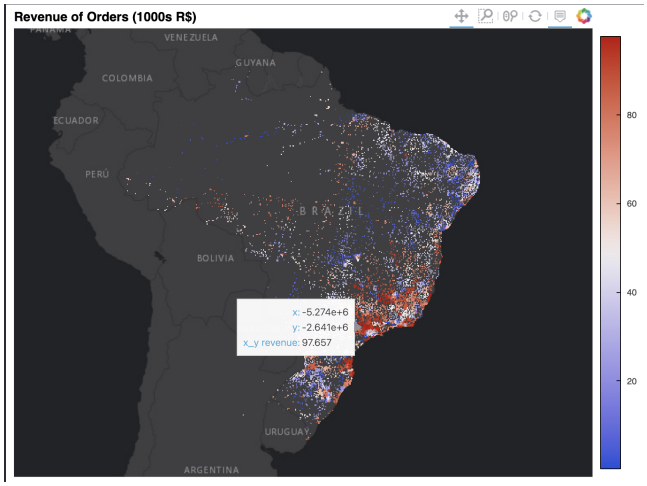

Figure 4: Revenue of Orders(1000s R$)

The revenue is represented in Thousands of Russian dollars for each location. We can observe that the majority of the revenue came from Brazil's South-East areas by plotting the sum of price values grouped by the customer zip code prefix. Furthermore, it is obvious that the cities in the North-West region contribute very little in terms of revenue. The revenue contribution from the South-East region is the highest.

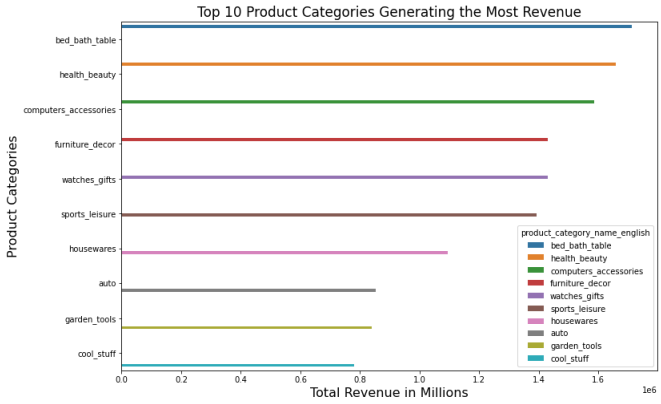

Figure 5: Top 10 product categories generating the Most Revenue

Fig 5. displays a bar plot of the top 10 product categories generating the most revenue. Here the products categories bed bath table, health and beauty and computer accessories are contributing more towards revenue.

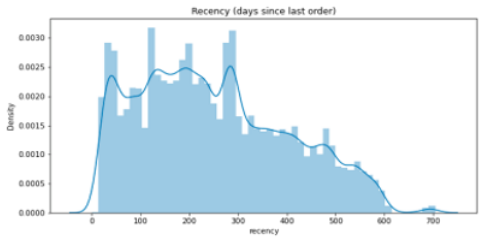

Figure 6: Recency Distribution of Customers

The above figure gives the recency distribution of the customers

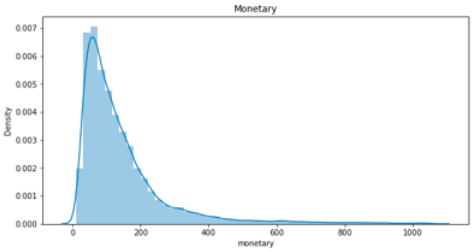

Figure 7: Monetary Distribution of Customers

The above figure gives the monetary distribution of the customers

Algorithms Used

We used the following algorithms which have different approaches to perform clustering

K-Means K-means is an unsupervised machine-learning algorithm that is used for clustering. The goal of k-means is to divide a dataset into a specified number of clusters, with each cluster represented by its centroid. This is achieved by iteratively assigning each data point to a cluster with the nearest centroid and then updating the centroids based on the data points assigned.
One of the key advantages of k-means is that it is computationally efficient, making it well-suited for large datasets. Additionally, because k-means is unsupervised, it can be used when the classes of the data are not known in advance. However, a major limitation of k-means is that it can only find clusters that are convex and of similar size, which may only sometimes be the case in real-world datasets. Additionally, the user should specify the number of clusters, which can sometimes be challenging.
The mathematical formula for the k-means algorithm is as follows:
Given a set of data points

$$\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$$

and a number of clusters k, the goal of the k-means algorithm is to partition the data into k clusters, such that the sum of distances between each data point and its assigned cluster centroid is minimized. This is mathematically expressed as follows:

$$\arg \min_{c_1,\ldots,c_k} \sum_{i=1}^{n} \min_{j=1,\ldots,k} |\mathbf{x}_i - c_j|^2$$

where $c_1, \ldots, c_k$ are the cluster centroids.

$$\arg \min_{c_1,\ldots,c_k} \sum_{i=1}^{n} \min_{j=1,\ldots,k} |\mathbf{x}_i - c_j|^2$$

Agglomerative Clustering Agglomerative Clustering is a hierarchical clustering algorithm used to cluster data points into clusters based on their similarity. It works by treating each data point as a separate cluster and then iteratively merging the closest pairs of clusters until a desired number of clusters is obtained. The similarity between two clusters can be measured using various metrics such as Euclidean distance, Manhattan distance, or cosine similarity.
One advantage of agglomerative clustering is that it allows the user to specify the desired number of clusters, unlike other clustering algorithms that automatically determine the number of clusters. Additionally, it can handle non-linearly separable data and produce nested clusters, which can be helpful for hierarchical data visualization. However, one disadvantage of agglomerative clustering is that it can be computationally expensive, especially for large datasets.

Given a set of data points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n \in \mathbb{R}^d$, the goal of the agglomerative clustering algorithm is to partition the data into k clusters, such that the sum of distances between each pair of data points in the same cluster is minimized. This is mathematically expressed as follows:

$$\arg \min_{c_1,\ldots,c_k} \sum_{i=1}^{k} \sum_{\mathbf{x}_j, \mathbf{x}_l \in c_i} |\mathbf{x}_j - \mathbf{x}_l|^2$$

where $c_1, \ldots, c_k$ are the clusters.

$$\arg \min_{c_1,\ldots,c_k} \sum_{i=1}^{k} \sum_{\mathbf{x}_j, \mathbf{x}_l \in c_i} |\mathbf{x}_j - \mathbf{x}_l|^2$$

K-Prototype k-prototypes is a clustering approach for mixed-type data that combines principles from the k-means and k-modes algorithms. The Brazilian E-Commerce dataset includes variables that are both numerical and categorical. The k-prototype algorithm is one of the finest strategies for clustering this type of data. The goal of the k-prototype method is to divide the dataset into k clusters while reducing the cost function. The k-prototypes algorithm combines the numerical "means" and categorical "modes" to create a hybrid Cluster Center prototype.
The Cost Function is mathematically represented by the formula:

$$F(U, Q) = \sum_{l=1}^{k} \sum_{i=1}^{n} U_{il} d(x_i, q_l)$$

On the basis of "prototype," it creates a Dissimilarity Coefficient formula and a Cost Function suitable for mixed-type data. The γ parameter is introduced to regulate the influence of the Categorical Feature and the Numerical Feature on the clustering process. The mixed-type dataset is expected to contain p numerical and$m$-p categorical features. The Dissimilarity Coefficient of k-prototypes is is theoretically stated in the formula for any $x_i, q_l \, \epsilon \, D$ as follows:

$$d(x_i, q_l) = \gamma \sum_{s=1}^{p} \delta(x_{i,s}^C - q_{l,s}^C) + \sum_{s=p+1}^{m} \sqrt{(x_{i,s}^C - q_{l,s}^C)^2},$$

where $\delta(x_{i,s}, q_{l,s}) \doteq \begin{cases} 0, x_{i,s} \doteq q_{l,s} \\ 1, x_{i,s} \neq q_{l,s} \end{cases}$

The Dissimilarity Coefficient is divided into two components for separate calculation. Categorical distances are determined by Hamming distances, and numerical distances are determined by the square of Euclidean distances.

DBSCAN DBSCAN, a density-based technique, is one of the most widely used clustering algorithms. A relevant customer segmentation can be achieved by applying DBSCAN (Density-Based Spatial Clustering of Applications with Noise). In density-based clustering, clusters of varying densities are generated. Data density refers to the number of points inside a certain region surrounding a point P. This region is often an N-sphere with a radius ϵ that indicates the greatest possible distance between any data point and P. Low density data points are typically treated as noise or border points as shown in the figure below.

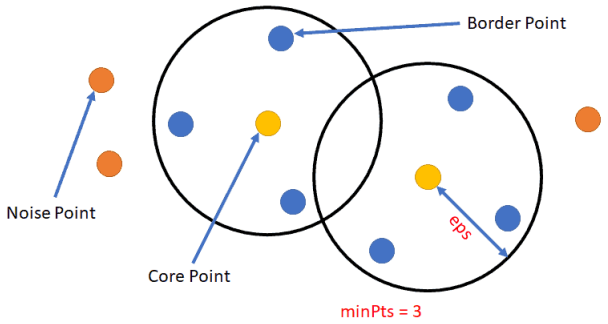

Figure 8: DBSCAN Clustering

DBSCAN finds clusters of arbitrary shape using the neighborhood principle, which states that points within ϵ radius are neighbors. Using the parameter *MinPts*, you can control the size of the cluster and how many minimum points are considered.

The DBSCAN algorithm has the advantage of not requiring the user to specify the total number of clusters to be formed. In order to decide the number of meaningful clusters, it uses two parameters: radius of neighborhoods ϵ, which is the radius within which points are considered neighbors to each other and *MinPts*, denotes the minimum number of points necessary to be considered a cluster. Compared to other centroid-based clustering techniques, DBSCAN outperforms them on noisy datasets. However, This method suffers from a significant disadvantage in that it depends on a density drop to detect cluster borders.

Empirical Results

As we are using unsupervised algorithms, we evaluate the performance using Silhouette coefficient and Calinski-Harabasz Index.
The Table 1 shows the results of applying different clustering algorithms to a dataset. The Agglomerative Clustering algorithm had the best performance, with a Silhouette coefficient of 0.807 and a Calinski-Harabasz Index of 37848.147. The K-Means algorithm had the second-best performance, with a Silhouette coefficient of 0.727 and a Calinski-Harabasz Index of 27398.637. The K-Prototype and DBSCAN algorithms had lower performance, with Silhouette coefficients of -0.036 and -0.817, respectively, and Calinski-Harabasz Indexes of 214.076 and 12.036, respectively.,

	Silhoutte coefficient	Calinski-Harabasz Index
Agglomerative Clustering	0.807	37848.147
K – Means	0.727	27398.637
K - Prototype	-0.036	214.076
DBSCAN	-0.817	12.036

Table 1: Performance Metrics for the Algorithms

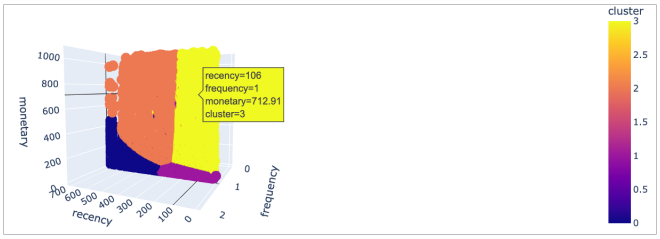

Figure 9: Cluster Visualization in 3D

From Figure 9 we can see that the clusters are well separated in 3 dimensions along the axes recency, monetary and frequency.
DBSCAN fails to perform segmentation as the customer's data density does not help. And, K-prototye works well if there are more categorical variables in the selected features that are being used to train the model and performs poorly in our case.

Conclusion

The clusters can be identified into any of the 11 segments using the rfm score. Once the segment is identified using the rfm score of the customers, a promotional campaign could be carried out for those customers.
For example, if there are At Risk customers in a cluster like the first cluster from figure 10, then we can send them personalised emails and provide support on the site. And, if there are many promising customers in a segment, their interaction can be improved by sending them discounts on products, etc., to retain them. Similarly, if the cluster has a majority of Potential Loyalists in a segment, send them membership programs to keep them engaged.
The granularity of clusters can be further broken down into states and cities and send curated promotions or emails to the identified segments of customers.

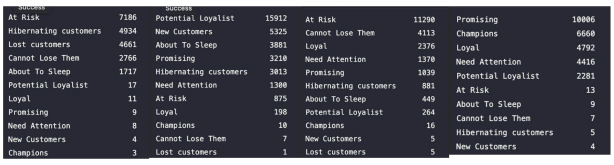

Figure 10: RFM segment counts in the clusters

Future Work

In the dataset, there are feedback comments on orders from customers. However, these comments are in Portuguese, In the future we plan to translate these to English and we can use text embedding techniques live GloVe etc., to assign a positive or negative sentiment on the comment. There are different Neural Networks as well like LSTM, which perform better on text data. We plan to implement these and can use this cluster positive and negative comments for further analysis.

References

[1] Abbasimehr, H.; and Shabani, M. 2021. A new framework for predicting customer behavior in terms of RFM by considering the temporal aspect based on time series techniques. *Journal of ambient intelligence and humanized computing*, 12(1): 515–531.

[2] Asllani, A.; and Halstead, D. 2011. Using RFM data to optimize direct marketing campaigns: A linear programming approach. *Academy of Marketing Studies Journal*, 15: 59.

[3] Dursun, A.; and Caber, M. 2016. Using data mining techniques for profiling profitable hotel customers: An application of RFM analysis. *Tourism management perspectives*, 18: 153–160.

[4] Hossain, A. S. 2017. Customer segmentation using centroid based and density based clustering algorithms. In *2017 3rd International Conference on Electrical Information and Communication Technology (EICT)*, 1–6.

[5] Hughes, A. M. 2005. *Strategic database marketing*. McGraw-Hill Pub. Co.

[6] Li, Y.; Chu, X.; Tian, D.; Feng, J.; and Mu, W. 2021. Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing*, 113: 107924.

[7] Zhang, Q.; Yamashita, H.; Mikawa, K.; and Goto, M. 2020. Analysis of purchase history data based on a new latent class model for RFM analysis. *Industrial Engineering & Management Systems*, 19(2): 476–483.