Proceedings of the 4th South American International Industrial Engineering and Operations Management Conference, Lima, Peru, May 9-11, 2023

Using Predictive Analytics for Mini-Grocery Store Characterization in Barranquilla (Colombia)

Yesenia Cruz Cantillo, PhD, MSIE, MECE
Industrial Engineering Department
Universidad Ana G. Méndez-Gurabo Campus
Gurabo, Puerto Rico
cruzy2@uagm.edu

Cristian Solano Payares, PhD Student
Industrial Engineering Department
Universidad del Atlántico, Colombia
cristianjsolano@mail.uniatlantico.edu.co

Carlos González Oquendo, MEEE
Engineering Technology and Aviation Department
Universidad Ana G. Méndez-Carolina Campus
Carolina, Puerto Rico
gonzalezc4@uagm.edu

Abstract

Mini-Grocery Stores are retail formats located closest to consumers' homes, where consumers' day-to-day needs are quickly satisfied. During the last 3 years, discount stores in Barranquilla (Colombia) have increased their market share, satisfying their requirements in similar conditions to mini-grocery stores. Nowadays, these stores' market share has fallen to 48%, which is still high but is further from its historical 50%. Previous studies have shown the store as a social phenomenon from the marketing point of view. However, before the incursion of hard discount stores, it is important to establish characteristics and understand their functional structures. The goal of this research is to build a decision model that allows us to identify the most representative channel variables and in turn to classify them according to those variables' performance. A survey was applied to a sample size of 341 stores. Predictive analytics will be used to extract information, and data mining techniques will be applied to detect unexpected patterns. This study will be able to define a logistical and operational intervention model that allows it to increase the store's performance and establish elements that lead to an increase in its competitiveness in the retail sector. This research is important for mini-grocery stores, as to face competition, they must be prepared to improve their competitive performance in several areas. Aligned to their organizational capabilities, which are defined according to the development context of each store based on a set of variables that must be identified and analyzed.

Keywords

Data Mining, Logistics, Predictive Analytics, Retailer and Supply Chain

1. Introduction

There are many definitions of the supply chain, including different approaches. However, for the development of this research, the definition employed is from the American Production and Inventory Control Society (APICS). APICS defined a supply chain as a set of processes that involve suppliers and their customers and connect companies from the raw material initial source to the finished product point of consumption (Pires 2012). In this sense, if you want to implement a management system through the supply chain, a strategic orientation towards corporate efforts is considered essential. To synchronize intra-company and inter-company operational capacities in a unified whole (Mentzer et al. 2001). Confirming that not all companies being part of a supply chain have the same management and

operation conditions. In Colombia, for mass consumption companies, a large part of the responsibility for product commercialization falls on the traditional channel, which integrates Mini-Grocery stores and Superettes or Mini Markets (Londoño et al. 2014).

K, According to The National Trade Federation of Colombia (FENALCO by its acronym in Spanish) (FENALCO 2021), stores capture more than 48% of the entire family basket market in large cities, while in small towns their participation increases to 62%. Despite this representativeness, the deficiency in some of its characteristics affects the supply chain's global performance (Londoño and Navas 2014). Including variables such as size, technology, and organization. These aspects show some opportunities for improvement that allow raising the channel competitiveness level.

Barranquilla is the capital of the Atlántico Department, located in Colombia. In Barranquilla, the stores have the same behavior. In a characterization study, responses to previously analyzed variables were established. However, it is important to characterize the stores to identify sources of work. Allowing balance in their strategic and operational performance within the channel and integrating their performance with other links in the chain. For this purpose, data mining and predictive analysis were used through classification trees that allowed the stores to be categorized into three large groups, according to common characteristics. Letting the work plans design focus on these groups' characteristics and improve the integral performance of the entire supply chain.

2. Problem Description

There are around 500,000 stores in Colombia, of which it is considered that 60% of this figure is legally constituted. The format is part of the national tradition, characterized by closeness, immediacy, and trust (PORTAFOLIO 2021). Additionally, it is considered that 65% of Colombians purchase basic products from the family basket in this channel. Therefore, it is considered that this format can represent up to 70% of sales for some mass-consumption companies. On the other hand, a store can have three (3) different service formats: counter sales, which represents 58.3%; self-service with 27.74% or a combination of the two previous ones with 14.02%. Representing different characteristics and requirements between them (Londoño and Navas 2014). Barranquilla has 58 stores per km^2, which places them among the top micro-businesses that boost commerce in the Atlántico capital (Grupo BIT Business Analytics 2020). However, in 2009, hard discount stores entered the aggressive market, becoming a competitor to the stores. They present their own characteristics such as large deployment within a specific geography, advanced logistics, marketing knowledge; as well as technological tools (Universidad EAFIT 2022).

An average consumer visits a store every two (2) days to stock up, while in hard discount stores on average they visit every two or three weeks. In this sense, it is important to consider what a product not available on that channel represents. Because it may affect the Mini-Grocery store or the product manufacturer. In this scenario, the power of purchasing decision is transferred to the consumer, who could base his conclusion based on his brand recognition. Therefore, product availability in Mini-Grocery stores is essential since they generate a purchase decision for the final consumer (Universidad EAFIT 2022).

GS1 Colombia is a member organization of the global GS1 network, it provides connectivity solutions to companies from different sectors of the country, making their product information visible through identification and communication standards. A study made by them indicates that when it comes to personal hygiene products, 27% of consumers prefer to change their place of purchase before changing product brands. Meanwhile, when it is food or a household cleaning product, only 16% make that decision. However, as was evidenced in the study, it is in the last supply chain link that the power of decision is given to the final consumer over the out-of-stock products. In this sense, since the store is part of the traditional channel, it has great challenges to catch up with its direct competitors and reach the effectiveness levels that they currently have (PORTAFOLIO 2021).

Mass consumption stores and companies must advance in the challenges that the channel demands of them, such as: optimizing the processes of the traditional channel through information and offering new options that satisfy the needs of the consumer, which are more demanding. In turn, it is necessary to evaluate the consumer based on evaluating their needs, their levels of loyalty, and satisfaction. Either due to product quality or service quality received by the store. In the same way, it is necessary to strengthen the virtual channels and the delivery service. In addition to conducting training processes to improve the store's performance, it is clear that the trade of the shopkeepers is based

on the empirical. Usually, they have little knowledge and experience in topics such as mass consumption, sales, accounting, distribution, and logistics marketing, among others (Grupo BIT Business Analytics 2020).

3. Related Research

3.1 Data Mining and Predictive Analytics in Retail

Data mining and classification trees have been used to identify the reorganization of the physical distribution process for brick-and-mortar retailers to integrate the online channel into their business model (Ishfaq et al. 2016).

Classification trees among other data mining techniques have been used to find brand exhibit classifications, brand sales relationships, and customer profiles. This knowledge was employed to suggest options for store layout and bundling management development and sales business process (Liao and Tasi 2019).

In other research, an author explains the importance of using decision trees rather than linear regression or logistic regression. Using Walmart store sales data, he grew classification trees to model a business outcome (Wright, n.d.).

3.2 Classification Trees

Decision tree building is one of the classification methods for data mining (Larose & Larose, 2015). Building decision trees in SPSS Software includes four statistical algorithms: Quick, Unbiased, Efficient Statistical Tree (QUEST), Classification and Regression Trees (CRT), Chi-squared Automatic Interaction Detection (CHAID), and Exhaustive CHAID (IBM, n.d.).

QUEST prevents predictors from favoring multiple categories by other methods. CRT separates data into homogeneous sectors considering the dependent variable. CHAID is based on χ^2 association test (Michael & Gordon, 2004), and at each step chooses the predictor variable that strongly interrelates with the dependent variable. Exhaustive CHAID is a CHAID adjustment examining each predictor's probable splits. (IBM, n.d.).

4. Methodology

4.1 Mini-Grocery Store Survey and Data Collection

For the methodological development of this research, a secondary information source is used. It contains information on the number of stores located in Barranquilla City, according to the data provided by the Chamber of Commerce (Urruchurto and Fuentes 2018). This document establishes a population of approximately 3,003 legally constituted stores. To know the sample size, Equation 1 was applied:

$$n = \frac{N * Z^2 * p * (1-p)}{(N-1) * e^2 + Z^2 * p * (1-p)} \quad (1)$$

where, n=size of the sample to be surveyed; N=size of the target population; Z=constant of the normal table corresponding to the desired confidence level; p=study success proportion we expect to find. As a general rule, it is used p=50% if no information is available about the value expected to find; and e= percent of allowed error. Stores in Barranquilla with renewed registration in the Chamber of Commerce from 2015-2016 were chosen as the target population, which corresponds to 3,003 entities. On the other hand, there is no knowledge of respondents' response behavior. Therefore, a 50% probability is assumed that the respondent will answer in a certain way.

The confidence level chosen to apply the survey was 95%, then the error rate corresponds to 5%. According to the estimates of the normal distribution table, the constant value for a confidence level of 95% corresponds to 1.96. Accordingly, the following data are available to calculate the sample size: N=3,003; Z=1.96; p=0.5; e=0.05. From equation (1): $n = 340.99$. Rounded to a sample size of 341 individuals is obtained for a confidence level of 95%.

The survey applied to 341 Barranquilla stores pursued to know aspects about: (1) Store General Information (store size, employees quantity, approximate number of customers, Stock Keeping Units (SKU) handled in each Mini-Grocery store and staff required training); (2) Product Portfolio, pursued to determine the quantity and variety there are in products categories and subcategories in Barranquilla stores (products categories and subcategories with the most variety, brands and SKUs in a store's portfolio); (3) Supply Chain section, its objective was the identification of products that are being delivered in a reliable and timely manner (supply means according to products category and

point of sale control over its inventory); (4) Out-of-stock products, we want to obtain relevant results that allow shopkeepers to make decisions in the presence of out-of-stock products at their points of sale (products category with the most out-of-stocks, negotiation's method with suppliers and in front behavior of out-of-stocks at their points of sale).

4.2 Data Mining and Predictive Analytics

Data set from the Mini-Grocery stores survey was employed for this research. Using data mining techniques allowed the discovery of three (3) store sizes according to their most representative variables such as storage space, SKUs, number of workers, and type of service in the store.

The classification tree method was applied to generate rules for data classification. Dependent and independent variables were carefully chosen to associate their behavior with the definition of three (3) store sizes: small, medium, and large. In model development, some variables were recoded to give it greater robustness following the results that were intended to be obtained.

Classification tree generation helped to predict responses on the categorical dependent variable (Mini-Grocery Store Size). Exhaustive CHAID was the growing method chosen for classification tree building. Exhaustive CHAID chose independent predictors which had the strongest interaction with dependent variables selected from the Mini-Grocery stores survey. If-Then statements were employed to describe the likelihood of the selected variables once the trees were built.

4.3 Finding Relationships Among Variables

Mini-Grocery Store Size was the chosen categorical dependent variable. Exhaustive CHAID analyzed all potential splits on each node, the splitting procedure does not end even if the optimal split is completed (Gunduz and Lutfi 2021).

Steps taken include (1) To create two categories, the pairs with the highest p-values were combined. The process of replicating will continue until two categories appear. (2) The variables with the lowest p-values were employed to split the nodes in the tree. (3) The robustness of the decision tree, the separation, and the least number of parent and child nodes presented (Novita et al 27–29 May 2015).

The risk estimate and its standard error determine the tree's predictive precision (IBM, n.d.). For this research, the dependent variable is categorical. Cases proportions incorrectly classified once adjusted for prior probabilities and misclassification costs are risk estimates (IBM, n.d.).

4. Results

The main predictor of "mini-grocery store size" is the "quantity of people that works in the store" (see Figure 1). The physical space of the stores had an impact on the definition of the store's size. 72.7% of the sampled stores have a physical space of less than 100 m^2, being considered small. Stores that have between 1 and 2 employees had a 79.4% likelihood of establishing themselves as small store. Additionally, those that have less space, between 1 and 2 employees, and less than 100 SKUs have a 91.7% likelihood of being considered a small store.

25.5% of the sampled stores have a physical space between 100 m^2 and 400 m^2, being considered medium-sized. Stores that have between 2 and 3 employees had a 39.2% likelihood of being considered a medium-sized store. On the other hand, a store has a 35.2% likelihood to be considered medium size if it has a space between 100 m^2 and 400 m^2, between 2 and 3 employees, and has a type of counter service or a combination between counter and self-service. In addition, there is a 41% likelihood of being considered a medium-sized store if it has up to 500 SKUs.

1.8% of the sampled stores belong to the large size category. Here it includes those stores with a space greater than 400 m^2. Stores that have more than 4 employees had a 14% likelihood of being considered a large-sized store. The last aspect that is considered for the definition of this type of store, not only includes a space greater than 400m^2, and more than 4 workers but also the type of customer service, which is 100% self-service.

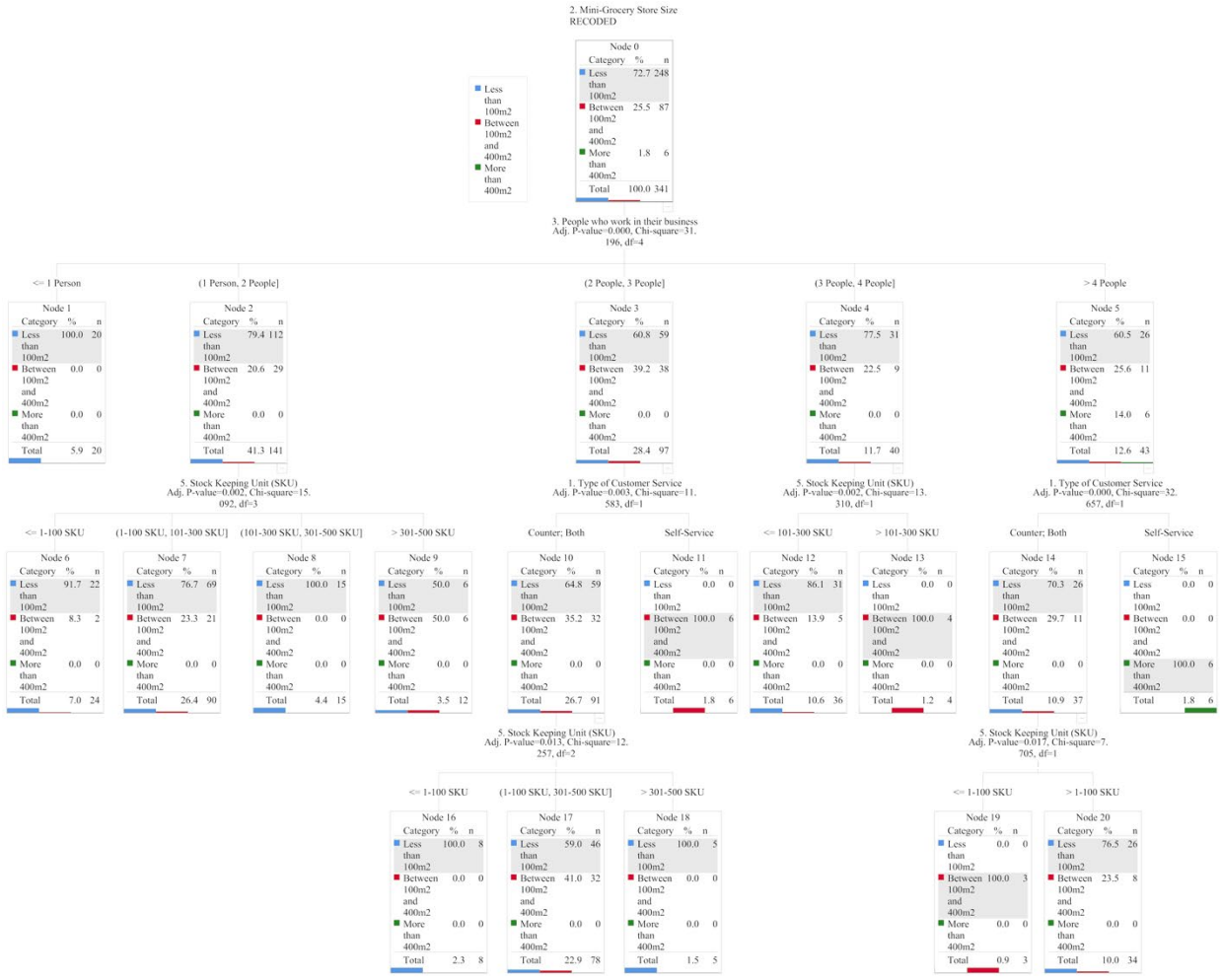

Figure 1. Mini-Grocery Store Size Predictor

Table 1 shows that this model has a risk estimate of 0.217. Suggesting that the category predicted by the model (less than 100 m^2, between 100 m^2 and 400 m^2, or more than 400 m^2) is wrong for 21.7% of the stores. Therefore, the "risk" of misclassifying a store is approximately 22%.

Table 1. Mini-Grocery Store Size Risk Table

Risk

Estimate	Std. Error
.217	.022

Growing Method: EXHAUSTIVE CHAID
Dependent Variable: 2. Mini-Grocery Store Size RECODED

Table 2 displays that the model classifies approximately 78.3% of Mini-Grocery stores correctly. The classification table exposes one possible dilemma with this model: for stores with a space size between 100 m^2 and 400 m^2, it predicts medium size for only 14.9% of them, which means that 85.1% of stores with a space size between 100 m^2 and 400 m^2 are inaccurately classified.

Table 2. Mini-Grocery Store Size Classification Table

Classification

Observed	Predicted: Less than 100m2	Predicted: Between 100m2 and 400m2	Predicted: More than 400m2	Percent Correct
Less than 100m2	248	0	0	**100.0%**
Between 100m2 and 400m2	74	13	0	**14.9%**
More than 400m2	0	0	6	**100.0%**
Overall Percentage	**94.4%**	**3.8%**	**1.8%**	**78.3%**

6. Conclusions

This research allowed us to establish a Mini-Grocery store categorization based on the analysis of four (4) important variables: physical space, number of employees, SKUs, and type of service. From there, three types of stores were identified.

The small store has a space of less than 100 m^2. It has between 1 and 2 employees serving the customers and has a maximum of 100 SKUs on display. The medium store has a space between 100 m^2 and 400 m^2. It has between 2 and 3 employees and uses a service that includes counter and self-service sales. Finally, they have a maximum number of 500 SKUs in their exhibition. The large store has a space greater than 400 m^2, has more than 4 workers, and has self-service sales.

Additionally, it can be defined that Barranquilla stores' profile tends to be small since their physical space (dependent variable) has a percentage of 72.7% in the category of less than 100 m^2.

Acknowledgments

The authors want to thank the students Karilyn Elena Urruchurto González and Laura Fernanda Fuentes Amaya. The students developed a degree project named Methodology Design for Measuring Out of Stock for the Store-to-Store Channel in Barranquilla City. Providing us with the data set used in this research.

References

Pires, S., *Gestión de la cadena de suministros,* 1st Edition, McGraw-Hill, Madrid, 2012.

Mentzer, J., DeWitt, W., Keebler, J., Min, S., Nix, N., Smith, C., and Zacharia, Z., defining supply chain management, *Journal of Business Logistics,* vol. 22, no 2, pp. 1-25, 2001.

Londoño Aldana, E., and Navas Rios , M. *Canal tradicional de productos de gran consumo.* 1st Edition, Editorial universitaria Universidad de Cartagena, Cartagena, 2014.

FENALCO, Available: https://www.fenalco.com.co/es/noticias/2021/12/16/la-tienda-de-barrio-sigue-siendo-la-joya-de-la-corona-para-los-productos-de-consumo-masivo/, Accessed on December 21, 2022.

PORTAFOLIO, Available: https://blogs.portafolio.co/negocios-e-inspiracion/tendero-nuevo-influenciador-del-consumo-masivo/, Accessed on December 21, 2022.

Grupo BIT Business Analytics, *Todo lo que necesita saber sobre el canal tradicional en Colombia y Latinoamérica*, 1st Edition, GRUPO BIT, Bogotá, 2020.

Universidad EAFIT, Available: https://mas.eafit.edu.co/sincategoria/tiendas-de-barrio-vs-hard-discount/, Accessed on December 26, 2022.

Portafolio Finanzas, Available: https://www.portafolio.co/economia/finanzas/productos-agotados-generan-grandes-perdidas-supermercados-tiendas-476344, Accessed on December 21, 2022

Ishfaq, R., Defee, C., and Gibson, B., Realignment of the physical distribution process in omni-channel fulfillment, *International Journal of Physical Distribution & Logistics Management,* vol. 46, no 6/7, pp 543-561, 2016.

Liao, S., and Tasi, Y., Big data analysis on the business process and management for the store layout and bundling sales. *Business Process Management Journal,* vol. 25, no 7, pp 1783-1801, 2019.

Wright, V., Machine Learning: Decision Trees in Retail, Available: https://www.wrightanalytics-mn.com/, Accessed on February 8, 2023.

Larose, D. T., and Larose, C. D, *Decision Trees*. En *Data Mining and Predictive Analytics*, Wiley, Hoboken, 2015.

IBM, Available: https://www.ibm.com/docs/en/spss-statistics/27.0.0?topic=trees-creating-decision, Accessed on January 25, 2023.

Michael, J. A., and Gordon, S. L., Data Mining Technique: For Marketing, Sales and Customer Support, Wiley Publishing, Indianapolis, 2004.

Urruchurto, K., and Fuentes, L., Diseño de una Metodología de Medición de Agotados del Canal Tienda a Tienda de la Ciudad de Barranquilla. *Proceedings of the International Conference on Industrial Engineering and Operations Management,* pp. 861 – 871, Bogota, Colombia, October 25-26, 2017.

Gunduz, M., and Lutfi, H. M., Go/No-Go Decision Model for Owners Using Exhaustive, *Sustainability,* vol. 13, pp. 24, 2021.

Novita, R., Sabariah, M., and Effendy, Identifying factors that influence student failure rate using Exhaustive CHAID (Chi-square automatic interaction detection). *Proceedings of the 2015 3rd International Conference on Information and Communication Technology (ICoICT),* Nusa Dua, Bali, May 27-29, 2015.

Biographies

Yesenia Cruz Cantillo, PhD, MSIE, MECE, is an Industrial Engineer with a Master's Degree in Industrial Engineering focused on Quality Systems and expertise in Quality Control Systems. With 6 years of experience in the development, evaluation, and implementation of quality management systems using ISO 9000 standards and statistical analysis applied to several production and service processes. She also holds a PhD in Civil/Transportation Engineering, with a Master's Degree in Transportation Engineering and strong abilities in simulation. In addition, she has over 17 years of experience in research and teaching and almost 8 years of experience in Transportation Consulting. Strong background in production engineering with experience in the improvement of information systems using informatics and production planning and control.

Cristian Solano Payares, PhD Student, is an Industrial Engineer with 21 years of experience in teaching and research and 2½ years in Logistic Consulting. He is a Business Logistic Specialist and has a Master's Degree in Management Engineering focused on Logistics. He also is a Doctoral student at Universidad Popular Autónoma del Estado de Puebla in the Logistics and Supply Chain Management Program. He is a professor of the Industrial Engineering program at Universidad del Atlántico in logistics. Researcher in Hospital Logistics and process improvement projects for logistics process of health services companies. Member of the research engineering group, Research and Innovation for Development (3i+d).

Carlos J. González-Oquendo, MEEE, EIT #16744, TEL-049780-SE, has a Master's Degree in Electrical Engineering with a specialization in Telecommunications, Signal, and Image Processing. He has a minor in advanced mathematics and possesses experience in Statistics, Stochastic and Random Processes, Estimation Theory, and Math Modelling; also, as in scientific programming and simulation.