Fourth North American International Conference on Industrial Engineering and Operations Management (IEOM)
Toronto, Canada. October 23-25, 2019

Implementation of Analytics Procedures to Predict Stock-Outs in store for a retailer. A case in Mexico.

Cinthya Yaresi Tamez Silva
Business Management Engineering
University of Monterrey
cinthya.tamez@udem.edu

Ana Patricia Sepúlveda González
Business Management Engineering
University of Monterrey
ana.sepulvedag@udem.edu

Martín Flores Maradiaga
Business Management Engineering
University of Monterrey
Martin.flores@udem.edu

Juan Ignacio González Espinosa
Business Management Engineering Department
University of Monterrey
juan.gonzaleze@udem.edu

Abstract

The term Big Data has been used to refer to the extensive data gathering that cannot be managed by traditional methods. This research applies data mining and analytics techniques to give a picture of the interaction of performance between stochastic and deterministic variables and store stock-outs through predictive models. These variables materialize in different types of information, from demographic data like age and customers' perception to operational features like shelf capacity and inventory. While these variables were previously analysed by isolated studies, this pioneering project joins this approach to provide an integral analytical solution.
This research is conducted through the application of logistic regression, and some others such as deviation analysis, clustering and Sigma, for selecting relevant family and sub-family products that were the focus of the models developed. Moreover, this study emphasizes on some recommended and specific actions aimed to reduce the in-store stock outs, based on the insights emerged from the models developed.

Keywords
Data Analytics, Business Intelligence, Logistic Regression, Predictive Analytics, Big Data

Biography
Cinthya Tamez is a Business Management Engineering Student at Universidad de Monterrey; with more than 2 years of work experience, Cinthya has contributed in different consultancy projects related with inventory turnover, strategic planning and predictive analytics. Moreover, Cinthya has earned several certifications in Customer Analytics, Operations Management and Process Improvement from universities such as University of Pennsylvania, University of Illinois, and University of California. Currently she is working at a Mexican e-commerce company as a Category Manager and her interests include sales, marketing, strategic planning, entrepreneurship and business development.

Ana Patricia Sepúlveda is a Business Management Engineering Student at Universidad de Monterrey. With more than 2 years of experience developing consultancy projects, she is currently a Project Manager specialized

in IT and Big Data initiatives for the Sales Direction at Grupo AlEn. Ana has participated in several projects related with lean manufacturing implementation, balancing workloads, standardization of Key Performance Indicators, strategic planning, and predictive analytics, among other initiatives of high relevance to the organization she works for. Thus far, Ana has several certifications such as Strategic Management and Innovation, Strategic Business Strategy, Business Analytics, Data Science at Scale, Design Thinking for Innovation and Foundations of Business Strategy; all of them certificated by global Universities such as ESSEC Business School, Duke University, University of Pennsylvania, University of Virginia, etc.

Martín Flores is a Business Management Engineering Student at Universidad de Monterrey '19, with more than 4 years of work experience in the metal-mechanical industry, public services, steel sector and retail industry. Throughout his experience Martin has gained knowledge of lean manufacturing, quality audits, industrial safety, organizational behaviour processes, commercial planning, business development and recently predictive analytics.

Juan Ignacio González Joined Universidad de Monterrey as a Full Professor in Engineering Management in 2017. He earned a Phd in Business Strategy from EGADE Business School and completed a specialization in quantitative methods and Structural Equation Modeling (SEM) in the Fisher College of Business, at Ohio State University. He earned also a B.S. in marketing management and a MBA from EGADE Business School. As a practitioner, he developed a productive, 15-year career on executive and managerial positions in relevant corporations in Monterrey and Saltillo industrial areas (Sigma Alimentos/Alfa Corporativo, Bokados/Arca Continental, Whirlpool Corporation, Vitromex/Grupo Industrial Saltillo, Axalta/DuPont). Besides, Dr Gonzalez has been consultant and teacher since 2000 at undergraduate and graduate level, focusing on analytics and business strategy, market and business intelligence, entrepreneurship and business development.

1. Introduction

This project is applied to an American retail chain operating internationally. Today, the chain operates in more than 300 stores with presence in United States and the north of Mexico. In this last country, the retail chain has 64 points of sale, from which 47 belong to the format "Hypermarket" and the rest from "Low Income".
The difference between formats is that "Hypermarket" is a grocery store that aims to sell a vast variety of products of the highest quality, giving the customer a full buying experience; on the other hand, "Low Income" is a store that looks forward to offer a low price scheme without compromising quality on the products sold, most of the products exhibited in this format are the high runners of the hypermarket. Low Income is located in towns with 100,000 or less habitants; Hypermarkets are focused on highly inhabited cities like Monterrey, Mexico.

The retail store has many distribution centres along the northern part of the country. Due to confidentiality, the stores in this research will be referred as "Store 1" and "Store 2" according to their formats.
Depending upon their formats, each store can manage up to 20,000 SKUs (stock keeping units) which are held under the retail scheme. Furthermore, they have a wide variety of imported and national products that range from perishables, groceries, services and entertainment.
Occasionally, and due to the big number of products the store manages, this retail chain's selling points and its different formats happen to have stock-outs which later, transform into lost sales.
The retail chain refers to "stock-out" as "any missing product on a shelf". It classifies its stock-outs as *shelf scanned stock-outs* which allude to the ones captured by Shelf Managers, and *in store stock-outs*, which are declared directly from the system.

The main objective of this project is, with more than 50% of effectiveness, predict the probability of stock-outs by a subfamily sample within the stores selected through the construction of predictive models.
This project applies data mining techniques to explore available sources of information from the retail store regarding different independent variables. Variables under the study are from two different nature types: stochastic and deterministic. On the one hand, *stochastic* variables are usually unknown until they are addressed, and their information source is not always available. On the other hand, *deterministic* variables are those that can be controlled, and their values are known at any time; commonly their information sources are available. Through business analytics and predictive models, relevant variables are going to be recognized and analysed to know their impact on the generation of stock-outs.

2. Methodology

To tackle the project, CRISP-DM (Cross Industry Standard Process for Data Mining) was the chosen methodology to address the research due to its versatile stages of study. Since data mining is a creative process

and today there's no standardized method to carry out these kinds of problems, it requires several different skills and knowledge that should be adapted depending on the scope of work. Because of the variability and dynamism of data mining projects, accurate methodologies are searched to provide good enough theoretical frameworks to follow through the research.
The authors Wirth, and Hipp, (2000) carried out different studies based on the CRISP-DM methodology to prove its replicability and functionality. A brief explanation of the different stages of this methodology is featured below.

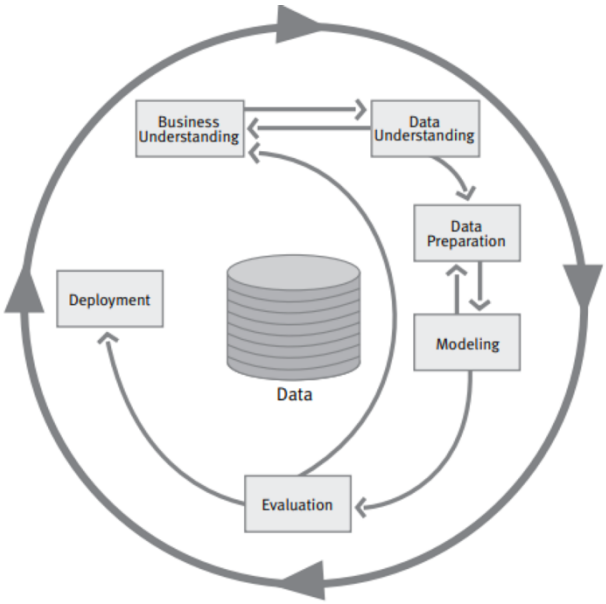

Figure 1. Stages of methodology CRISP-DM

2.1 Business Understanding

This phase focuses in understanding the project's objectives and the requirements according to the business perspective. This knowledge is to be translated into data mining to define the issue. For purposes of this project, information from the stores was obtained to understand every involved process, people, terms used across the industry and established metrics. In this stage objectives and work methods should be set for both, researchers and the client.
Around 7 interviews were made with the client to determine general guidelines and needs. In this stage it was determined that only *in store stock-outs* were to be addressed, and at the same time, the intervention would only take place for product subfamilies of "Fresh" and "Dry" department from the retail's stores.
In the same way, it was defined that the project should be carried out in one store from both formats: "Hypermarket" and "Low income".
Likewise, team sessions were conducted with different areas from the American retail's corporate: Supply, Business Intelligence, Shelf Management and Planograms. Alongside, the research team visited the stores under study with the purpose of knowing the staff and the stock-out scanning process. Corporate staff also attended to these visits.

2.2 Data Understanding

This phase consisted on gathering the first databases used to work, management of general information and understanding data's behaviour. It is important to mention that in this stage no data was despised, but the team comprehend the available sources of information, which were the procedures to obtain them and how to manage data at the same level. To illustrate the process of data collection, below is presented a flow chart showing each step.

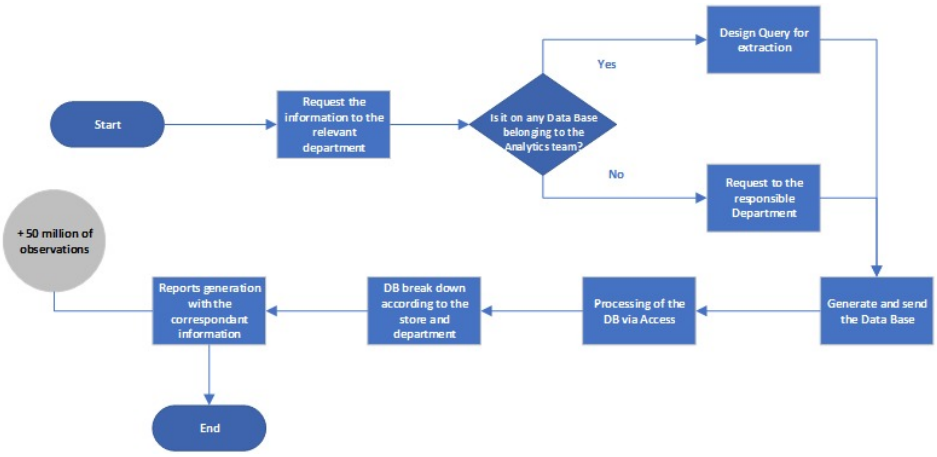

Figure 2. Data collection flowchart

Once the client's needs, scope of the project and actual retail situation were known, the team conducted a very extensive bibliography inquiry. This consisted in an extensive review of several documents and scientific articles related to topics such as inventory management, data mining, root causes of stock-outs and more. As a result, there were more than 30 final selected articles from 20 institutions in 9 countries.

Afterwards, the variables and *drivers* selected to introduce to each predictive model were defined. In the first instance, a list of suggested variables was presented to the client to receive feedback prior to gathering the information. From this list, variables without available or enough historic information were discarded; in the second place, several important variables to the client were added to the list even though they were not found in the literature. The final variable list used in the predictive models is presented in the following table:

Variable	Scale	Source	Type
Product Freshness and Quality	Interval	Best Place to Shop - Retail Chain	Stochastic
Available Checkout lines	Interval	Best Place to Shop - Retail Chain	Stochastic
Help within the buying process	Ratio	Best Place to Shop - Retail Chain	Stochastic
In-store Order and arrangement	Interval	Best Place to Shop - Retail Chain	Stochastic
Packer´s attention	Interval	Best Place to Shop - Retail Chain	Stochastic
Sensory aspects of the store	Interval	Best Place to Shop - Retail Chain	Stochastic
Buyer´s Age	Interval	Best Place to Shop - Retail Chain	Stochastic
Cashier´s attention	Interval	Best Place to Shop - Retail Chain	Stochastic

Paying time	Interval	Best Place to Shop - Retail Chain	Stochastic
The Store´s exterior aspect	Interval	Best Place to Shop - Retail Chain	Stochastic
Gentleness and qualification of the employees	Interval	Best Place to Shop - Retail Chain	Stochastic
Payment speed at checkout line	Interval	Best Place to Shop - Retail Chain	Stochastic
General perception of the store	Interval	Best Place to Shop - Retail Chain	Stochastic
Experience of the buying process	Interval	Best Place to Shop - Retail Chain	Stochastic
Perceived payed value	Interval	Best Place to Shop - Retail Chain	Stochastic
Product availability	Interval	Best Place to Shop - Retail Chain	Stochastic
Perception of promotions	Interval	Best Place to Shop - Retail Chain	Stochastic
Product´s selling price	Ratio	Retail Chain	Stochastic
Product´s selling price in competitor´s store	Ratio	Nielsen	Stochastic
Years since store´s opening	Ratio	Retail Chain	Stochastic
Average temperature	Ratio	The Weather Channel	Stochastic
Rain	Nominal	The Weather Channel	Stochastic
Promotion	Nominal	Retail Chain	Stochastic
Average daily sale of the product	Ratio	Retail Chain	Stochastic
Sales of other categories	Ratio	Retail Chain	Stochastic
Shelf capacity for the product	Ratio	Retail Chain	Deterministic
Shelf facings	Ratio	Retail Chain	Deterministic
Restocking head count	Ratio	Retail Chain	Deterministic
Restocking schedule	Ratio	Retail Chain	Deterministic
Distance to distribution center	Ratio	Retail Chain	Deterministic
Supply frequency	Ratio	Retail Chain	Deterministic

Inventory days	Ratio	Retail Chain	Deterministic
On shelf inventory days	Ratio	Retail Chain	Deterministic

2.3 Data Preparation

Due to the project's reach and scope, the agreement with the customer included the selection of subfamilies of products that represent the highest impact based in 4 criteria: Units sold, sales margin, lost sales, and lost sales margin. Depending on its importance, each of these criteria was weighted to get a final global score that considered the four perspectives daily. The weights defined are 35% for units sold, 30% for sales margin, 25% for lost sales, and 10% for lost sales margin.

Once the results were obtained, the next step consisted on clustering, using a quintiles method developed by the team in charge of the project. In this methodology, the maximum and minimum value was established to define a range to be used as a base to define the 5 clusters that will divide the subfamilies. Once the method was implemented the last 4 quintiles were selected, the selection for every subfamily made almost 80% of the total score. The selection is presented in the following table:

Store 1		*Store 2*	
Dairy	*Drinks and snacks*	*Dairy*	*Drinks and snacks*
Juices	Seeds and snacks	Liquid Cream	Family Snacks
Vegetable Origin	Regular Cola Drinks	Animal Origin	Juices
Speciality	Mainstream	Eggs	Regular Soda Drinks
Eggs	Chips	Margarine	Chips
Greek Yoghurt	Diet Cola Drinks	Drinkable	Regular Cola Drinks
Butter		Infantile	Popcorn
Refrigerated Dough		Solids	Mainstream
Animal Origin			Flavoured Drinks
Spread Cheese			Diet Cola Drinks

The next step consisted in making sure that the sisgnificance of the model is good enough to be representative for the sample of SKUs to analyse per model; therefore, the sample of SKUs was chosen based on their behaviour. The first step to analyse the behaviour consisted on plotting the behaviour of the subfamilies selected, aiming to observe a normal probability behaviour. As a next step, those families that did not show a normal behaviour where put through a standardizing process using the normal logarithm.

A program was developed to analyse the database of the subfamilies and select the SKUs that have a similar behaviour and that are meaningful for the sample. The program considers the daily information with the criteria mentioned before, structures the information by subfamily and breaks down as a monthly behaviour for every SKU. The next step gathers the average and the standard deviation of the whole subfamily, with this information the program defines the upper and lower limits for the first sigma, and it marks the SKUs that, on an annual average, meet with the requirements of the limits established by the result of measuring the first sigma of the sample. With, the program had as an output the selection of a total of 33 models for both stores.
Once the sample for every subfamily was identified, the next step involved the gathering of the information of all the variables for every SKU, with the purpose of creating the database for every model. These databases were made using Microsoft Office Excel.

2.4 Modelling

This phase corresponds to modelling data through predictive algorithms to obtain the significant variables for each model. The research team decided to use the statistical software Rapid Miner with the objective of giving the client replicable models in a software for which the Analytics team has knowledge about. Each model introduced in the software is made up of data from each variable according to the selected subfamily. At the same time, several reviews were made in SPSS software to match the results.

Before introducing all the data into the software, a different procedure was made to determine from the initial variable list, all the important ones for every analysis. The purpose of this process consisted in obtaining all the correlation coefficients for each variable of each model and despise all the ones with the strongest correlation. It is important to mention that having variables with strong correlation can skew the model's iterations; this correlation analysis was carried in Excel spreadsheets using the Data Analysis tool.
Once this exercise was performed for every model and the variables were obtained, every model database was introduced into the Rapid Miner software and were submitted to logistic regression analysis. It is crucial to mention that the alpha value was set to 0.1. ($\alpha = 0.1$)

Statistical results from logistic regression analysis are used to decide which variables are *significant* to the model and which are not. For this project, the statistical evidence analysed to take these decisions were: Beta Coefficient, Standard Coefficient, Standard Error, Z-Value, P-Value, Confusion Matrix and ROC Curve (Receiver Operating Characteristic).

As mentioned before, the alpha value was set to 0.1 so every time a model was submitted to the software and the iteration was performed, the first analysis made was to note which variable's P-Value was less than 0.1. P-Value is the first important result in this type of projects. Once the variables above 0.1 were noted, the model was submitted one more time but now despising these variables. This exercise should be performed until every variable fulfil the requirement of alpha.

2.5 Evaluation
This phase corresponds of the evaluation of every model created as mentioned in the last stage. Once this exercise was performed on every one of the 33 models, a thorough review was realized in every run of the software to verify that the established objectives were achieved.
From the initial models, 3 of them were despised: *Fermentados ("*Fermented" Store 1), *Carne Seca* (Dried Meat, Store 1), and *Naranjadas (Orange-flavoured beverages, Store 2).*

3.6 Deployment

This stage gathers the complete information obtained during the analysis including the results and the final deliveries for the client. During this phase several preventive actions where defined for each of the stores analysed, based on the significant variables. Because of the project's scope, it was agreed with the customer that the final models where going to be deployed in a replicable software, in this case it was Rapid Miner. In addition, a dashboard was created as a graphic and simulating tool, where the probabilities of going out of stock could be measured considering the different values that the variables can take. At last, a training manual is also part of the final deliveries, this manual has the objective of becoming the main tool during the training sessions, the tool is designed to explain step by step how to develop a logistic regression model with clear examples, the goal is that the client can consider the training when addressing new projects.

3. Conclusions
The insights developed from the data mining research are results of a subfamily approach to determine the variables and drivers of *in store stock-outs.* It is of great significance to combine variables with different sources and nature in one whole system to develop more accurate solutions. Based in a logistic regression analysis and predictive models, the conclusions and insights can be summarized as follows:

- Demand: With these variables, the customer's perception was analysed regarding different factors such as order and accommodation, employee's kindness and general experience from the store; the team was able to explain how these perspectives have a direct impact in the generation of stock-outs. So far, no study with this approach using the survey's *Best Place to Shop* variables was realized in the store; for this reason, it is recommended to carry out this type of projects in the store using different approaches and involving these variables. The store has currently potential information from their customers that can help to achieve indicators.

- In-store operation: These variables belong to the execution of the operation inside the store, such as the defined spaces for the product exhibition. Compared to other retail stores, the one under study is in charge and in control of the definition of exhibition spaces and the structure of the shelves for most of the departments. As a highlight, the variables of Price, Promotion, Shelf

Capacity, Shelf facings, and On-shelf inventory days have a direct impact in the prediction of an out of stock. The retail chain manages these variables and can adjust them considering the results of the models and the analysis that the company will make in a future, to optimize the chances of ending up with a stock out.

- Supply Chain: Supply Frequency and Inventory Days are variables that are controlled by the whole supply chain system. Incorporating this type of variables is of utmost importance since they are the backbone for the availability of a product in time and form in the store. It is concluded that entailing efforts together with the commercial area is of paramount importance to work at the same level and benefit from the demand impulses, such as the variables previously mentioned.

To conclude the project and as a way of verification, each model with its significant variables was submitted to an effectiveness evaluation tool in the software Rapid Miner. The aim of the tool was to sustain that the logistic regression method is the best way to carry out this kind of projects. It is important so mention that the analysis consisted in evaluating different methods such as Random Forest, Deep Learning, and Naive Bayes, through effectiveness percentages and simulations.

On the whole, the project interprets each of the variables transforming them into preventive actions that impact directly to 60% of the lost sales in the previous selection of sub families; therefore, delivering a functional, replicable and scalable methodology to address out of stock problems in retail stores.

4. References

[1] Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0 Step-by-step data mining guide

[2] Carvalho, J. y Patino, C. (2015). "What does the P Value really mean?" Jornal Brasileiro de Pheumologia. Brasil.

[3] Domínguez, E. y González, R. (2002). "Análisis de las curvas Receiver – Operating Characteristic: Un método útil para evaluar procederes diagnósticos." Revista Cubana de Endocrinología. La Habana.

[4] Escobar, N. (2013). "Análisis de Regresión Logística para Investigación de Mercados." Escuela de Administración y Contaduría Pública No. 18. Universidad Nacional de Colombia. Bogotá.

[5] Minitab (2018). ¿Qué es el Valor Z?". Soporte de Minitab 18.

[6] Narkhede, S. (2018). "Understanding AUC-ROC Curve." Medium. New York.

[7] SAS (2015). "La Minería de Datos de la A la Z: Cómo descubrir conocimientos y crear mejores oportunidades." SAS Institute Inc. México.

[8] Usman, K. (2008). "Determination of Drivers of Stock-Out Performance of Retail Stores Using Data Mining Techniques." Massachusetts Institute of Technology.

[9] Villardón, JLV. (2007). "Introducción al análisis de clusters". Universidad de Salamanca.

[10] Wirth, R. y Hipp, J. (2000). "CRISP-DM Towards a Standard Process Model for Data Minig". Germany.