Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Department of Mathematics and Physics
Master of Data Science

Maximizing the expected revenue: The use of machine learning models for the case of a soccer ball company.

THESIS to obtain the **DEGREE** of
MASTER OF DATA SCIENCE

A thesis presented by:
Fátima A. García Meléndez

Thesis Advisor:
Dra. Rocío Carrasco Navarro

Tlaquepaque, Jalisco, December 2022

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física
Maestría en Ciencia de Datos

Maximización del ingreso esperado: El uso de modelos de inteligencia artificial para el caso de una empresa de balones de futbol en México.

TESIS para obtención de **GRADO** de la
MAESTRÍA EN CIENCIA DE DATOS

Tesis presentada por:
Fátima A. García Meléndez

Asesor de Tesis:
Dra. Rocío Carrasco Navarro

Tlaquepaque, Jalisco, December 2022

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Department of Mathematics and Physics

Master of Data Science Approval Form

Thesis Title: **Maximizing the expected revenue: The use of machine learning models for the case of a soccer ball company.**

Author: **Fátima A. García Meléndez**

Thesis Approved to complete all degree requirements for the Master of Science Degree in Data Science.

Thesis Advisor, **Dra. Rocío Carrasco Navarro**

Thesis Reader, **Dr. Juan Diego Sánchez-Torres**

Thesis Reader, **Dr. Riemann Ruiz Cruz**

Tlaquepaque, Jalisco, December 2022

Instituto Tecnológico y de Estudios Superiores de Occidente

Reconocimiento de validez oficial de estudios de nivel superior según acuerdo secretarial 15018, publicado en el Diario Oficial de la Federación del 29 de noviembre de 1976.

Departamento de Matemáticas y Física

Formulario de Aprobación para Maestría en Ciencia de Datos

Tı́ulo de Tesis: **Maximización del ingreso esperado: El uso de modelos de inteligencia artificial para el caso de una empresa de balones de futbol en México.**

Autor: **Fátima A. García Meléndez**

Tesis aprobada para completar todos los requisitos de grado para la Maestría en Ciencias de Datos.

Asesor de Tesis, **Dra. Rocío Carrasco Navarro**

Lector de Tesis, **Dr. Juan Diego Sánchez-Torres**

Lector de Tesis, **Dr. Riemann Ruiz Cruz**

Tlaquepaque, Jalisco, December 2022

Maximizing the expected revenue: The use of machine learning models for the case of a soccer ball company.

Fátima A. García Meléndez

Abstract

Every company begins with a fundamental questions and re-asks this questions several times throughout the company´s life, what is the product that will create enough value for it´s customers so that enough money can be charged to make a profit and keep on creating more value? So basically, a company´s strategy begins with a great product design and a price tag that customers are willing to pay that will maximize revenue. In past years, pricing has been so unattended because usually the responsibility tends to fall under different areas of the company and due to it´s complexity, task associated with setting prices are often not on top of the incumbency list. Ergo, prices are not varied enough for different product items, market segments and purchased occasion, impacting the demand, sales and perceived value of the product and brand. Therefore, this study aimed to determine how can machine learning models help create value and maximize revenue by determining the best product and price for a soccer ball company in Mexico. As a result, this research was able to determine that there are 3 different customer segments and that each of them values different characteristics of the soccer ball. Also, that a random forest model was the best model to calculate the purchase probabilities compared to a naive bayes model, a general linear model with logit link and a support vector machine model. Given those probabilities, the expected revenue was calculated for all the different product profiles, or combinations of the ball, and concluded that a price discriminated model with 3 balls; 1 targeted for each customer segment, can increase the expected revenue from an approximate of $166 to $1,572 dollars, proving that machine learning models and information-based decision making processes should be a must for every company.

Maximización del ingreso esperado: El uso de modelos de inteligencia artificial para el caso de una empresa de balones de futbol en México.

Fátima A. García Meléndez

Resumen

Cada empresa comienza con una pregunta fundamental y vuelve a hacerse esta pregunta varias veces a lo largo de su vida activa, ¿cuál es el producto que creará suficiente valor para sus clientes de modo que se pueda cobrar suficiente dinero para obtener ganancias y continuar creando más valor? Básicamente, la estrategia de una empresa comienza con un gran diseño de producto y un precio que los clientes están dispuestos a pagar para maximizar los ingresos. En los últimos años, los precios han estado sumamente desatendidos porque por lo general la responsabilidad tiende a recaer en diferentes áreas de la empresa y debido a su complejidad, las tareas asociada con la fijación de precios a menudo no está en la parte superior de la lista de responsabilidades. Debido a esto, los precios no varían lo suficiente para los diferentes productos, segmentos de mercado y ocasiones de compra, lo que afecta la demanda, las ventas y el valor percibido del producto y marca. Por lo tanto, este estudio tuvo como objetivo determinar si los modelos de aprendizaje automático pueden ayudar a crear valor y maximizar los ingresos al determinar el producto y precio adecuado para una empresa de balones de fútbol en México. Como resultado, este trabajo pudo determinar que existen 3 segmentos de clientes y que cada uno de ellos valora diferentes características del balón de fútbol. Además, que un modelo de bosque aleatorio fue el mejor modelo para calcular las probabilidades de compra en comparación con un modelo naive bayes, un modelo lineal general con enlace logit y un modelo de máquina de vector de soporte. Dadas esas probabilidades, se calculó el ingreso esperado para todos los diferentes perfiles de productos, o combinaciones de pelota, y se concluyó que un modelo discriminado por precio con 3 pelotas; cada uno dirigido a cada segmento de clientes, puede aumentar los ingresos esperados de un aproximado de $166 a $1,572 dólares, lo que demuestra que los modelos de aprendizaje automático y los procesos de toma de decisiones basados en información deben ser imprescindibles para todas las empresas.

Contents

Page

List of Figures

List of Tables

Dedicated to my wife and kids who supported and encouraged me every moment along this process.

Dedicado a mi esposa e hijos quienes me apoyaron y motivaron en cada momento de este proceso.

1 Introduction

Contents

1.1 Background and business context

Every company begins with a fundamental question and re-asks this question several times throughout the company´s life, what is the product that will create enough value for it's customers so that enough money can be charged to make a profit and keep on creating more value?

Is it all this simple? Depending on who you ask, a business may have hundreds of different purposes. Some may say that companies exist to provide jobs, others, to produce goods and services and even pay taxes to help support public goods, or just to make a profit. But let´s put it as simply as: businesses exist to provide value. Sounds obvious, but if you take a look at your business, think about all the individual efforts that happen. You need to mix creativity, with allocation of cash, with sales and marketing, with logistics and fulfillment of goods and all the other parts, just to create enough value in the world that another human being is willing to pay x amount of dollars in exchange for all of that, cobbled together. Because no one buys a product because they want to give the company money; customers buy and use products because those products address their needs and therefore are willing to pay for it.

So broadly, a company´s strategy begins with a great product design but still you want to exchange that value for compensation, so you need to put a price tag on that package of value created. And would any company still exist if they sold that product at the cost it took to assemble all that together without making a profit?

So then, the real questions is: How much is your customer willing to pay for everything you already did so *they* don´t have to?

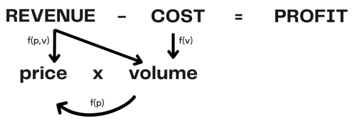


Figure 1.1: Profit formula: Revenue is a function of price and volume; volume is a function of price and the competitors price; variable costs is a function of volume, which as mentioned is a function of price. Hence profit is all about pricing.

Since most companies already have created a product that they believe creates enough value, and as seen in Figure 1.1, pricing is the main lever companies have to create profit, therefore, pricing will be discussed first.

Pricing strategies fall largely into 3 categories: Cost-based pricing, competitors-based pricing and value-based pricing.

Cost-based pricing, the most common strategy used. It involves adding a markup on top of the cost it takes to produce one unit of a product.[1] The resulting number is the selling price of the product. This pricing strategy focuses on internal factors like production cost rather than external factors like consumer demand and competitor prices. It´s commonly used by retail stores to set prices.

Competition-based prices are decided relevant to those of competitors. After researching competitors' pricing, companies determine whether to match or have higher or lower prices then the relevant competitors. Such a method may well apply to medium-share companies competing against high-share competitors (such as local hotels competing with international hotel chains) or for products with low differentiation.

Value-based prices are primarily based on a consumer's perceived value of a product or service. This mainly applies to markets where possessing an item enhances a customer's self-image or facilitates unparalleled life experiences. To that end, this perceived value reflects the worth of an item that consumers are willing to assign to it, and consequently, directly affects the price the consumer ultimately pays.

Pricing has been so unattended on past years, the responsibility usually tends to fall under different areas of the company such as finance, sales or marketing. Although Price is one of the 4P´s of marketing, when marketers talk about their role, tasks associated with setting prices are often not on the top of the list. Consequently, prices are not varied enough for different product items, market segments and/or purchase occasions. And since price is the amount a customer pays for the value received, then adjusting the price has a profound impact on the marketing strategy, and depending on the price elasticity of the product it will affect the demand, sales and the perceived value of the product and consequently the brand.

As mentioned, everything begins with Product design, the process

[1] Unit cost is a total expenditure incurred by a company to produce, store, and sell one unit of a particular product or service. Also known as Cost of goods sold (COGS).

of imagining, creating, and iterating products that address specific needs in a given market. It requires the understanding of the market behaviors, habits, wants and value to create products that are aimed specifically to a certain segment and are adopted flawlessly because their needs and usage are anticipated.

You may have noticed that pricing and product design go hand in hand because ultimately a company has to create a product desired by the customer which in turn will make it easier for the company to price that product at levels higher than the cost and make a profit of it.

This study aims to determine how can machine learning methodologies help create value by determining the best product and improving the pricing strategy decision making process through an example of a data set collected from a choice based survey for soccer balls in Mexico.

1.2 Objective

1.2.1 General

Machine learning models can help marketers establish better pricing and value decisions. This study pretends to understand what are the product characteristics and price that will maximize revenue for a soccer ball company.

1.2.2 Specific goals

With the information on hand, can it be determined who should the company be selling to, what product should they be selling, and at what price should they sell it that will generate a win-win strategy?

To that end, this paper intents to answer the following questions:

1. How many different customer segments can be identified in the soccer ball industry?

2. What are the most relevant characteristics consumers analyze in their purchasing decision?

3. What is the purchase probability of the different types of balls for each segment according to the value obtained?

4. What is the optimal soccer ball for each targeted segment that will maximize the expected return for the company?

To achieve the previous stated goals, first a clustering analysis will be performed comparing K-means and Hierarchical clustering algorithms to segment and group different types of customers; afterwards a decision tree will be used to establish which are the most relevant characteristics that customers take into account when making the purchase decision. Thereafter, since the aim is to forecast the demand probability based on the previous characteristics, 4 different classification algorithms will be examined and the one that performs the best will be used. Finally, with the purchasing probabilities at hand, a recommendation will be made choosing the portfolio, meaning the product with specific characteristics and the price that will maximize expected revenue.

2 Data

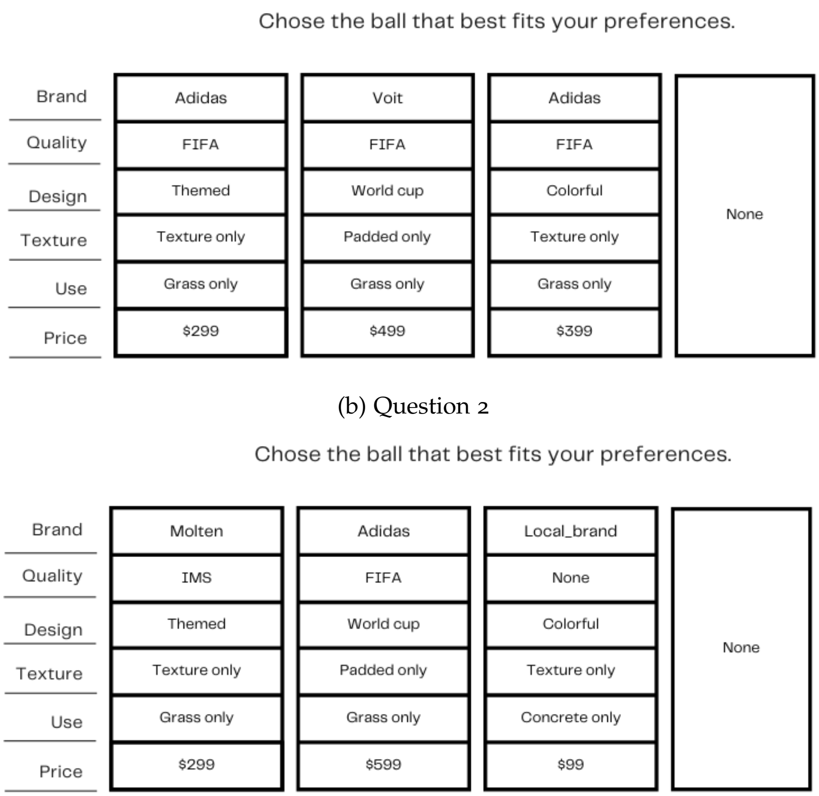
Contents

For this research, a field choice-based experiment was conducted over the web in a computer aided personal interviewing system examining 115 people to study their purchasing behavior. Each questionnaire first asked about demographic characteristics such as: age, gender, location, highest level of education. Then, it inquired more on typical purchase conducts like frequency of purchase, number of balls purchased in the last year, main motive of purchase, and frequent establishment were balls are purchased. And finally, the experimentation part which will be discussed in detail in the next section.

2.1 Experiment

This type of study aims to explore about the respondents' preferences for a combination of features also known as attributes, and each attribute has a specific number of levels that make up the product, which will now be referred to as attribute-level.

The respondent was asked to answer 10 questions, each with 3 sets of concepts or products, plus a none of the above option, and he or she chooses among those concepts. Concepts are changed for each question based on the attributes-levels. See example below:



(a) Question 1

Chose the ball that best fits your preferences.

Brand	Adidas	Voit	Adidas	None
Quality	FIFA	FIFA	FIFA	
Design	Themed	World cup	Colorful	
Texture	Texture only	Padded only	Texture only	
Use	Grass only	Grass only	Grass only	
Price	$299	$499	$399	

(b) Question 2

Chose the ball that best fits your preferences.

Brand	Molten	Adidas	Local_brand	None
Quality	IMS	FIFA	None	
Design	Themed	World cup	Colorful	
Texture	Texture only	Padded only	Texture only	
Use	Grass only	Grass only	Concrete only	
Price	$299	$599	$99	

Figure 2.1: Question examples

2.2 *Pre-processing*

Considering that the respondents had the alternative of choosing *none* of the options, and since all the variables included in the experiment are treated as categorical variables, those "none" records will be removed to avoid bias in the model. Leaving 3,171 records to create and test the models where 33% of the records are True. [1]

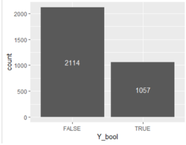


Figure 2.2: True/False count

2.3 *Variables*

Demographics

- Age (5 levels) = Less than 18, between 18 and 30, between 31 and 45, between 46 and 60, older than 60.
- Location = 32 states in Mexico

[1] The ratio between true and false maintains even after the values where removed, causing no impact on the data set.

- Education Level (6 levels) = Elementary, Middle School, High school, Associate´s Degree, Bachelor´s Degree, Professional Degree.
- Gender (3 levels) = Female, Male, Neutral

Recent purchase information

- Number of balls purchased in last year (7 levels) = None, 1, 2, 3, 4, 5, or more than 5
- Place of purchase preference (5 levels) = Sports shop, retail shop, flea Market, online, other.
- Purchase main reason (5 levels) = Gift, fun, training, game, none

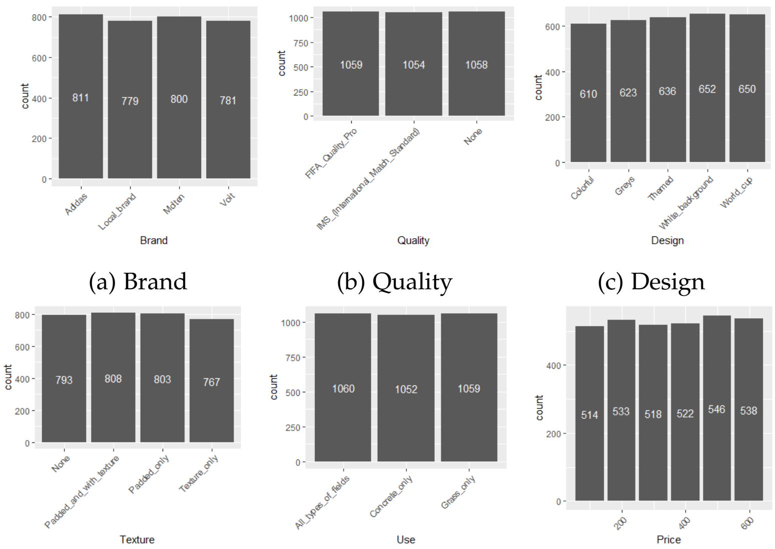
Experimental

- Y (2 levels) = Chosen (1), Not chosen (0)
- Brand (4 levels) = Local Brand, Molten, Voit, Adidas
- Quality (3 levels) = FIFA quality pro, International match, none
- Design (5 levels) = colorful, grey, white background, themed, world cup
- Texture (4 levels) = Padded and with texture, padded only, texture only, none
- Use (3 levels) = All types, grass only, concrete only
- Price (6 levels) = $99, $199, $299, $399, $499, $599

2.4 *Balanced Data*

This is a balanced data set, each attribute-level appears approximately the same number of times across the entire study. For 1 respondent the brand Molten may appear twice while for others it may appear 3 times, but overall brand Molten will have appeared 800 times while Adidas 811, Local brand [2] 779, and Voit 781. And this rule will apply for every attribute. (Figure 2.3)

[2] Due to confidential restrictions, LOCAL brand will maintain it's anonymity across the study.



(a) Brand (b) Quality (c) Design

(d) Texture (e) Use (f) Price

Figure 2.3: Balanced attribute-levels

2.5 *Demand vs price*

Microeconomics theory has studied 2 types of goods (Figure 2.4). Normal goods state that the demand of the product will increase as the price of the product decreases and vice-versa. And, Veblen or luxury goods express that the demand of the product will increase as the price of the product increases and vice-versa.

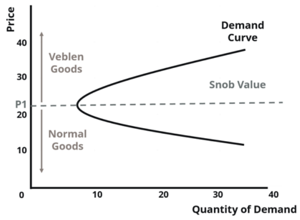


Figure 2.4: Demand curve: Normal goods vs Luxury Goods. This concepts are under the ceteris paribus idea, meaning that if everything else is constant the previous effects should be observed.

In this research, one of the first findings is that the soccer balls do not follow the theory behind microeconomics shown in Figure 2.5, where the expectation is that the higher the price the lower the demand should be. Hence, implying that all the other attributes are relevant for consumers when making a purchasing decision. Therefore, it becomes significant to review the demand for each attribute-level, and determine first hand, which attribute will most likely have the highest impact in the demand due to the variability [3] observed within each attribute.

As seen in Figure 2.6, the brand and the quality have the most variability, implying that those two attributes most likely will be the most important for consumers in their purchase decision.

[3] Measure of variability used is Range = Highest value minus lower value

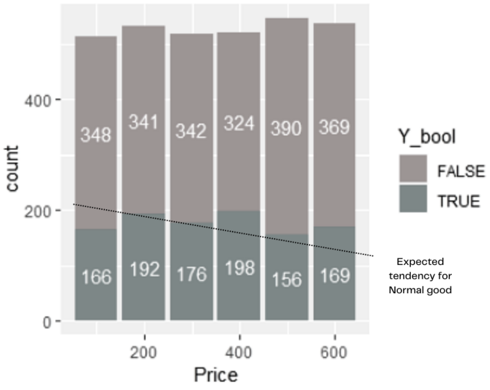


Figure 2.5: Quantity demanded by price vs expected

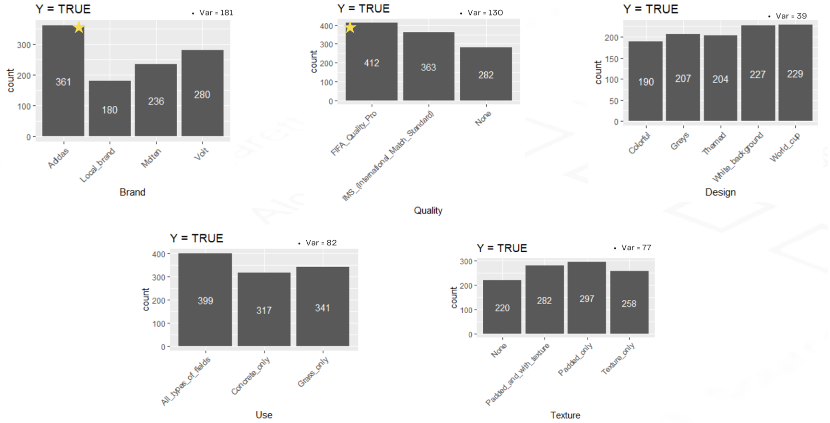


Figure 2.6: Variance per attribute

3 Methodology

Contents

To accomplish the goal of this paper and establish the right product, to the right customer, at the price that maximizes revenue, the first topic that will be addressed is regarding segmentation and the different algorithms used. Once market segments are defined, demand forecasting will be discussed and the the best model will be chosen. After a demand forecasting model is selected, the expected revenue will be calculated for each type of ball and the best portfolio will be determined based on the expected revenue for each target segment.

3.1 Segmentation

Segmentation refers to dividing the market into homogeneous groups of prospective buyers with common needs and who respond similarly to a marketing action. McDonald and Dunbar mention 4 types of segmentation:

Demographic segmentation involves breaking the market into customer demographics as age, income, gender, race, education, or occupation. This market segmentation strategy assumes that individuals with similar demographics will have similar needs.

Geographic segmentation groups customers by physical location, assuming that people within a given geographical area may have similar needs. This strategy is more useful for larger companies seeking to expand into different branches, offices, or locations.

Psychographic segmentation strives to classify consumers based on their lifestyle, personality, opinions, and interests. This may be more difficult to achieve, as these traits may change easily and may not have available objective data.

Behavioral segmentation groups consumers based on how they have previously interacted with markets and products. This type of segmentation relies heavily on market data, consumer actions, and decision making patterns.

Since the interested of this study is to understand the features that will impact the demand including price, then, behavioral segmentation vs demographic segmentation will be explored.

3.1.1 Clustering

Cluster analysis is a method commonly used in many disciplines to categorize entities into groups that are homogeneous along a range of observed characteristics. Once those homogeneous groups are formed then the researcher can focus on small number of groups rather than the large original entities.

Clustering can be achieved through different types of algorithms depending on it´s very own understanding of what constitutes a cluster and their similarity or distance parameters. Therefore, there is no objectively correct clustering algorithm, and the most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally. For this work, 2 clustering algorithms will be included.

Clustering is considered an unsupervised learning method since there is no ground truth to compare the output of the clustering algorithm to the true labels to evaluate its performance, meaning, no previous information of clusters is known to compare with.

K-Means is a centroid-based clustering algorithm, were each cluster is represented by a central vector (not necessarily included in the data set). This algorithm requires that the number of "k" clusters be specified in advanced. It tries to make the the point within each cluster as similar as possible while keeping the clusters as different or far as possible. It assigns data points to a cluster such that the sum of the squared distance between the data points and the cluster's centroid, the mean of all the data points that belong to that cluster, is at the minimum.

According to Wu kmeans follows the next steps:

1. Specify the number of K clusters.

2. Initialize centroids by first shuffling the dataset and then randomly selecting K data points to be used as the centroids.

3. Measure the distance between each data point and the centroids.

4. Assign the data point to the cluster with the nearest centroid. Using the Euclidean distance in Figure 3.2.

   Once all data points are assigned to a cluster, compute the new centroids by using the arithmetical mean.

5. Asses the quality of each cluster by calculating the Within-Cluster Sum of Squares (WCSS) [1] to quantify the variance needed to minimize.

6. Iterate or repeat steps 3 to 5 with the new centroids calculated until it converges or no more changes in centroids are found.

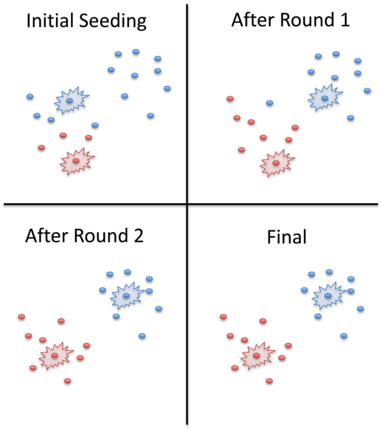


Figure 3.1: K-Mean Iteration

Hierarchical Cluster is another unsupervised machine learning algorithm, which is used to group the unlabeled datasets into a cluster. In this algorithm clusters are formed in a tree-shaped structure known as dendrogram. Results in K-Means and Hierarchical clustering may be similar but they both differ on how they work. Hierarchical clustering for one, does not require to know the numbers of cluster before hand, and also the hierarchical clustering may use different similarity of distance measures and linkage between those measures or from where the distance is computed.

[1] WSCC = $\sum_{i=1}^{Nk} \sum_{x \epsilon Ki} d(x, x\bar{_{Ki}})^2$

The distance metric should be computed based on the expertise of the domain, the most common metric is called Euclidean, see Figure 3.2. Other distance measures can be used such as: maximum, manhattan, canberra, or gower for categorical variables. For more information read Gower´s paper. [2]

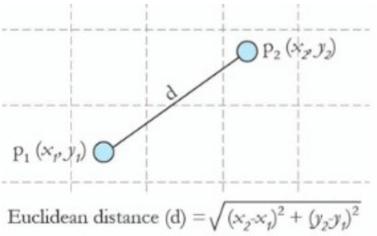


Figure 3.2: Euclidean Distance: is the length of a line segment between the two points. It can be calculated from the Cartesian coordinates of the points using the Pythagorean theorem.

The linkage criteria or from where the distance is computed can be between the two most similar parts of the cluster called single-linkage, or the 2 least similar data points of the cluster called complete-linkage, the center of the clusters called average-linkage, or some other criteria can be developed. Where there are no clear theoretical justifications for the choice of linkage criteria, Ward's method is the default. This method works out which observations to group based on reducing the sum of squared distances of each observation from the average observation in a cluster.

This algorithm has 2 main approaches: Agglomerative is a bottoms-up approach in which the algorithm starts with taking all data points as single clusters and merging them until one cluster is left. And Divisive algorithm is the reverse of the agglomerative algorithm as it is a top-down approach.

1. Compute the proximity matrix using a distance metric.
2. Each data point is assigned to a cluster.
3. Merge the clusters based on a metric for the similarity between clusters.
4. Update the distance matrix.
5. Repeat steps 3 and 4 until only a single cluster remains forming a dendrogram (Figure 3.3)

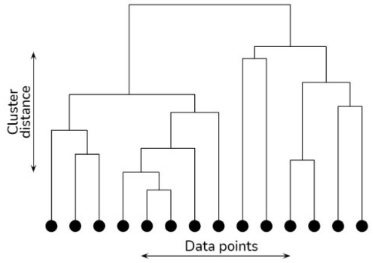


Figure 3.3: Dendogram

3.2 *Demand forecast*

Demand forecasting is the process that uses historical data to estimate and predict customers' future demand for a product or service. Companies spend huge amount of time and money trying to estimate demands because this helps companies improve production lead times, increase operational efficiencies, launch new products, provide better customer experience, and also formulate a competitive pricing strategy by understanding the market potential and business opportunities.

For this specific case, the demand will not be forecast in a numerical volume quantity, rather the probability of purchase based on the characteristics of the product, and this will be used further to estimate the expected revenue. (see Figure 3.4)

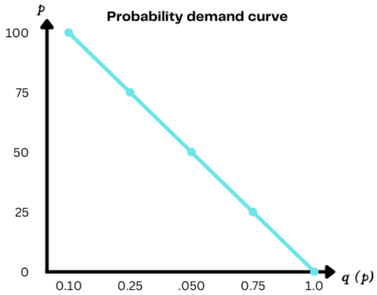


Figure 3.4: Probability demand curve

[2] Godfred Owusu-Bempah, Ebenezer Bennet, Eugene Okyere-Kwakye, and Dennis Amoako. A general coefficient of similarity and some of its properties. *JSTOR*, 27(4):857–871, 1971

3.2.1 Random Forest

Random Forest model belongs to the category of Classification and regression (CART) algorithms [3], where trees are created for binary classification, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Schonlau and Zou explain how decision trees can be used to visually and explicitly represent decisions and decision making as they seek to find the best split to subset the data. Metrics, such as Gini impurity, information gain, or mean square error (MSE), can be used to evaluate the quality of the split.

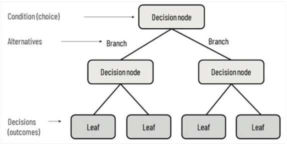


Figure 3.5: Elements of a decision tree

Leo Breiman and Adele Culter trademarked the name Random forest [4] by combining multiple output decision trees to reach a single result and therefore remove any over-fitting to the model.

Random forest algorithms have 3 main hyper-parameters that need to be set before testing the model like the node size, the number of trees, and the number of features sampled.

The algorithm is made up of a collection of decision trees, and each tree is comprised of a data sample drawn from the training set with replacement. Of the training data, one portion is set aside for cross validation.

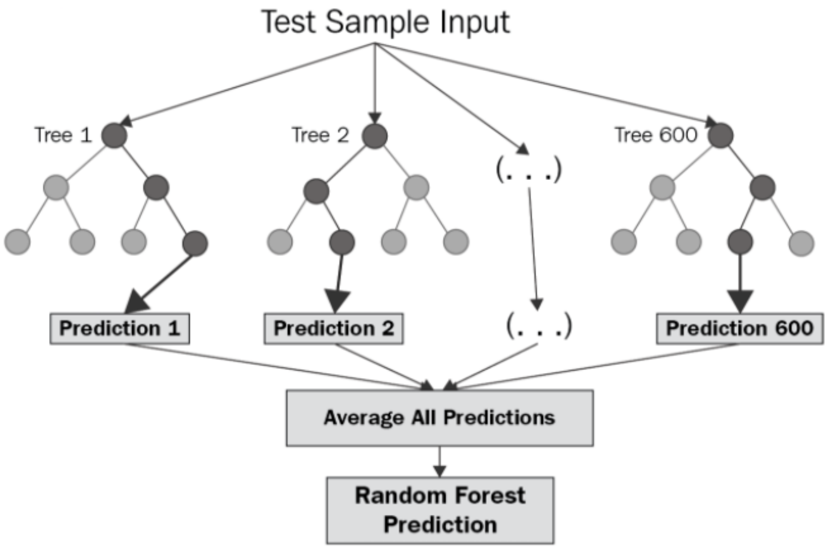


Figure 3.6: Random Forest Algorithm

[3] Pratap Dangeti. *Statistics for Machine Learning*. Packt Publishing, 2017. ISBN 9781788295758

[4] IBM Cloud Education. Random forest. URL https://www.ibm.com/cloud/learn/random-forest

3.2.2 Näive Bayes Clasifier

Näive Bayes Clasifier is a probabilistic machine learning algorithm based on the Bayes Theorem conditional probability. [5]

Conditional probability [6] is a measure of the probability of an event occurring given that another event has occurred, or in others words is the probability of an event occurring based on the prior knowledge times the likelihood of the event, divided by the evidence at hand. A fundamental assumption of this model is that each features are independent and equal contributors to the outcome.

The Naive Bayes classifier combines the Bayes theorem with a decision rule. One common rule is to pick the hypothesis that is most probable so as to this minimize the probability of misclassification, known as the maximum a posteriori or MAP decision rule.

$$\hat{y} = argmax\ p(C_k) \prod_{i=1}^{n} p(x_i \mid C_k) \tag{3.1}$$

Bernoulli Naive bayes is a model used where binary term occurrence features are used rather than term frequencies shown in 3.2. This is the model that will be used in this study.

$$p(\mathbf{x} \mid C_k) = \prod_{i=1}^{n} p_{ki}^{x_i} (1 - p_{ki})^{(1-x_i)} \tag{3.2}$$

3.2.3 Generalized Linear Model

Generalized Linear Models or GLM models, is a flexible generalization of the ordinary linear regression by allowing the response variable to have other exponential distributions instead of normal distributions. Using a link function GLMs allow the output to be related with the linear model through a link function. [7] GLM uses the maximum likelihood estimation of the model parameters for the exponential family and least squares for normal linear models.

Linear model

$$Y = b_0 + b_1 X_1 + \ldots + b_n X_n + e \tag{3.3}$$

$$Y \sim Bern(p) \tag{3.4}$$

Logistic regression measures the relationship between the dependent variable and one or more independent variables by estimating

[5] P(A| $B) = \frac{P(B|A)*P(A)}{P(B)}$

[6] Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning.*, 29(1):31–163, 1997

[7] Giorgio Alfredo Spedicato, Christophe Dutang, and Leonardo Petrini. Machine learning methods to perform pricing optimization. a comparison with standard glms. *Hal open science.*, 2021

probabilities using the logit function, also known as the link function mentioned above. [8]

$$\text{logit}(p) = \ln(\frac{p}{1-p}) = b_0 + b_1 X_1 + \ldots + b_n X_n + e \tag{3.5}$$

The inverse of the logit function is the sigmoid function. The sigmoid function maps probabilities to the range [0, 1] – and this makes logistic regression as a classifier.

$$P = \frac{e^{logit(P)}}{e^{logit(P)} + 1} \tag{3.6}$$

Therefore, the GLM algorithm with logit link estimates the probability of a positive response as seen in equation 3.7.

$$Y(P = 1) = \frac{e^{b_0 + b_1 X_1 + \ldots + e}}{e^{b_0 + b_1 X_1 + \ldots + e} + 1} \tag{3.7}$$

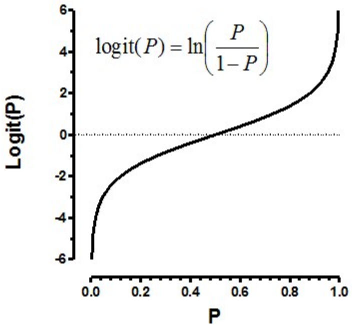


Figure 3.7: Logit Function

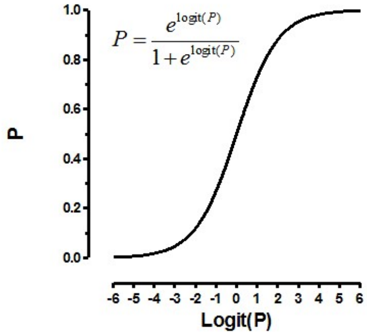


Figure 3.8: Sigmoid Function

3.2.4 *Support Vector Machines*

Support Vector Machines or SVM is a supervised machine learning algorithm used for both classification and regression. The objective of SVM algorithm is to find an optimal hyperplane in an N-dimensional space that distinctly classifies the data points. The dimension of the hyperplane depends upon the number of features and a trick called Kernel.[9]

The optimal hyperplane is the one that maximizes the margins from each class. In other words, the best hyperplane will maximize the distance of the nearest element of each tag or also called the margin (see Figure 3.10).

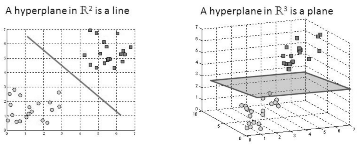


Figure 3.9: Hyperplane 2D vs 3D

$$M = max \frac{1}{\|w\|} \tag{3.8}$$

or also can be written as

$$M = min\|w\| \tag{3.9}$$

and since this becomes an optimization problem, l2 optimization are often more stable then l1, therefore the above equation can be written as:

$$min \frac{\|w\|^2}{2} \tag{3.10}$$

Abe explains in his books that for this optimization problem, 2 classification models exist; hard margin, where there should be no

[8] P McCullagh and J.A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability 37. Champman and Hall

[9] SVM algorithms use a set of mathematical functions that are defined as the kernel. The function of kernel is to take data as input and transform it into the required form.

misclassification, or a soft margin that some misclassifications can be accepted.

The primal equation for the soft margin problem is:

$$\min_{w,b,\zeta} \frac{1}{2} \|w\|^2 + c \sum_i \zeta_i \tag{3.11}$$

subject to

$$y_i(w^T x_i + b) \geq 1 - \xi_i \tag{3.12}$$

and

$$\xi_i \geq 0 \tag{3.13}$$

Using the lagrange multiplier the above equation is transformed into the dual equation, that will maximize then the margin due to the constraints in the same problem.

$$\max_{\lambda \geq 0} -\frac{1}{2} \sum_i \sum_j \lambda_i \lambda_j y_i y_j x^{(i)t} x^{(j)} + \sum_i \lambda_i \tag{3.14}$$

subject to

$$\sum_i \lambda_i y_i = 0 \tag{3.15}$$

and

$$c \geq \lambda_i \geq 0 \tag{3.16}$$

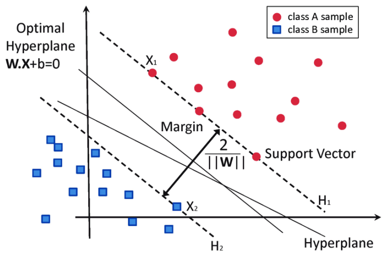


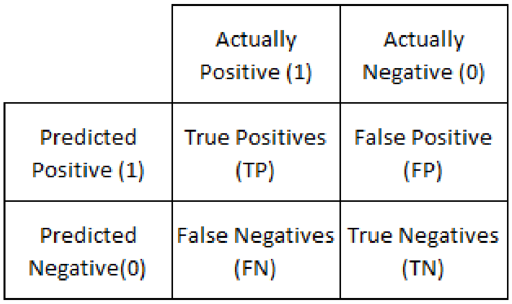
Figure 3.10: Best Hyperplane

One of the main advantage that SVM with dual optimization has is that it supports different Kernel functions or even custom made kernels that best fit the data. The most common kernels are: Gaussian, Polynomial, Radial, Sigmoid and Gamma. The best kernel depends on the data and the domain in study.

3.2.5 Metrics for model selection

The most important task in building any machine learning model is to evaluate its performance. According to Hossin and Sulaiman evaluation metric have been employed into two stages, training stage (learning process) and testing stage. In training stage the evaluation metric is used to optimize the algorithm. Meanwhile, in the testing stage, the evaluation metric is used as the evaluator to measure the effectiveness of produced classifier when tested with the unseen data.

In binary clasifications problems, such is the case, the evaluation of the clasification algorithm can be done based on the confusion matrix in Figure 3.11. The row of the table represents the predicted class, while the column represents the actual class, where TP and TN denote the number of positive and negative instances that are correctly classified. Meanwhile, FP and FN denote the number of misclassified negative and positive instances. Several metrics can be generated to evaluate the performance from the confusion matrix shown in Table 3.1.



	Actually Positive (1)	Actually Negative (0)
Predicted Positive (1)	True Positives (TP)	False Positive (FP)
Predicted Negative(0)	False Negatives (FN)	True Negatives (TN)

Figure 3.11: Confusion Matrix

Metric	Formula
Accuracy	$\frac{TP+TN}{TP+FN+TN+FP}$
Error Rate	$\frac{FP+FN}{TP+FP+TN+FN}$
Sensitivity	$\frac{TP}{TP+FN}$
Specificity	$\frac{TN}{TN+FP}$
Precision	$\frac{TP}{TP+FP}$

Table 3.1: Metrics from confusion matrix.

When evaluating models, the best metrics to choose from may vary according to the data set. In his article, Zeljko Vujovic [10] distinguishes that metrics like accuracy, precision and recall are good ways to evaluate classification models for balanced datasets, but if the data is imbalanced then other methods like ROC/AUC perform better in evaluating the model performance.

This research has balanced data and the intention is to estimate the probability of a positive results. Therefore, the best metric should be precision, which measures the number of accurate positive predictions, and it is mostly useful where False Positive is a higher concern than False Negatives, although all metrics will be calculated, precision will be the tie-breaker.

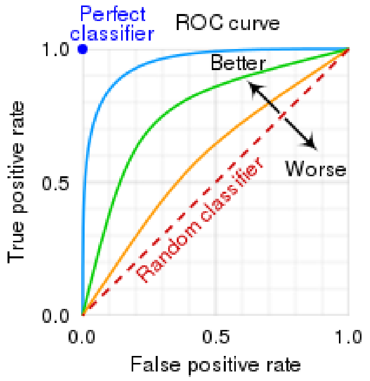


Figure 3.12: Expected Revenue

[10] Zeljko Vujovic. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, 12(6):599–606, 2021

3.3 Expected revenue

Expected value [11] is the weighted average of possible values of a random variable. It uses probability to predict what to expect in an experiment in the long run after many trials. It´s also called the mean of the probability distribution. (See Figure 3.17)

$$E[X] = \mu = \sum_{i=1}^{n} x_i P(x_i) \tag{3.17}$$

For this research the expected value will represent the expected revenue to achieve. The previous models will calculated the probability of a customer purchasing the ball with certain characteristics, including price, and the expected revenue [12] is the probability of selling the ball at the price times the price.

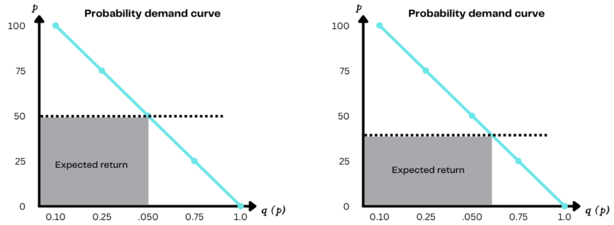


Figure 3.13: Expected Revenue

3.4 Portfolio planning

Price discrimination is practiced based on the seller's belief that customers in certain groups can be asked to pay more or less based on certain demographics or on how they value the product or service in question. Doraszelski and Draganska mentioned that it is most valuable when the profit that is earned as a result of separating the markets is greater than the profit that is earned as a result of keeping the markets combined. This is only possible because different customer segments have different characteristics and different price points that they are willing to pay. If everything were priced at say the average cost, people with lower price points could never afford it. And on the other hand, those with higher price points could hoard it.

The first step of creating a portfolio that discriminates pricing and maximizes the company´s earnings, requires identifying the target groups and exploiting their differences to create targeted marketing

[11] Brian Knutson, Jonathan Taylor, Matthew Kaufman, Richard Peterson, and Gary Glover. Distributed neural representation of expected value. *Journal of Neuroscience.*, 25(19):4806–4812, 2005

[12] Revenue = Price x Quantity

plans. Recalling, a marketing plan entails product development, pricing, product placement and promotion; for this specific case, placement and promotion will be set aside.

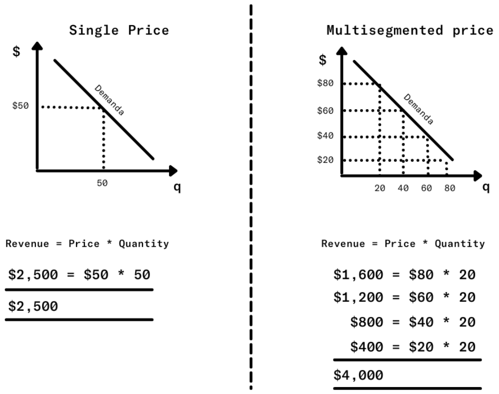


Figure 3.14: Segmented Pricing

4 Results

Contents

4.1 Clustering

Several models of clustering were performed. K-means with only the experimental variables and also adding the demographic variables. Likewise, a hierarchical clustering model was created using the experiment variables plus the demographic variables. In the end, both of the clustering had similar results, therefore, and in line with the parsimony principle [1], a K-means with the experimental variables was chosen. Besides, considering that the intention of this paper is to understand the behavioral aspects of consumers purchasing decisions, experimental variables are the only variables needed to achieve this goal and answer the first question of the objectives.

The first thing to address is the optimal number of clusters using an elbow analysis as mentioned before calculating the Within-Cluster Sum of Squares (WCSS). Seeing Figure 4.1, there is no clear number of clusters needed.

When theory and data are not the answer, then previous knowledge can help determine the right number of cluster. Experts in pricing have discussed that usually one can find [2] 4 types of brand scales based on the brand value benefits and each within 3 levels of pricing as seen in

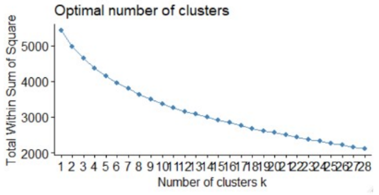


Figure 4.1: Elbow Method

[1] the most acceptable explanation of an occurrence, is the simplest, involving the fewest entities, assumptions, or changes

[2] Cesar Perez-Carballada. What is price premium and what is the difference with premium brand? URL http://englishmarketisimo.blogspot.com/2018/03/what-is-price-premium-and-what-is.html

Figure 4.2. Considering that this study already contains 4 brands, 3 clusters were chosen.

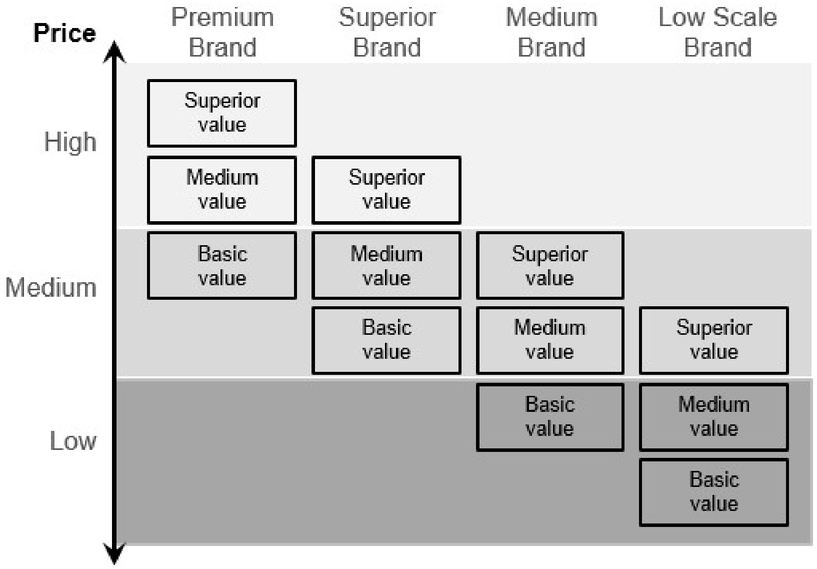


Figure 4.2: Brand Scales and pricing levels

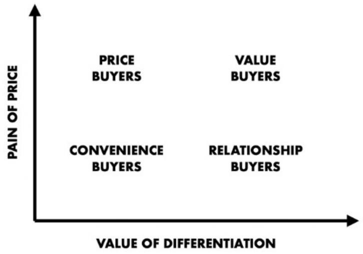
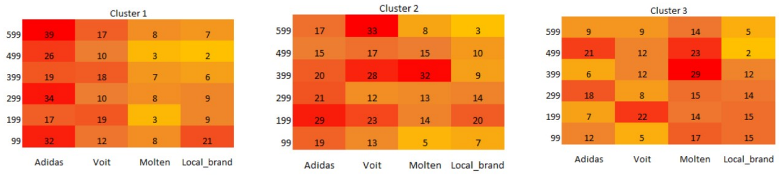


Figure 4.3: Buyer Types

Taking into account the frequency each brand was chosen and the level of price, Clusters 2 and 3 can be labeled as value buyers, whereas value is relevant but the customers are also seeking for a fair price, meanwhile, cluster 1 can be designated as brand buyer or also called relationship buyer, such that most buyers prefer a premium Adidas brand at any level of price than any other brand. To understand better the buyers types please refer to Nagel et al. in his book and see Figure 4.3.



(a) C1 - Brand buyer

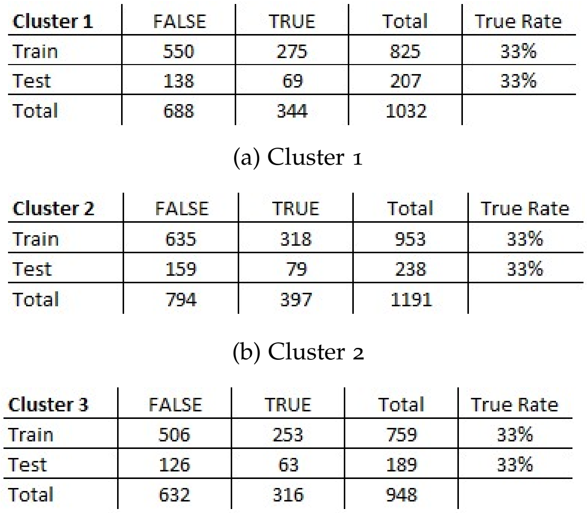
(b) C2 - Value buyer

(c) C3 - Value buyer

Figure 4.4: Type of buyers according to brand and price

4.2 Split data

Separating data into training and testing sets is an important part of evaluating the data models and determining the confusion matrix. Typically, most of the data is used for training, and a smaller portion of the data is used for testing. In this case, 80% was for training and 20% for testing, each cluster was divided as seen below:



Cluster 1	FALSE	TRUE	Total	True Rate
Train	550	275	825	33%
Test	138	69	207	33%
Total	688	344	1032	

(a) Cluster 1

Cluster 2	FALSE	TRUE	Total	True Rate
Train	635	318	953	33%
Test	159	79	238	33%
Total	794	397	1191	

(b) Cluster 2

Cluster 3	FALSE	TRUE	Total	True Rate
Train	506	253	759	33%
Test	126	63	189	33%
Total	632	316	948	

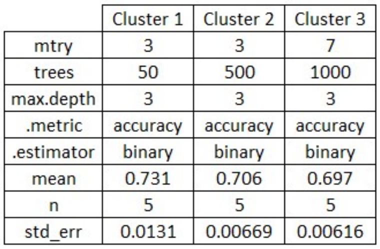
(c) Cluster 3

Figure 4.5: Split training vs test by cluster

4.3 Random Forest

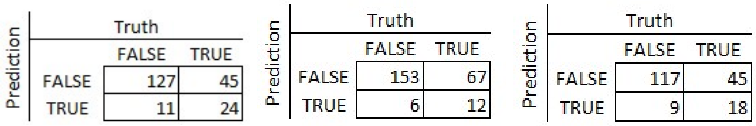
A random forest model was executed for each of the 3 denominated clusters. Hyper-parameters in Figure 3.9, were tuned for each model and confusion matrices were calculated as seen below in Figure 4.7.

The results obtained for each metric in evaluation can be referenced in Figure 4.8.



	Cluster 1	Cluster 2	Cluster 3
mtry	3	3	7
trees	50	500	1000
max.depth	3	3	3
.metric	accuracy	accuracy	accuracy
.estimator	binary	binary	binary
mean	0.731	0.706	0.697
n	5	5	5
std_err	0.0131	0.00669	0.00616

Figure 4.6: Random Forest Model Hyper-parameters



		Truth	
		FALSE	TRUE
Prediction	FALSE	127	45
	TRUE	11	24

(a) Cluster 1

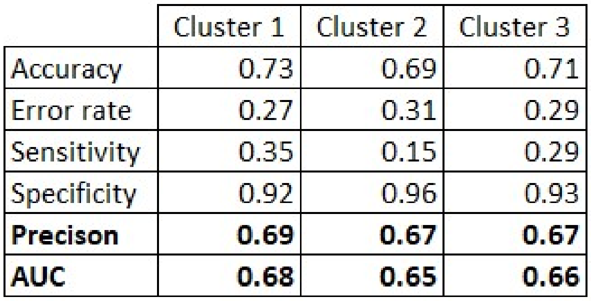
		Truth	
		FALSE	TRUE
Prediction	FALSE	153	67
	TRUE	6	12

(b) Cluster 2

		Truth	
		FALSE	TRUE
Prediction	FALSE	117	45
	TRUE	9	18

(c) Cluster 3

Figure 4.7: RF testing data confusion matrix by cluster

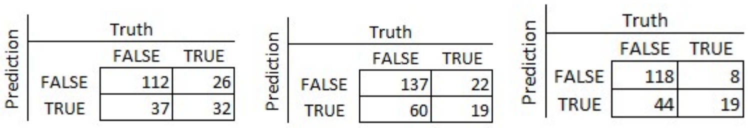


	Cluster 1	Cluster 2	Cluster 3
Accuracy	0.73	0.69	0.71
Error rate	0.27	0.31	0.29
Sensitivity	0.35	0.15	0.29
Specificity	0.92	0.96	0.93
Precison	**0.69**	**0.67**	**0.67**
AUC	**0.68**	**0.65**	**0.66**

Figure 4.8: Random Forest Model Results by Cluster

4.4 *Naïve Bayes*

Naïve Bayes confusion matrix for each cluster can be seen in the next Figure 4.9. And the results for the metrics in question in Figure 4.10.



Prediction	Truth FALSE	Truth TRUE
FALSE	112	26
TRUE	37	32

(a) Cluster 1

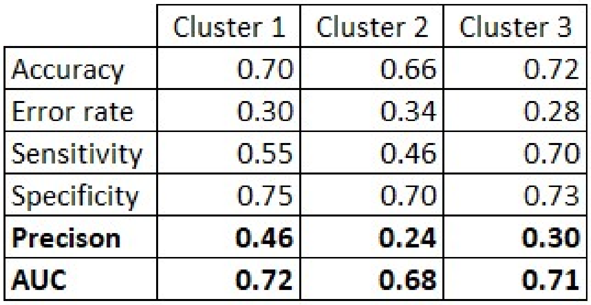
Prediction	Truth FALSE	Truth TRUE
FALSE	137	22
TRUE	60	19

(b) Cluster 2

Prediction	Truth FALSE	Truth TRUE
FALSE	118	8
TRUE	44	19

(c) Cluster 3

Figure 4.9: NB testing data confusion matrix by cluster

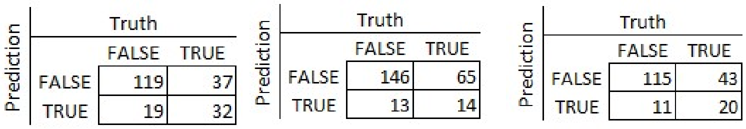


	Cluster 1	Cluster 2	Cluster 3
Accuracy	0.70	0.66	0.72
Error rate	0.30	0.34	0.28
Sensitivity	0.55	0.46	0.70
Specificity	0.75	0.70	0.73
Precison	**0.46**	**0.24**	**0.30**
AUC	**0.72**	**0.68**	**0.71**

Figure 4.10: Naive Bayes Model Results by Cluster

4.5 *Generalized Linear Model*

The confusion matrices for the generalized linear model with link logit can be found in 4.11 and the results for the main metrics in Figure 4.12.



Prediction	Truth FALSE	Truth TRUE
FALSE	119	37
TRUE	19	32

(a) Cluster 1

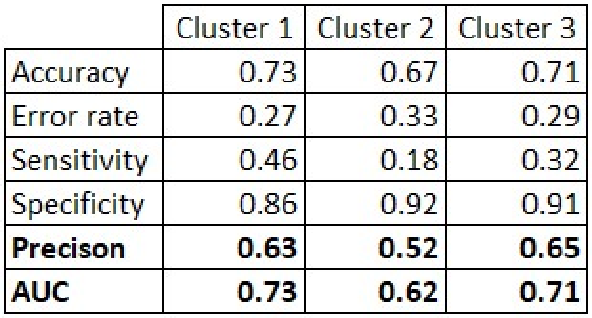
Prediction	Truth FALSE	Truth TRUE
FALSE	146	65
TRUE	13	14

(b) Cluster 2

Prediction	Truth FALSE	Truth TRUE
FALSE	115	43
TRUE	11	20

(c) Cluster 3

Figure 4.11: GLM testing data confusion matrix by cluster

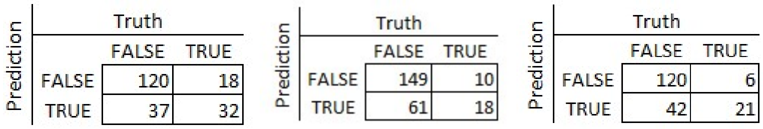


	Cluster 1	Cluster 2	Cluster 3
Accuracy	0.73	0.67	0.71
Error rate	0.27	0.33	0.29
Sensitivity	0.46	0.18	0.32
Specificity	0.86	0.92	0.91
Precison	**0.63**	**0.52**	**0.65**
AUC	**0.73**	**0.62**	**0.71**

Figure 4.12: Generalized Linear Model Results by Cluster

4.6 *Support Vector Machines*

And finally, the confusion matrices and metrics results for the support vector machine models can be found below:



Prediction \ Truth	FALSE	TRUE
FALSE	120	18
TRUE	37	32

(a) Cluster 1

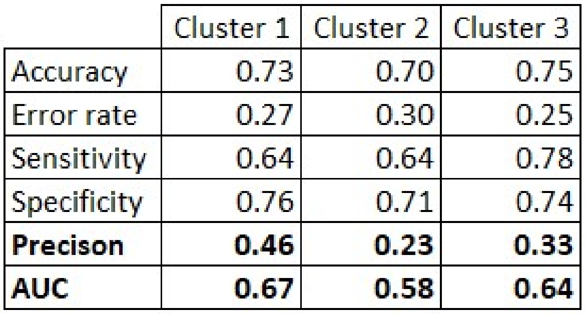
Prediction \ Truth	FALSE	TRUE
FALSE	149	10
TRUE	61	18

(b) Cluster 2

Prediction \ Truth	FALSE	TRUE
FALSE	120	6
TRUE	42	21

(c) Cluster 3

Figure 4.13: SVM testing data confusion matrix by cluster



	Cluster 1	Cluster 2	Cluster 3
Accuracy	0.73	0.70	0.75
Error rate	0.27	0.30	0.25
Sensitivity	0.64	0.64	0.78
Specificity	0.76	0.71	0.74
Precison	**0.46**	**0.23**	**0.33**
AUC	**0.67**	**0.58**	**0.64**

Figure 4.14: Support Vector Machines Results by Cluster

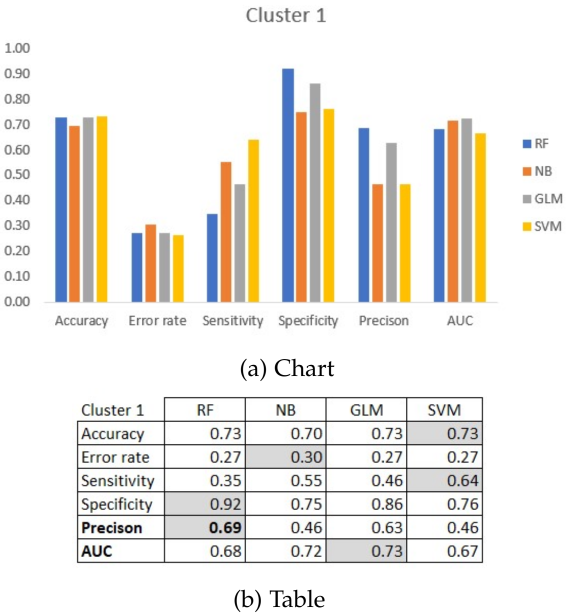
4.7 *Model comparison*

To answer question number 3 and predict the purchasing probability of a ball according to each cluster, each of the previous algorithms were executed for the 3 clusters and evaluation metrics were calculated as seen below.

Cluster 1 For this group the best algorithm can be determined as Random Forest, although Support vector machine has a great power in accuracy, random forest also shares that same power and has a better precision, which means that it´s better at predicting the true responses, or the probability that a consumer will purchase the ball.

Cluster 2 This conglomerate shares similar results as cluster 1, random forest has a good ability to predict the true real choices and therefore also has been chosen for this group.

Cluster 3 Lastly, this cluster also has similar results as previous clusters, therefore, random forest will also be used to predict the purchase probability for all 3 clusters.

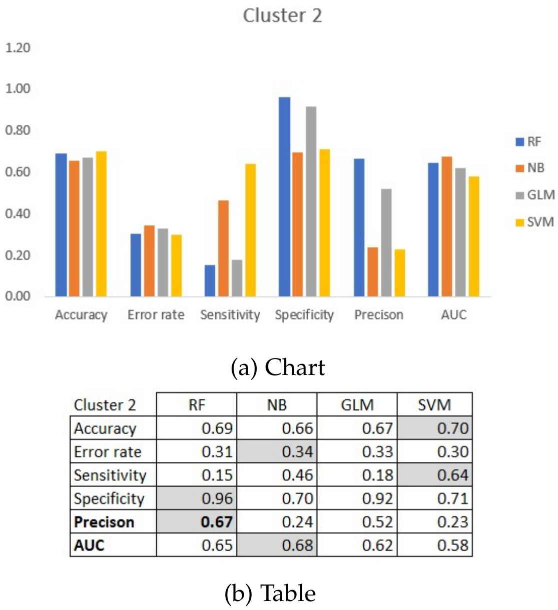


(a) Chart

Cluster 1	RF	NB	GLM	SVM
Accuracy	0.73	0.70	0.73	0.73
Error rate	0.27	0.30	0.27	0.27
Sensitivity	0.35	0.55	0.46	0.64
Specificity	0.92	0.75	0.86	0.76
Precison	**0.69**	0.46	0.63	0.46
AUC	0.68	0.72	0.73	0.67

(b) Table

Figure 4.15: Cluster 1 Model Comparison

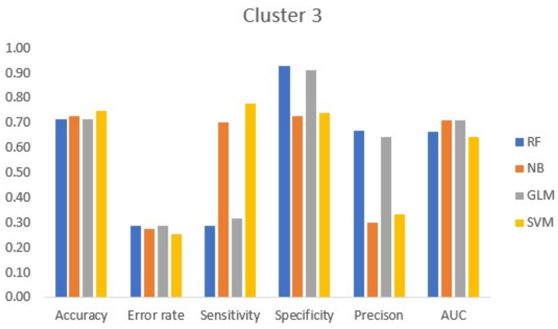


(a) Chart

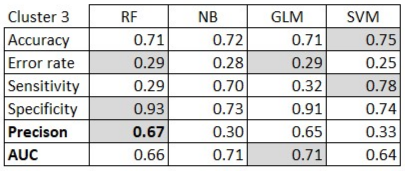
Cluster 2	RF	NB	GLM	SVM
Accuracy	0.69	0.66	0.67	0.70
Error rate	0.31	0.34	0.33	0.30
Sensitivity	0.15	0.46	0.18	0.64
Specificity	0.96	0.70	0.92	0.71
Precison	**0.67**	0.24	0.52	0.23
AUC	0.65	0.68	0.62	0.58

(b) Table

Figure 4.16: Cluster 2 Model Comparison



(a) Chart



Cluster 3	RF	NB	GLM	SVM
Accuracy	0.71	0.72	0.71	0.75
Error rate	0.29	0.28	0.29	0.25
Sensitivity	0.29	0.70	0.32	0.78
Specificity	0.93	0.73	0.91	0.74
Precison	**0.67**	0.30	0.65	0.33
AUC	0.66	0.71	0.71	0.64

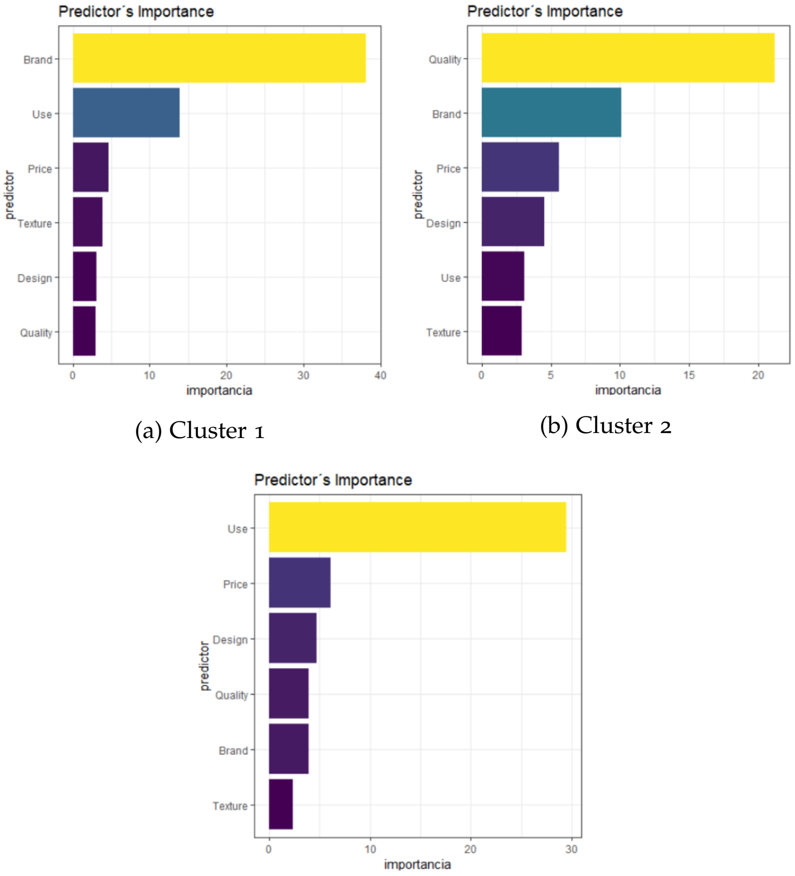
(b) Table

Figure 4.17: Cluster 3 Model Comparison

4.8 Attribute relevance

As mentioned before, random forest model has the capability to determine the relevance of each attribute when splitting the branches and therefore, the importance for choice making.

As hypothesised in the data prepossessing analysis, clusters´ 1 main drive is the brand and has a significant difference in relevance with all other variables including price. For cluster 2, quality is the main offender and brand is the second most relevant attribute followed by price. Regarding cluster 3, the most relevant attribute is use with a wide difference with the second most relevant, price.



(a) Cluster 1

(b) Cluster 2

(c) Cluster 3

Figure 4.18: Attribute relevance per cluster

5 Conclusion and future work

Contents

With the previous sections, questions 1, 2 and 3 have been answered and only question 4 is pending up to this point. Question 1 and 2 were answered with the clustering algorithm determining that there are 3 significant groups, and that group 1 purchase decisions are mostly based on the brand, while the other two clusters center more on other characteristics such as use and quality, but overall none of the 3 clusters purchasing choices are based on pricing, confirming what was hypothesized from the pre-processing. Additionally, the outcome for question three was gathered from the random forest machine learning algorithms that captured the procuring probability. The final question pending to respond is: what is the optimal product to sell to each cluster and the price that will maximize the overall expected revenue for the company.

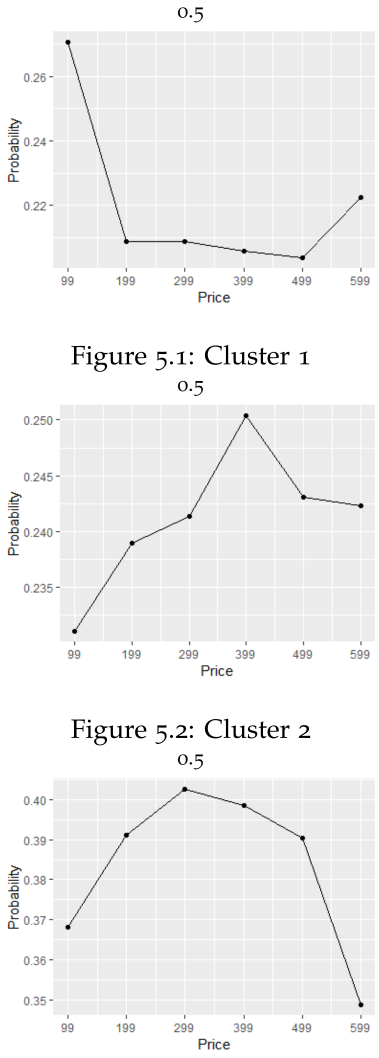


Figure 5.1: Cluster 1

Figure 5.2: Cluster 2

Figure 5.3: Cluster 3

Figure 5.4: Typical ball purchase probabilities

5.1 Portfolio and Pricing

Before making any recommendations on the optimal ball, the first step is to obtain the acquisition probability of the current typical existing ball for each cluster and the expected revenue at the different price levels. The typical ball is shown in Figure 5.5; it´s a local brand ball, without any quality assurance, colorful design, used in all types of fields, padded only and at a price of $199.

Expected return for this typical ball for cluster 1 is 41.55, for cluster 2 is 47.56 and for cluster 3 is 77.83. This means that overall the company´s expected return for this ball is 166.94. Reference the purchase probabilities in Figure 5.4.

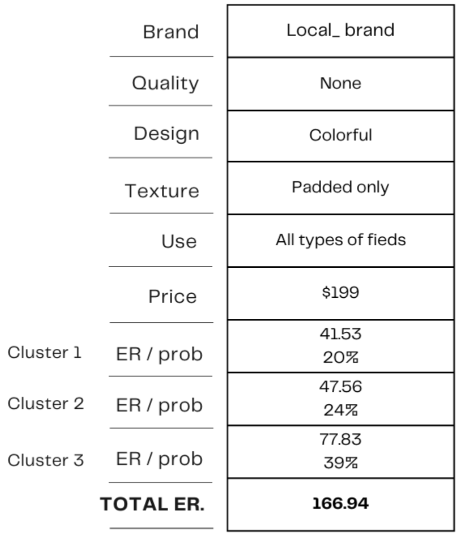


Figure 5.5: Typical ball

	Brand	Local_ brand
	Quality	None
	Design	Colorful
	Texture	Padded only
	Use	All types of fieds
	Price	$199
Cluster 1	ER / prob	41.53 20%
Cluster 2	ER / prob	47.56 24%
Cluster 3	ER / prob	77.83 39%
	TOTAL ER.	**166.94**

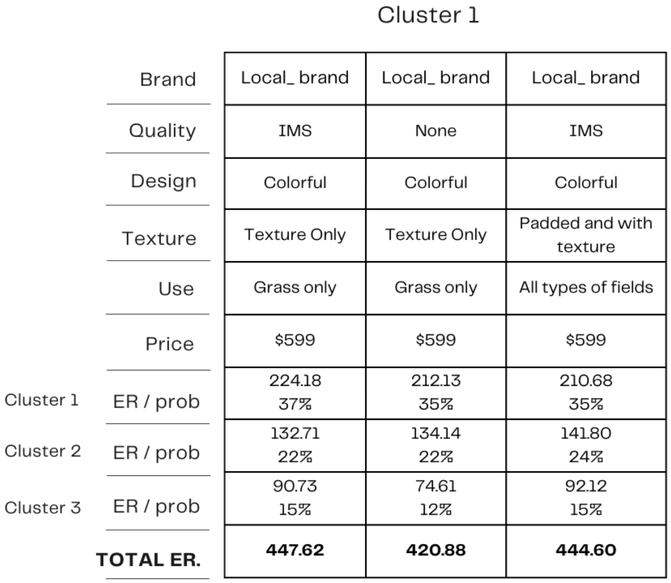
Optimal portfolio Choosing the optimal portfolio may become an art and science at the same time. Although the main decision is to optimize the expected revenue as mentioned before, also, one must consider the different choices for each cluster to avoid cannibalization. [1]

The first step is to determine what are the highest expected revenue balls for each cluster, but also what is the expected revenue and impact each of those balls has on the others clusters. Seeing Figure 5.8, you can notice that the expected revenue of selling to cluster 3 is much higher than the other clusters, therefore, other questions arise, such as, does the company want to target all the segments?

For this study, the assumption is that the company wants to target all 3 segments to expand market share and wants to maximize revenue, therefore, the final recommendations lies in the balls that maximizes the total expected revenue for each segment taking into account all other segments. This means, for example, the highest revenue ball for cluster 2 only is not padded and has a 53% probability of being purchased by that segment, but taking into account all other segments, then the best ball is the one that is padded and has a 52% probability that cluster 2 buys it, meanwhile has a higher probability of being purchased also by cluster 1 and 3, without cannibalizing the highest revenue balls for each of those clusters.

And the final assessment made for the final recommendation is, what can help segmentation even further? For cluster 1, the highest revenue

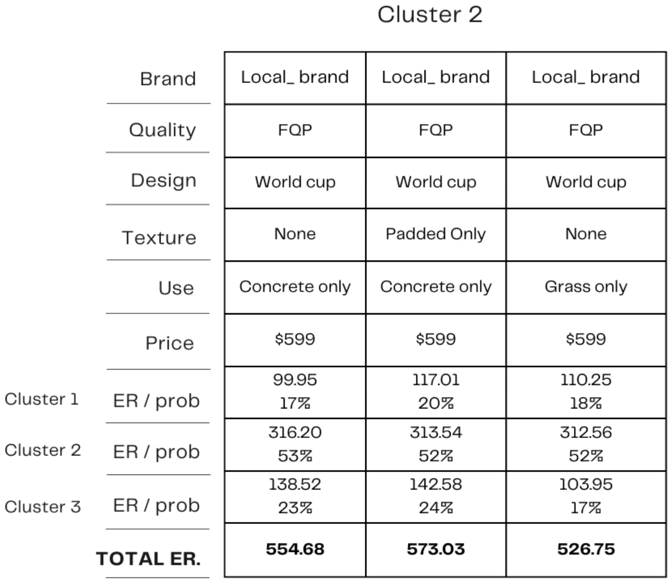
[1] is a reduction in sales volume, sales revenue, or market share of one product when the same company introduces a new product



Cluster 1

	Brand	Local_ brand	Local_ brand	Local_ brand
	Quality	IMS	None	IMS
	Design	Colorful	Colorful	Colorful
	Texture	Texture Only	Texture Only	Padded and with texture
	Use	Grass only	Grass only	All types of fields
	Price	$599	$599	$599
Cluster 1	ER / prob	224.18 37%	212.13 35%	210.68 35%
Cluster 2	ER / prob	132.71 22%	134.14 22%	141.80 24%
Cluster 3	ER / prob	90.73 15%	74.61 12%	92.12 15%
	TOTAL ER.	**447.62**	**420.88**	**444.60**

Figure 5.6: Cluster 1 - Highest expected revenue

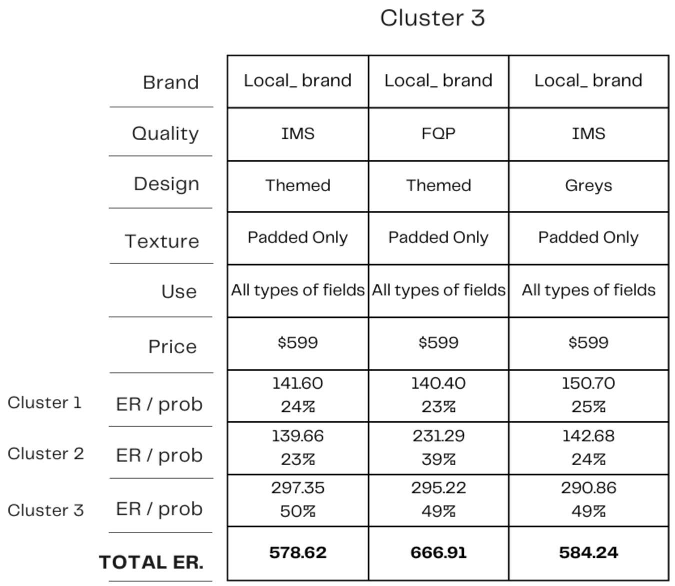


Cluster 2

	Brand	Local_ brand	Local_ brand	Local_ brand
	Quality	FQP	FQP	FQP
	Design	World cup	World cup	World cup
	Texture	None	Padded Only	None
	Use	Concrete only	Concrete only	Grass only
	Price	$599	$599	$599
Cluster 1	ER / prob	99.95 17%	117.01 20%	110.25 18%
Cluster 2	ER / prob	316.20 53%	313.54 52%	312.56 52%
Cluster 3	ER / prob	138.52 23%	142.58 24%	103.95 17%
	TOTAL ER.	**554.68**	**573.03**	**526.75**

Figure 5.7: Cluster 2 - Highest expected revenue

ball is the one that has IMS or International Standard Match quality, but remembering that for cluster 2, the main attribute is quality, then the decision to maintain a low quality ball for cluster 1 might be a good choice if revenue is not highly impacted, which is the case for cluster 1. On that account, the final recommendation that will segment customer



Cluster 3

	Brand	Local_ brand	Local_ brand	Local_ brand
	Quality	IMS	FQP	IMS
	Design	Themed	Themed	Greys
	Texture	Padded Only	Padded Only	Padded Only
	Use	All types of fields	All types of fields	All types of fields
	Price	$599	$599	$599
Cluster 1	ER / prob	141.60 24%	140.40 23%	150.70 25%
Cluster 2	ER / prob	139.66 23%	231.29 39%	142.68 24%
Cluster 3	ER / prob	297.35 50%	295.22 49%	290.86 49%
	TOTAL ER.	**578.62**	**666.91**	**584.24**

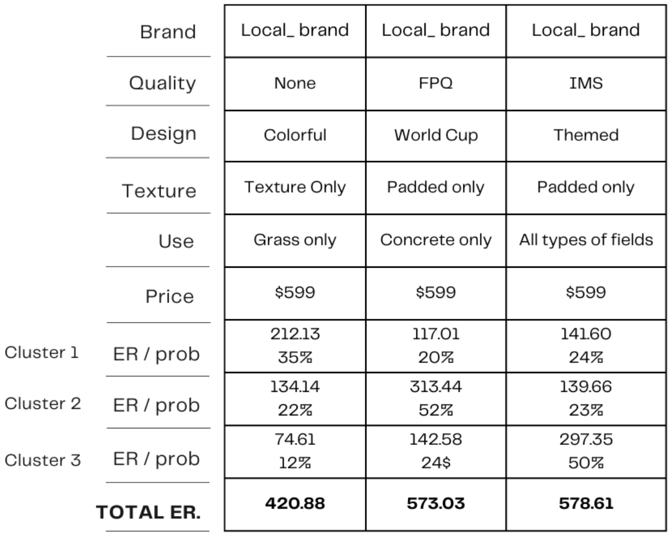
Figure 5.8: Cluster 3 - Highest expected revenue

and maximize expected revenue without cannibalization can be seen below in Figure 5.9, with an overall expected revenue of 1,572.53 vs the original expected revenue of 166.94.

Given this results, and going back to the original question of this study, can machine learning provide information that will help a company maximize value? Then, without a doubt, this work just proved that revenue could be maximized with the science behind and an information-based decision making process.

5.2 *Future Work*

Given the conclusion, 3 things must be considered for future work. For starters, customers seem not to be price sensitive, then it would be recommended to expand the survey into a higher range of possible prices, instead of limiting the choices to $599. Also, it is known that International standard match quality and FIFA quality pro are 2 very expensive processes to obtain, thus, cost information about product modifications would also be favored to be added, and instead of focusing on maximizing expected revenue, ideally, the focus should be in maximizing profit, and ultimately, considering that the cluster 2 and 3 have higher purchasing probabilities and expected revenue, should the company make the choice to focus only on those segments and become more of a niche market instead? That is something that could



	Brand	Local_ brand	Local_ brand	Local_ brand
	Quality	None	FPQ	IMS
	Design	Colorful	World Cup	Themed
	Texture	Texture Only	Padded only	Padded only
	Use	Grass only	Concrete only	All types of fields
	Price	$599	$599	$599
Cluster 1	ER / prob	212.13 35%	117.01 20%	141.60 24%
Cluster 2	ER / prob	134.14 22%	313.44 52%	139.66 23%
Cluster 3	ER / prob	74.61 12%	142.58 24$	297.35 50%
	TOTAL ER.	**420.88**	**573.03**	**578.61**

Figure 5.9: Final Recommendation

be also considered and asked about the companies overall competitive strategy for a better product decision making.

Bibliography

Shigeo Abe. *Support Vector Machines for Pattern Classification.* Springer, 2010. ISBN 978-1-84996-097-7.

Pratap Dangeti. *Statistics for Machine Learning.* Packt Publishing, 2017. ISBN 9781788295758.

Ulrich Doraszelski and Michaela Draganska. Market segmentation strategies of multiproduct firms. *The Journal of Industrial Economics.*, 54(1):125–149, 2006.

IBM Cloud Education. Random forest. URL https://www.ibm.com/cloud/learn/random-forest.

Nir Friedman, Dan Geiger, and Moises Goldszmidt. Bayesian network classifiers. *Machine Learning.*, 29(1):31–163, 1997.

Mohammad Hossin and MN Sulaiman. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining Knowledge Management Process*, 5(2):1, 2015.

Brian Knutson, Jonathan Taylor, Matthew Kaufman, Richard Peterson, and Gary Glover. Distributed neural representation of expected value. *Journal of Neuroscience.*, 25(19):4806–4812, 2005.

P McCullagh and J.A. Nelder. *Generalized Linear Models*. Monographs on Statistics and Applied Probability 37. Champman and Hall.

Malcolm McDonald and Ian Dunbar. *Market Segmentation*. Elsevier Butterworth-Heinemann, 2004. ISBN 0750659815.

Thomas T. Nagel, John E Hogan, and Joseph Zale. *The Strategy and Tactics of pricing: A guide to growing more profitably.* Prentice Hall, 2010. ISBN 978-0-13-910681-4.

Godfred Owusu-Bempah, Ebenezer Bennet, Eugene Okyere-Kwakye, and Dennis Amoako. A general coefficient of similarity and some of its properties. *JSTOR*, 27(4):857–871, 1971.

Cesar Perez-Carballada. What is price premium and what is the difference with premium brand? URL http://englishmarketisimo.blogspot.com/2018/03/what-is-price-premium-and-what-is.html.

Matthias Schonlau and Rosie Yuyan Zou. The random forest algorithm for statistical learning. *The Stata Journal.*, 20(1):3–29, 2020.

Giorgio Alfredo Spedicato, Christophe Dutang, and Leonardo Petrini. Machine learning methods to perform pricing optimization. a comparison with standard glms. *Hal open science.*, 2021.

Zeljko Vujovic. Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications.*, 12(6):599–606, 2021.

Junjie Wu. *Advances in K-means Clustering. A Data Mining thinking*. Springer, 2012. ISBN 978-3-642-29807-3.

Index