UNIVERSIDADE CATÓLICA PORTUGUESA

Product Portfolio Management:

What are Sonae MC's Star Products?

José Leopoldo Soares Almeida Carvalhaes

Católica Porto Business School

2022

UNIVERSIDADE CATÓLICA PORTUGUESA

Product Portfolio Management:

What are Sonae MC's Star Products?

Relatório de Estágio apresentado à Universidade Católica Portuguesa

para obtenção do grau de mestre em Gestão

por

José Leopoldo Soares Almeida Carvalhaes

sob orientação de

Professor Leonardo Costa

Católica Porto Business School

2022

Acknowledgments

I thank Sonae MC for the opportunity granted to develop my master's final work in such a dynamic and demanding environment, enabling me to grow as a person and as a professional. A special thanks to Telma Paulo, Ana Freitas and all my team, for their guidance and support, which not only made me feel welcome but strongly contributed to my development.

I also thank Católica Porto Business School for the opportunity provided and above all for the guidance of Prof. Leonardo Costa, critical not only for the conclusion of this dissertation, but also in all stages of this journey.

A final thanks to my parents and brother Filipe, to my girlfriend Mariana, and to the rest of my family and friends, such as Gonçalo and Pedro, without whom this experience would not be as exciting and enriching as it was.

Resumo

A gestão do portefólio de produtos desempenha um papel fundamental no desempenho de uma empresa que opera num ambiente muito competitivo, como é o caso da Sonae MC. Conseguir identificar quais os produtos estrela e quais os fatores que fazem com que um produto seja estrela adquire uma especial importância para a empresa. Neste Trabalho Final de Mestrado é proposto um modelo de classificação de produtos baseado numa combinação de diferentes segmentações de clientes, para a identificação de produtos estrela da Sonae MC. Com o suporte de modelos de Regressão Logística e de Árvores de Decisão, identificam-se os principais fatores que afetam a probabilidade de um produto ser estrela. Os resultados mostram que a probabilidade de um produto ser estrela é influenciada pela frequência média de vendas, pela percentagem de vendas em promoção, pelo nível de preço e pela percentagem de clientes que se encontra num dos 3 seguintes segmentos: Qualidade, Família e Preço/Promoção. Comparando os modelos de Regressão Logística e de Árvores de Decisão, a maior capacidade preditiva deste último fornece melhores métricas de classificação.

Palavras chave: Produtos estrela, Modelo de classificação, Regressão logística, Árvores de decisão, Gestão do portefólio de produtos

Palavras: 6265

Abstract

Product portfolio management plays a key role in the performance of a company operating in a very competitive environment, as it is the case with Sonae MC. Being able to identify which products are Star and which factors make a product a Star is of paramount importance for the company. In this Master Final Assignment, a product classification model based on a combination of different customer segmentation is proposed to identify Sonae MC's Star products. A Logistic Regression and Decision Tree models are then used to verify main factors that affect the probability of a product being a star. Results show that the probability of a product being a star is influenced by the average frequency of sales, the percentage of sales when on sale, the price level and the percentage of customers in the following 3 segments: Quality, Family and Price/Promotion. Comparing Logistic Regression and Decision Tree models, the greater predictive capacity of the latter provides better classification metrics.

Keywords: Star products, Classification model, Logistic regression, Decision trees, Product portfolio management

Words: 6265

Index

Figures Index

Tables Index

Chapter 1

Introduction

The main objectives of this Final Master's Work (MFA) are: i) to identify which products are Star products for Sonae MC; and ii) to assess which product and customer attributes increase the probability of a product being a Star for the company.

The best-selling products have already been identified by Sonae MC, but it is necessary to identify which products have the greatest potential not only in sales, but also in attracting new customers to the stores. This MFA proposes a new classification model to close this gap and/or identify which products are Star for the company, with the available data. In addition, it uses and compares two alternative methods - logistic regression and decision trees - to assess how the different attributes of products and customers affect the probability of a product being a Star for the company, allowing for the comparison of the two methods.

As the leading food retailer in Portugal, Sonae MC has a history of over 30 years in the Portuguese market and a multi-format business. In 2021, with more than 1342 stores and 4 million families as customers, the company had a business turnover of 5362 million euros (Source: Sonae MC. 2022, available at https://mc.sonae.pt/sobre-nos/; 16H 00M).

Competing in such a fierce environment requires companies to extract the full potential of their assets. With its multi-format business portfolio, Sonae MC offers a wide range of products and services, having a constant need to monitor and manage this product portfolio. Understanding which products have the greatest potential, both in terms of sales and in attracting new customers to the stores, creating an opportunity to strengthen the relationship between the company and customers, is key to Sonae MC's success. The MFA unfolds as follows. After this introductory chapter, chapter 2 presents a literature review on

what Star products are, store brands, the logistic regression, and the decision trees. Chapter 3 presents the empirical model, the results and their discussion. Finally, in chapter 4, the main conclusions are presented.

Chapter 2

Literature review

2.1. The BCG Matrix

Created by the Boston Consulting Group (which gave rise to its name) in the 1960s, the BCG Matrix is a tool that allows looking at an organization as a set of activities/products, where each of the products has a different contribution to companies profitability and growth (Hax & Majluf, 1983).

To assess the role that each of the products plays in the organization, the matrix creates two axes and four quadrants. On the horizontal axis we have the relative (to competitors) market share for each product, while on the vertical axis we have the average market growth rate in recent years. Products are usually represented by circles, the size of the result being total sales. According to the quadrant of the matrix in which a product is inserted, it is designated as: "Dog", "Question mark", "Cash cow" or "Star", with this last quadrant being the one with the most promising perspective for the companies' results, both in terms of relative market share and growth.

The BCG Matrix has been popular since its creation, mainly due to its conceptual simplicity, the objective way in which the data is presented and the ease of interpretation of its graphical representation. The main drawbacks of this tool are related to its scope. The BCG Matrix includes only existing products, assumes that a company's competitiveness depends exclusively on its relative market shares, not embracing other generic factors such as, for example, human and financial resources, and its nature is contingent (Lindon *et al.*, 2004).

Once products are classified within the BCG Matrix, understanding which factors affect the classification of a product as "Star" can be key for a good management of the product portfolio. The importance of product portfolio and activity

management has been a growing concern for academic literature and models that optimize the trade-off between risk and return (Adams et al., 2006).

2.2. Brand Equity and Market Share

Due to the growth of private labels in the last two decades, the specialized literature has focused on the market share of private labels and how product attributes affect this share (Yagüe & Rubio, 2009).

Brand equity, as the combination of customer's behaviors and associations that allow for a brand to earn greater volume (Keller 1998), is positively related to market share. It is therefore critical to understand why consumers may be willing to pay a premium for manufacturer brands over store brands, or the other way around. This "premium" value that consumers will pay for a product is mainly affected by perceived quality (Sethuraman, 2001).

Investment in store brand is positively associated with the growth of store brand market share in a product category, while an increased price differential between manufacturer and store brands incentives customers to choose the store brand (Yagüe & Rubio, 2009).

2.3. Store Brands

The retail industry has seen store brands as an important factor in retail differentiation and a key contributor in promoting store loyalty. Often seen as extensions of the image of each store, store brands can be the determining factor for positioning strategies among retailers (Collins-Dodd & Lindley, 2003).

Store-branded products may have a greater chance of success whenever the consumer's prior uncertainty about the product is low and the store-branded product provides consistent levels of attributes over time. This information will

be used by consumers to make a diagnosis when they are more price sensitive, less quality sensitive and less risk averse (Erdem & Valenzuela, 2004).

In addition, the country of origin of the store's branded product can leverage consumers' perception of quality and moderate other product attributes. There is also a link between price and perceived quality, and between perceived quality and perceived value (Teas & Agarwal, 2000).

2.4. Logistic regression

Already a fundamental part of any analysis aimed at describing the relationship between a response variable and one or more predictors, Logistic regression can be distinguished from Linear regression by the type of dependent variable assumed. The Logistic regression model is a probabilistic model that assumes a binary or dichotomous dependent variable ($Y = 1$ or $Y = 0$), while Linear regression assumes a continuous dependent variable Y. The assumptions of each model affect its form, but both regression models follow similar principles (Hosmer, 2000).

The simple logistic model assumes the following form:

$$\pi = E(Y = 1\backslash X) = \frac{e^{X\beta}}{1+e^{X\beta}} \qquad (1)$$

π is the probability of the outcome $Y = 1$ to happen for given outcome X. Solving equation (1) for the odds ratio $\left(\frac{\pi}{1-\pi}\right)$ and taking the natural logs we get:

$$ln\left(\frac{\pi}{1-\pi}\right) = X\beta \qquad (2)$$

$ln\left(\frac{\pi}{1-\pi}\right)$ is called the logit, which is the natural logarithm of an odds ratio.

The expression in (2) can be treated much like multiple linear regression. X represents the set of explanatory variables or predictors, and β are the coefficients to be estimated typically by the maximum likelihood method (Peng *et al.*, 2002).

2.5. Decision Trees

Like machine learning sequential models, Decision Trees combine a sequence of tests in which a numerical attribute is compared to a threshold. The rules that this Decision Trees follow are of easy interpretation and closely resemble to human reasoning. The way they work, when a certain value falls into a region that was partitioned, they classify it as belonging to the class, part of the region or node in which more frequently falls (Kotsiantis, 2013).

Decision Trees have been particularly popular since they were introduced electronically as a general-purpose prediction and classification mechanism. This popularity is strongly linked to the ease of understanding provided by this method, as its graphical structure contains a set of attributes that allow users to focus on the most relevant ones. Another facilitator of understanding is that its structure is based on hierarchies, providing the user with the relative importance of attributes. An attribute is more relevant the closer it is to the root (Freitas, 2014).

Chapter 3

Empirical Model

3.1. Classification Model

Market Share information and data was not available at the level required to perform a BCG Matrix, as it would be necessary to have that information at the product level. Without information that deep into the company's (and its competitors) product mix, it was necessary to create a different classification model.

For this new model, with the goal of understanding what are Sonae MC's Star products and considering the definition of star products as products that are relevant for the majority of types of customers, a combination of types of segments of clients created by the company was used, together with a new *Key Performance Indicator* (KPI), to see what products were across the greatest number of possible combinations of segments of customers within the top 20% products, with this new metric.

The *Structured Query Language* (SQL) language was used to extract data from Sonae MC's databases, while the Python programming language was chosen for data analysis.

3.1.1. Data

The process of data collection and preparation for our classification model was divided in the following steps:

(a) filter the object of our analysis – for the first step of data collection, it was defined that the data collected would be regarding only Private Labels of Sonae MC for a period of 12 months – year 2021. Also, only products from the Food & Fresh area were collected. Given the extent

of Sonae MC's product portfolio, it was necessary to proceed with this filtering of the data, since it was a requirement from Sonae MC's team.

(b) Identify and combine the chosen customer segmentation data – after defining the scope of the data, there was a necessary decision on what types of customer segmentation to use. Three of the most popular types of customer segmentation within Sonae MC are "Lifestyle", "Life Stage" and "Price Sensitivity". Lifestyle segmentation divides customers into 8 different groups according to what motivates them to buy. Life Stage segmentation divides customers into 5 different segments according to the stage the customers are at in their lives, while Price Sensitivity segmentation separates customers into 8 groups based on their price sensitivity. The combination of the three types of customer segmentation yields a customer segmentation with 320 groups of customers or segments: 8 x 5 x 8 = 320.

(c) Extract customer and product transactional data – after deciding the data scope, it was necessary to extract all relevant data from the customers and products transactions.

3.1.2. Average expense KPI

To rank the products, a new KPI was created:

$$Average\ Expense = \frac{Average\ Frequency * 0{,}70 + Average\ Ticket * 0{,}30}{2} \quad (3)$$

This new metric uses the average frequency which represents the rate to which a certain product is bought over the period in analysis and the average ticket, which is average value of the ticket a customer gets when making a purchase.
This KPI can be described as average value of each customer, for the entirety of our period, which in this case is the year 2021.

3.1.3. Model creation

For the model creation, each of the products from Sonae MC's food and fresh private labels appeared 320 times, one per new combined segment, as aforementioned. The total values for each attribute were calculated as an aggregate of the entire period. Then, the Average Expense for each of those rows was calculated using the formula in (3). Finally, for each combination of segments, the top 20% products in Average Expense were extracted.

3.1.4. Star products

From a business standpoint, the more a product repeats in these top 20% products across all 320 combined segments, the better the penetration of the product among different segments, and, therefore, the better the performance of the product is. Figure 1 relates the number of times a specific product is repeated across the new combined segments and the accumulated percentage of sales for each product with the number of products.

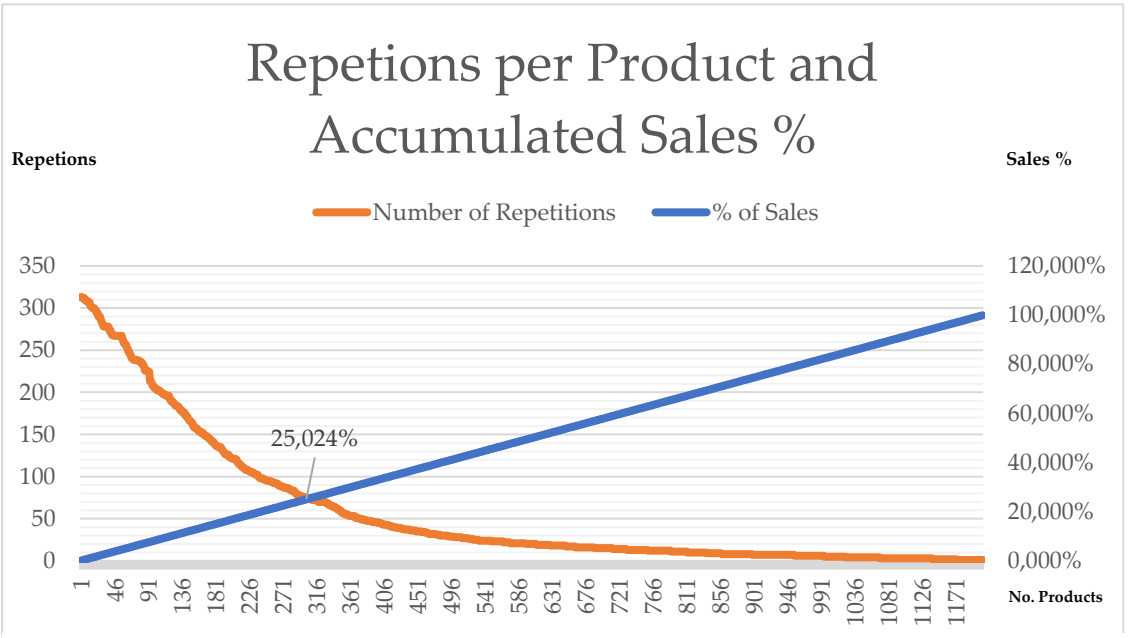

Figure 1 - Repetitions and Accumulated Sales % per Product. Source: Author

Looking at the interception between the two lines in Figure 1, around 295 products repeating 77 times or more across the new segments of consumers considered account for 25% of total sales. These 295 products are classified as Star products. A new variable or column was created, where the value of 1 was assigned to these star products and the value of 0 to the other products. This variable will be the dependent variable for the upcoming prediction models.

3.2. Prediction models

3.2.1 Data description

Dependent variable

As explained above, 295 products were classified as star products and the remaining 938 products as non-star products. With that, the dependent variable is a binary variable, where 1 represents a product being a star and 0 the product being a non-star.

Independent variables

For the independent variables, data regarding customers, products and transactions were used. The literature highlighted that perceived quality and price are key attributes for the success of store brands (Erdem & Valenzuela, 2004). Additionally, from a business standpoint, consumers' attitudes towards a purchase can also be related to their household as a whole, with an emphasis on the family.

With that in mind, the 320 different combined segments aforementioned where grouped in three main groups: "Family", "Quality" and "Price/Promotion". This grouping was based on customers attitudes towards their purchases. The "Family" group consists of customers that when go shopping are thinking about the family, whether they are grandparents buying food for their grandchildren,

or parents looking for a quick and easy solution for their kids snacks at school and families with young adults. As for the "Quality" group of segments, it includes more "premium" seeking clients, who are more financially stable and because of that they are more aware both environmentally and in quality of the products they buy. "Price/Promotion" is the group that defines customers whose main goal in shopping is to get the lower possible price, buying products that are cheaper or by taking closer attention to the promotions. So, what separates these groups of segments is their intention when making a purchase, whether it is a purchase for their family, or if they are looking for quality or price.

For the products variables, the Price Level (as a categorical variable, dividing products into 3 subgroups) of the products was considered, meaning that the products were attributed one out of 3 levels, accordingly to how expensive they are. The top 25% products were assigned as 2, the bottom 25% as 1, and the rest as 0.

Also, the provenance of the products was evaluated, by determining whether a product was manufactured in Portugal or not, as a binary variable.

In what regards the transactions, the percentage of sales in promotion and the average frequency to which a product has been purchased, were the variables introduced.

Variable	Observations	Type
Output Variable (Y)	Is the product a Star?	Binary
Average Frequency	Average of times a product has been purchased, within our period	Numeric
% SPromo	Percentage of sales in promotion	Numeric
PVPL	Price Level 2 to 0	Categorical

Family	% of customers that belong to this group of segments	Numeric
Quality	% of customers that belong to this group of segments	Numeric
Price/Promo	% of customers that belong to this group of segments	Numeric
PT	Is the product Portuguese? 1 – Yes; 0 – No	Binary

Table 1 - Variables considered. Source: Author

Numerical variables' description

Table 2 provides a description of the numerical variables of the dataset being considered:

	Average Frequency	**%SPROMO**	**Quality**	**Family**	**Price/Promo**
mean	1,480	0,524	0,218	0,133	0,081
std	0,450	0,180	0,068	0,031	0,034
min	1,022	0	0	0	0
25%	1,159	0,408	0,174	0,118	0,057
50%	1,316	0,487	0,209	0,136	0,077
75%	1,636	0,606	0,254	0,150	0,101
max	3,591	1,032	0,500	0,333	0,250

Table 2 - Numerical variables description. Source: Author

The mean for the Average Frequency indicates that the products were purchased, on average, 1,48 times during 2021. The Price/Promo was the one with the lowest average percentage of customers, while Quality was the highest.

Correlations between the numerical values

Table 3 shows the Pearson pairwise correlations between the numeric variables.

	Medium Frequency	%SPROMO	Quality	Family	Price/Promo
Medium Frequency	1.000000	-0.091062	0.438828	-0.432276	-0.263963
%SPROMO	-0.091062	1.000000	-0.147073	0.311073	-0.137002
Quality	0.438828	-0.147073	1.000000	-0.885447	-0.768720
Family	-0.432276	0.311073	-0.885447	1.000000	0.383576
Price/Promo	-0.263963	-0.137002	-0.768720	0.383576	1.000000

Table 3 - Numerical variables correlations. Source: Author

From the correlations illustrated in Table 3, it is possible to see that Quality is negatively correlated with Family and Price/Promo, while it seems that there are no further correlations across numeric variables in this data set.

Binary and scale variables

Table 4 is a frequency table for the binary and categorical variables considered:

Classes	# of instances
Star	
0	**937**
1	**296**

PT	
0	**268**
1	**965**
PVPL	
0	**619**
1	**308**
2	**306**

Table 4 - Frequency table for binary and scale variables. Source: Author

From Table 4 it is possible to see that most of the products are non-star products, and Portuguese. It is also possible to confirm that the Price Level was divided into 3 classes with 1 and 2 being close to 25% each, as intended.

3.2.2. Data preparation

Handling missing values

As the first step of data preparation, the data cleaning started with analyzing all the missing values within our data set:

VARIABLE	# OF MISSING VALUES
PT	90
%SPROMO	19
STAR	36

Table 5 - Missing values. Source: Author

Starting with the missing values in the 'PT' variable, all the missing values were assigned the value 1 (meaning the product is Portuguese). This was done considering that products with a foreign origin are correctly classified, because of business practices, therefore, the non-classified products, have a much higher chance of being 'PT', as stated by the business team. As for the %SPROMO and Star variables, the value assigned was 0, meaning respectively that a product had no sales in promotion (as if it had, it would be specified) and that a product was not a star product.

Training and test data split

Since two different machine learning algorithms will be used to make predictions for this data set, the train-test split procedure will allow to estimate the performance of each of these algorithms, allowing for a comparison.

This procedure involves dividing the dataset into two subsets, where the first one will be used to fit the model (the Training subset) and the other one will be later provided to the model, so it makes predictions within that subset, comparing them with the actual values observed (the Test subset).

To set the split, the one parameter needed is the size attributed to each subset in percentage. The split used in this dataset was the most commonly used by the analytical team at Sonae MC, 80/20, meaning 80% of the data was assigned to the Training subset and 20% to the Test subset.

For that purpose, the 'train_test_split' method from the 'sklearn' package was used, by passing the desired percentage (80/20) and setting a 'random_state' which ensures that the results would be reproducible, meaning that each time the code is ran the same examples are included in the respective subset. The final result of this split is as follows:

NUMBER OF INSTANCES	
TRAIN SET	**TEST SET**
986	247

Table 6 - Number of instances per subset. Source: Author

Oversampling

Plotting the class of the products, using the 'Star' variable, we can see that this is an imbalanced data set, with almost three times the number of products not being a star.

In order to balance out the training data set, there are two common approaches: i) the first one is the Random Oversampling, where random examples of the training dataset are duplicated, in the minority class (the one with fewer number of cases); ii) the second one is the Random Undersampling, which deletes some of the examples from the majority class.

Given the dataset, the Oversampling of the minority class, in this case the class '1' with the attribute 'Star', seems to be the best option, as choosing the Undersampling would result in a dataset considerably smaller, given that the minority class is much smaller than the majority one.

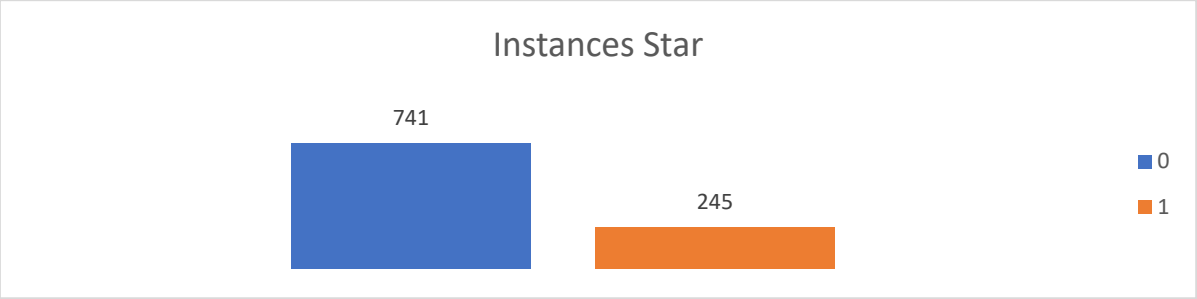

Figure 2 - Instances non star (0) and star (1). Source: Author

3.3. Logistic regression

As opposite to the traditional linear regression, logistic regression is an econometric model appropriated for the study of binary variables. The main reason for logistic regression popularity, and the reason why it is a model appropriate for the problem of this work, lies within the purposes of such econometric model. A logistic regression allows one to predict the probability of a given outcome, categorize predictions and outcomes, and show how the model's predictors affect the odds of the outcome (Hilbe, 2011).

3.3.1. Estimation of the Model

In this MFA, the objective of the logistic regression is to assess how well it can predict the probability of a product being a Star product and to understand how the different predictors associated with the model affect the chances of a product being a Star product. The first model (MODEL 1) logistic regression equation is the following:

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1\, Medium\ Frequency + \beta_2\ \%\ of\ Sales\ in\ Promotion + \beta_3\, Price\ Level + \beta_4\ Quality + \beta_5\ Family + \beta_6\ Price/Promotion + \beta_7\ Portugal \quad (4)$$

As π goes from 0 to 1, the logit $ln\left(\frac{\pi}{1-\pi}\right)$ goes from $-\infty$ to $+\infty$. The logit is linear in X while the probabilities are not. Logistic regression results can be presented showing the beta coefficients or equivalently the related 'odds ratios'. A positive (negative) beta coefficient implies the corresponding 'odds ratio' being greater (less) than one. That is, if the value of the specific predictor increases (decreases) the odds that the product is a start product increase (decrease).

By estimating the model (MODEL 1) for the training subset, the following results were obtained.

Logit Regression Results

Dep. Variable:	Estrela	No. Observations:	986
Model:	Logit	Df Residuals:	978
Method:	MLE	Df Model:	7
Date:	Tue, 26 Apr 2022	Pseudo R-squ.:	0.2471
Time:	14:39:14	Log-Likelihood:	-416.21
converged:	True	LL-Null:	-552.81
Covariance Type:	nonrobust	LLR p-value:	3.162e-55

	coef	std err	z	P>|z|	[0.025	0.975]
const	388.8690	107.898	3.604	0.000	177.392	600.346
Medium Frequency	-2.6932	0.619	-4.350	0.000	-3.907	-1.480
%SPROMO	2.3241	0.592	3.924	0.000	1.163	3.485
PVPL	0.3480	0.133	2.622	0.009	0.088	0.608
Quality	-779.5803	217.336	-3.587	0.000	-1205.551	-353.610
Family	-1164.4451	327.671	-3.554	0.000	-1806.668	-522.223
Price/Promo	-782.1697	217.155	-3.602	0.000	-1207.785	-356.554
PT	-0.3132	0.204	-1.534	0.125	-0.713	0.087

Table 7 - Logistic regression results for the training subset (MODEL 1). Source: Author

For the analysis, a confidence interval of 95% was considered. Therefore, for a variable coefficient to be statistically significant, its p-value must be less than 0.05. From the first summary of the model, the PT variable appears to be the only variable that actually has a p-value greater than 0.05 and therefore should be considered non-statistically significant.

In order to improve our model, we have decided to apply the stepwise search strategy, which uses the Akaike Information Criterion (commonly known as AIC). After applying the stepwise search strategy, we can see that it dropped 'PT', the only variable whose coefficient was not significant. Excluding this variable, the second logistic regression model (MODEL 2) was the following:

$$ln\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1\, Medium\ Frequency + \beta_2\ \%\ of\ Sales\ in\ Promotion + \beta_3\, Price\ Level + \beta_4\ Quality + \beta_5\ Family + \beta_6\ Price/Promotion \quad (5)$$

By estimating the model (MODEL 2) for the training subset, the following results were obtained.

Logit Regression Results

Dep. Variable:	Estrela	No. Observations:	986
Model:	Logit	Df Residuals:	979
Method:	MLE	Df Model:	6
Date:	Tue, 26 Apr 2022	Pseudo R-squ.:	0.2450
Time:	14:54:49	Log-Likelihood:	-417.37
converged:	True	LL-Null:	-552.81
Covariance Type:	nonrobust	LLR p-value:	1.410e-55

	coef	std err	z	P>|z|	[0.025	0.975]
const	390.3499	108.183	3.608	0.000	178.315	602.385
Medium Frequency	-2.7535	0.622	-4.428	0.000	-3.972	-1.535
%SPROMO	2.3067	0.591	3.905	0.000	1.149	3.464
PVPL	0.3484	0.132	2.634	0.008	0.089	0.608
Quality	-782.9058	217.914	-3.593	0.000	-1210.009	-355.803
Family	-1169.9063	328.561	-3.561	0.000	-1813.874	-525.938
Price/Promo	-784.4546	217.711	-3.603	0.000	-1211.161	-357.748

Table 8 - Logistic regression results for the training subset (MODEL 2). Source: Author

This time, for the 95% confidence interval considered, none of the variables had coefficients with a *p-value* lower than 0,05. In other words, all variables have statistically significant coefficients and estimated MODEL 2 is our final model.

$$ln\left(\frac{\pi}{1-\pi}\right)$$
$$= 390.35 - 2.75\ Medium\ Frequency$$
$$+ 2.31\ \%\ of\ Sales\ in\ Promotion - 1169.91\ Family - 782.91\ Quality - 784.46\ Price/Promotion + 0.35\ Price\ Level$$

This final model (estimated MODEL 2) has a very low Pseudo R-squared (0.245), which is the percentage of variation of the dependent variable explained by the predictors. The variable coefficients were all statistically significant, and so, looking at the Average Frequency example, an increase of 1 unit in the Average

Frequency would represent a decrease of 2.75 % in the odds of a product being a Star product.

3.3.2. Confusion matrix (Logistic regression model)

After creating the final Logistic regression model, it is now necessary to understand the quality of the model in its predictions. For that, the model estimated with the Training subset was applied to the Test subset, and a confusion matrix computed for better visual comprehension.

	Predicted No	***Predicted Yes***	***Total***
Actual No	190	6	196
Actual Yes	35	16	51

Table 9 - Confusion Matrix. Source: Author

What is possible to see from the confusion matrix is that out of 196 that were not Star products, 190 (97%) were correctly classified. In the case of the actual Star products, only 16 out of 51 (31%) were predicted correctly. This analysis will serve to compare this logistic regression model with the decision trees model that we will discuss next, in terms of its relative predictive capacity.

3.4. Decision trees

Decision trees, as a general-purpose prediction and classification engine, have been particularly popular since they were introduced electronically. This popularity is strongly linked to the ease of understanding provided by this method, as its graphical structure contains a subset of attributes, allowing users to focus on the most relevant ones. Another understanding facilitator is its

structure, which is based on hierarchies, providing the user with the relative importance of attributes. An attribute is more relevant the closer it is to the root. On the other hand, some of the branches may have to be included in the tree, just in order to preserve its structure, even in those cases where a branch is being associated with an attribute of irrelevant value. This can lead to a misunderstanding of the rating for the user and lead to overfitting.

3.4.1 Variables included

The same variables used in the final model of the Logistic regression (MODEL 2) were used: Average Frequency, % of Sales in Promotion, Family, Quality, Price/Promotion, Price Level.

3.4.2. Preliminary classification tree

With the data already divided into training and testing subsets, and setting a random state so that the model can be replicated in the future without changing the results, the training data were fitted in the 'Decision Tree Classifier' method, of the sklearn tree module. Plotting the tree resulting from that, we get:

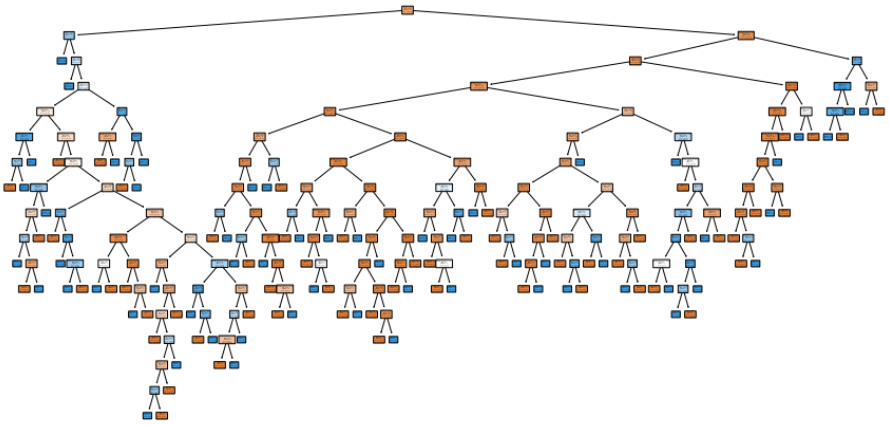

Figure 3 - Preliminary classification tree plot. Source: Author

Visually, the tree is too wide and deep to be easily understood. But with the data now trained for this model, it is possible to predict for the test subset and create a confusion matrix:

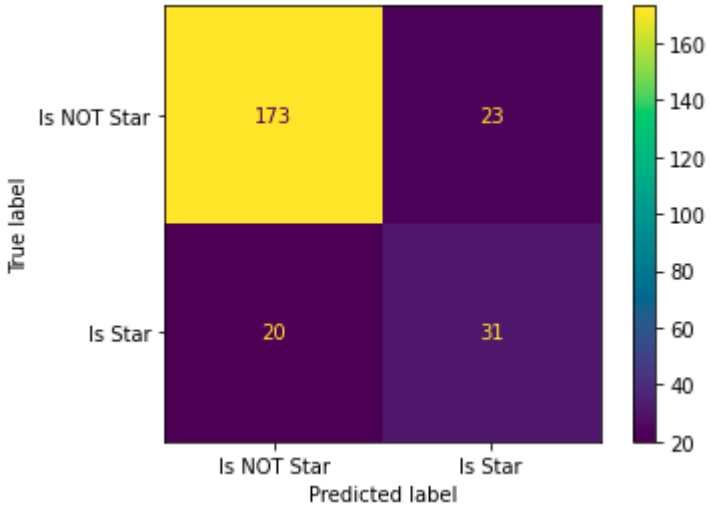

Figure 4 - Confusion matrix (Decision tree model). Source: Author

Analyzing the confusion Matrix, out of the 196 products that were not Star, 173 (~88%) were classified correctly, whereas for the Star products, 31 out of 51 (~61%) were predicted correctly.

One aspect that must be taken into account is that Decision trees are prone to overfitting and therefore it is common to prune the Decision tree to try to solve this problem, while improving visual capabilities, by reducing the width and length of the tree.

3.4.3. Cost complexity pruning

A common pruning tool from the 'sklearn package' is cost complexity pruning, which gives the option to control the size of the tree by parameterizing the cost

complexity parameter, also known as 'ccp_alpha'. The parameter works as follows: the higher the value of 'ccp_alpha', the more pruned nodes are expected. And so, to use this tool effectively it is critical to find the correct pruning parameter, the 'ccp_alpha' value that prunes the tree to the size that provides the best model. This pruning is done for the training subset.

The first step is to extract all possible alpha values from this tree, so that a pruned tree can be created for each of those distinct values.

After that, computing the accuracy of the created trees in a graph, making use of our Training and Test subsets, allows to find the maximum possible alpha value for the Test dataset.

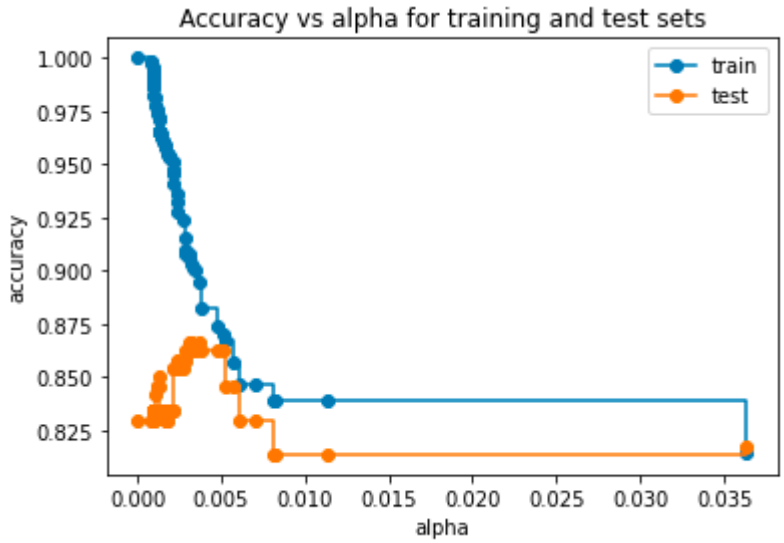

Figure 5 - Cost complexity pruning. Source: Author

From the visualization of Figure 5, the Test data reaches its maximum accuracy around 0.0025 and until 0.005, at which point it starts to drop in value, suggesting that the alpha value should be within this range. For the purpose in view, 0.0025 is considered in this stage.

3.4.4. Cross validation

Even though the previous graph suggested that an optimal value for alpha could be found around 0.0025, if we were not working with a defined random state, another combination of subsets could suggest another optimal value. To cross validate and using the 'cross_val_score' function, different Training and Test subsets were created and applied to the tree.

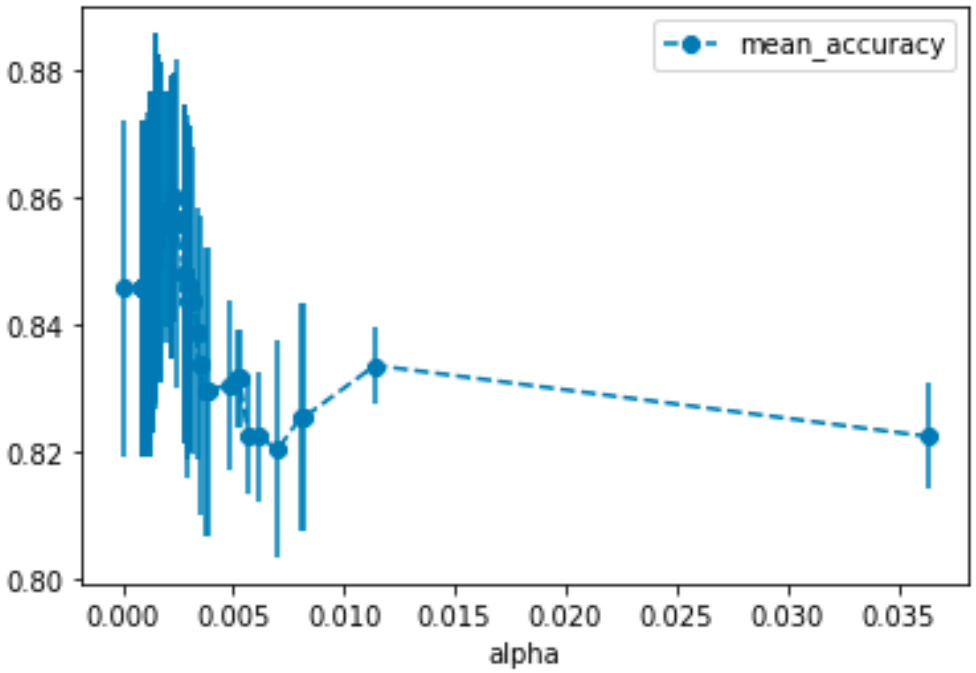

Figure 6 - Cross validation of alfa. Source: Author

As we can see in Figure 6, the same alpha applied to distinct datasets produced different accuracies, which indicates that the value alpha value is sensitive to the choice of datasets.

To find what should be the optimal value of alpha, a 5-fold cross validation was performed, with each candidate value for alpha, storing the standard deviation and means of each attempt. Although somewhat visually confusing, this plot allows to understand that maybe the best alpha value is actually slightly lower

than 0.0025, and so, once the visuals can be confusing, Table 9 with the respective values for alpha between 0.0020 and 0.0025 can be helpful.

	alpha	mean_accuracy	std
34	0.002164	0.857032	0.022291
35	0.002184	0.857032	0.022291
36	0.002188	0.857032	0.022291
37	0.002197	0.857032	0.022291
38	0.002340	0.860068	0.019715
39	0.002359	0.860068	0.019715
40	0.002434	0.856027	0.025885

Table 10 - The optimal value of alpha. Source: Author

From Table 10, an alpha with a value of 0.023 is the one with the best mean accuracy and, therefore, the alpha value to be used for pruning the tree with the cost complexity method.

3.4.5. Pruned tree

The full pruned tree can be found in the Appendix, as it is still very large. An example of a visual cut for understanding the rationale behind Decision trees is done in Figure 7.

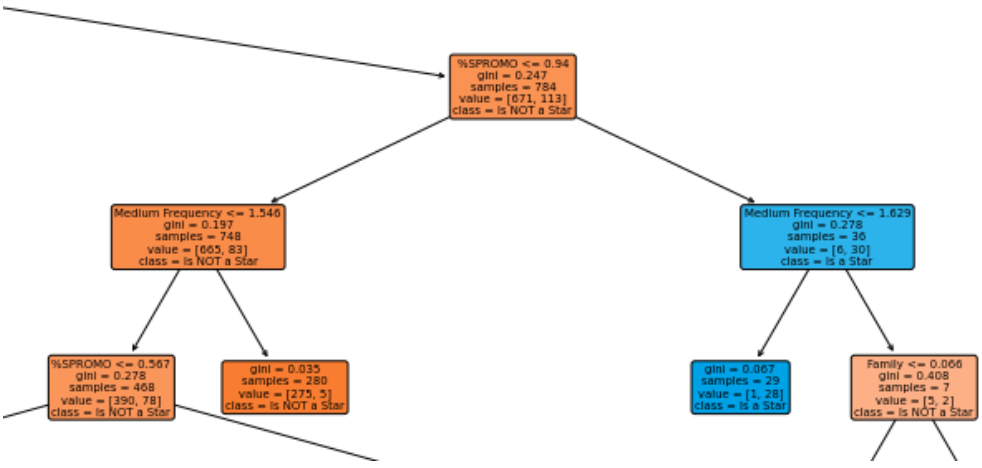

Figure 7 - Visual cut of the pruned tree. Source: Author

Looking at the top node, we used the %SPromo observation to split the data, where all products on the left of it will be less than or equal to 0.94 in %SPromo, while values greater than that will be on the right. The gini respects the gini index of that node, while the sample line represents the number of products in that node, with the value dividing them according to their class. In this first node, there are 784 products, of which 671 are non-star and 113 are star, on the training dataset. The last line represents the majority class within the node, in this case, Not Star has the highest frequency, therefore, it was considered the majority class for the node.

3.4.6. Confusion matrix with the pruned

With the pruned Tree, it is possible to calculate the new confusion matrix and to compare it to the previously obtained for the first Tree:

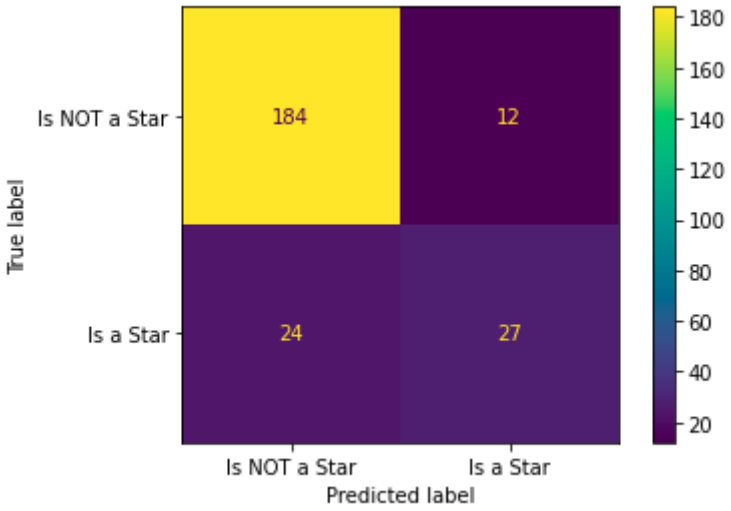

Figure 8 - Confusion matrix (pruned tree). Source: Author

This new confusion matrix (Figure 9) shows that 184 non-Star products were rated correctly, out of 196 (~94%), which represents a nearly 6% improvement over the full-size tree. But in the case of Star products, it decreased by 8%, with only 27 out of 51 products being correctly predicted (53%). From a business point

of view, correctly identifying that a product is not a star was actually more useful, and, in that sense, this pruned tree was more suitable, while avoiding the overfitting problem of a full-size tree.

3.5. Results and Discussion

As the two models were applied a Classification Report was created for both of them, with resort to the 'sklearn' package, which creates a text report showing the main classification metrics. This classification metrics are:

(a) Precision – calculates, among a class, how many samples were correctly classified. Ability to not label positive a negative sample

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

(b) Recall – calculates, among a class, how many samples of a class can be found over the whole number of samples. Ability to find all the positive samples

$$Recall = \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

(c) F1-Score – calculates the harmonic mean between precision and recall

$$F1 - Score = \frac{2 * (Precision * Recall)}{Precision + Recall}$$

(d) Support – number of occurrences of the class

These ranking metrics allow you to compare the predictive ability of the model, so that the most suitable model is found. The computed rating is reported in Table 11.

	Logistic regression			*Decision tree*			
Class	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Support
0	0.84	0.97	0.90	0.88	0.94	0.91	196
1	0.73	0.31	0.44	0.69	0.53	0.60	51
accuracy			0.83			0.85	247
macro avg	0.79	0.64	0.67	0.79	0.73	0.76	247
weighted avg	0.82	0.83	0.81	0.84	0.85	0.85	247

Table 11 - Logistic regression and Decision tree comparison. Source: Author

Looking at Class 0 (Non-star products) in Table 11, the Precision and the F1-Score are higher in the Decision tree model, whereas the Recall is higher in the Logistic regression model.

Regarding Class 1 (Star Products) the Precision is actually higher in the Logistic regression model, whereas the recall and F1-Score are higher in the Decision tree model.

The F1-Score metric, which combines the other two metrics, seems to be indicating that the Decision tree model is better fit to make predictions in this dataset. From a business standpoint, the ability not to label as Star a non-Star product (Class 0) is more important than the ability not to label as non-Star a Star product (Class 1). Relative to Class 0, Precision is higher in the decision tree model than in the logistic regression model.

In this case, where we have an imbalanced dataset, the weighted average is a better metric for model comparison than the macro average, as it calculates the

weighted average of precision, recall and F1-Score between classes, respecting how many samples there are in each class. Looking at the weighed averages of the models, the Decision tree seems to be better fit in all the metrics.

Chapter 4

Conclusion

It was the intention of this work to be able to identify the class of products that are Star products for Sonae MC and to assess which customer and product attributes increase or decrease the probability of a product being a Star product.

For the first objective, a model was created based on the top products in terms of Average Expense, for the different types of customer segments. This rating labeled 296 products out of 1233 as "Star". These 1233 products, already labeled, were then divided into training and test subsets, so that the predicting capability of the models to come could be evaluated and compared. The training subset also seen a random oversampling of its minority class, so the data would be better balanced.

For the second objective, to evaluate which attributes increase or decrease the probability of a product being a star, logistic regression and decision tree models were used. Both models saw the training data fitted to them, being trained to predict the results for the Test data subset. Probably due to the lack of available variables, the logistic regression model had very low explanatory power, but it served to see how the possible predictors to be collected interfered with the probability of a product being a star and to make predictions. Statistically significant variables in the logistic regression model were used in the decision tree model. To avoid overfitting the data, the latter was pruned.

Both Trained models were used to predict for the Test subset whether their products were star or not and the results plotted in confusion matrices. Using classification metrics applied to both confusion matrices, the data tree model seems to be a more suitable model for this type of prediction, since its F1-Score value is higher. However, depending on the context, accuracy or recall may be of greater importance in selecting the model, and may lead to a different choice.

When the weighted averages are computed, Decision tree is a better fit in all metrics. One of the main interesting findings for the business is to deep dive on the false positive products, since they possess qualities (product characteristics and/or transactional behavior) that classify them as Star products although the company did not have this information beforehand

Initially, this work intended to use the Growth-Share or BCG Matrix, created by the Boston Consulting Group, as a model for classifying Sonae MC's star products. But during the data collection phase, the author became aware that available data was narrower in scope than desired regarding the information available on different products. The market share data was only available at products' category level, and not at product level. Therefore, the BCG had, for the time being, to be abandoned. If made available market shares at product level, doing the BCG Matrix would actually be beneficial, even to compare it to the classification model used in this work.

In addition, there were some data access issues, with long periods with no data available to work with. These problems with data access did not facilitate the collection of new variables that could be inserted in the models, especially in the Logistic regression model, in order to increase its explanatory power. There is potential in exploring this analysis with more and newer variables, especially with variables related to the products and clients. There is also value in the possibility of applying different econometric models to this dataset. As logistic regression and Decision tree models are simpler and more easy use in an organizational context, other more complex techniques were left out.

The approach to handle the imbalanced dataset resorted to the random oversampling, where random examples of the training dataset were duplicated in the minority class. However, there is a downside to this technique since doing this random oversampling of the data might lead to overfitting. In that regard, a future possible approach would be through the SMOT (Synthetic Minority

Oversampling Technique), which similarly adds data to the minority class. However, in this technique, instead of duplicating, synthetic examples are created. This different technique, although somewhat similar to the one that was used, allows to collect more information, since there is a wider range of representations of an object.

Finally, the scope of the dataset was actually reduced. This was necessary for data processing, since the range of products in the organization was very wide and is desirable on a proof of concept approach such as this one. But for future work, either using different scopes, or making it a broader scope, may allow a better understanding of Soane MC's product portfolio.

References

Adams, R., Bessant, J., & Phelps, R. (2006). Innovation management measurement: A review XX Innovation management measurement: A review. International Journal of Management Reviews, 8, 21–47.

Collins-Dodd, C., & Lindley, T. (2003). Store brands and retail differentiation: the influence of store image and store brand attitude on store own brand perceptions. Journal of Retailing and Consumer Services, 10(6), 345–352. https://doi.org/10.1016/S0969-6989(02)00054-1

Erdem, T., Zhao, Y., & Valenzuela, A. (2004). Performance of Store Brands: A Cross-Country Analysis of Consumer Store-Brand Preferences, Perceptions, and Risk. Journal of Marketing Research, 41(1), 86–100. https://doi.org/10.1509/jmkr.41.1.86.25087

Freitas, A. (2014). Comprehensible classification models: a position paper. SIGKDD Explor. Newsl. 15, 1 (June 2013), 1–10. https://doi.org/10.1145/2594473.2594475

Hax, A. C., & Majluf, N. S. (1983). The Use of the Growth-Share Matrix in Strategic Planning. Interfaces, 13(1), 46–60. https://doi.org/10.1287/inte.13.1.46

Hilbe, J. M. (2011). Logistic Regression. International encyclopedia of statistical science, 1, 15-32.

Kotsiantis, S.B. (2013). Decision trees: a recent overview. Artif Intell Rev 39, 261–283 https://doi.org/10.1007/s10462-011-9272-4

Lindon, D., Lendrevie, J., Lévy, J. Dionísio, P. Rodrigues, J. (2004). Mercator, X. X. I.: Teoria e prática d<fo Marketing, 10. a edição. Lisboa: Dom Quixote

Peng, C., Lee, K., & Ingersoll, G. (2002) An Introduction to Logistic Regression Analysis and Reporting, The Journal of Educational Research, 96:1, 3-14

Sethuraman, Raj. (2001). What Makes Consumers Pay More for National Brands than for Store Brands - Image or Quality?. Marketing Science Institute Paper Series. 00. 10.2139/ssrn.310883.

Teas, R. K., & Agarwal, S. (2000). The Effects of Extrinsic Product Cues on Consumers' Perceptions of Quality, Sacrifice, and Value. In Journal of the Academy of Marketing Science (Vol. 28, Issue 2).

Yagüe, María & Rubio, Natalia. (2009). The Determinants of Store Brand Market Share - a Temporal and Cross-sectional Analysis. International Journal of Market Research. 51. 501-520. 10.2501/S1470785309200700.

Appendix

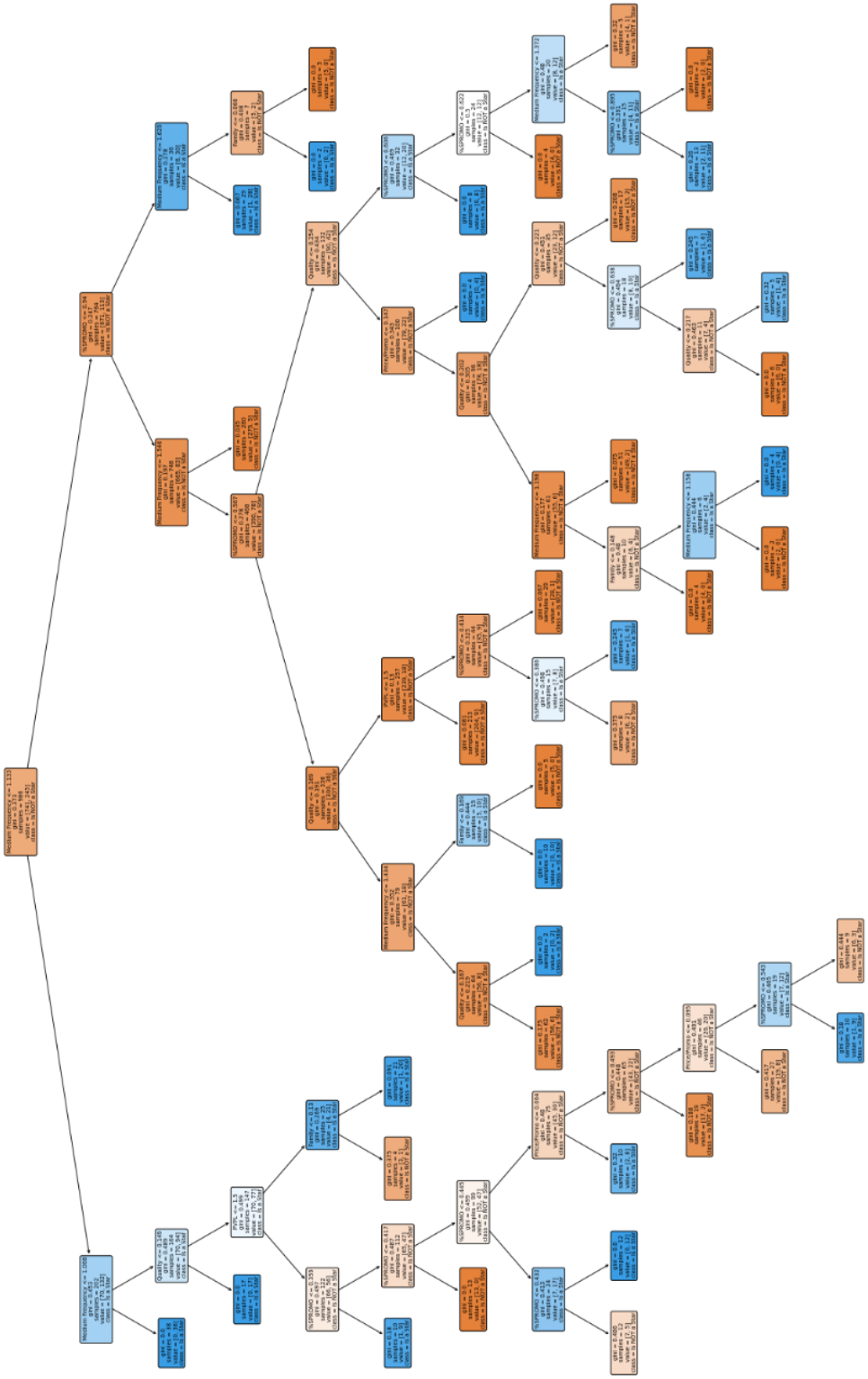

Figure 1A - Full-size Tree. Source: Author