Final Report H-Farm Project

209590 - Business Analytics - H-Farm

Group 2: Aliya Davletshina, Angelantonio Dilengite, Rocco Gazzaneo & Jakob Schlierf

Contents

List of Figures

List of Tables

Chapter 1

Executive Summary

The insurance industry of today faces enormous challenges in how insurance products can be sold in the future. The traditional sales channels, relying mainly on direct contact with agents, work less and less, not only in the current environment impacted by COVID-19, but also on a more long term scale. This creates the challenge of how insurance companies can effectively target customers. One approach for this could be the reliance on data analytics. In the present case, supermarket shopping data was used to try to understand whether an increased likelihood to buy a specific insurance product could be identified. For this, a multitude of insurance products was considered, before ultimately being reduced to four contenders.

From the data perspective, the chosen approach was to categorize the data to reduce the complexity from 53,157 unique products to only 38 categories. This was done part manually, part automated, resulting in approximately 85% of items categorized. Following the categorization, the data was clustered using several approaches as well as techniques. First, general behavior was clustered using both DBSCAN and K-Means. These data-driven approaches however failed to provide enough actionable insight to proceed, so further analysis was determined to be necessary.

In an ensuing theory-driven approach utilizing several assumptions, customers were scored using linear combinations of average consumption level of each food category. This approach yielded more promising results, and was therefore continued. It quickly emerged that 'healthy' eating behaviors would pose an interesting cluster to further investigate. This was done initially using K-Means both on the category space, as well as on the macronutrient composition. After these proved to be unsuccessful, a gradient-descent based method was used to maximize the difference in health score. This proved to be valuable in creating a 'health' score along which customers were separated into 'healthy' and 'unhealthy' clusters. Based on this, the theory that customers with healthy eating habits were more likely to buy longevity insurance was created.

To test this theory, a survey was designed, around 120 responses collected and the results evaluated. These seem to confirm the hypothesis. To further validate the assumptions, an experiment is proposed which could alleviate some of the inherent flaws present in a survey.

Chapter 2

Problem Statement

The channels for selling insurance have remained broadly similar since the early 1900s[1]. Even though introduction of online purchasing throughout the 1990s has added another way through which insurances can be purveyed, this has only been true for some insurance products, that customers seemingly feel more comfortable buying online.

First among these insurances, are those that are the most standardized, such as automobile insurance or term life insurance, where there nowadays even exist several companies whose business model is built on comparing and recommending these insurance products to customers online.

Yet the majority of insurance products are still sold through the traditional channels of either directly-employed agents, independent agents, or intermediate partners such as banks [2]

In the US life insurance market in 2019 for example, approximately 89% of the total market share for life insurance was sold through either independent (53%) or affiliated agents (36%) [3].

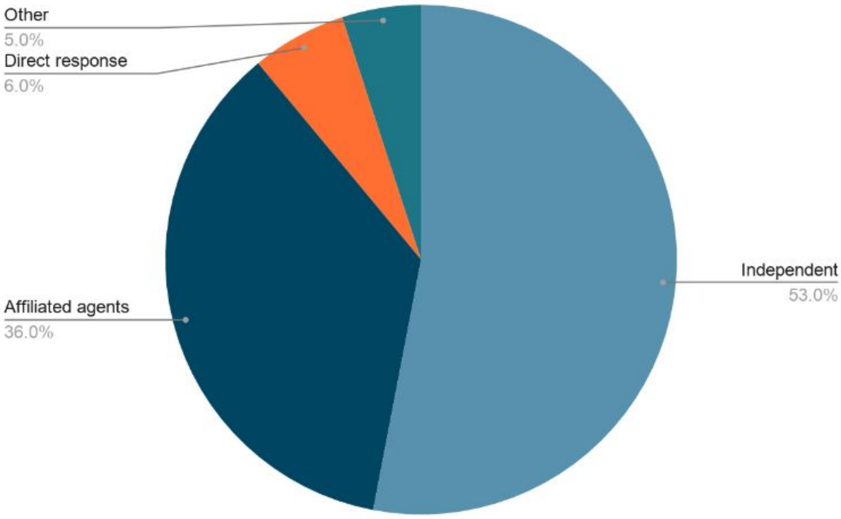


Figure 2.1: Life insurance distribution channels in the US 2019 by market share (in %)

Even though this is certainly shifting to a more online-based-distribution (not the least

[1] Background on: Buying Insurance. (n.d.).

[2] Kaesler, S., Leo, M., Varney, S., & Young, K. (2020).

[3] Insurance Information Institute. (2020)

due to COVID-19), agents, as well as the relationships and personalization options they bring to the table, will still play an important role[4].

For insurance providers, this presents an enormous challenge: They have to increase their digital business while also retaining some of the values that have made the personal distribution channels of the past so successful[5].

One way this could be approached is through personalizing and targeting specific products to customers showing a strong likelihood to be interested in them through the use of data analytics.

Insurance companies have long used sophisticated data analytics algorithms in underwriting, this is the core of every insurance's business model: Estimating risks for insured events and managing the expected amount in insurance payments through collecting adequately sufficient premiums[6].

But until recently, limited computing resources, as well as a general lack of data presented limits as to how much analysis was done to target and actually sell insurance. While these are not issues any longer, the relative scarcity of true data experts, the lack of connectivity between data analytics and business acumen, low integration of analytics into the overall business processes as well as the missing of a clear, company-wide analytics strategy have remained challenges that stopped an increase in the use of analytics[7].

Nevertheless, those challenges can be overcome. The following project shall serve as an example for how data analytics in the insurance industry can help with both market segmentation and targeting. Its goal was to use purchase data recorded at two separate supermarkets to create a testable theory according to which customers could be offered insurance products, resulting in a higher likelihood of sales. To achieve this, two simultaneous actions were undertaken:

The preliminary analysis of the provided data, which shall be discussed in the following chapter, as well as the identification of customer profiles for specific insurance products (which will be discussed in detail in the chapter following the preliminary data analysis).

Consequently, these two groups were evaluated against each other to find the most stable and presumed to be most profitable matching. Out of this match, a theory in conjunction to a testable hypothesis was developed. This hypothesis served as the starting point for a survey that was consequently run. The results as well as suggested further analysis will be described in the remainder of this analysis.

[4] Kaesler, S. et al. (2020)

[5] Kaesler, S. et al. (2020)

[6] Collignon, R., Dekkers, J., van Veen, J., & Scheeren, D. (2018)

[7] Collignon, R. et al. (2018).

Chapter 3

Insurance Product Selection

3.1 Selection Criteria

Within this chapter, the process that the team followed to select the insurances to be analyzed, as well as the insurances that were taken into further consideration will be introduced. This allows the authors to explain why specific insurances were deemed more likely to be successfully matched to supermarket purchase data. Lastly, another factor that was considered, was the projected market opportunity for the insurance that would be targeted to identified customers. Early in the project, the team decided to focus on a low number of insurance products. This was done for several reasons:

1. Since the team started with a wide range of insurance products, reducing these down to only a few would allow it to be careful when selecting finalists, presumably ensuring a more successful match.

2. From this limited number of customer profiles, specific key identifiers of those customers could be selected. Searching for a limited number of identifiers within the categorized data should prove to be easier yet more effective.

3. A low number of insurance products would increase the depth that the team could explore the validity of the matches proposed. This means that not only could a survey be proposed and run, further analysis following the results could be performed.

Moreover, another key decision was to focus on insurances whose customer profile would be comparatively simple to identify within the presented data. While this might appear to be a fairly obvious conclusion, from the team's perspective, this criterion alone eliminated several insurance products, such as insurance products for personal products not typically bought at supermarkets, such as computers, boats, bicycles. While a subset of owners of these products could probably be identified (for example automobile owners through the purchase of automobile-related products such as car cleaners, automotive air purifiers or the like), a customer profile for which a large addressable market could be identified using just the supermarket data at hand.

Starting with a curated list of 30+ different insurance products, categorized by the object or event to be insured, customer profiles were developed to understand who would buy this insurance. Using these profiles, the most sellable solutions identified to be suitable to be targeted were: Life insurance, Health insurance, and Longevity insurance[8]. Each of those shall be shortly introduced to the reader in the following.

3.2 Overview of selected insurance products

3.2.1 Life insurance

Life insurance is a contract between an insurer and a policyholder in which the insurer guarantees payment of a death benefit to named beneficiaries when the insured dies[9]. It supports the family of the departed by covering at least funeral costs, and depending on its coverage also some living expenses like mortgage payments, outstanding debts, taxes, child care and future education. Previous studies have shown that certain life events are reliable predictors of the likelihood to purchase life insurance. Specifically, individuals who got married, had a child, became a homeowner, or retired were more likely to have purchased life insurance relative to the average person. By contrast, renting a home and/or being single decreased the likelihood of having purchased life insurance[10].

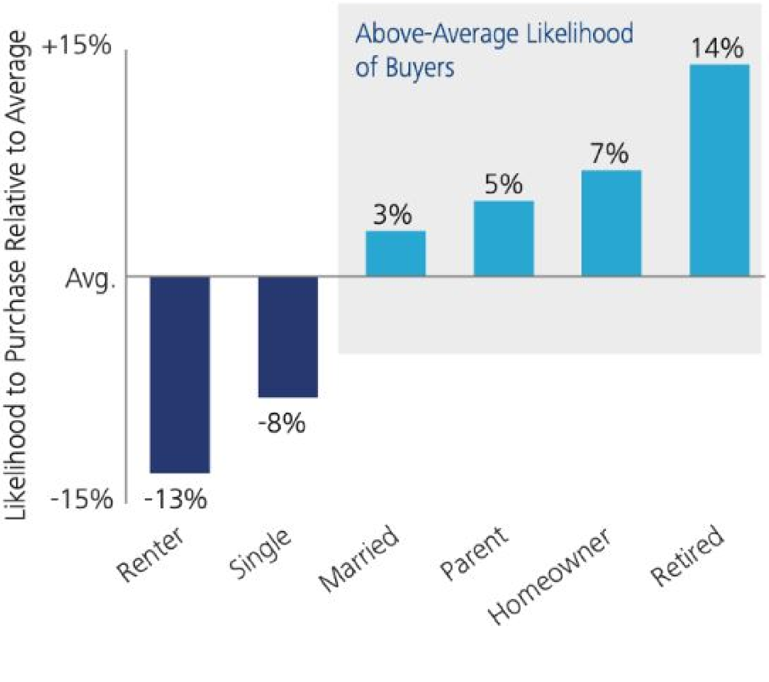


Figure 3.1: Likelihood of Life Insurance Ownership by life event vs average

[8] McMaken, L. (2020, August 28).

[9] Fontinelle, A. (2020, November 16)

[10] Sharps,K., Hitsky, D., Hodgins, S., Ma, C. (2015)

3.2.2 Health insurance

Health insurance is a contract between a policyholder and an insurer who agrees to cover the whole or a part of the risk of a person incurring medical expenses. Healthcare can be dramatically expensive when a health policy is absent in the insurance portfolio of the injured. According to a survey made in 2019 by the American Journal of Public Health, medical bills and income losses due to illness were the major cause of personal bankruptcy filing in the US between 2013 and 2016[11]. Therefore, in countries with a more expensive healthcare system, health insurance should be considered to be an important safeguard against vast amounts of debt.

Regardless of a country's health policy, owning some form of health insurance constitutes an important buffer to disease-induced poverty. Irrespective to nationality and personal attitudes, buying health insurance is something that mostly depends on wealth. This was shown by a study over the Malaysian peninsular population, where a collection of data about the personality, health status and socioeconomic conditions of public sector workers led to the conclusion that income is the only reliable predictor of health insurance purchase, notwithstanding health risks exposure[12]. Therefore, one could expect that the probability of buying this type of insurance is higher for wealthier individuals.

3.2.3 Longevity insurance

Also known as Qualifying Longevity Annuity Contract (QLAC), longevity insurance entitles the policyholder to receive predefined payments for life starting at a pre-established future age. It is comparable to a "reverse-life insurance" because a large premium is paid upfront while annuities are collected once a certain turning point is reached, which here is an age threshold rather than death of the insured.

Longevity insurance is not something anyone can afford, but it surely is a nice-to-have among the assets of every retiree. Apparently, willingness to buy this type of insurance is strongly correlated with one's expectations regarding one's lifespan. The more optimistic these expectations are, the higher is the likelihood to buy a longevity insurance.

[11] Himmelstein, D. U., Lawless, R. M., Thorne, D., Foohey, P., & Woolhandler, S. (2019).

[12] Husniyah, A. R., Norhasmah, S., & Mohamad O. A. (2017)

Chapter 4

Analysis of Carrefour Data

4.1 Categorization

An overview of insurance products and their target audience gave an idea of what types of customers should be identified based on the data analysis. Therefore, it was decided to leverage a data-driven approach and look for different groups of target audience through clustering customers according to their purchasing habits. Consumer behaviour is believed to be highly determined by an individual's lifestyle. A number of research have been carried out to study how characteristic patterns of living can influence consumers' motivation to purchase products and brands. For example, Verain, M., Sijtsema, S., & Antonides, G. (2016) identified consumer segments based on importance consumers attach to a range of food-category attributes and concluded that the identified segments differ in their general food choice motives as well as in their perception of synergy between healthiness and sustainability of food products [13].

Thus it can be reasonably assumed that analysis of shopping baskets indeed can tell a lot about customer lifestyle and, consequently, about their propensity to buy a particular kind of insurance. The dataset at disposal had the following columns:

- id
- mall: 1 or 2
- date: date and time of purchase
- customer
- desc: short description of a product
- net am: price of a product item
- n unit: quantity of a product purchased

Since the data had 580,437 unique 'id' values (which assumably stand for a unique visit) and 65,934 'customer' values (which are believed to represent a unique loyalty card), it may be assumed that on average a person had about 8 visits to the shop within the period when the data had been collected. Further analysis showed that most customers indeed had multiple

[13] Verain, M., Sijtsema, S., & Antonides, G. (2016).

visits to the shop, however, time trends could not be taken into consideration because of the anomalies related to customer loyalty cards described in the next paragraph. Customer profiles identification consisted of the following steps:

Preliminary data analysis to check for missing values and outliers.

Exploratory analysis of the data revealed that there were items with zero or negative price, which could be either system errors or promotion of some products or a return. At any rate, these outliers accounted for only 0.81% of the data, therefore their elimination can be considered negligible, given the dataset with more than 6.7 million rows.

Furthermore, a few anomalies were observed: there were 17,383 customers who made more than 10 visits in less than 24 hours and one customer made 49 visits in a single day. One plausible explanation might be that a single loyalty card was shared among multiple people, in this particular case it could be a corporate loyalty card.

Creating categories of products to reduce complexity

One approach to clustering customers could be to treat the data as transactional data and apply an appropriate algorithm. However, the data had 53,157 unique products, therefore proceeding with such high-dimensional data would be a cumbersome and time consuming process. Therefore, another approach was taken to address the problem of high-dimensionality.

First, the 2000 most frequently purchased products in the dataset were classified into 38 different categories (e.g. fruits, vegetables, red and white meat, fish, junk food, ect.). According to preset classification rules, each product could fall into no more than five categories. As a result, 2000 products got labels of the categories they belonged to and the words used in product description were set as keywords for a corresponding category. Additionally, other relevant keywords were added for each category to increase classification precision rate.

Classifying all the products in the dataset into the categories.

The rest of the dataset was labeled based on the following logic: classify the product into the category if its description contains one of the most frequent keywords of a category. This approach resulted in 82% of products being labeled. Thus, on average every category had around 180 products.

Category	Number of products	Percentage
Dairy	266	11.7
Vegetables	175	7.7
High in sugar products	170	7.4
Refined	156	6.9
Red meat	144	6.3
Drinks	143	6.3
Junk	124	5.4

Table 4.1: 7 largest categories out of total 38

Creating an estimator for a margin of classification error.

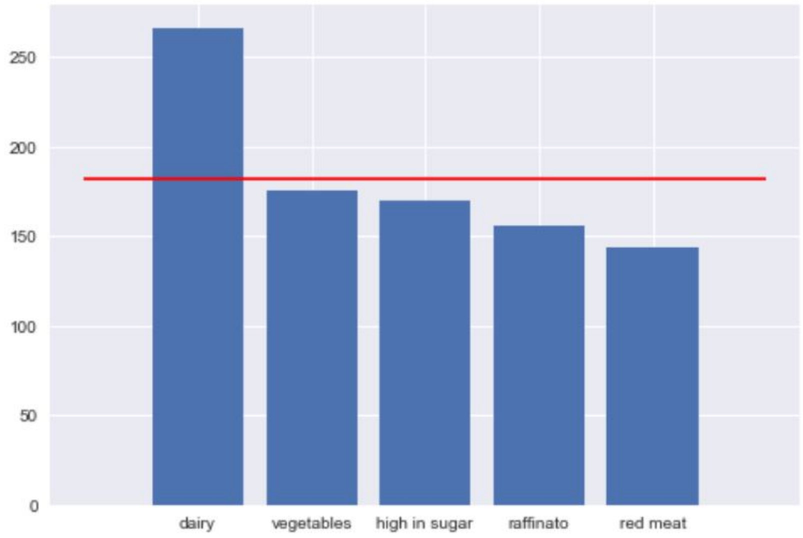


Figure 4.1: Distribution of products across categories

Since it could be the case that the same keyword appeared in different categories (e.g. ‘aceite’ was one of the keywords for multiple categories as there were ‘tortas de aceite’ in ‘processed/refined food’, ‘rosquillas aceite’ in ‘high in sugar’ and ‘atun claro aceite’ in ‘fish’ categories), a simple metric was calculated to serve as a rough estimate for the classification precision:

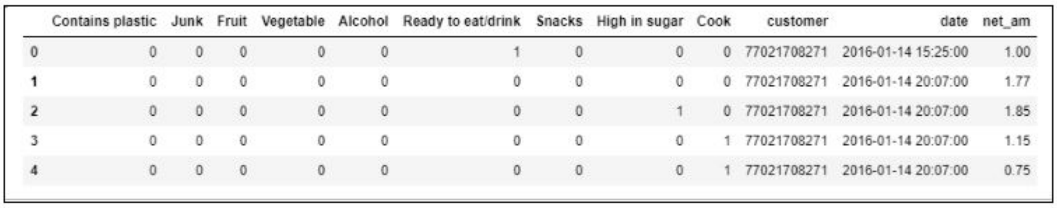
$$\text{Score} = \frac{\text{Number of products which were classified into more than 5 categories}}{\text{Sum of products in all the categories}}$$

Using this score the margin of error was estimated to be around 7.8%.

4.2 Clustering

4.2.1 Metrics to analyze customer purchasing behavior

The categorization of the products augmented the dimensionality of the dataframe, increasing the number of variables to 49, with 38 categories and 11 initial columns. The table below represents a snapshot of the processed dataset.



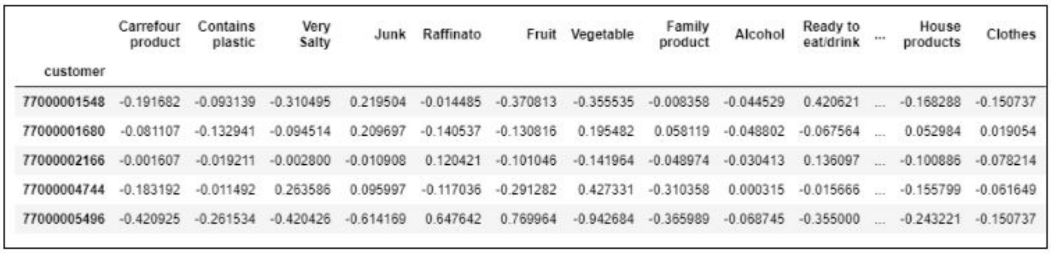
	Contains plastic	Junk	Fruit	Vegetable	Alcohol	Ready to eat/drink	Snacks	High in sugar	Cook	customer	date	net_am
0	0	0	0	0	0	1	0	0	0	77021708271	2016-01-14 15:25:00	1.00
1	0	0	0	0	0	0	0	0	0	77021708271	2016-01-14 20:07:00	1.77
2	0	0	0	0	0	0	0	1	0	77021708271	2016-01-14 20:07:00	1.85
3	0	0	0	0	0	0	0	0	1	77021708271	2016-01-14 20:07:00	1.15
4	0	0	0	0	0	0	0	0	1	77021708271	2016-01-14 20:07:00	0.75

Figure 4.2: Snapshot of the dataset with created food categories

Since the analysis was primarily concerned with comprehending customer behavior, the next step was to group the dataset by customer that could be identified by a unique loyalty card number. Grouping the dataset by customer implies the establishment of a metric which effectively summarizes data about consumption of food categories for every customer visit. Having considered multiple metrics that could describe customer behavior, the two most suitable ones were selected:

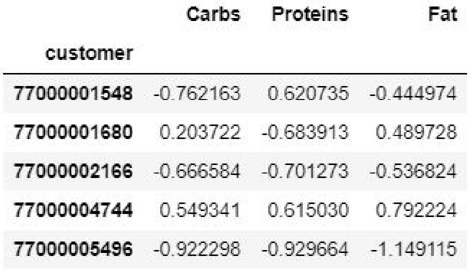
1. Average consumption by category for each customer
2. Macronutrient composition of the average consumption by category for each customer

As a result of this grouping operation, each row of the new datasets represented a unique customer.



	Carrefour product	Contains plastic	Very Salty	Junk	Raffinato	Fruit	Vegetable	Family product	Alcohol	Ready to eat/drink	...	House products	Clothes
customer													
77000001548	-0.191682	-0.093139	-0.310495	0.219504	-0.014485	-0.370813	-0.355535	-0.008358	-0.044529	0.420621	...	-0.168288	-0.150737
77000001680	-0.081107	-0.132941	-0.094514	0.209697	-0.140537	-0.130816	0.195482	0.058119	-0.048802	-0.067564	...	0.052984	0.019054
77000002166	-0.001607	-0.019211	-0.002800	-0.010908	0.120421	-0.101046	-0.141964	-0.048974	-0.030413	0.136097	...	-0.100886	-0.078214
77000004744	-0.183192	-0.011492	0.263586	0.095997	-0.117036	-0.291282	0.427331	-0.310358	0.000315	-0.015666	...	-0.155799	-0.061649
77000005496	-0.420925	-0.261534	-0.420426	-0.614169	0.647642	0.769964	-0.942684	-0.365989	-0.068745	-0.355000	...	-0.243221	-0.150737

Figure 4.3: Snapshot of resulting dataframe when grouping through metric 1



	Carbs	Proteins	Fat
customer			
77000001548	-0.762163	0.620735	-0.444974
77000001680	0.203722	-0.683913	0.489728
77000002166	-0.666584	-0.701273	-0.536824
77000004744	0.549341	0.615030	0.792224
77000005496	-0.922298	-0.929664	-1.149115

Figure 4.4: Snapshot of resulting dataframe when grouping through metric 2

4.2.2 General behavior identification

4.2.2.1 DBSCAN and K-means on average consumption of categories

After grouping the dataset a general cluster analysis on customers was conducted to see whether there were groups of people with similar purchasing behaviour. For this purpose KMeans and DBSCAN (Density-based spatial clustering of applications with noise) were chosen as the algorithms most suitable for the problem. The first one, that is a partitioning algorithm (assigns a label to all data points), works well on large datasets (as the one that was at hand) and is robust to outliers, although it requires the number of clusters to be set prior to clustering. DBSCAN (a clustering algorithm) on the other hand, builds clusters based on dense regions and therefore can create clusters of different shapes and sizes, possibly excluding some data points from the cluster, which seemed highly likely given the dataset. Before proceeding with algorithms the data was subject to further processing: first, all the categories were standardized to have a mean of zero and a standard deviation of one. Next, given the distribution of the categories was skewed, a square root was taken to bring it closer to normal distribution. The figure below shows how the DBSCAN clustering algorithm efficiently divides the space among consumers that consume more or less Vegetable and Junk food on average.

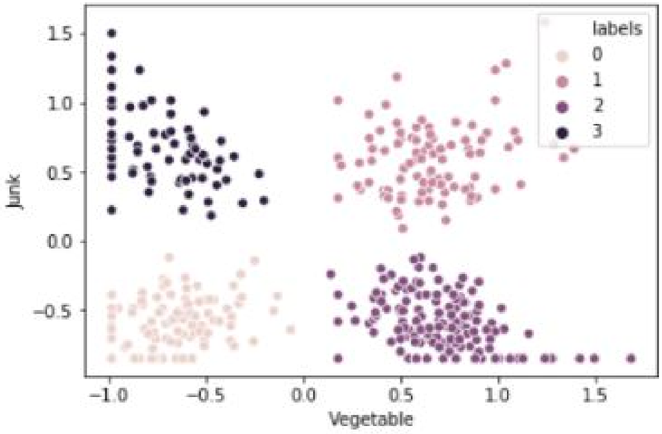


Figure 4.5: DBSCAN clustering on Junk and Vegetable categories

Although the clustering algorithm successfully clustered data points, it was not giving any meaningful insights about customer attributes, which could only be inferred by changing the space where data points are represented and applying the clustering algorithm on that space. K-Means algorithm produced somewhat similar results, therefore it was decided to proceed with another approach.

4.2.2.2 Linear combinations of average consumption by category

The second approach relied heavily on a few prior assumptions regarding customer profiles. Some of these assumptions were built on a number of previous research works while others were based on common sense. For example, the findings of a past study showed that income has a positive and significant relationship with life insurance ownership[14]. Another study claims that married and employed people are more likely to purchase private insurance than their counterparts [15]. Thus, different customer characteristics such as income level, marital- & employment status and a number of others are likely to be correlated with the propensity to buy a particular insurance product. Therefore, instead of building customer profiles right away, the plan now was to identify such customer characteristics and then build profiles through a combination of those. Table 5 shows an abstract of a list of characteristics to identify and corresponding ways to do so. In general, for each characteristic either a special score was calculated as a combination of certain categories of products or a rule was created to assign the right label.

[14] Smith, T. A. (2019).

[15] Liu, T., & Chen, C. (2002)

Customer characteristics	Categories used to infer the characteristics
income level	Median price of the item per bucket
uses a car	1 if number of items bought per visit is more than 30, 0 otherwise
cooking / not cooking	A linear combination of 'Fish', 'White meat', 'Red meat', 'Legumes' and 'Egg' categories.
has a family	A linear combination of 'Family products' and products 'For children'.
environmentally friendly	A combination of 'one-time use', 'contains plastic', 'packed in glass or aluminium' categories.
employed / unemployed	1 if purchase is made on weekend or after 5 pm on weekday, 0 otherwise
healthy eating habits	A weighted sum of 'high sugar', 'junk food', 'vegetables' and 'fruits'.

Table 4.2: Characteristics of customers to be identified

Following these rules, customers were assigned either a label or a score for a particular characteristic. Next, k-means clustering was performed using the scores to find out the number of distinct groups for each characteristic. The following graphs represent the resulting clusters for some of the customer characteristics.

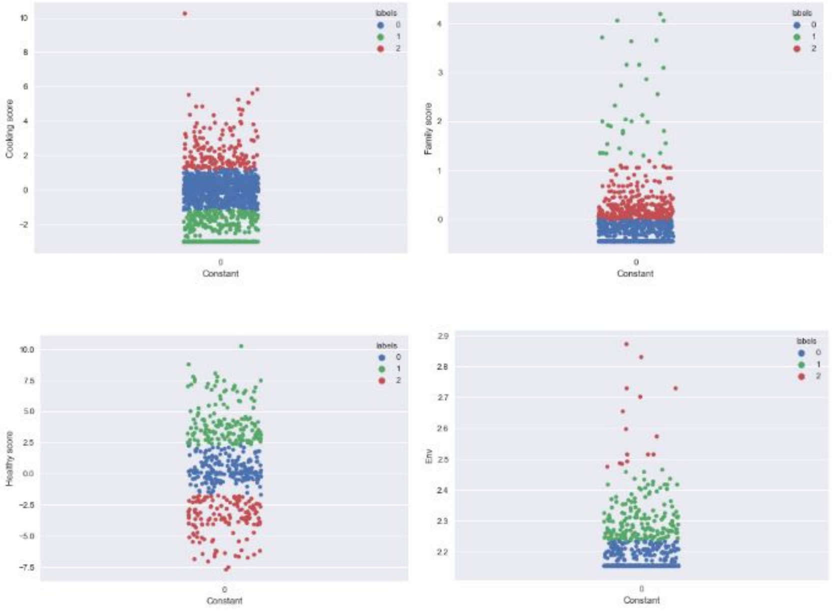


Figure 4.6: Results of K-Means clustering based on customer characteristics

4.2.3 Health eating behavior identification

A closer look into created clusters revealed that identification of people with healthy eating habits produced the best result in terms of the difference between clusters created. Such a result was quite expected as eating habits are more correlated with health than with any other characteristic of an individual. Therefore, it was more reliable to use Carrefour data to recognize a group of 'healthy' people rather than speculating, for instance, about them owning a house or a car.

4.2.3.1 K-Means on the category space

Although clustering results looked promising, still there was further room for improvement since the goal was to find the way to get as distinct clusters as possible. First, a subset of food categories was chosen to evaluate to what extent one's eating habits could be considered healthy. The selection relied primarily on the team's assumptions regarding the food constituting a healthy diet. The initial attempt made to distinguish between consumers with healthy eating habits (hereinafter referred to as 'healthy' people) and those with unhealthy eating habits (hereinafter referred to as 'unhealthy' people) was to apply K-Means on the space of the relevant food categories ('Fruit', 'Vegetable', 'Red meat', 'Fish', 'Egg', 'Snacks', 'Junk', 'Ready to eat/drink', 'Alcohol', 'Contains plastic'), *by including our prior opinion.* The way this is done is not letting the algorithm randomly select the centroids, but *setting the cluster centroids based on prior opinion*, by initializing the centroids at the highest value for food categories which 'healthy' people supposedly consume more, and at the minimum for the ones 'healthy' people consume less. The following figure suggests the algorithm yielded a hardly satisfactory result: although the algorithm identified a group of individuals eating on average more fruit and vegetables, the same group appeared to consume more junk food and snacks.

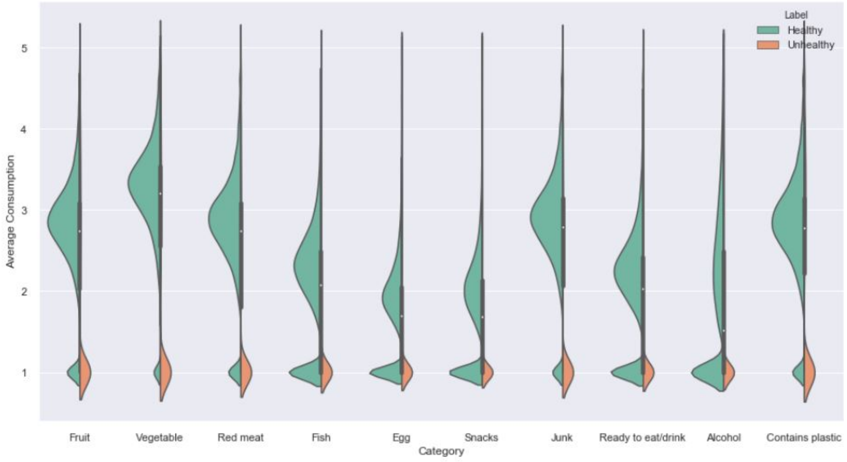


Figure 4.7: Comparing groups produced using K-means on the category space

4.2.3.2 K-Means on the macronutrient space

With the intention of improving the previous clustering results, another approach was undertaken. The categories related to 'healthy' habits were described in terms of macronutrients composition, i.e. proteins, carbohydrates and fats, which allowed to reduce the dimensionality of the data down to 3 features each representing average consumption of the macronutrients by a customer. In fact, the process fell into three consecutive steps:

1. Taking the mode product for each category
2. Assigning the macronutrient composition of that product to its category
3. Calculating the average consumption of each nutrient as a dot product of average consumption of a category and its macronutrient composition

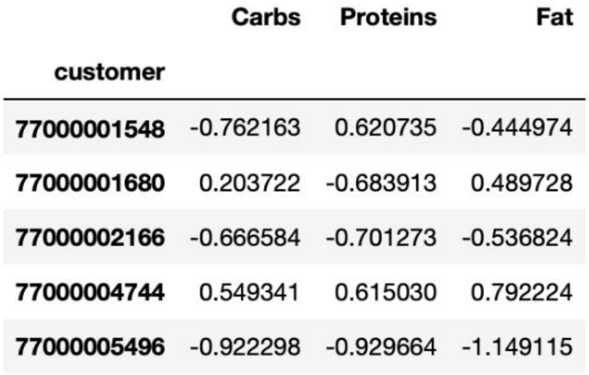
For example, the most frequent product in the 'Fruit' category was 'freson tarrina'. Therefore, its macronutrient composition [16], (85.4% carbs, 6.9% protein, 7.7% fat) was

[16] Fragola: calorie e valori nutrizionali.

assigned to the whole Fruit category. The rest of the categories were described accordingly. Thus, customer A consuming on average 3 fruits and 5 vegetables across visits (macronutrient composition (75%, 12%, 11%)) would have the following macronutrient consumption levels:

1. Carbs: 0.85 * 3 + 0.75 * 5 = 6.3
2. Proteins: 0.069 * 3 + 0.12 * 5 = 0.0757
3. Fat: 0.077 * 3 + 0.11*5 = 0.781

The following table represents the resulting dataframe



	Carbs	Proteins	Fat
customer			
77000001548	-0.762163	0.620735	-0.444974
77000001680	0.203722	-0.683913	0.489728
77000002166	-0.666584	-0.701273	-0.536824
77000004744	0.549341	0.615030	0.792224
77000005496	-0.922298	-0.929664	-1.149115

Figure 4.8: Average macronutrient consumption per customer

The data was once again standardized and the K-Means algorithm was applied to produce 8 clusters represented in the following figure.

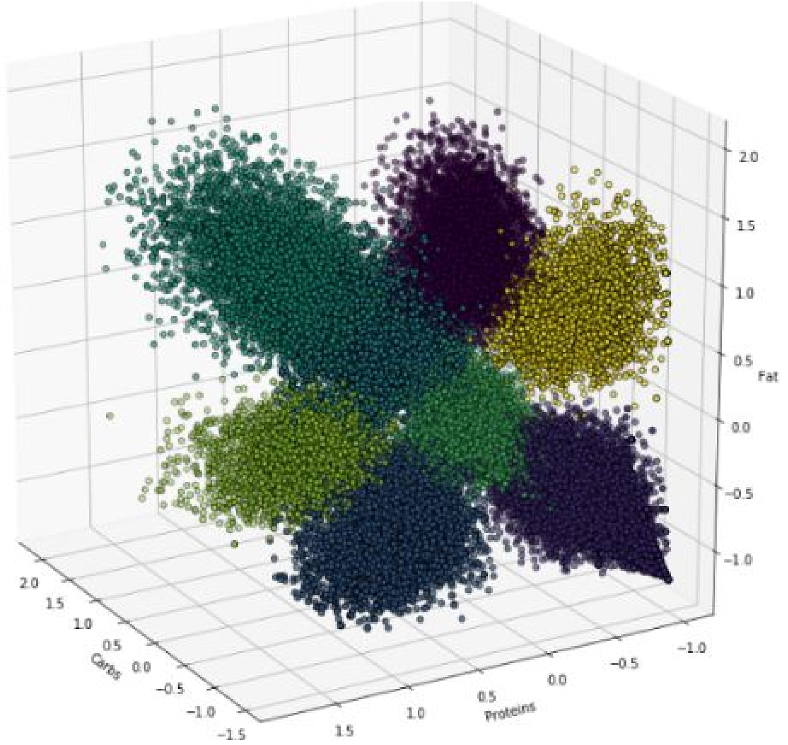


Figure 4.9: Results of clustering based on average macronutrient consumption

The graph shows that customers unconsciously follow consumption patterns which could easily be clustered using macronutrient composition. Indeed, this is a perfect example of high

intra-cluster similarity and high **between-clusters** dissimilarity. One can clearly see that there exists a group which on average consumes more carbohydrates and less proteins and fats, and such a group exists for every possible permutation of the macronutrient, together with all the customers that actually follow a balanced diet. However, the main objective being to identify healthy eating habits, it was difficult to associate a 'healthy' label to one of the clusters in particular. For instance, fruits contain as many carbohydrates as chocolate, as well as red meat which contains as much protein as white meat, but one can hardly assume that an individual which consumes a lot of red meat and chocolate is as healthy as the one who consumes more white meat and fruit. Indeed, if there was a clear separation between a group of people following a balanced diet and another group consuming more fat with respect to other macronutrients, an assumption on healthy habits would have been more realistic. This indicated a need to look for another method to identify people with healthy eating habits.

4.2.4 Gradient Descent to maximize difference in health score

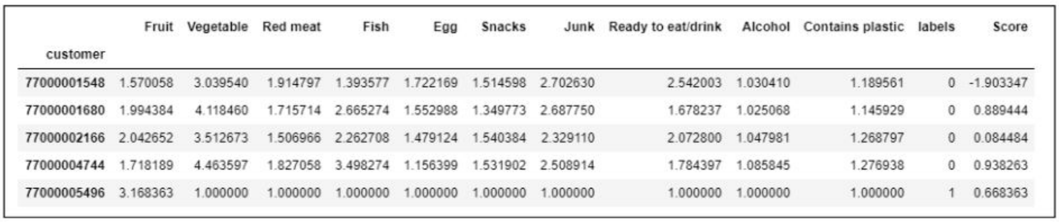
Since none of the aforementioned methods produced satisfactory results, another approach was developed and later approved to be the final clustering tool. It implied transporting each data point into a 'health space', thus allowing the group to infer whether a customer's consumption pattern is adjacent to healthy eating habits. The process can be described as follows:

- Create a **healthy score** based on consumption of the food categories selected earlier.
- Observe how the difference in health score variates when dividing the 'healthy' group from the 'unhealthy' group in different ways.
- Find the optimal split between 'healthy' and 'unhealthy' that maximize the difference in health score using gradient descent minimization technique

The way to separate the 'healthy' from the 'unhealthy' group is to look for some cutoffs in all categories. Consider that all the categories have been scaled from 1 to 5. For instance, if the cutoff for Fruit is 3.4, and for vegetables is 3.2, all the customers purchasing on average a level above 3.4 for fruit and above 3.2 vegetable, are classified as 'healthy'. The objective therefore is to find optimal cutoffs by means of a 'healthy' score defined as follows:

$$HealthScore_i = x_{fruit,i} + x_{vegetable,i} - a * x_{redmeat,i} + a * x_{fish,i} + a * x_{egg,i} - x_{snack,i} - x_{junk,i} - a * x_{ready-to-eat,i} - a * x_{alcohol,i}$$

Where $x_{k,i}$ is the average consumption of category k for individual i, and a is the normalizing weight to adjust for less relevant categories in inferring 'healthy' score.



	Fruit	Vegetable	Red meat	Fish	Egg	Snacks	Junk	Ready to eat/drink	Alcohol	Contains plastic	labels	Score
customer												
77000001548	1.570058	3.039540	1.914797	1.393577	1.722169	1.514598	2.702630	2.542003	1.030410	1.189561	0	-1.903347
77000001680	1.994384	4.118460	1.715714	2.665274	1.552988	1.349773	2.687750	1.678237	1.025068	1.145929	0	0.889444
77000002166	2.042652	3.512673	1.506966	2.262708	1.479124	1.540384	2.329110	2.072800	1.047981	1.268797	0	0.084484
77000004744	1.718189	4.463597	1.827058	3.498274	1.156399	1.531902	2.508914	1.784397	1.085845	1.276938	0	0.938263
77000005496	3.168363	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1.000000	1	0.668363

Figure 4.10: Resulting dataset after adding a health score

The table above shows a dataset with a 'healthy' score calculated for each customer. Next step was to find the optimal non-linear split in health score which would generate the highest between-class differences. For this reason a function was created that maps all the possible permutations of different cutoff levels to a difference in health score. For instance, the cutoffs of [3,4,2,3,4] for five product categories would generate two groups of consumers with a difference of 0.2 in mean health score. The figure below shows such discrete mapping for only fruit and vegetables due to visualization limitations.

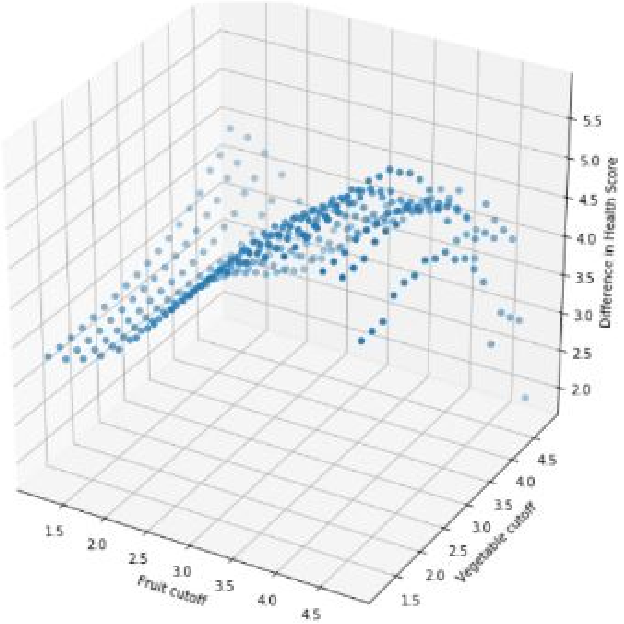


Figure 4.11: Scatterplot mapping of fruit and vegetable cutoff levels to difference in mean health score

The mapping function is discrete, and discrete optimization problems are mostly NP-Hard[17]. Therefore, the function is smoothed using Polynomial Regression of third order, and then maximized using Stochastic Gradient Descent from scipy.optimize.minimize.

$$x_{cutoffs,t+1} = x_{cutoffs,t} - \alpha * \nabla \text{difference in health score}(x_{cutoffs,t})$$

As a result, the following optimal cutoffs were found for the subset of food categories used in the 'healthy' score function:

[17] Sergienko, I. V., & Shylo, V. P. (2006).

Category	Cutoff
Fruit	2.73
Vegetable	2.41
Snacks	1.84
Junk	4.15
Alcohol	3
Red meat	3
Fish	1.84

Table 4.3: Final cutoffs after maximizing through Gradient Descent

It should be stressed once again that whilst the cutoff levels were obtained using a *data-driven approach*, the categories included were instead chosen using a *theory-driven approach*, by only selecting features which, under certain assumptions, well-defined healthy eating habits.

4.3 Closer look at 'healthy' and 'unhealthy' consumers

Once the cutoffs were defined, all the customers with a loyalty card were classified as either 'healthy' (those with healthy eating habits) or 'unhealthy' (those with less healthy eating habits). Overall, there appeared to be 50,276 (92% of the total number of consumers) 'unhealthy' and 4,373 (8%) 'healthy' consumers. Further analysis was conducted to learn more about these two clusters. No significant differences in terms of preferred shopping time have been found. For both groups there are two peak times - afternoon and evening hours

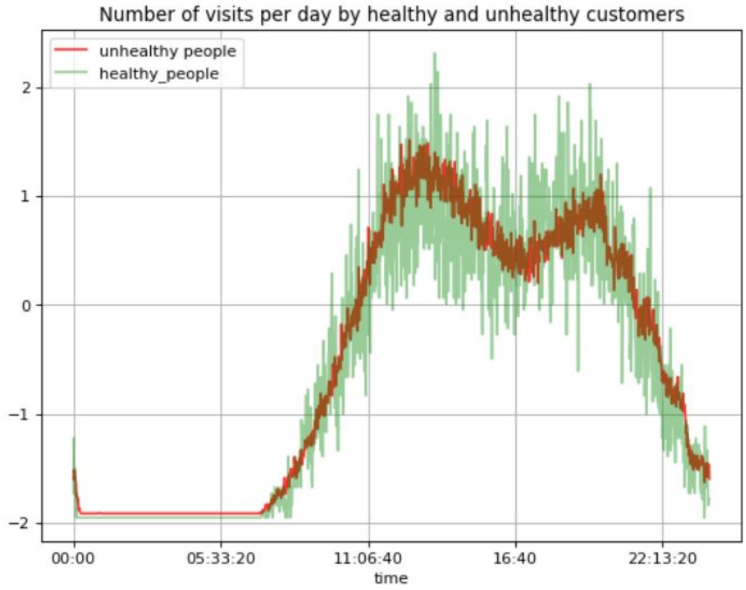


Figure 4.12: Distribution of visits throughout a day

Removing products most commonly bought by both groups, one can see that 'healthy' people tend to opt for more fruits and vegetables whereas 'unhealthy' consumers consume more meat and other animal products.

	'Healthy' consumers	**'Unhealthy' consumers**
1	fuji apple	coca cola
2	mandarin	pork ribs
3	strawberries	beef steak
4	spinach	pork tenderloin chunks
5	oranges	espetec tarradella
6	smoked salmon	beef mince
7	melon	turkey breast
8	cooked prawns	4 grated cheese
9	champinon	chiquilin cookies
10	parsley	condensed milk

Table 4.4: Most frequently purchased products by two groups

Regarding the number of items and total consumer spending per visit, no significant differences were found: for both groups, the median number of items per visit is about **11**, the median receipt is **33 euros**.

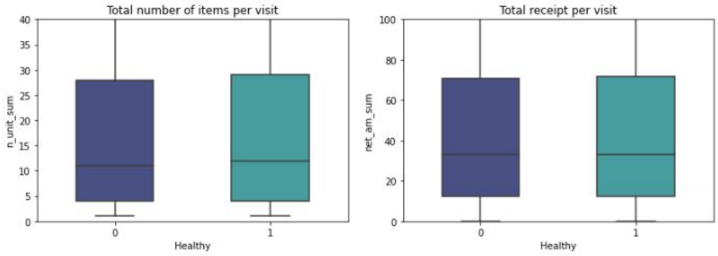


Figure 4.13: Total number of items and total receipt for 'healthy' and 'unhealthy' consumers

Apparently average consumption of various product categories differs between two groups. 'Healthy' people consume more fruit, vegetables and eggs compared to the average level of consumption of these categories. They also buy far less junk food, snacks and red meat.

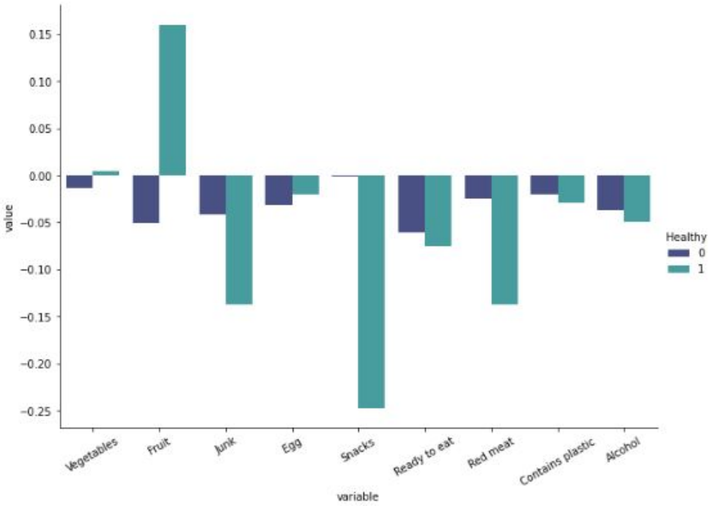


Figure 4.14: Average consumption of different product categories by two groups

Chapter 5

Theory & Hypothesis

Having identified a distinct cluster of people with healthy eating habits, customer profiles of different insurance products were once again reviewed in order to match the cluster to a particular type of insurance.

Longevity insurance was the one for which potential customers could be identified based on their perception of self well-being. The latter is believed to be positively correlated with a healthy lifestyle and, consequently, with healthy eating habits. Here it can be claimed that the identified cluster of people with healthy eating habits can represent an appropriate target audience of longevity insurance.

A person with unhealthy eating habits and lifestyle might reasonably expect not to live a long enough life to enjoy the elderly benefits of longevity insurance Indeed, choosing healthy food has a huge impact in reducing the risk of premature death, as shown in a recent study led by the Harvard T.H. Chan School of Public Health, where people eating healthy low-carbohydrate or low-fat diets resulted to be 27 percent less likely to die prematurely compared to others who did not follow any of those diets[18].

Taking all of these into consideration, the following theory was developed:

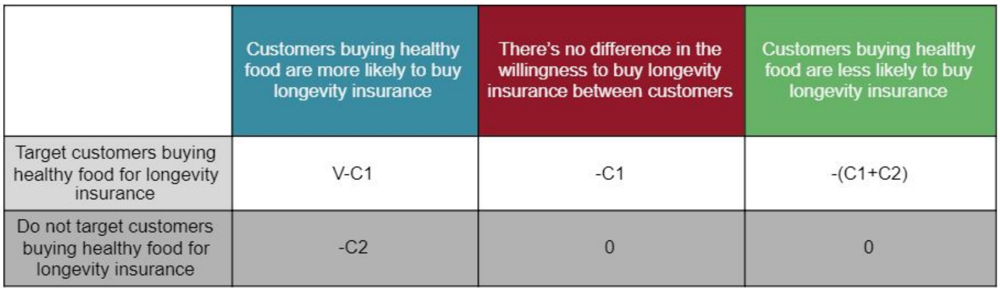
People with healthy eating habits, unlike those without them, tend to believe that they will live a long life. Therefore they are more likely to be concerned about their financial well-being during retirement, which increases the probability of them buying longevity insurance.

It is worth saying that this theory relied on the assumption that healthy eating habits being positively correlated with health in general can impact one's lifespan. Next, in order to verify our theory, a hypothesis to be tested in the following was built:

People who prefer healthy food are more willing to buy longevity insurance than those opting for less healthy food.

Finally, a scenario-action map was built to serve as a guidance for further decisions.

[18] Roeder, A. (2020).



	Customers buying healthy food are more likely to buy longevity insurance	There's no difference in the willingness to buy longevity insurance between customers	Customers buying healthy food are less likely to buy longevity insurance
Target customers buying healthy food for longevity insurance	V-C1	-C1	-(C1+C2)
Do not target customers buying healthy food for longevity insurance	-C2	0	0

Figure 5.1: Scenario-Action pair matrix

- **V**: Additional revenue in case correct target audience is identified
- **C1**: Targeting costs
- **C2**: Opportunity costs

Chapter 6

Testing & Results

6.1 Survey introduction

The next fundamental step was to test the hypothesis by means of a survey following the approach defined below:

1. Gather sample data
2. Learn about their consumption behavior
3. Measure their inclination to buy longevity insurance

One might argue that a survey is an ineffective way to test willingness to buy, since people tend to fail acting upon their intentions. Nevertheless, given the scarcity of our resources, it was decided to proceed with the survey assuming that the intention-behaviour gap at the moment is negligible and the sample is fairly representative. The survey consisted of 3 question blocks gathering information on:

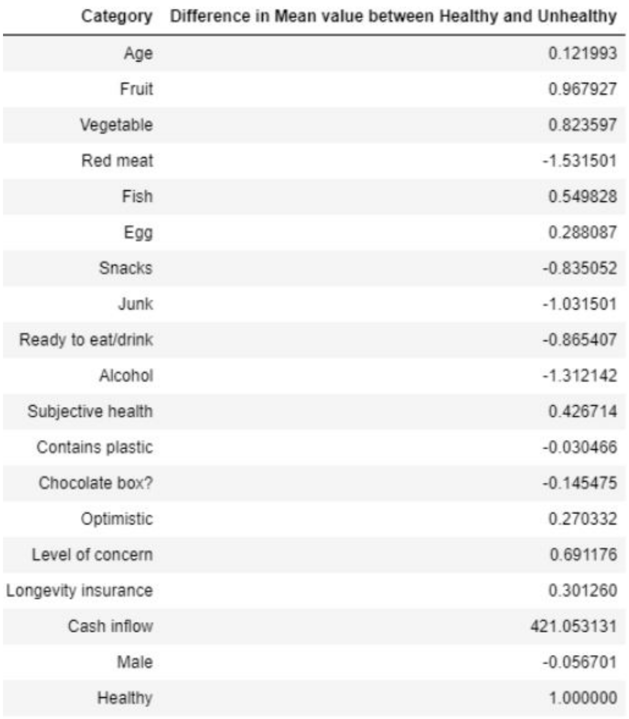
1. General: gender, age, income
2. Level of consumption of the food categories present in the clustering model (Fruit, Vegetable, Meat, Fish, Eggs, Snacks, Fast food, Ready-to-eat, Alcohol), measured on a scale from 1 (low) to 5 (high). This was later used as an estimate for a respondent's average consumption of each food category
3. Propensity to buy insurance estimated as a combination of 4 factors: level of concern, level of optimism, preference to postpone pleasure and willingness to buy longevity insurance among other options, which together corresponded to the following questions in the survey:
    - Estimate your level of concern regarding your financial well-being at an old age (1 for low, 5 for high)
    - Do you expect to live more than or less than 80 years?
    - In a box of chocolates, do you prefer to eat the best or the worst chocolate last?
    - After a brief introduction of some relevant financial instruments (among which pension funds, investments, retirement funds, longevity insurance) respondents

were asked whether they would obtain longevity insurance to ensure their financial well-being at old age.

Asking questions about consumption is necessary to label the respondents into the a priori 'healthy' and 'unhealthy' categorization. At this stage of the scientific approach, theory and testing actually communicate with each other. Before the testing, it is already known who the 'healthy' and the 'unhealthy' respondents are going to be, as being chosen thanks to the cutoffs that maximize the difference in health score across the two groups.

6.2 T-test for the mean difference in likelihood to buy insurance between 'healthy' and 'unhealthy' customers

The collected data was used to classify respondents into either 'healthy' or 'unhealthy' according to the clustering method chosen earlier, i.e. each respondent being assigned a 'health score'. Next, the difference in mean consumption level of different food categories of both clusters was observed.



Category	Difference in Mean value between Healthy and Unhealthy
Age	0.121993
Fruit	0.967927
Vegetable	0.823597
Red meat	-1.531501
Fish	0.549828
Egg	0.288087
Snacks	-0.835052
Junk	-1.031501
Ready to eat/drink	-0.865407
Alcohol	-1.312142
Subjective health	0.426714
Contains plastic	-0.030466
Chocolate box?	-0.145475
Optimistic	0.270332
Level of concern	0.691176
Longevity insurance	0.301260
Cash inflow	421.053131
Male	-0.056701
Healthy	1.000000

Figure 6.1: Difference in mean consumption of different food categories between 'healthy' and 'unhealthy' respondents

One can see that the 'healthy' people in the survey (identified by the model) consume more fruit, vegetables and fish but less junk, ready-to-eat food and alcohol. They are also slightly older and wealthier. These mean differences are not meant to be quantified, still, they provide an idea of what is the sign of the relationship between the variables defining healthy eating habits and the analyzed controls.

Next, a t-test was performed to see whether a significant difference in propensity to buy insurance between ‘healthy’ and ‘unhealthy’ people exists.

$$H_0 : \mu_{healthy} - \mu_{unhealthy} = 0 \text{ vs } H_1 : \mu_{healthy} - \mu_{unhealthy} \neq 0$$

where μ is the mean propensity to buy longevity insurance.

The mean propensity to buy insurance appeared to be 0.88 for the ‘healthy’ group, while it was 0.58 for the ‘unhealthy’ group. The standard error was calculated as follows:

$$\sqrt{\mu_{healthy} \cdot \left(1 - \mu_{healthy}\right) / n_{healthy} + \mu_{unhealthy} \cdot \left(1 - \mu_{unhealthy}\right) / n_{unhealthy}}$$

Figure 6.2: T-Test standard error

With test statistics of 3.37, the resulting p-value calculated on the t-student with (n-2) degrees of freedom was 0.001, meaning that the probability that data had been generated under a process in which the two means were equal for the groups was extremely low. Nonetheless, further analysis was needed to conclude that there existed a significant difference in the propensity to buy longevity insurance.

6.3 Power analysis

Since the survey collected 120 responses, a natural question to ask is whether there are enough data points to conclude with some satisfactory degree of certainty that there exists a significant difference across the two groups.

To confirm this, a power analysis is run to understand what is the minimum required sample size to conduct the analysis, given an effect size (difference in means across the two groups), a desired power of the test and an alpha threshold.

With commonly-used research values of alpha equal to 0.05 (5% chance to reject H0 when true), and power equal to 0.75 (25% chance to fail to reject H0 when false), the required sample size is 150, while the total number of respondents was 120.

Given the extremely low p-value and the somewhat satisfactory sample size, it can be concluded, for the moment only for the survey data, that splitting data space according to the thresholds obtained by maximizing the difference in health score, actually creates 2 distinct groups with both different health level and willingness to buy longevity insurance.

The figure below shows the mapping from the number of observations to the power of the test given a test level of 5% and three different effect sizes.

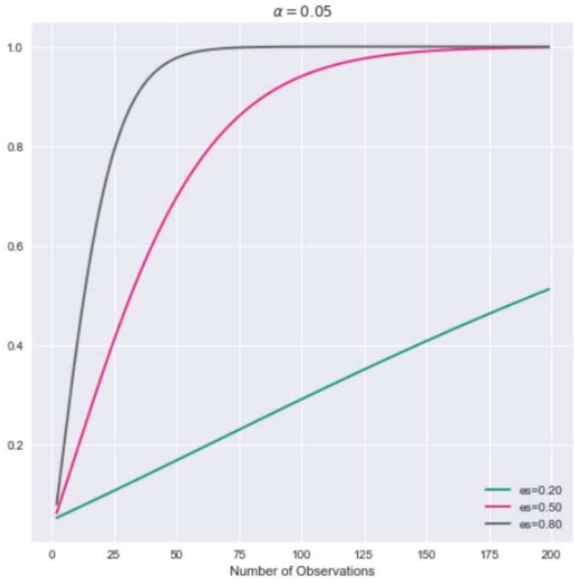
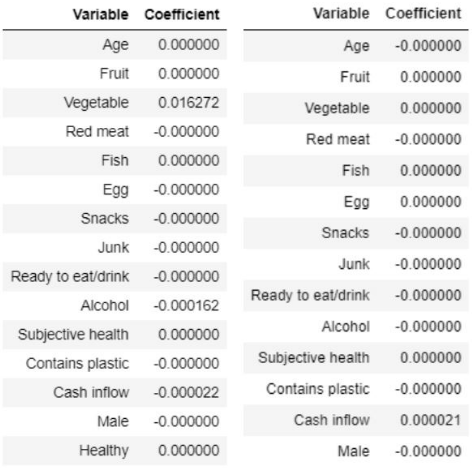


Figure 6.3: Mapping of number of observations to power of test given 5% alpha

6.4 PDSLasso to quantify the effect of health level on likelihood to buy insurance

In order to claim a *causal relationship* between healthy eating habits and likelihood to purchase longevity insurance, a PDSLasso was applied. This particular method was selected for two reasons: first, it allows to interpret the variable of interest (Health) as causal, assuming its exogeneity, second, it addresses the problem of high number of regressors (20) given a small sample size (120) which could be a potential threat to predictive performance.

Indeed, PDS Lasso well handles the *bias-variance trade-off*, as being an effective compromise between consistency in parameters and predictive performance. Running PDSLasso regression yielded the following results.

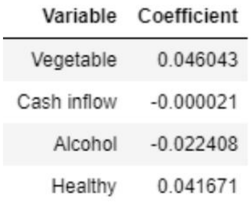


Variable	Coefficient
Age	0.000000
Fruit	0.000000
Vegetable	0.016272
Red meat	-0.000000
Fish	0.000000
Egg	-0.000000
Snacks	-0.000000
Junk	-0.000000
Ready to eat/drink	-0.000000
Alcohol	-0.000162
Subjective health	0.000000
Contains plastic	-0.000000
Cash inflow	-0.000022
Male	-0.000000
Healthy	0.000000

Variable	Coefficient
Age	-0.000000
Fruit	0.000000
Vegetable	0.000000
Red meat	-0.000000
Fish	0.000000
Egg	0.000000
Snacks	-0.000000
Junk	-0.000000
Ready to eat/drink	-0.000000
Alcohol	-0.000000
Subjective health	0.000000
Contains plastic	-0.000000
Cash inflow	0.000021
Male	-0.000000

Figure 6.4: Result of first and second step of PDS Lasso Regression

In the first step of PDSLasso where Propensity to buy insurance is regressed on controls, the selected controls are Vegetable and Income Level. In the second step of regression where Health is regressed on controls, the selected ones are Vegetable, Alcohol and Income Level. The final step is an OLS regression of Propensity to buy insurance on ‘healthy’ with the union of the selected controls in the first 2 steps. It is important to recall that it is possible to interpret only the variable of interest, which is ‘healthy’ in this case.



Variable	Coefficient
Vegetable	0.046043
Cash inflow	-0.000021
Alcohol	-0.022408
Healthy	0.041671

Figure 6.5: Coefficients of the selected controls when regressing Longevity Insurance on the union of regressors

The coefficient of ‘healthy’ is significant and equal to 0.04, meaning that, ceteris paribus, being classified as ‘healthy’ increases the probability of buying longevity insurance by 4%.

One could argue that it does not make sense to add controls such as Fruit, Vegetable and Snacks when regressing Longevity insurance on the ‘healthy’ variable, since the latter was engineered using exactly those variables. However, the variables are *far* from having a **perfect linear relationship** due to the non-linear optimization implemented through Gradient Descent (as shown by the correlation matrix).

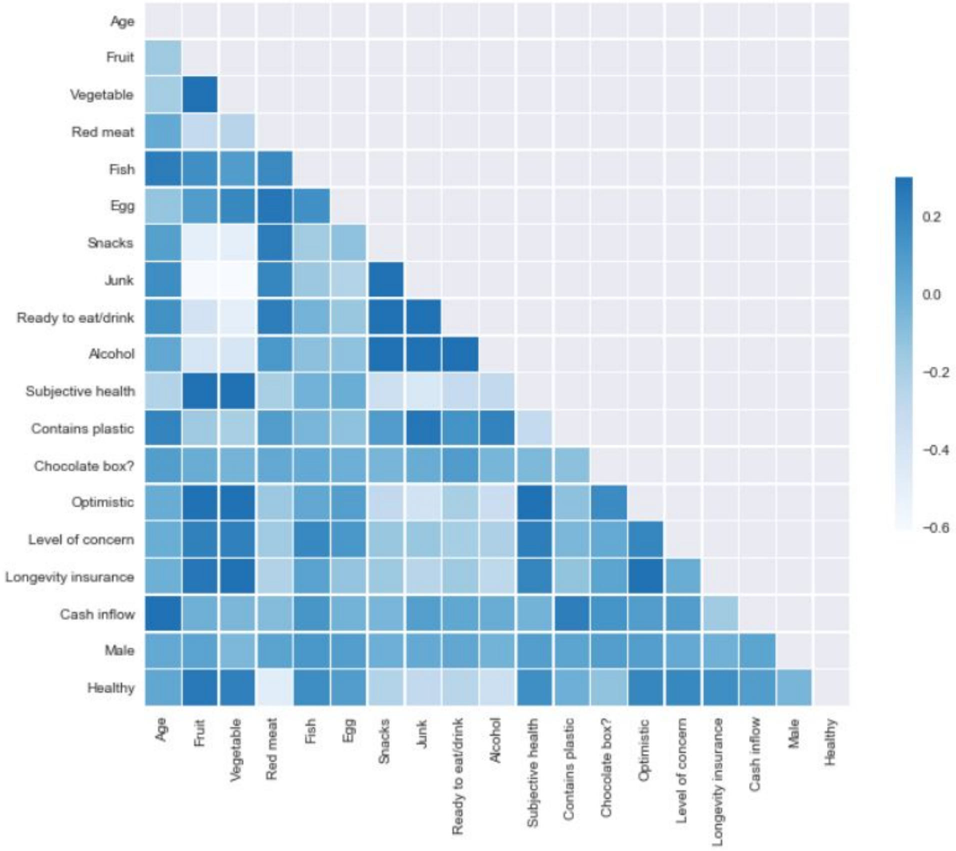


Figure 6.6: Correlation matrix

6.5 Reducing omitted variable bias by means of IVs

Using PDSLasso to comprehend the relationship between the level of Health and the propensity to buy insurance, revealed the contamination of biases into the analysis. With the purpose of avoiding this, the consumption of some categories (fruit, for instance) could be used as *instrumental variables* when regressing Longevity Insurance on Health. Indeed, both IV requirements (relevance and validity) might be satisfied. One caveat is that the variable 'health', because it is constructed using cutoffs from fruit, cannot be used. To solve this issue, the variable 'Subjective Health', obtained by asking the respondents to estimate their level of health, is used as a replacement. The latter holds when assuming that the respondent's estimate is equal to the true value, which is the assumption one analyst makes when using a survey in the first place.

Testing for **relevance**

$$Health_i = \pi_0 + \pi_1 * Fruit_i + \epsilon_i$$

$$H_0 : \pi_1 = 0 \text{ vs. } H_1 : \pi_1 \neq 0$$

reveals a resulting p-value of 0 (which is expected due to the 30% correlation between health level and fruit consumption). This guarantees the relevance of the variable 'Fruit' when inferring the level of Health.

Concerning the **validity** assumption, it is required that consumption of Fruit impacts Longevity insurance *only via 'Health'*, i.e. that the instrument impacts the explanatory variable only via the regressor. In other words, to destroy the validity assumption, one should argue that there exists another reason other than the increase in health when justifying the increase in propensity to buy longevity insurance as a result of a higher consumption of fruit.

Validity allows the calculation of the IV coefficient:

$$\beta_{IV} = (Z'X)^{-1}X'Y$$

which is equal to 0.14. This coefficient, although it may be subject to an upward bias, suggests that there is a **substantial and worth-to-explore** possibility that health status affects the likelihood of purchasing longevity insurance.

6.6 Back to Carrefour data

6.6.1 Distance from the customer persona

To connect survey results to the Carrefour dataset it is useful to provide an instrument which will give a quantitative estimation of how consumption patterns of a customer in the dataset differ from those of the customer persona.

In order to give that estimate, it is necessary to:

- Learn about the customer persona that is more inclined to buy insurance based on the survey results
- Use a metric which identifies the distance between each customer in the Carrefour data and the customer persona

The following figure shows the distribution of consumption of each food category of respondents who said they would buy longevity insurance. This set of information describes the 'customer persona'.

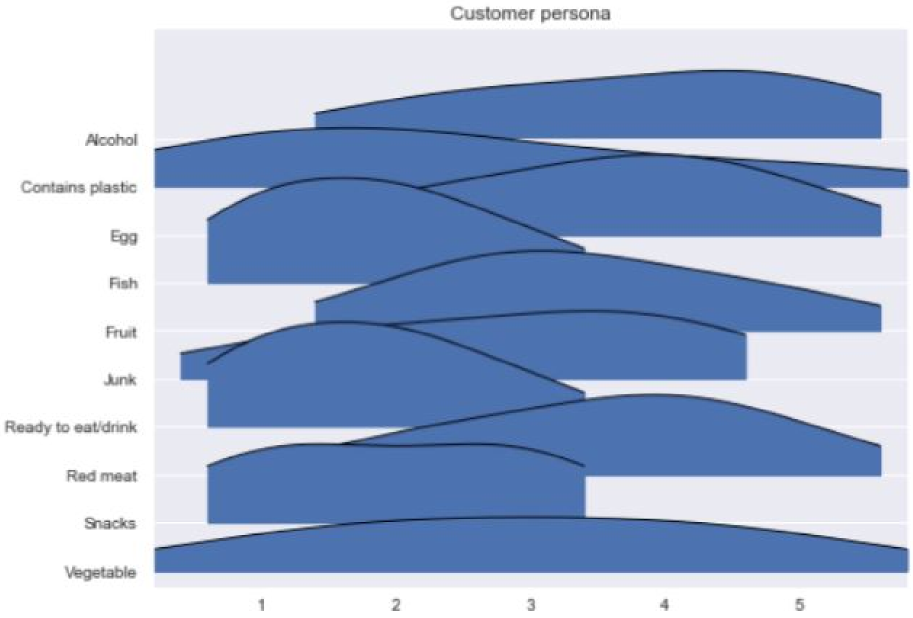
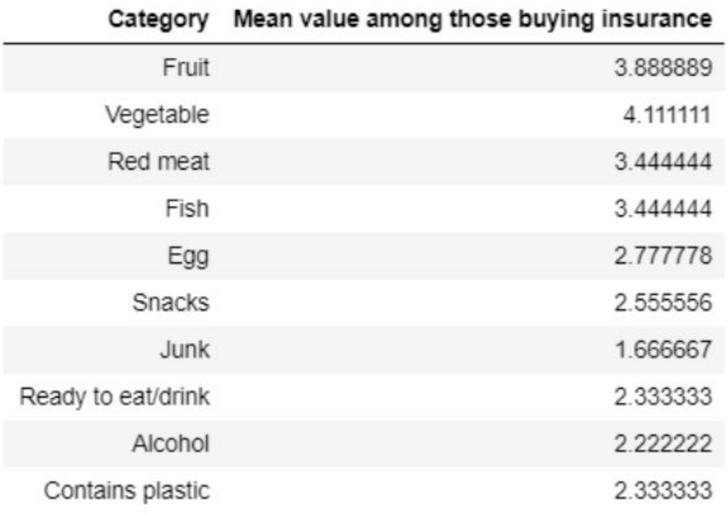


Figure 6.7: Distribution of consumption of categories for the individuals willing to purchase longevity insurance



Category	Mean value among those buying insurance
Fruit	3.888889
Vegetable	4.111111
Red meat	3.444444
Fish	3.444444
Egg	2.777778
Snacks	2.555556
Junk	1.666667
Ready to eat/drink	2.333333
Alcohol	2.222222
Contains plastic	2.333333

Figure 6.8: Customer Persona characteristics

However, this information is not for its own sake, but is to be used in the Carrefour data as this 'ideal customer' with some 'ideal characteristics' that are to be compared with normal customers.

Indeed, an estimate of the propensity to buy insurance should be a measure of "distance" of each customer from our customer persona. The most immediate solution is to calculate the difference in average consumption of each category between each consumer and the customer persona. Ideally, it would be preferred to penalize large differences, therefore it is a good idea to calculate the squared differences for each category and then sum the result across categories. The figures below show the distribution of the "loss" as a distance between each consumer and the customer persona. The higher the loss, the higher the "distance" from the customer persona.

$$Loss_i : \sum_{i=0}^{k}(X_{i,c} - X_{customerpersona,c})^2$$

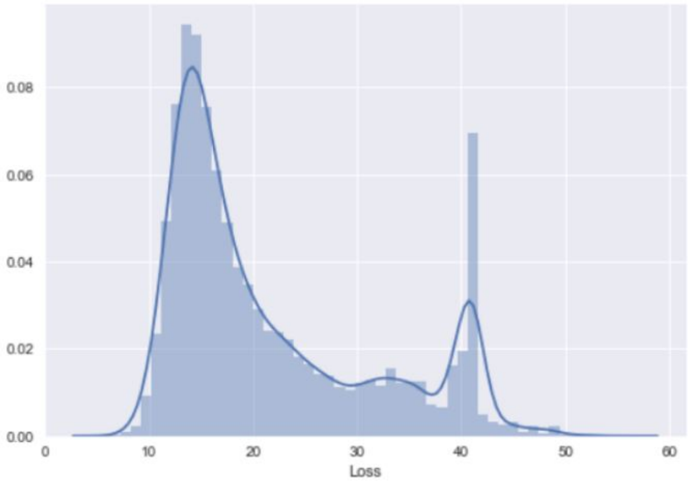


Figure 6.9: Loss distribution for customers in the Carrefour Dataset

For every individual i and relevant category c, where k is the number of total relevant categories.

Finally, it would be interesting to compare the a priori 'healthy' clusterization with the independently generated loss metric to check at a larger scale (with many more data points), to verify that there exists in the carrefour data a significant difference in distance from the customer persona between the 'healthy'/'unhealthy' categorization obtained through Gradient Descent.

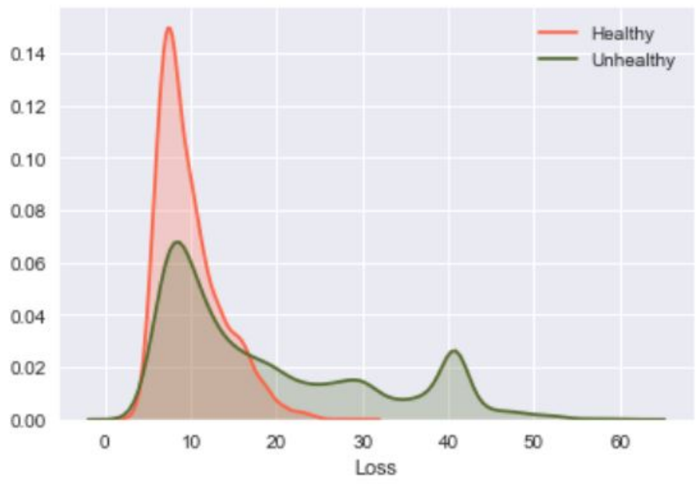


Figure 6.10: Distance distribution from customer persona comparing healthy and unhealthy customers in the Carrefor dataset

A mean loss value of 19.05 for the 'unhealthy' group and of 9.93 for the 'healthy' group generate a large test statistics of 120, which implies that the resulting probability that the data were generated in a process in which 'healthy' people are as distant as 'unhealthy' people from the ideal persona willing to buy insurance is zero. This makes us certain about the existence of a possibility that the health level of a certain individual might have an impact on its likelihood to purchase longevity insurance.

Chapter 7

Recommendations

7.1 Potential issues with the analysis

Despite the fact that the analysis produced good results in terms of connecting healthy lifestyle to propensity to buy longevity insurance, there are several reasons why it cannot be claimed to have quantified a ceteris paribus relationship between the two variables.

The first problem comes from our inability to judge the "validity" of the Carrefour dataset. Because we are trying to infer behavior from consumption, a coherent collection of data points homogeneously spread across age, income, demographics, gender and other controls would be necessary, but no such information on customers is currently available. This puts the analysis at risk of a sample bias which may be a threat to **external validity**, when trying to apply the results to a different market/area than the one the analysis was applied on.

Another assumption was concluding that consumers which have healthy purchasing habits, actually follow a healthy lifestyle. While this is still an assumption, it shall be noticed that among the set of all possible attributes of an individual one could infer from consumption data, the 'healthy' one is definitely the closest and most natural association one can make, given that a healthy lifestyle is very correlated to the food one consumes.

Other threats to the analysis are the assumptions made using a survey when making inference such as:

- Negligible intention-behavior gap
- Level of representativeness of the sample
- Homogeneity in pre-treatment characteristics across units

7.2 Internal validity

When trying to establish a ceteris-paribus effect of the level of health on the likelihood to purchase insurance, a naive econometrician would argue that a ceteris paribus effect is granted. However, many factors shall be taken into account before concluding such relationship is valid:

- When analysing consumer consumption over a long time span, there is the possibility that the singular individual's predisposition to a healthy lifestyle changes over time. These changes would be not controlled by the algorithm, which would average out all variations and assign a label at the end. This variation could not be controlled for within the given dataset, as it doesn't describe analyzable time trends on which one could make inference.
- **Partial compliance**. When an individual is assigned the 'healthy' label, there is the possibility the individual may not 'comply'. This is meant in the sense that a customer may very likely shop at different supermarkets and if one were to consider the aggregate consumption among the shopping at all supermarkets, the assigned label may be different. This is a strong source of biases, especially because no dataset can actually summarize the overall consumption of a given individual.

However, because the "treatment" is assigned by the algorithm, and the consumers are totally unaware of whether they are categorized as 'healthy' or not, there is no possibility of *spillover* (contamination of the 'healthy' people on the 'non-healthy'), nor of **The Hawthorne and John Henry Effects**, as individuals would never change eating habits as a result of an experiment of any kind.

7.3 Proposed experiment

In light of the promising results of the analysis and with the purpose of mitigating the above described threats, it is highly recommended to also undertake an experiment which would yield more credible results in terms of consumers' decision to buy longevity insurance. As mentioned earlier, people tend to fail acting upon their intentions, therefore a properly designed experiment would show if people who prefer healthy food are in fact more likely to buy longevity insurance than those opting for less healthy food.

An experiment could be run using customer data in one of the supermarkets and could include the following steps:

1. Collect consumer purchasing data for two months.
2. Identify 'healthy' people using the proposed clustering approach.
3. Offer longevity insurance to customers by reaching out to them through the personal information provided in a loyalty scheme.
4. Run a Logit Model testing the hypothesis: *People who prefer healthy food are more willing to buy longevity insurance than those opting for less healthy food*

$$\log \frac{p_i}{1-p_i} = \beta_0 + \beta_1 healthy + \beta_2 controls_i$$

Where p_i is the probability of purchasing longevity insurance for individual i

Following the analysis of the above designed experiment, one of the three following options should be considered, provided that the coefficients are significant. They depend on the outcome of the experiment, as well as the underwriting for the actual insurance product that is to be offered:

1. If the difference in average predicted probability for 'healthy' and 'unhealthy' is bigger than 50%, a recommendation can be made to roll out this targeting approach across the entire customer base of the supermarket chains that are currently partners. This is dependent as well on further research showing that an attractive product can be offered to these customers at a reasonable price while still maintaining a profit. Additional thought could be given as to how much targeting should be done, and through which channels, such as direct mail, e-mail, or flyers in the supermarket.
2. Should the experiment show that the difference in average predicted probability for 'healthy' & 'unhealthy' be between 25-50%, or there was a significant problem in underwriting the policy, strong enough to deem the project unprofitable in case it were to be launched at scale, then the project should not be launched without serious re-evaluation. This evaluation could take the form of a redesign of the hypothesis in that either the customer cluster or the insurance product (or both) could be changed. This change could be marginal or quite substantial, depending on the detailed result of the experiment. Overall it would still prove that even though far less than optimal, there still is an addressable market for longevity insurance among healthy customers.
3. Lastly, if the difference in average predicted probability is less than 25%, the null-hypothesis could not be rejected. This would mean that while there may be a higher likelihood to buy longevity insurance for customers with a healthy eating habit, it is too small to create a viable business model based on this information. It could also happen that there is in fact a lower likelihood to buy longevity insurance. In that case, the data gathered in the experiment could be reexamined for another cluster displaying a higher likelihood to buy longevity insurance.

In summary, this project showed that determining whether or not product could be marketed to customers based on seemingly unrelated purchase data that, at first glance, would not indicate which specific product should be marketed to which customer could prove possible given the following assertions:

First, the data at hand would need to be classifiable, as well as clusterable, resulting in clearly distinct groups indicating different behavior. In fact, the clearer the difference, the better it would be. Secondly, the product, or product category to be sold should be split up into the atomic products that could be offered. After performing a logical grouping, customer profiles should be created. Following, matchings could be created by trying to cluster for some of the identified characteristics that a potential customer should exhibit in the data. Lastly, the proposed matching should be tested by means of an experiment, to avoid several complications that could otherwise endanger the validity of the test, such as the aforementioned intention-behavior gap.

Bibliography

[1] Background on: Buying Insurance. (n.d.). - Retrieved November 20, 2020 from *https://www.iii.org/article/background-on-buying-insurance* .

[2] Collignon, R., Dekkers, J., van Veen, J., & Scheeren, D. (2018), Insurance Analytics: Organizing Analytics capabilities to get value from Data Analytics solutions. *https://www2.deloitte.com/content/dam/Deloitte/nl/Documents/financial-services/deloitte-nl-fsi-insurance-data-analytics-within-the-insurance-industry.pdf.* .

[3] Dragos, S. L., Dragos, C. M., & Muresan, G. M. (2020). From intention to decision in purchasing life insurance and private pensions: Different effects of knowledge and behavioural factors. , *ournal of Behavioral and Experimental Economics, 87, 101555. doi:10.1016/j.socec.2020.101555* .

[4] Fontinelle, A. (2020, November 16). Guide to Life Insurance. Retrieved November 22, 2020, *from https://www.investopedia.com/terms/l/lifeinsurance.asp.*

[5] Fragola: calorie e valori nutrizionali., *http://www.dietabit.it/alimenti/frutta/fragola/* .

[6] Himmelstein, D. U., Lawless, R. M., Thorne, D., Foohey, P., & Woolhandler, S. (2019). Medical Bankruptcy: Still Common Despite the Affordable Care Act. American Journal of Public Health, 109(3), 431–433 *https://doi.org/10.2105/ajph.2018.304901* .

[7] Husniyah, A. R., Norhasmah, S., & Mohamad O. A. (2017). Assessing Predictors for Health Insurance Purchase Among Malaysian Public Sector Employees https://link.springer.com/chapter/10.1007/978-3-319-54112-91_7#: :text=The%20likelihood%20of%20investment%20involvement,health%20risk%20among%20th

[8] Hopkins J. (2015). 6 Reasons To Consider Buying Longevity Insurance *https://www.forbes.com/sites/jamiehopkins/2015/04/06/6-reasons-to-consider-buying-longevity-insurance/?sh=3e15e6215ed4*

[9] Insurance Information Institute. (August 31, 2020). Life insurance distribution channels in the United States in 2019 [Graph]. In Statista. Retrieved November 19, 2020, from, *https://www.statista.com/statistics/377638/life-insurance-distribution-channels-usa/*

[10] Kaesler, S., Leo, M., Varney, S., & Young, K. (2020, June 12). How insurance can prepare for the next distribution model. Retrieved from,

https://www.mckinsey.com/industries/financial-services/our-insights/how-insurance-can-prepare-for-the-next-distribution-model .

[11] Liu, T., & Chen, C. (2002). An analysis of private health insurance purchasing decisions with national health insurance in Taiwan. Social Science & Medicine, 55(5), 755-774. doi:10.1016/s0277-9536(01)00201-5

[12] McMaken, L. (2020, August 28). 4 Types of Insurance Everyone Needs. Investopedia., *http://www.investopedia.com/financial-edge/0212/4-types-of-insurance-everyone-needs.aspx.* .

[13] Miguéis, V., Camanho, A., & Cunha, J. F. (2012). Customer data mining for lifestyle segmentation. , *Expert Systems with Applications, 39(10), 9359-9366. doi:10.1016/j.eswa.2012.02.133* .

[14] Roeder, A. (2020, January 28). Healthy diets may reduce risk of premature death. Harvard Gazette. , *https://news.harvard.edu/gazette/story/newsplus/healthy-diets-may-reduce-risk-of-premature-death/.* .

[15] Scanlon, J., Leyes, M., & Terry, K. (2018). 2018 Insurance Barometer. LIMRA., *https://www.gpagency.com/wp-content/uploads/2018-Insurance-Barometer-Study.pdf.*

[16] Sergienko, I. V., & Shylo, V. P. (2006). Problems of discrete optimization: Challenges and main approaches to solve them. *Cybernetics and Systems Analysis, 42(4), 465–482. https://doi.org/10.1007/s10559-006-0086-3.*

[17] Sharps,K., Hitsky, D., Hodgins, S., Ma, C. (2015), *https://www2.deloitte.com/content/dam/Deloitte/us/Documents/strategy/us-cons-life-insurance-consumer-study.pdf.*

[18] Smith, T. A. (2019)., *A buyer behavioural model for associating personality traits with likelihood to buy life insurance. Journal of Customer Behaviour, 18(1), 61-78. doi:10.1362/147539219x15633616548524* .

[19] Verain, M., Sijtsema, S., & Antonides, G. (2016), *Consumer segmentation based on food-category attribute importance: The relation with healthiness and sustainability perceptions. Appetite, 101, 217. doi:10.1016/j.appet.2016.02.049.*