Universidad Pontificia de Comillas - ICADE

Data Analytics and Machine Learning Applied to Insurance. Leveraging Machine Learning to Enhance customer acquisition Strategy.

Author: Inés Gómez Fortis
Director: María Coronado Vaca

MADRID | April 2023

Abstract

The project employs a range of data analytics and machine learning techniques for a comprehensive analysis of Getlife's client portfolio, intending to optimize the customer acquisition strategy of the insurtech.

So far, the startup has used its resources to acquire customers, gain popularity and collect data. However, the insurtech has found that this technique is inefficient due to a lack of targeted marketing efforts and a failure to fully understand the needs and preferences of the customer base. As a result, the current plan is to optimize the customer acquisition strategy by focusing on customer profiles that represent a lower acquisition cost so that the startup can allocate its resources more efficiently and effectively to acquire customers more likely to become loyal and long-term customers.

For departments such as marketing and product to get the most out of the available resources and thus provide greater benefits to the company, it is necessary to identify the profile that attracts the life insurance industry. Among other things, the idea is to acquire a complete understanding of the industry to personalize marketing campaigns, offer special deals, and customize products and services, as well as to focus the business strategy on a specific audience.

First, a descriptive analysis of the variables that characterize the profiles attracted by the service has been carried out to identify potential customers. On the other hand, to attract customers more effectively, a series of clustering analyses have been developed to classify customers and to be able to personalize advertising campaigns (a specific analysis of the case of Facebook is included) or products and services. In addition, the relationship between the existence of certain pathologies and the purchase of life insurance has been studied. In addition, we have included an analysis of survival curves around the time from when a lead is created and when it converts. Finally, the profiles associated with cancellations have been evaluated descriptively.

This research underscores the importance of market research, customer segmentation, and targeted marketing strategies in the life insurance industry. Understanding customer profiles, preferences, and behaviors can help insurers better serve their target audience and design effective business strategies for growth and customer retention.

Keywords: life insurance industry, insurtech, customer acquisition strategy, client portfolio, optimization, personalization, advertising campaigns, marketing, product, lead conversion, cancellations, clustering analyses, K-prototypes, survival analysis.

Resumen

El proyecto emplea una serie de técnicas de análisis de datos y *Machine Learning* para realizar un análisis exhaustivo de la cartera de clientes de Getlife, con el objetivo de optimizar la estrategia de captación de clientes de la *insurtech*.

Hasta el momento, la startup ha optado por emplear sus recursos en adquirir clientes de forma masiva con el fin de ganar popularidad y recopilar datos. No obstante, esta estrategia resulta poco eficiente debido a la falta de esfuerzos específicos de marketing y a la incapacidad de comprender plenamente las necesidades y preferencias de la base de clientes. Como consecuencia, se busca optimizar dicha estrategia centrándola en perfiles de clientes que representen un menor coste de adquisición, de forma que la startup pueda emplear sus recursos de manera más eficiente y eficaz y adquirir así clientes con más probabilidades de convertirse en fieles y duraderos.

Para que departamentos como marketing y producto puedan sacar el máximo rendimiento de los recursos disponibles y proporcionar así mayores beneficios a la compañía resulta necesario identificar el perfil que atrae el sector de los seguros de vida. Entre otras cosas, la idea es adquirir un conocimiento completo de la industria para personalizar las campañas de marketing, ofrecer ofertas especiales y customizar los productos y servicios, así como centrar la estrategia empresarial en un público específico.

En primer lugar se ha realizado un análisis descriptivo de las variables que caracterizan los perfiles atraídos por el servicio para identificar potenciales clientes. Por otro lado, se han desarrollado una serie de análisis clustering para clasificar a los clientes y poder personalizar campañas publicitarias (se incluye un análisis concreto del caso de Facebook) o bien productos y servicios que atraigan clientes más fácilmente. Unido a esto se ha estudiado la relación entre la existencia de determinadas patologías con adquirir un seguro de vida. Además, se ha incluido un análisis de curvas de supervivencia en torno al tiempo desde que se crea un lead y este convierte. Por último, se han evaluado los perfiles asociados a las cancelaciones de manera descriptiva.

Este estudio destaca la importancia de investigar el mercado, conocer el perfil de los clientes del sector y desarrollar estrategias de marketing específicas para mejorar el servicio ofrecido, y diseñar estrategias comerciales eficaces para el crecimiento de la empresa y cumplir con el objetivo de retención de clientes.

Palabras clave: seguros de vida, insurtech, estrategia de adquisición de clientes, cartera de clientes, optimización, personalización, campañas publicitarias, marketing, producto, conversión de clientes, cancelaciones, clustering, K-prototypes, análisis de supervivencia.

Table of Contents

Index of Figures

Index of Tables

Chapter 1. Introduction

1.1. OBJECTIVES

First of all, I would like to highlight that this thesis is carried out in collaboration with the Spanish insurtech Getlife[1].

Getlife is a life insurance startup that was founded in 2021 and is revolutionizing the insurance industry. This insurtech offers life insurance to people between 18 and 75 years of age through 100% online processes that require only a few minutes and ensure instantly without the need for any medical examinations. It acts as a mediator between the client and large insurance companies with which it has an agreement to offer the services of these companies in a personalized way to its clients. Getlife provides its products for Spanish residents but also the company presents a recent international expansion in the French market. However, at the moment of writing this research, there was no available information about this latter market. As a consequence, real data only from the Spanish market will be used to answer a series of questions that arise within the framework of the company, which will be detailed below.

The objective of the research is to analyze the characteristics of customers and potential customers (leads), as well as to study the type of cancellations that occur in the company to better adapt products, provide new ones, or launch specific customer acquisition campaigns.

The idea is to find which profile the life insurance industry attracts so that departments like marketing and product can get the most out of their resources to provide a better overall performance of the company. That is, customizing marketing campaigns, offering special offers, and personalizing products and services, as well as focusing the business strategy on targeting the specific profile so that it is optimized.

In addition to this, we will carry out a survival analysis for which the outcome variable of interest is time until an event occurs, often referred to as a failure time, survival time, or event time. In our particular case, this variable of interest is the time to purchase a product by customers. By examining the times and the relevant characteristics of the individual or group it is possible to estimate the probability of survival at different time points, which can be useful for making strategic decisions about marketing actions and product development.

Last, another important aim is to study the main reasons why policies are canceled and when they occur, in order to understand the needs of clients and thus reduce the churn rate.

[1] www.getlife.es

1.2. MOTIVES

The life insurance industry represents a growing market that offers a combination of development potential, stability, profitability, and the opportunity to make a positive impact on people's lives. The industry has seen significant changes over the past decade, with developing economies being the main drivers of growth (Bernard et al., 2020).

New technologies and developments are transforming the insurance industry, providing opportunities for innovation and disruption. These technologies are enabling continual underwriting and innovative products that reflect changing customer needs (Krishnakanthan et al., 2021). According to this, it illustrates a highly profitable area that can generate significant returns for its shareholders.

“The Global Insurance Report 2023” by Agrawal et al. (2022) states that life insurance companies must adapt to changing customer needs and preferences, embrace new technologies and partnerships, and focus on promoting overall wellness and sustainability to remain relevant in the future.

It is necessary to examine the trends and challenges facing the life insurance industry and propose solutions to help it adapt and grow in the future. Moreover, analyzing the needs and preferences of different types of its clients, and identifying ways in which the industry can better serve is essential to succeed in this industry (Duncan et al., 2016).

For the reasons mentioned above, it appears worthwhile to gain a better understanding of the industry and how it operates, as well as the various products and services it offers.

1.3. METHODOLOGY AND STRUCTURE

To meet the objectives set in the previous section, the work is divided into five different sections.

First, the data visualization tool, DataStudio, will be used to develop a descriptive analysis of the characteristics of Getlife's customers through visualizations. The idea of this section is to identify the profile of customers that the life industry attracts through a series of graphs representing the main characteristics of actual clients. Some of the variables considered are age, gender, nationality, body mass index, etc.

Next, we will proceed with a customer segmentation phase. To do this we will focus on the functions performed by the marketing and product departments, departments of vital importance when it comes to attracting customers. Each of these analyses follows the focus of the corresponding department, i.e., how to launch marketing campaigns that are more attractive from a marketing point of view and identify what kind of products are currently demanded by the market in the product case. As a consequence, the data and information handled in each of the analyses are different.

To continue, we will analyze the time that elapses from the moment a lead is created until the purchase of a policy. To do so, we will apply survival analysis techniques (Sullivan, 2016). In particular, we will determine groups of survival curves based on purchase times using the algorithm proposed by Villanueva et al. (2019).

Lastly, our attention will be directed towards cancellations. For life insurance companies, cancellation represents a crucial metric, as comprehending the underlying reasons behind customer cancellations may aid in the enhancement of the organization's products and services. The process requires the categorization of cancellations and addressing pertinent inquiries such as: what factors lead to customers cancelling their policies, what are the primary causes of cancellation, is there a recurring profile among those who cancel, and how long is the average policyholder's tenure before cancelling their policy?

Sharps et al., (2015) state that the majority of lapses are tied to a change in or concern about financial standing. This suggests that carriers can benefit from emphasizing the importance of life insurance and how it can prevent financial hardship.

To do this, we will develop a series of visualizations to find the characteristics of the profiles who cancel. It is vital to understand why a company is losing its clients. Having an answer to the previous questions can help develop strategies to increase the retention rate.

We will also calculate the churn rate of Getlife. The idea is to compare the numbers obtained with those of the industry to see the company's position in the market. The churn rate is defined as follows:

$$churn\ \ rate = \frac{\#\ of\ \ canceled\ policies}{\#\ of\ \ total\ policies}$$

It should be noted that each of these parts is associated with an initial phase of preparation and data processing that consumes a large amount of time.

Chapter 2. State of the Art

The life insurance industry has been around for centuries and has evolved significantly over the years. It has grown over time to become a sizable global industry, offering a variety of plans, goods, and services, including term life insurance, permanent life insurance, and annuities, to meet the various requirements of people and families.

Based on "The Global Insurance Report 2023" by Agrawal et al. (2022), the industry has also experienced continual change due to sociological and economic changes and technological trends, displaying its adaptability to the fast-paced environment of the modern world.

At the present time, the life insurance industry plays an essential role in the economy by providing protection against financial loss, encouraging savings and investment, and contributing to economic growth (Bernard et al., 2020).

During the last few years, the sector has undergone a substantial transition spurred by technological advancements, including data analytics and machine learning.

Analytics and machine learning techniques have been used in the life insurance industry for many years. According to the "Insurance Trends 2019" report by O'Hearn et al., insurance companies are investing more in technology and data analytics to gain a competitive advantage. The report highlights that data analytics can help companies improve customer engagement, speed up underwriting and claims processes, and develop new products and services.

For example, some algorithms can automatically assess risks and calculate premiums based on data inputs, leading to more accurate and efficient underwriting decisions. Similarly, it can expedite claims processing by automating claims assessment, reducing costs, and improving customer satisfaction.

Additionally, the ability to use massive amounts of data to acquire insights into consumer behavior, mortality trends, and risk factors has made data analytics an essential tool for life insurance businesses. Following the report Harnessing the power of digital life insurance (2016), insurance companies can produce better products and make more educated decisions regarding underwriting, pricing, and lifestyle choices by evaluating data from various sources, including medical records, lifestyle choices, and social media. An example would be the usage of a Fitbit or Apple Watch to provide customized insurance depending on lifestyle preferences.

One of the challenges the life insurance industry faces is the need to personalize its products and services to meet the specific needs of each client. Customers need more straightforward insurance products with concise coverage explanations that can assist them to gain their trust and comprehension while seeking customized coverage (Adamova et al., 2018).

Machine learning algorithms can analyze large amounts of data to identify patterns and make predictions, which can be particularly useful in the insurance industry, where data analysis can help companies develop more accurate risk models and tailor products to individual customers. For instance, machine learning algorithms can analyze medical records to assess an individual's health risks and mortality probabilities more accurately.

One method of using data analytics and machine learning to tailor goods and services is telematics, which involves sensors and other data-collection devices to gather information about a customer's driving behavior. With the help of this information, usage-based insurance solutions can be developed, with prices based on actual driving behavior as opposed to more traditional risk factors like age, gender, and location. Moreover, it can increase safety by giving drivers feedback on their driving habits, as noted by NAIC (National Association of Insurance Commissioners).

Another area where data analytics and machine learning are being used in the life insurance industry is fraud detection, one of the major issues for businesses, costing billions of dollars annually. According to data by the Coalition Against Insurance Fraud[2] , fraud accounts for about 10% of all property and casualty insurance losses and thus represents a main issue for businesses. Thanks to different machine learning algorithms, it is possible to evaluate data and spot trends that indicate fraud, assisting insurance firms in reducing losses and detecting fraud more rapidly (Severino & Peng, 2021). Some popular algorithms used for insurance fraud detection are Support Vector Machine (SVM), Random-Forest (RF), and Gradient Boosting (Rukhsar et al., 2022).

However, despite all the above-mentioned, the life insurance industry faces significant challenges due to changing customer expectations, regulatory pressures, and technological disruption (Bernard et al., 2020). Customers now expect more personalized products and services that adjust to their specific needs, which is a challenge for the industry to meet. Companies are also under regulatory pressure to abide by stringent rules and data protection legislation, which can be difficult to navigate. Technological disruption is forcing the industry to adapt to new digital and mobile advances, and to make progress in advanced analytics and artificial intelligence (Bernard et al., 2020). The industry is also facing harsh economic environments and record-high inflation, which are significant challenges.

In conclusion, technological trends, including data analytics and machine learning, are reshaping the life insurance sector, driving innovation and efficiency. These advancements enable insurers to leverage data for better risk assessment, automate processes, and enhance customer engagement. As technology continues to evolve, the life insurance industry is likely to further adapt and transform, shaping the way insurers operate and serve their customers in the future.

[2] https://insurancefraud.org/fraud-stats/

Chapter 3. Empiric Analysis

3.1. DATABASES

As mentioned above, the data used in the different analyses run in the research correspond to actual data from Getlife.

Having an organized database is vital to understand the information available for reporting and data analysis, as well as facilitating decision-making.

In Getlife, the data is stored in BigQuery[3], a highly scalable Google data warehouse associated with a Google Cloud Platform (GCP)[4] project that allows you to load data, create tables, and perform data queries using Structured Query Language (SQL)[5].

The principal function of this tool is to unify the multitude of information coming from the company's various data sources, such as Google ads[6], Google Analytics[7], Facebook ads[8], Stripe[9] (a tool used to manage customer charges and returns), Cloudtalk[10] (the platform used to contact leads by phone) and the IT department's databases.

I would like to mention that many of the tables used were made specifically for the analyses developed in this study. For this purpose, we select the desired variables from the original tables as appropriate, create new variables and join the desired datasets using different types of joins as needed.

The data preparation is performed in SQL. However, the data processing and subsequent analysis are carried out in Python[11] (Van Rossum and Drake, 2009) or R[12] (v4.1.3; R Core Team 2022). To access the data, it is necessary to connect the corresponding tool to BigQuery. The database is downloaded directly through the connection, without the need to load the data into a local Excel file. Annexes II and III detail how to connect to BigQuery from each of these tools.

It should be highlighted that due to confidentiality issues, the queries related to databases and data downloading have been removed from the project, as well as sensitive information related to the business or customers of Getlife.

---

[3] Big Query: https://cloud.google.com/bigquery
[4] GCP: https://cloud.google.com/gcp?hl=es-419
[5] SQL: https://cloud.google.com/sql-server?hl=es
[6] Google ads: https://ads.google.com/intl/es_ES/home/?pli=1
[7] Google Analytics: https://analytics.google.com/analytics/web/provision/#/provision
[8] Facebook ads: https://es-es.facebook.com/business/ads
[9] Stripe: https://stripe.com/es
[10] Cloudtalk: https://www.cloudtalk.io/es/
[11] Python: https://www.python.org/
[12] R: https://www.R-project.org/

3.2. METHODOLOGY

As per the research, one of the primary objectives is to analyze and understand the type of clients that the life insurance industry attracts. This analysis will help insurance companies and insurtechs like Getlife to optimize their resources and provide better services to their clients. Companies can adjust their goods and services to match the unique needs of their target market by determining the characteristics of their customers. Better client satisfaction and overall company performance will follow from this.

3.2.1. CUSTOMER PROFILE ANALYSIS IN THE INSURANCE INDUSTRY

To start, we are going to use DataStudio[13] to run a descriptive analysis of the characteristics of the customers of the life insurance industry. The idea of this section is to identify the profile of customers that Getlife attracts through a series of visualizations representing the main characteristics of a sample of its actual clients.

In the first place, we use a set of pie charts in which we represent different categorical variables like gender, whether a lead has a risky job or not, health status, whether a lead smokes or not, the type of coverage acquired, and the purchase intention (Figure 1. Relative frequencies of categorical variables.).

Figure 1. Relative frequencies of categorical variables.

According to the pie charts, most clients are Spanish males who don't have risky jobs and don't smoke. The majority of clients either have a health condition or have a body mass index (BMI) that is not the recommended one.

According to studies from the National Center for Biotechnology Information (Klatsky et al., 2017), a healthy BMI is considered to be between 18.5 to 24.9 kg/m2, and it is also associated with the lowest rates of all-cause mortality. Thus, the BMI represents a crucial parameter for life insurance companies to evaluate risk and set premiums.

We can also see that the most purchased product is the one that includes coverage in case of total permanent disability (TPD or IPA in Spanish) and that the most common reason to purchase a policy is family followed by a loan.

To evaluate the distribution of continuous variables like age, income, and BMI, we have had to use bar plots instead of histograms due to the limitation of DataStudio in creating

[13] DataStudio: https://datastudio.google.com/

histograms for continuous variables (Figure 2. Relative frequencies for grouped ages, income, or BMI.).

Figure 2. Relative frequencies for grouped ages, income, or BMI.

In Figure 2. Relative frequencies for grouped ages, income, or , it can be observed that the average policyholder is 47 years old, earns about 26.5k a year, and is slightly overweight, according to standard BMI categorizations.

Finally, we examined the beneficiary variable. This variable is of great interest and importance since, together with intention, it allows us to know why someone decides to take out life insurance. For this representation, we again use sector diagrams (Figure 3. Beneficiaries' distribution).
According to Figure 3, more than 86% of the clients choose a physical person. As this number represents a great number of the policyholders it is interesting to dive into this data.

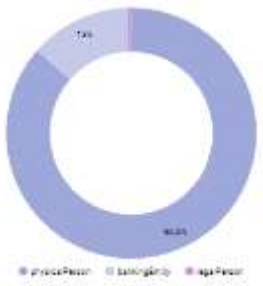

Figure 3. Beneficiaries' distribution

To do this we create two new variables. First, we create the variable type of beneficiary in which we indicate what type of relationship the insured person and the beneficiary of the policy have. There are three possible situations: 1. Both the beneficiary and the policyholder share two last names (siblings). 2. Only one last name is shared (relatives). 3. Or there are no last names in common (any other type of relationship).The second variable indicates the number of beneficiaries designated in the policy.

The distribution of these variables is shown in Figure 4. Physical person: beneficiaries' distribution(left and right panels, respectively).

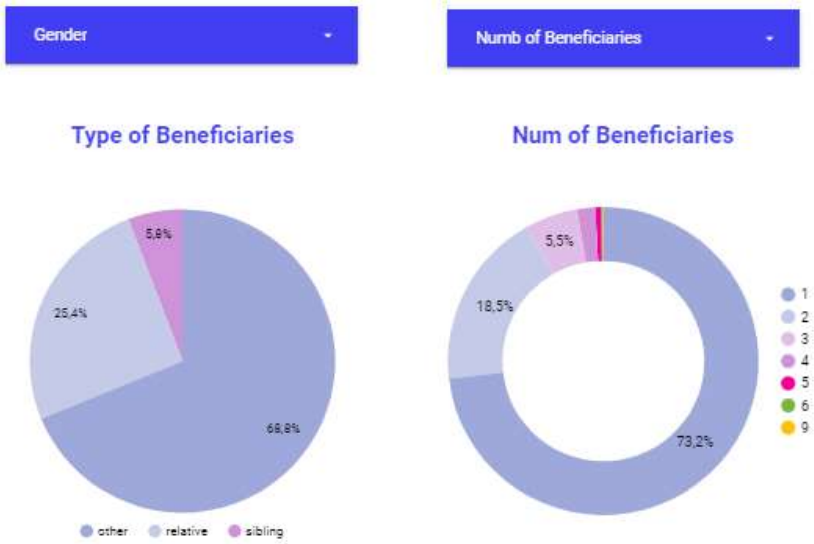

Figure 4. Physical person: beneficiaries' distribution

Figure 4 indicates that most people designate a person with whom they do not share a surname as a beneficiary. This leads us to believe that the most common type of beneficiary is a significant other. We also see that the second highest number is associated with family members, for example, children. Additionally, if we look to the right, we see that most people designate a single beneficiary, which again leads us to think that the most common tendency is to leave a partner as a beneficiary.

Inference Analysis

After analyzing the data in a descriptive way, it seems interesting to perform an initial analysis around personal data such as age, capital, and body mass index (BMI). To this end, the most common partitional clustering techniques are applied.

On the one hand, the K-means algorithm is one of the most popular iterative clustering methods and it is appropriate for situations in which the variables are of quantitative type (Macqueen, 1967). It is based on the squared Euclidean distance. The idea behind this algorithm is constructing clusters, given a number of K so that the total within-cluster sum of squares (i.e., variance) is minimized.

On the other hand, the K-medians, which is a variation of the K-means algorithm where, instead of calculating the mean for each cluster to determine its centroid, the median is obtained (Kaufman & Rousseeuw, 2009).

The algorithm chosen to run the analysis in this section is K-Means since it is a popular and effective method for clustering data that has continuous variables, such as age, capital, and BMI.

Before proceeding with the algorithm, it is important to select the desired variables for the analysis (age, capital, and BMI). The data is obtained by connecting the data available in BigQuery to Python and downloading the data from the desired database using SQL.

Once we have selected and processed the data, we must decide the optimal number of clusters. For this, there are several approaches such as the elbow method, which plots the sum of squared errors (SSE) for each value of K (number of clusters) and looks for an "elbow" in the plot, indicating the best value for K.

The assignment of the points to the three groups can be seen in Figure 5. Inference Analysis. Intra-cluster evolution in which the algorithm classifies the data into the corresponding cluster after choosing the number optimal of groups, i.e., K=3.

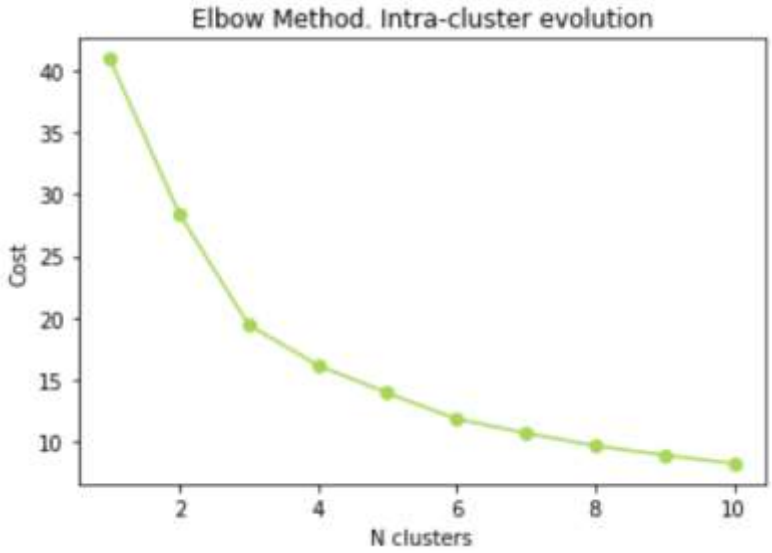

Figure 5. Inference Analysis. Intra-cluster evolution

The next step is to run the algorithm setting 3 as the number of desired clusters. The centroids for each cluster are shown in Table 1.

	capital	BMI	age
0	125195.945946	24.319054	36.927928
1	104053.299492	27.222640	51.040609
2	345079.365079	25.278095	40.444444

Table 1. Inference Analysis. Cluster's centroids

Table 1. illustrates three very distinct clusters. In the first place, there is a group formed by the youngest people with low capital. Next, there is the oldest group, characterized by the lowest capital and the highest BMI. The last group represents older people than the ones in the first group but with higher capital.

In addition to this, a 3D plot of the clusters has been included in the analysis to reflect the distribution of the clusters around the variables used in the model, BMI, capital, and age.

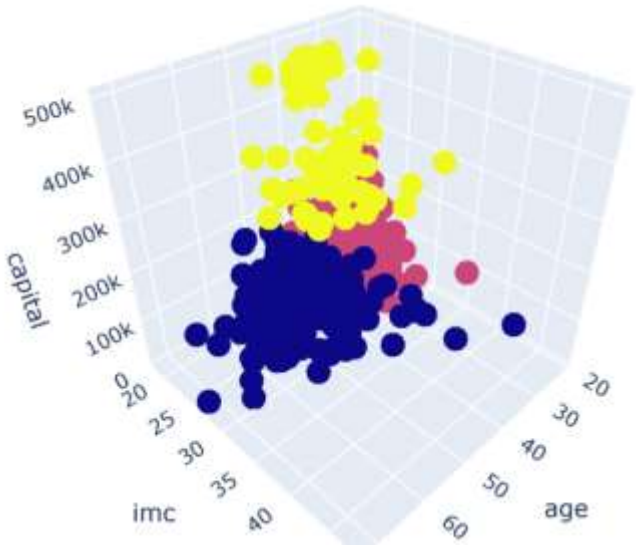

Figure 6. Inference Analysis. Clusters' 3d representation

These are some of the variables that help us understand the motive behind a life insurance purchase, which is useful when customizing marketing campaigns or a particular product. Let's now continue with this last idea.

There are typically two departments in every company specially focused on attracting clients, marketing, and product. To facilitate their duties, the research includes two different client approaches, one from each department's point of view.

3.2.2. PROFILING CLIENTS FOR MARKETING PURPOSES

The goal of a marketing department is to develop and put into action plans that will assist a business in boosting sales, increasing brand recognition, and reaching its target market. This entails planning marketing initiatives, producing content, running social media pages, conducting market research, and gathering and evaluating data.

Overall, the strategic role of marketing in driving company success includes building customer value, using data and analytics to guide decision-making, working with other departments, and taking responsibility for outcomes (Morgan, 2012).

Social media analysis

One of the most popular tools to reach potential clients is social networks. A marketing department can benefit greatly from social media in many different ways. They can be utilized to reach a larger audience, develop relationships with clients, raise brand recognition, and produce leads. Social networks can be used to track client feedback and discover patterns in consumer behavior. Additionally, they can also be applied to create targeted campaigns and advertise certain deals and discounts (Tiago & Veríssimo, 2014).

An example of this is Facebook. This social platform has had a tremendous impact on our lives. It has changed the way we communicate, share information, and stay connected with friends and family. It has also become an important platform for businesses to reach their target audiences. In short, Facebook has become an integral part of our lives and its influence is only growing. Consequently, it can help the marketing department reach potential clients in a variety of ways.

First, it can be used to create targeted ads that are tailored to the interests and demographics of the desired audience. Additionally, it can be used to create engaging content that will draw potential customers in and encourage them to interact with the brand. Finally, Facebook can be used to build relationships with potential customers by responding to comments and messages, likewise providing customer service.

As a result, it seems interesting to run an analysis around other personal data available on Facebook such as gender, age, and location.

The algorithm chosen for the analysis is K-Prototypes (Huang, 1998) as the analysis involves numerical variables as well as categorical. K-Prototypes is an unsupervised machine learning algorithm that combines the K-Means clustering algorithm with the concept of prototypes used to cluster data that contains both categorical and numerical features. Given a number of groups K, the algorithm first assigns each data point to a cluster based on its numerical features, then allocates the different data points to a cluster based on its categorical features. The groups are then refined by iteratively reassigning data points to the aggregations based on their distance from the cluster's prototype. The prototype for each group is a combination of the most common values for each feature in the collection. The algorithm continues until it converges on a solution where no further reassignments are necessary.

In this case, the data includes information about the gender, age, and location of a sample of leads and clients. Again, it is obtained by connecting BigQuery to Python and downloading the information from the desired database using SQL.

After selecting and processing the data it is necessary to choose the numbers of clusters. As done previously we represent the elbow curve and look for the optimal value of k.

According to the graph shown in Figure 7, there are six possible groups.

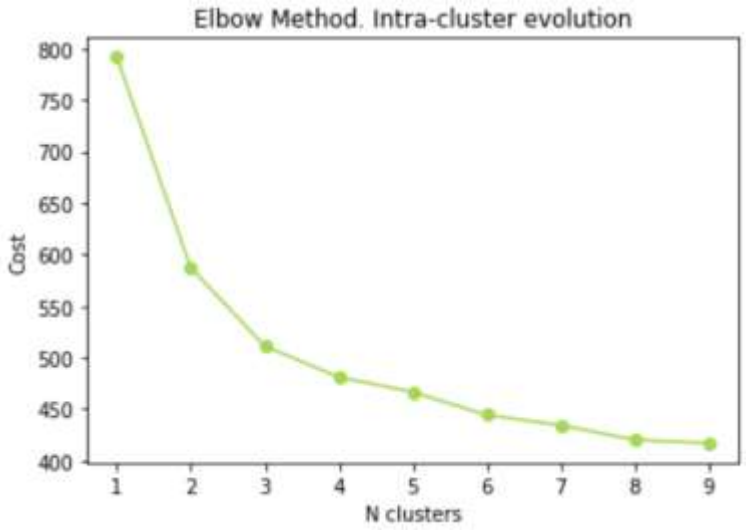

Figure 7. Social media analysis. Inference Analysis. Intra-cluster evolution

The next step is to run the algorithm setting 6 as the number of desired clusters. If we categorize the groups according to the variables used in the model, we obtain the results shown in Table 2.

Segment	Total	province	gender	age
First	1756	Sevilla	male	43.248292
Second	2734	Barcelona	male	48.120337
Third	2110	Barcelona	female	52.633649
Fourth	2986	Madrid	male	57.653382
Fifth	2646	Madrid	female	39.818972
Sixth	2229	Madrid	male	34.131898

Table 2. Social media analysis. Cluster's centroids

It is significant to highlight that in general the clusters have a similar number of leads.

From Table 2 we can see that the most common province is Madrid as it appears as the dominant one in three out of six clusters, followed by Barcelona. We can also appreciate that most groups are formed by a higher percentage of males than females.

We can complete this information with visualizations around the given variables (Figure 8 and Figure 9).

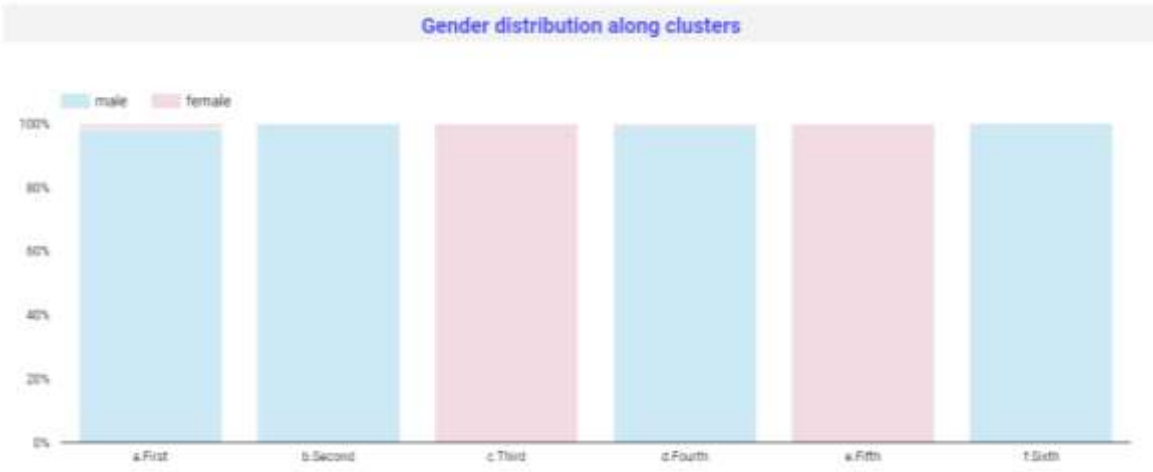

Figure 8. Social media analysis. Gender distribution along clusters

Figure 8 shows how most of the data corresponds to men. In addition, almost 100% of the data in all groups corresponds to a single gender. This is of great interest given that from a marketing perspective, gender can affect how a product or service is marketed (Wolin, 2003). For example, marketers may use different language, imagery, and messaging when targeting male and female audiences (Bui, 2021). Additionally, marketers may tailor their approach to targeting different genders in order to better appeal to their target audience (Moss et al., 2006).

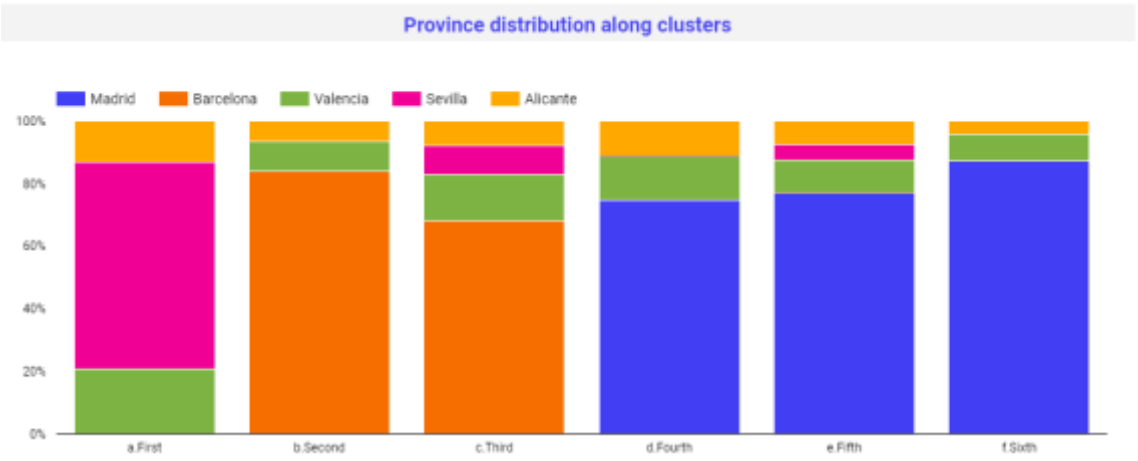

Figure 9. Social media analysis. Province distribution along clusters

Figure 9 shows the distribution of leads by province. We can see how Madrid appears with the highest frequency in half of the groups, followed by Barcelona. We also see that Valencia is positioned as a common province; however, it does not comprise the majority in any of the clusters.

It is important to remark that the information analyzed considers both potential clients, also known as leads, and actual clients. Up to this point, we need to decide which cluster to target first. In order to make this decision we are going to focus on the conversion rate, a measure that indicates how many people purchase a policy after registration.

Thus, it helps to analyze the effectiveness of an online marketing campaign or website in terms of generating.

It should be mentioned that at the time of downloading the data a binary variable was included, which was not included in the model, to indicate whether the lead had converted or not.

We define the conversion rate as follows:

$$conversion\,rate = \frac{\#\ of\ clients}{\#\ of\ total\ leads}$$

In order to choose the optimal cluster, we compute these measures for each of the clusters, the numbers are not included as they are confidential data. The results of this analysis show that it is possible to order and select the best clusters to target first as it has the highest conversion rate. As a consequence, the recommendation to a marketing team to try to reach customers through Facebook would be a campaign designed for a specific target audience.

An example of the K-Prototypes algorithm developed is shown in Annex I.

Marketing Segments

Continuing with the customer acquisition strategy from a marketing perspective, the study now focuses on profiling the actual clients of the company. The idea behind this is to better understand what leads a person to get life insurance.

Understanding the customers' identity and preferences is vital for various reasons. First, it enables the company to customize their products and services to satisfy the specific requirements of their target audience. By comprehending the needs of customers, the firm can create products and services that are high in demand and therefore gain more success in the marketplace (Dougherty, 1990). This approach can differentiate the business from its competitors and build up customer loyalty.

Second, knowing the audience of a certain industry can help find new development and expansion prospects. Successful business owners are aware of what their customers want and how best to deliver their goods or services. Finding new market possibilities or consumer categories to target is made simpler by having greater knowledge of customers' interests, preferences, and purchasing habits (Simonson, 1993).

Customers need to feel valued and need to clearly understand the value of what a company has to offer them. Overall, understanding customers is crucial for identifying new opportunities for growth and expansion, optimizing the customer experience, and fostering customer loyalty.

Finally, improved interactions with customers can be facilitated by understanding them. Building trust and loyalty with consumers can result in higher sales and higher rates of customer retention by demonstrating that the company is aware of their needs and is prepared to go above and beyond to meet them (Hoffman et al., 1999).

The variables included in this part of the research are linked to personal conditions and behavioral attitudes. These are age, beneficiary type, purchase intention, the time needed to convert into a client after the registration, whether they have a risky job or not, number of relatives, and number of pathologies.

On this occasion, the data processing is somewhat more complex than in the previous section. There is no dataset containing all the information corresponding to the different pathologies.

There is a code associated with each of the possible pathologies, but this code varies depending on the version of the agreement with the insurers offering the life insurance contract and the insurer itself. That is why a table unifying all the codes and pathologies was made before running the algorithm. In addition, the variable considered in the analysis is the number of pathologies, not the pathology itself, so this variable must be created first. Once this has been done, it must be considered that there is a specific code associated with "no pathology", so the number of zero pathologies must be associated with it.

Finally, it is worth mentioning that this time the information corresponds to "quotes", i.e., those leads that have answered all the marketing questions in the funnel.

Once all the data preprocessing is done, we proceed to the elaboration of the model. The algorithm used is K-prototypes, since the dataset again combines both numerical and categorical variables. In this case, the elbow curve is as follows:

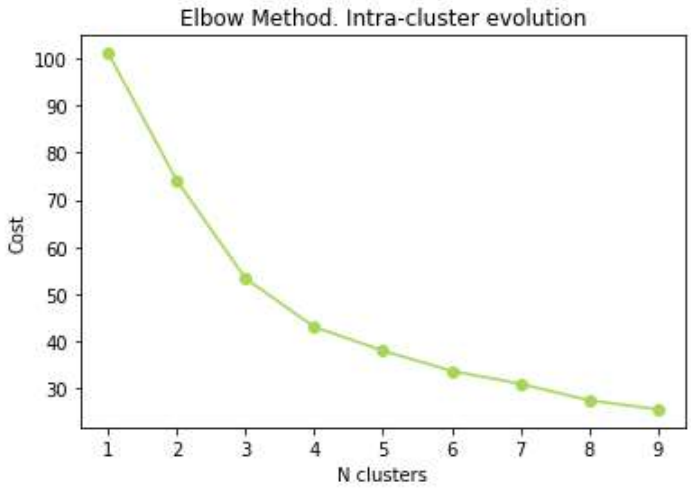

Figure 10. Marketing segments. Intra-cluster evolution

On this occasion, the number of groups to be created is set to 7 and then the algorithm is launched. The characteristics of the clusters are the following:

	Segment	Total	age	beneficiary_type	risk_job	intention	conversion_time	num_relatives	num_pathol
0	First	206	38.432039	physicalPerson	no	loan	7.305825	0.067961	0.053398
1	Second	163	50.840491	physicalPerson	no	loan	8.288344	0.104294	0.098160
2	Third	92	47.163043	physicalPerson	no	family	13.065217	2.315217	0.206522
3	Fourth	68	59.455882	physicalPerson	no	family	12.205882	0.544118	0.161765
4	Fifth	190	48.015789	physicalPerson	no	family	5.621053	0.042105	0.131579
5	Sixth	31	46.322581	physicalPerson	no	loan	153.193548	0.064516	0.258065
6	Seventh	216	35.962963	physicalPerson	no	family	5.537037	0.319444	0.129630

Table 3. Marketing Segments. Cluster's centroids

Table 3 shows that the prevailing profile is characterized by a job with low risk, with a physical person designated as the beneficiary and either family or loan as the intended purpose. Additionally, most of the leads seem to be in good health without any medical ailments (number of pathologies around zero). Furthermore, only the third cluster displays a notable feature of having more than one relative involved.

The variable related to the number of relatives is intriguing. Initially, it may seem reasonable to assume that the primary beneficiaries of the insurance are children. However, this outcome suggests that a considerable number of individuals name their partner as the beneficiary (number of relatives around zero). A possible explanation for this finding is that since there is an option to indicate "legal heir" as the beneficiary in the funnel, it becomes difficult to count the number of children.

To continue, further research about pathologies has been developed.

Pathologies analysis

Researching prevalent diseases can be highly beneficial as it enables the identification of potential customers. By categorizing customer profiles based on illnesses, products can be personalized to be more appealing to consumers.

According to the data reflected in the table above, it appears that taking out life insurance is not closely linked to having a disease. This can be seen in the cluster centroids; the mean number of pathologies is less than 1 in all groups.

Another way of looking at this is from Figure 11.

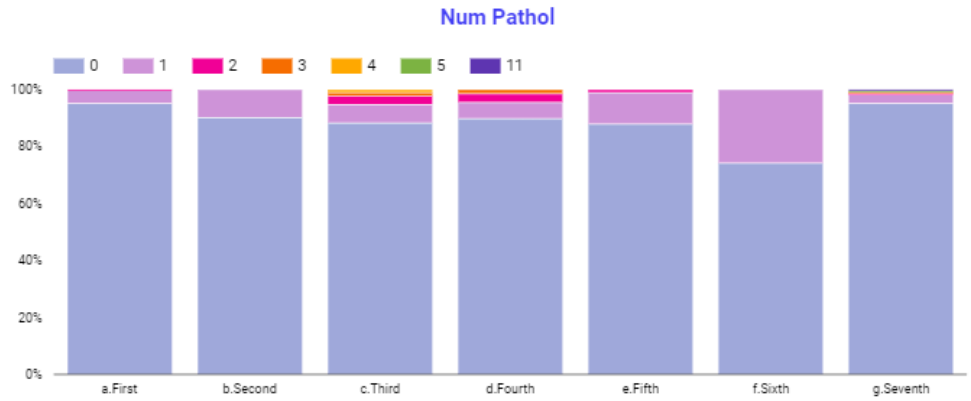

Figure 11. Frequency of the number of pathologies per cluster

Figure 11 shows that the most repeated number, zero, is related to not having any type of pathology. However, it is interesting to examine how often possible pathologies appear for what has been previously mentioned.

The following graph has been prepared for this purpose.

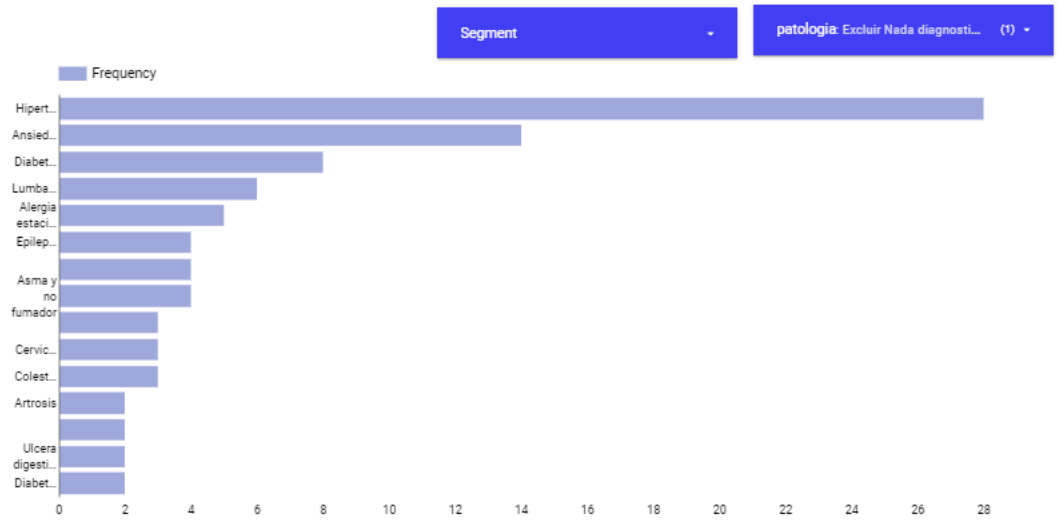

Figure 12. Frequencies of the different pathologies

According to Figure 12 the most common pathologies are hypertension and anxiety, followed by diabetes.

It should be noted that all these graphs are part of a dashboard prepared with Data Studio so that they are all connected to each other and by means of filters it is easy to obtain the characteristics of each group.

For example, in Figure 13 it can be seen that in the case of the second cluster, the dominant pathology is hypertension.

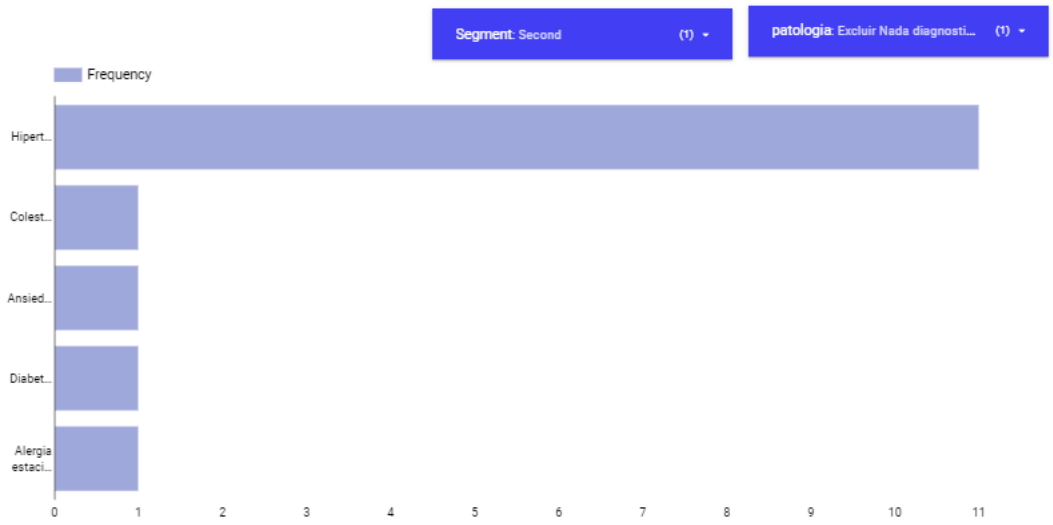

Figure 13. Frequencies of the different pathologies for a certain segment

In short, this section shows that whether or not a person suffers from a certain pathology is not a determining factor when purchasing life insurance. However, there are some dominant pathologies, a fact that could represent a business opportunity for insurance companies.

Following the previous idea, it is important to remark that there are many individuals, such as diabetics, who encounter a greater number of problems when it comes to finding a suitable life insurance policy for them, as there are several factors that can make getting life insurance more difficult for specific groups of people.

One explanation is that patients with specific medical conditions, such as diabetes, may be seen as a high-risk client for the insurance provider. This means that if they were to insure this person, they might be more likely to incur charges. Due to the increased risk, the insurance company might be less likely to offer coverage or may impose higher premiums (Low et al., 1998).

It's important to remember that getting life insurance is not impossible, even though it could be more challenging for some people with diabetes (Ramlau-Hansen, et al., 1987).

This aspect is precisely one of Getlife's value propositions, offering insurance for particular diseases such as diabetes.

The strategy of including and covering particular disease groups in a business's scope can be beneficial for many motives. For instance, it may aid in expanding the pool of possible customers. A company may be able to reach clients that might not have otherwise been accessible by focusing on diseases. This may encourage an increase in sales and revenue.

Additionally, focusing on diseases may expand the diversity of the consumer base. This may result in a more inviting and inclusive environment, which may appeal to a broader spectrum of clients. Moreover, it can support the development of a positive brand image and reputation. A company can stand out from rivals and come seen as more socially conscious and progressive by demonstrating a commitment to inclusion and diversity.

Lastly, it can encourage creativity. A firm may be able to spot new opportunities and find more inventive solutions to issues by bringing in fresh viewpoints and experiences. This can assist in achieving growth and competitiveness.

In order to investigate structures between leads and diseases through the use of networks and graph theory a social network analysis (SNA) was carried out.

Social network analysis (SNA) is the process of investigating social structures through the use of networks and graph theory, and it can be used to identify important components of the network, measure patterns, predict future behavior, identify key players, communities, and patterns in a network, and identify different types of social networks (Wasserman & Faust, 1994).

A sample of the leads is shown in Figure 14, in which the overall relational structure is characterized in terms of nodes (leads and diseases colored in blue and pink, respectively) and edges that connect them. It is clear that most of the leads are completely healthy or have had a common urinary infection.

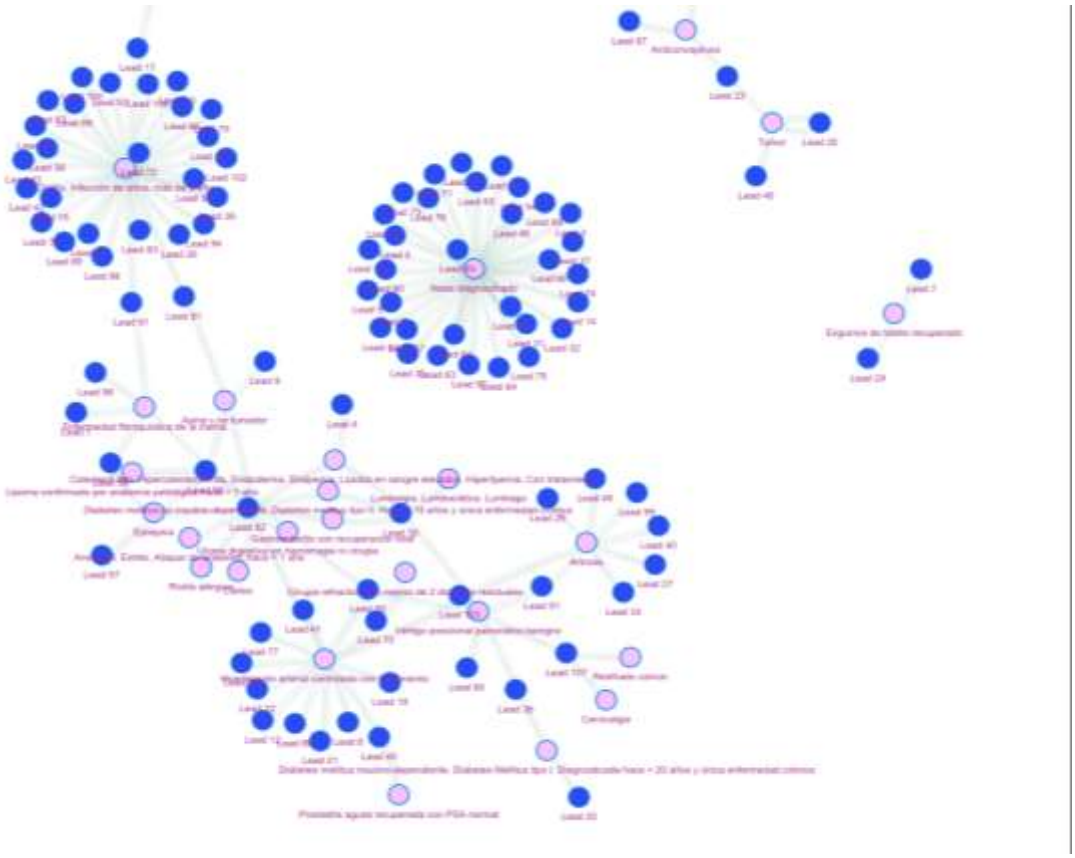

Figure 14. Structures between leads and diseases

3.2.3. *PROFILING CLIENTS FOR PRODUCT PURPOSES*

The product department is crucial to a company's success since it is in charge of identifying, creating, and bringing to market a product, which is frequently the principal source of income.

One of its central functions involves defining the features and specifications of the product, which plays an essential role in ensuring that the product meets the customer's needs and preferences. The product management function involves settling on the overall strategy for the product, which includes deciding on target markets, price, and a development schedule. Additionally, the department is in charge of promoting the product to potential customers and partners, as well as working with sales teams to drive adoption and revenue. Finally, it carries out product analysis and reporting, which involves tracking the performance of the product, analyzing customer feedback, and making data-driven decisions to improve the product and its positioning in the market (Ulrich et al., 2008).

Overall, it contributes significantly to a company's success by creating and delivering goods that satisfy customers' wants and brings in money, which is why the research's focus in this section is on creating customer profiles from the perspective of the product department.

Product Segments

As previously discussed, while facing the marketing perspective analysis, we are going to run a clustering algorithm to try to identify what characterizes the profile of policyholders.

This section concentrates on the purpose behind purchasing life insurance in this instance, as well as its intended beneficiaries. Understanding customers is essential since it can result in customized advertising and products that appeal to the general public more successfully, as previously discussed.

The variables chosen in this analysis are age, beneficiary type, intention, and conversion time, which is the time it takes a lead to purchase the products since their creation. These variables are believed to help understand the reason behind buying a life insurance policy and help the company to develop specific products to attract more clients. The article "A Study on Factors Affecting the Purchase of Life Insurance" by Madhumathi & Dr. K. Balamurugan (2014), and the report "Life insurance consumer purchase behavior Tailoring consumer engagement for Today’s middle market" by Sharps et al., (2015), discuss the relevance of these variables in understanding the reasons behind buying a life insurance policy and how they can be used to develop targeted products for different segments of the population.

It is crucial to note that since the goal of this study is to discover what motivated policyholders to purchase insurance, we will only work with existing clients in this instance; information on potential clients will be eliminated.

In the first place, we download the desired data just like we did in the other cases. Once we have obtained the data, we plot the Elbow curve to find the optimal number of clusters (Figure 15). According to Figure 15, the elbow lies around three clusters.

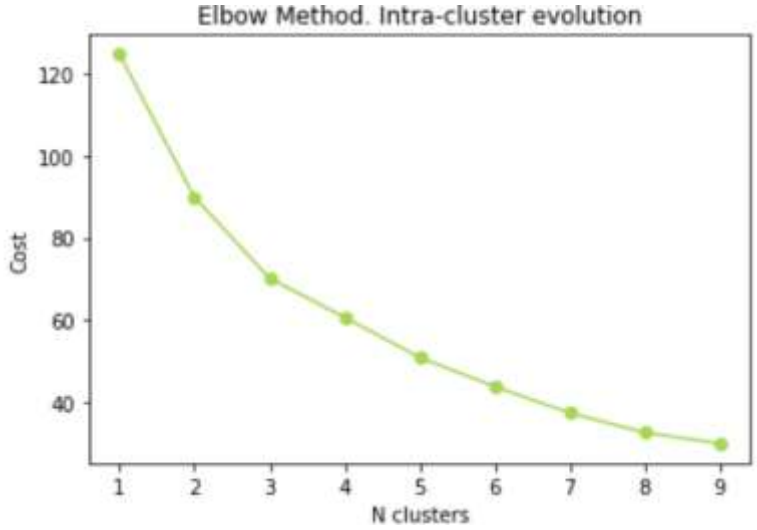

Figure 15. Product segments. Intra-cluster evolution

Given this result, we can identify each of the groups that we obtained in the previous analysis with the new ones so that the marketing and product departments can work together on personalized campaigns linked to specific products for each of the aggregations based on the purchase intention and the beneficiary of the policy.

We now run the algorithm for the number of clusters equal to 3 and the results obtained are shown in Table 4. Product segments. Clusters' centroids

	Segment	Total	age	beneficiary_type	intention	conversion_time
0	First	413	44.619855	physicalPerson	loan	14.196126
1	Second	420	38.033333	physicalPerson	family	5.945238
2	Third	432	50.25	physicalPerson	family	14.039352

Table 4. Product segments. Clusters' centroids

From Table 4, it can be observed that the majority of people assign a physical person as the beneficiary. Moreover, most people pursue life insurance with their family or a loan in mind.

Referring to the intention variables, it seems reasonable to believe that many clients may acquire life insurance to ensure that the surviving family members can maintain their standard of living. In addition, if we consider the variable age as a relevant factor, we can picture the possibility of people with children.

If we recall the marketing analysis, we saw that all of the groups were mostly female or male, which helps when launching a marketing campaign (Figure 8).

It might be noticeable that the variable gender has not been included in this analysis. The reason behind this is that from a product perspective, the variable gender is not relevant. Thus, it should not be considered a variable that causes a major effect on the groups. Nevertheless, a representation of the gender distribution among the clusters is shown in Figure 16.

Figure 16. Product segments. Gender distribution among clusters

In addition to this, we also include a few representations considering the distribution of intention and beneficiary type in the clusters (Figure 17).

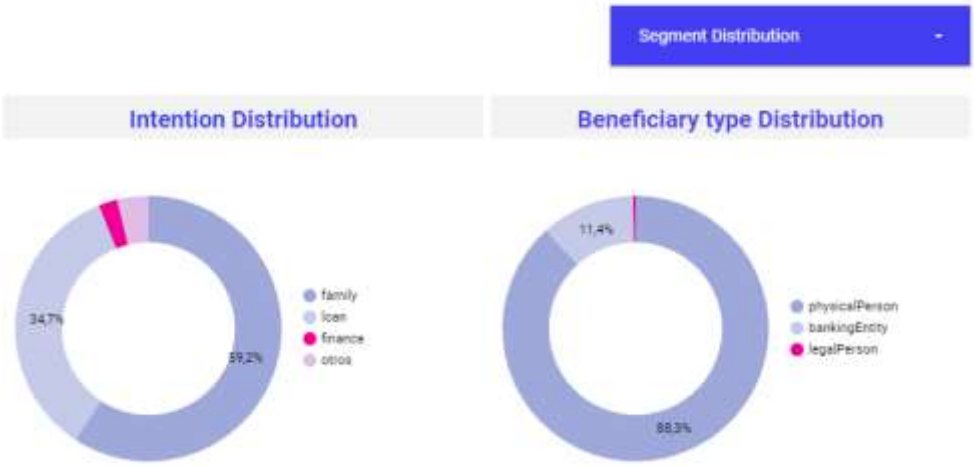

Figure 17. Product segments. Intention and beneficiary type distribution

These pie charts (Figure 17) show that most people take out a life insurance policy with some intention related to their family and that in almost 90% of the cases, the beneficiary is an individual, as we had previously intuited.

It should be noted that these graphs are linked to a filter, which allows us to obtain these values for a specific segment or group of segments.

Next, we represent the conversion time distribution in Figure 18Figure 17, and as it can be seen most of the leads convert at the initial purchase trigger, something that indicates that the first 24 hours are critical.

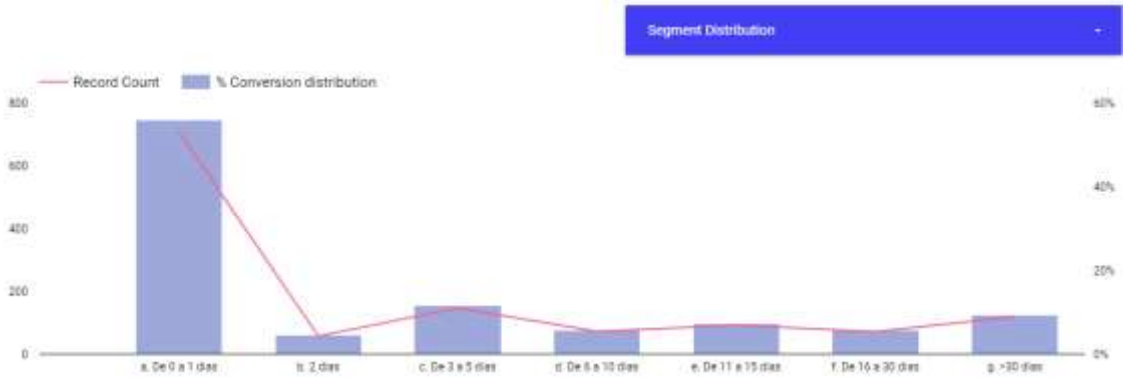

Figure 18. Product segments. Conversion time distribution among clusters

Finally, to complete the analysis, two other clustering models have been carried out, but this time not only customer data have been considered. In the first one, data from initial quotes are analyzed, i.e., those leads that have answered all the initial demographic questions. In the second, upsell data is considered. The idea of these two analyses is to use the K-Prototypes algorithm to determine the characteristics of the groups with the best conversion rate.

The results from the second analysis are shown in Table 5:

segment	total	age	gender	capital	province	intention	income_ranges	smoke	product	imc	price
First	888	40.62	male	139734.23	Madrid	family	range6	no	premium	25.97	20.74
Second	723	54.48	male	90605.12	Barcelona	family	range6	no	basic	29.29	42.22
Third	546	47.00	female	120945.97	Madrid	family	range6	no	basic	24.07	25.40
Fourth	501	43.32	male	135523.23	Madrid	family	range6	yes	premium	25.44	24.19
Fifth	751	44.30	male	110303.66	Barcelona	loan	range6	no	premium	25.56	21.60
Sixth	457	41.03	male	417621.44	Madrid	family	range6	no	premium	25.99	62.63

Table 5. Upsell data clusters' centroids.

Note that if we calculate the conversion rate for each cluster we can determine the best-converting profiles which have an important impact in the business.

In addition, if we compare the results obtained in the analysis of initial quotes with those of upsell, we obtain a higher conversion rate in the groups of the second analysis, as expected. The further you advance in the funnel, the more likely it is that the lead converts.

Comparing initial quotes to upsell allows us to confirm the initial hypothesis of relevant segments. Nevertheless, caution should be undertaken. Continuously optimizing and prioritizing demographics that convert better can lead to a cold-start problem for the company as chances are limited for other segments to thrive and grow too.

3.2.4. SURVIVAL ANALYSIS

Survival analysis is a collection of statistical methods focusing on the response of interest is, in this case, time until some specified event (Schober & Vetter, 2018). The event may be the failure of some electronic component in the engineering field (Smith, 2017), but also may be the purchase of some product (Prinzie & Van den Poel, 2007), customer churn (Larivière, & Van den Poel, 2004), or load default (Pelaez-Verdet & Loscertales-Sanchez, 2021). Although this survival analysis is quite often used in the biomedical field in which the event may be death and the time that an individual survived over some period is analyzed to measure the effectiveness of treatments or to predict the likelihood of an individual surviving a particular condition (Kasza et al., 2014).

The survival function is the probability that an individual or group will survive for a certain period of time (Jenkins, 2005).

There are several conclusions that can be drawn from survival curves. One of the most basic conclusions that can be drawn is the overall survival rate of a group or population. By examining the curve, it is possible to determine the percentage of individuals who survive for a given period of time (Sullivan, 2016).

Survival curves can be used to compare the survival rates of different groups, such as a group of cancer patients who received a particular treatment and a group of cancer patients who did not receive the treatment. The comparison of the two curves can determine the effect of the treatment on survival. If the curve for the treated group is steeper, meaning there is a greater probability of survival, it can be concluded that the treatment had a positive effect on survival (Berkson & Gage, 1952).

In addition to comparing survival rates, survival curves can also be used to identify factors that may influence survival. For example, if a survival curve is created for a group of cancer patients, and the curve is stratified by age, it may be possible to identify an association between age and survival. If the curve for younger patients is steeper than the curve for older patients, it can be concluded that age is a factor that influences survival (Liu et al., 2019).

Finally, they can be used to predict the likelihood of survival for an individual or group. By examining the curve and the relevant characteristics of the individual or group (e.g., age, treatment status, underlying risk factors), it is possible to estimate the probability of survival at different time points. This can be useful for forecasting purposes and for making decisions about treatment or other interventions.

In this section we will use survival curves to analyze the time it takes to purchase a product. Thus, it is very important to have data on the length of time it takes for each lead to make a purchase. We will define the survival time as the difference between the date when the lead was created, and the purchase date associated with the lead.

Apart from the survival time, an indicator of the purchase status (censoring indicator) and a covariate for the range salary are both used.

The estimation of the survival function can be performed through the most frequently used Kaplan-Meier estimator (Kaplan & Meier, 1958).

Model

In order to perform the analysis, we are going to use the clustcurv R package (Villanueva et al., 2021). Particularly, this package allows estimating survival curves and grouping them if they are not equal.

To estimate survival curves with the survclustcurves function (Villanueva, 2022), it is necessary to provide survival time, censoring indicator, and a covariate as input to the function.

The function will then fit the specified model to the data and provide a survival curve for each cluster. The resulting plot will show the survival curves for each cluster, with the curves for different clusters distinguished by different colors or symbols.

The survclustcurves function can be useful for analyzing survival data when the observations are not independent, as it allows you to take the clustering structure into account when fitting the model and creating the survival curves. By examining the survival curves for different clusters, you may be able to identify factors that influence survival and compare the survival rates of different groups.

The data used to elaborate the model is obtained by connecting the data available in BigQuery to R and downloading the data from the desired database using SQL. The information that we are going to use includes leads and clients.

In the first place, we need to prepare the dataset so that it has the variables needed to elaborate the model. The first thing is to create survival time. For this analysis, the variable time consists of the difference between the creation date of the lead and the purchase date. If a lead has not converted to a client, the time would be the difference between its creation date and the date “30/11/2022” set as the end date of the analysis.

Next, we proceed to declare the censoring indicator of the process; 0 if the total time is censored and 1 otherwise. Note that the total time will be censored if the lead has been converted.

Last, we must set a categorical variable indicating the population to which the observation belongs. For the analysis, we have decided to choose the variable income, which has six levels related to different income ranges.

Once the data has been prepared, we can proceed with the algorithm and present the results (Figure 19).

```
# Download the data
data_raw <- bq_project_query(project_id,sql)
df <- bq_table_download(data_raw)

# Inspect the data
df <- df[!is.na(df$income),]

table(df$income)
df$income <-as.factor(df$income)

# Model
res <- survclustcurves(time = df$event_time, status = df$purchase_event,
      x = df$income,algorithm = 'kmeans', cluster=TRUE, nboot = 200)

# Curves representation
p <-autoplot(res)
ggplotly(p)
```

The algorithm generates six survival curves, one corresponding to each of the different income levels, and classifies them into five distinct groups, each of them identified by a unique color. The estimated survival curves can be observed in Figure 19.

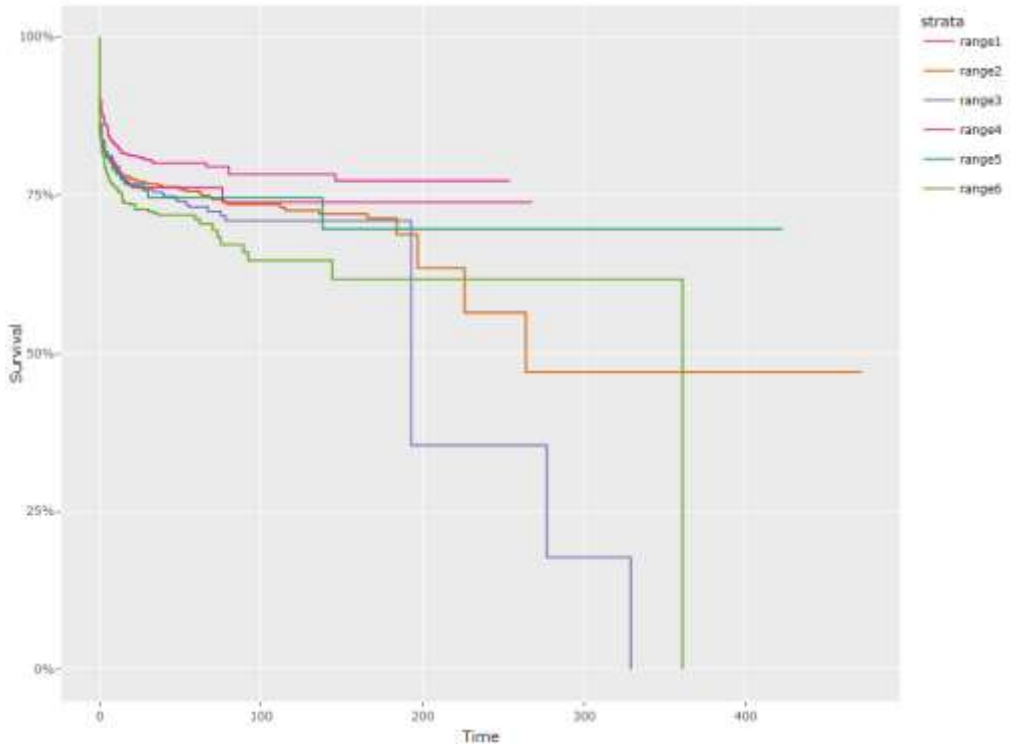

Figure 19. Estimated survival curves for each of the levels of the variable "income".

From Figure 19 it can be inferred that each curve is assigned a specific color that corresponds to the group it belongs to. In this case, there are five groups denoted by K=5. Each curve portrays the survival rate based on the salary range, while the color represents the associated cluster for each level. Analyzing the survival curves allows us to infer the duration it takes to complete a purchase.

The x-axis represents the time in days that a lead could take to convert into a client whereas the y-axis indicates the survival probability at any particular time.

If we stare at the figure, we see that the curve is steepest at the beginning and then flattens out over time in most cases. This may indicate that most leads make a purchase relatively quickly after being contacted, with a smaller percentage making a purchase later on. It should be highlighted that the survival curve with range6, which experiences the highest purchases in the first few days after lead creation.

Note that for the level range 3 and range 6 the latest time is not a censoring achieving this survival probability to zero, unlike the range1, range2, range4, and range5.

This is very useful because direct instructions can be given to the advisors to contact this type of lead as soon as possible since it is not likely that they will purchase life insurance after a certain time from their creation.

Finally, I would like to mention that this algorithm has been applied considering categorical variables (e.g., location,) indicating the population to which the observation belongs. However, all the results reflected the existence of a single cluster, i.e., the equality of the survival curves.

We also ran a different survival analysis but considering the cancelation date as the event time (measure as the difference between the cancelation date and the purchase date), the censoring indicator as 0 if canceled or 1 otherwise, and the categorical variables such as location, intention, or income. In all cases, the results reflected one single cluster. Thus, we can conclude that none of the categorical variables has a significant effect on survival that could have been used to identify which populations are more likely to cancel their policies.

Since the survival analysis did not show any relevant conclusions about canceled policies, we have decided not to include the results in this work. However, a descriptive analysis of canceled policies has been carried out.

3.2.5. CANCELATIONS

We have selected the data linked to canceled policies and developed a dashboard in DataStudio showing the relevant characteristics of cancellations.

In the first place, we computed several statistical measures to determine which position faced Getlife versus the market. The most important one is the cancellation rate, which represents the number of canceled policies divided by the total number of sold policies.

This measure is of great interest, as it can be used to measure customer loyalty, assess customer retention strategies, and evaluate the competitiveness of insurance companies in the market.

The paper "Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry" by Ahn et al. (2006), highlights the importance of understanding churn rate in different industries, such as telecommunications and insurance, and uses various data analysis techniques to predict and understand churn. Additionally, it can be used to compare the performance of a company to the whole industry.

After calculating the cancellation ratio, we move on to look at the cancellations' most important features. First, we plot the time elapsed from the time of purchase to the date of cancellation.

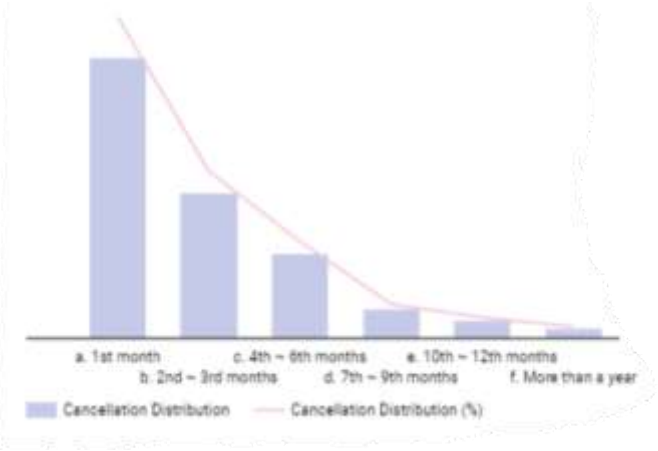

Figure 20. Cancellations' distribution

The Figure 20 shows that most cancellations take place within the first month or during the following two months. This leads us to think that there may be some relationship between the end of a promotion period, such as paying only 50% of the policy during the first two months, and the cancellation of the policy. We also see how, as time goes by, the number of cancellations decreases, so we can intuit that the clients are satisfied with the product offered.

Given that the highest number of cancellations occurs during the first month, it is interesting to examine this data in greater depth. The following graphs have been prepared for this purpose (Figure 21).

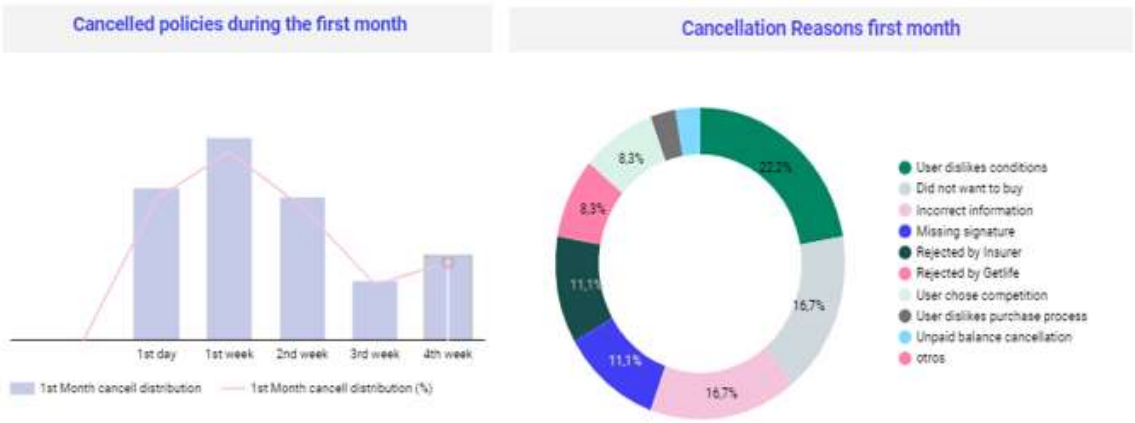

Figure 21. Left panel: Cancelled policies during the first month after purchase. Right: Cancellation reasons' frequencies during the first month

The first representation shows that most of the cancellations that take place during the first month occur in the first few weeks, which suggests that some type of error may be a common cause of cancellation. The second graph shows the most common reasons for policy cancellations. As expected, a large percentage corresponds to errors (purchase error, incorrect information, or lack of information). We also see that a considerable % corresponds to disappointment with the conditions.

This last point may be of great value, given that if the option of modifying the contract conditions for a number of users were to be considered, it could increase the retention rate.

Additionally, we examined the distribution of leads according to the status prior to cancellation.

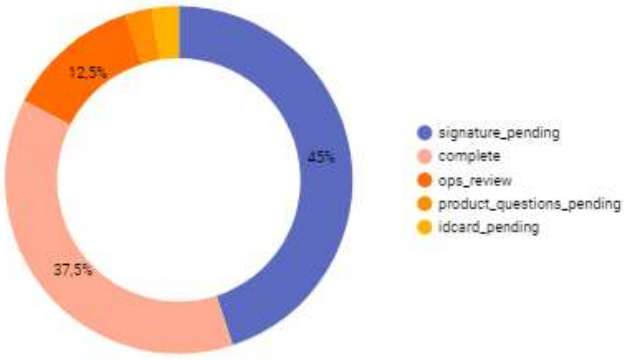

Figure 22. Lead status frequencies before cancellation

In Figure 22 it can be observed that most of the cancellations are due to policies that were not properly formalized because the client's signature was missing.

Finally, regarding the issue of cancellations, I would like to point out that insurance companies have the right to reject a potential client for different reasons, such as, for example, an individual engaging in high-risk activities. Thus, this could also be a reason why a policy gets canceled and an intermediary such as Getlife cannot control it.

Chapter 4. Conclusions and further research

Analytics and machine learning techniques are having a significant impact on the life insurance industry and its products and services.

These technologies can help life insurance companies make more accurate risk assessments, which can result in more precise underwriting, leading to more customized policies and lower premiums for low-risk customers.

In addition to this, they can be used to identify fraudulent activity in real-time, allowing life insurance companies to quickly detect and prevent fraud. Particularly in this industry, it is vital to be careful about this, as the demographic characteristics of clients and potential clients can be strong indicators of risk. Factors such as age, gender, health status, and lifestyle choices can all affect a person's likelihood of dying during the term of a life insurance policy. By using demographic information to assess risk, life insurance companies can better understand the likelihood of a claim being made and set premiums accordingly.

For example, a younger, healthy person is generally considered to be a lower risk than an older, unhealthy person. As a result, a younger, healthy person would typically be offered a lower premium for a life insurance policy than an older, unhealthy person. By considering demographic characteristics, life insurance companies can better assess risk and price policies more accurately, which can help them stay financially stable over the long term.

This is precisely the idea that is developed throughout the project, considering the available demographic information to segment the customer base and target specific groups with tailored products and services.

Different analyses and algorithms have been run along the research to identify and segment customers based on their demographic information, behaviors, and other factors to help life insurance companies tailor their products and services to better meet the needs of different customer groups.

Having completed a thorough examination of the industry and its customers, it is appropriate to now summarize the major results of our study and emphasize their importance.

In the first place, finding the type of customers an industry attracts is a relevant aspect of market research and business strategy. It helps companies to better understand their target audience, develop more effective marketing campaigns, and improve their products and services to better meet the needs of their customers.

In terms of marketing strategies, there is a distinction between the profiles of males and females, allowing for the creation of targeted campaigns aimed at attracting new customers. Launching a marketing campaign for women is dissimilar to launching one for men as women and men have unique tastes, requirements, and behaviors that must be considered when designing a marketing campaign.

When it comes to illnesses, there appears to be no correlation between having a specific condition and purchasing a life insurance policy. However, some illnesses are more prevalent, such as hypertension, anxiety, and diabetes, which could be considered when thinking of business expansion.

Regarding the time between the creation of a lead and the purchase of a policy, people usually buy the product in a small number of days concerning the day of lead creation. As the days go by, it is less likely that there will be a conversion. On the other hand, leads with lower revenue are those that consume less as the days go by.

Lastly, it was not possible to identify groups of clients who have canceled their policies in the past. However, the analysis revealed that the majority of policy cancellations take place within the first month after purchase, largely due to issues with the contract terms or errors made during the purchasing process.

Further research could be done in this area by studying the use of machine learning techniques to improve claims processing by automating routine tasks, reducing processing times, and reducing the risk of errors, as this could help improve the service offered, attract more clients, and increase the retention rate.

Overall, the use of analytics and machine learning in the life insurance industry is helping companies to improve their operations, better understand their customers, and offer more relevant and personalized products and services. However, it is important not to focus all resources on a specific public, since the opportunity to attract new consumer profiles may be lost.

Bibliography

Adamova, M., Boudet, J., Kalaoui, H., & Segev, I. (2018). How traditional insurance carriers can disrupt through personalized marketing. *McKinsey & Company*. https://www.mckinsey.com/industries/financial-services/our-insights/how-traditional-insurance-carriers-can-disrupt-through-personalized-marketing

Agrawal, V., Balasubramanian, R., Gestal, A., Bernard, P.-I., de Combles de Nayves, H., Cummings Cook, K., & Kotanko, B. (2022). Global insurance report 2023: Reimagining Life Insurance. *McKinsey & Company.* https://www.mckinsey.com/industries/financial-services/our-insights/global-insurance-report-2023-reimagining-life-insurance

Ahn, J. H., Han, S. P., & Lee, Y. S. (2006). Customer churn analysis: Churn determinants and mediation effects of partial defection in the Korean mobile telecommunications service industry. *Telecommunications policy*, *30*(10-11), 552-568.

Berkson, J., & Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, *47*(259), 501-515.

Bernard, P.-I., Ellingrud, K., Godsal, J., Kotanko, B., & Reich, A. (2020). The Future of Life Insurance: Reimagining the industry for the Decade Ahead. *McKinsey & Company*. https://www.mckinsey.com/industries/financial-services/our-insights/the-future-of-life-insurance-reimagining-the-industry-for-the-decade-ahead

Browne, M. J., & Kim, K. (1993). An international analysis of life insurance demand. *Journal of Risk and Insurance*, 616-634.

Bui, V. (2021). Gender language in modern advertising: An investigation. *Current research in behavioral sciences*, *2*, 100008.

Dougherty, D. (1990). Understanding new markets for new products. *Strategic management journal*, 59-78.

Duncan, E., Fanderl, H., Maechler, N., & Neher, K. (2016). Creating value through transforming customer journeys. *Mckinsey, USA*.

Evans, N. D. (2003). *Business innovation and disruptive technology: Harnessing the power of breakthrough technology... for competitive advantage*. FT Press.

Hoffman, D. L., Novak, T. P., & Peralta, M. (1999). Building consumer trust online. *Communications of the ACM*, *42*(4), 80-85.

Huang, Z. (1998). Extensions to the k-means algorithm for clustering large data sets with categorical values. *Data mining and knowledge discovery*, *2*(3), 283-304.

O'Hearn, S., Mukhopadhyay, A., Carr, M., Trowbridge, E., & De Vido, L. (2019). Insurance trends 2019: Digital transformation shifts from threat to opportunity. *Part of PwC's 22nd CEO Survey trend series.* https://www.pwc.com/cl/es/publicaciones/assets/2019/pwc-2019-ceo-survey-insurance-report.pdf

Jenkins, S. P. (2005). Survival analysis. *Unpublished manuscript, Institute for Social and Economic Research, University of Essex, Colchester, UK*, *42*, 54-56.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American statistical association, 53(282),* 457-481.

Kasza, J., Wraith, D., Lamb, K., & Wolfe, R. (2014). Survival analysis of time-to-event data in respiratory health research studies. *Respirology*, *19*(4), 483-492.

Kaufman, L., & Rousseeuw, P. J. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Klatsky, L., Zhang, J., Udaltsova, N., Li, Y., & Tran, H. N. (2017). Body mass index and mortality in a very large cohort: is it really healthier to be overweight? *Permanente Journal*, *21*(3).

Krishnakanthan, K., McElhaney, D., Milinkovich, N., & Pradhan, A. (2021). How top tech trends will transform insurance. *McKinsey & Company (Issue September)*.

Larivière, B., & Van den Poel, D. (2004). Investigating the role of product features in preventing customer churn, by using survival analysis and choice modeling: The case of financial services. *Expert Systems with Applications*, *27*(2), 277-285.

Liu, W. X., Shi, M., Su, H., Wang, Y., He, X., Xu, L. M., ... & Li, Y. X. (2019). Effect of age as a continuous variable on survival outcomes and treatment selection in patients with extranodal nasal-type NK/T-cell lymphoma from the China Lymphoma Collaborative Group (CLCG). *Aging (Albany NY)*, *11*(19), 8463.

Low, L., King, S., & Wilkie, T. (1998). Genetic discrimination in life insurance: empirical evidence from a cross sectional survey of genetic support groups in the United Kingdom. *Bmj*, *317*(7173), 1632-1635.

MacQueen, J. (1967, June). Classification and analysis of multivariate observations. In *5th Berkeley Symp. Math. Statist. Probability* (pp. 281-297). Los Angeles LA USA: University of California

Madhumathi, P., & Balamurugan, K. (2014). A Study on Factors Affecting the Purchase of Life Insurance. *Asian Journal of Research in Business Economics and Management*, *4(3),* 72-82.

Mckinsey & Company (2016). Harnessing the power of digital life insurance. https://www.mckinsey.com/~/media/McKinsey/Industries/Financial%20Services/Our%20Insights/Harnessing%20the%20power%20of%20digital%20in%20life%20insurance/Harnessing-the-power-of-digital-in-life-insurance.ashx

Morgan, N. A. (2012). Marketing and business performance. *Journal of the Academy of marketing science*, *40*, 102-119.

Moss, G., Gunn, R., & Heller, J. (2006). Some men like it black, some women like it pink: consumer implications of differences in male and female website design. *Journal of Consumer behaviour*, *5*(4), 328-341.

NAIC (2023). Telematics/usage-based insurance. https://content.naic.org/cipr-topics/telematicsusage-based-insurance

Pelaez-Verdet, A., & Loscertales-Sanchez, P. (2021). Key ratios for long-term prediction of hotel financial distress and corporate default: Survival analysis for an economic stagnation. *Sustainability*, *13*(3), 1473.

Prinzie, A., & Van den Poel, D. (2007). Predicting home-appliance acquisition sequences: Markov/Markov for discrimination and survival analysis for modeling sequential information in NPTB models. *Decision support systems*, *44*(1), 28-45.

R Core Team (2022). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Ramlau-Hansen, H., Chr. Bang Jespersen, N., Kragh Andersen, P., Borch-Johnsen, K., & Deckerf, T. (1987). Life insurance for insulin-dependent diabetics. *Scandinavian Actuarial Journal*, *1987*(1-2), 19-36.

Rukhsar, L., Bangyal, W. H., Nisar, K., & Nisar, S. (2022). Prediction of insurance fraud detection using machine learning algorithms. *Mehran University Research Journal of Engineering & Technology*, *41*(1), 33-40.

Schober, P., & Vetter, T. R. (2018). Survival analysis and interpretation of time-to-event data: the tortoise and the hare. *Anesthesia and analgesia*, *127*(3), 792.

Severino, M. K., & Peng, Y. (2021). Machine learning algorithms for fraud prediction in property insurance: Empirical evidence using real-world microdata. *Machine Learning with Applications*, *5*, 100074.

Sharps, K., Hitsky, D., Hodgins, S., & Ma, C. (2015). Life insurance consumer purchase behavior Tailoring consumer engagement for today's middle market. *Deloitte Center for Financial Services*. https://www2.deloitte.com/content/dam/Deloitte/us/Documents/strategy/us-cons-life-insurance-consumer-study.pdf

Simonson, I. (1993). Get closer to your customers by understanding how they make choices. *California Management Review*, *35*(4), 68-84.

Smith, P. J. (2017). *Analysis of failure and survival data*. CRC Press.

Sullivan, L. (2016). Survival analysis. *Boston University, School of Public Health*.

Tiago, M. T. P. M. B., & Veríssimo, J. M. C. (2014). Digital marketing and social media: Why bother? *Business horizons*, *57*(6), 703-708.

Ulrich, K. T., Eppinger, S. D., & Yang, M. C. (2008). *Product design and development* (Vol. 4, pp. 1-3). Boston: McGraw-Hill higher education.

Van Rossum, G., & Drake, F. L. (2009). *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Villanueva, N. M. (2022). Clustering of nonparametric curves by the clustcurv package. *work*.

Villanueva, N. M., Sestelo, M., & Meira-Machado, L. (2019). A method for determining groups in multiple survival curves. *Statistics in Medicine*, *38*(5), 866-877.

Villanueva, N., Sestelo, M., Machado, L. M., & Roca-Pardiñas, J. (2021). clustcurv: an R package for determining groups in multiple curves.

Wasserman, S., & Faust, K. (1994). Social network analysis: Methods and applications.

Wolin, L. D. (2003). Gender issues in advertising—An oversight synthesis of research: 1970–2002. *Journal of advertising research*, *43*(1), 111-129.

Annex I. K-Prototypes code

The K-Prototypes algorithm shown in this annex corresponds to the customer segmentation considering the marketing department's point of view previously explained.

```
pip install kmodes

import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import plotly.graph_objects as go
import math

from google.cloud import bigquery
from kmodes.kprototypes import KPrototypes
from google.oauth2 import service_account
import pandas_gbq
from sklearn.preprocessing import MinMaxScaler

import random
random.seed(10)

#NOTE: special permission might be needed to access the data
credentials =
service_account.Credentials.from_service_account_file('credentials.json')
project_id = 'project_id'

# Select the data we want to download from BigQuery
sql = """
      SQL query to access the data
"""

# Store the data in a dataframe to operate with
df = pandas_gbq.read_gbq(sql, project_id=project_id,
credentials=credentials)

# Check the variables' types and correct the mistakes
df.dtypes
df['num_pathol'] = df['num_pathol'].astype(int)

# Drop missing values and delete duplicates (each lead is unique)
df = df.dropna()
df = df.reset_index()
df = df.drop(['index'],axis=1)
df = df.drop_duplicates()

# Copy the original data into a new dataframe and explore the data
datos = df.copy()
df.head()
```

```
# Delete variables that will not be included in the model
df = df.drop(df.columns[[0,1,6]], axis='columns')
df.head()

# Normalize the numerical columns
scaler = MinMaxScaler()

df_numerical = df.loc[:,['age','conversion_time','num_pathol',
'num_relatives']]
normalized_data = scaler.fit_transform(df_numerical)

df['age'] = normalized_data[:,0]
df['conversion_time'] = normalized_data[:,1]
df['num_pathol'] = normalized_data[:,2]
df['num_relatives'] = normalized_data[:,3]
df.head()

# Identify the categorical columns
catColumnsPos = [df.columns.get_loc(col) for col in
list(df.select_dtypes('object').columns)]
print('Categorical columns           :
{}'.format(list(df.select_dtypes('object').columns)))
print('Categorical columns position  : {}'.format(catColumnsPos))
# Convert dataframe to matrix
dfMatrix = df.to_numpy()
dfMatrix

# Choose optimal K using Elbow method
cost = []
for cluster in range(1, 10):
  try:
    kprototype = KPrototypes(n_clusters = cluster, init='Huang',n_jobs=-
1,n_init=10,verbose=1)
    kprototype.fit_predict(dfMatrix, categorical = catColumnsPos)
    cost.append(kprototype.cost_)
  except Exception as e:
    print(e)
    break

plt.plot(range(1,10),cost,'o-',color='#A6D854')
plt.grid(b=False)
plt.title('Elbow Method. Intra-cluster evolution')
plt.xlabel('N clusters')
plt.ylabel('Cost')
plt.show()

# Fit the cluster
kprototype = KPrototypes(n_jobs = -1, n_clusters = 7, init = 'Huang',
random_state = 0)
kprototype.fit_predict(dfMatrix, categorical = catColumnsPos)

# Check the iteration of the clusters created
```

```
print("Iteration of the clusters: "+ str(kprototype.n_iter_))
# Check the cost of the clusters created
print("Cost of the clusters created: "+ str(kprototype.cost_))

# Add labels to the copy of the original dataframe
datos['cluster'] = kprototype.labels_

# Add the cluster label to the dataframe
datos['Segment'] = datos['cluster'].map({0:'First', 1:'Second',
2:'Third',3:'Fourth', 4:'Fifth', 5:'Sixth',6:'Seventh'})
# Order the cluster
datos['Segment'] = datos['Segment'].astype('category')
datos['Segment'] =
datos['Segment'].cat.reorder_categories(['First','Second','Third','Fourth
','Fifth','Sixth','Seventh'])

datos

# Cluster interpretation
datos.rename(columns = {'cluster':'Total'}, inplace = True)
datos.groupby('Segment').agg(
    {
        'Total':'count',
        'age': 'mean',
        'beneficiary_type': lambda x: x.value_counts().index[0],
        'intention': lambda x: x.value_counts().index[0],
        'risk_job': lambda x: x.value_counts().index[0],
        'conversion_time': 'mean',
        'num_relatives': 'mean',
        'num_pathol': 'mean'

    }
).reset_index()

# Save the data into an excel file to represent the results in a
dashboard
datos = datos.drop(['Total'],axis=1)
datos.to_excel("data_mkt.xlsx",index = False, encoding="latin1")
```

I would like to point out that all K-prototype models follow the scheme shown above, the only thing that changes is the dataset and consequently, the variables used in the analysis. Since the scheme is the same in all cases, only one case is shown.

Annex II. Connecting data from BigQuery to Python

This annex overviews how to connect to a database stored in bigQuery and how to download its information using python. It needs to be mentioned that the starting point is an existing bigQuery project and a local python script from Jupyter (ipynb). If any of these requirements are not met, you can review the following guides:

- How to create a cloud project: https://cloud.google.com/resource-manager/docs/creating-managing-projects?hl=en
- Install and use Jupyter: https://docs.jupyter.org/en/latest/install.html

1. Create a service account

In the first place we need to create a service account and link it to the project which contains the data that you'd like to download. The steps to do so can be found at: https://cloud.google.com/docs/authentication/getting-started#auth-cloud-implicit-python

Note that the only section that needs to be followed is "creating a service account".

Once all the steps have been completed, you'll obtain a .json file containing all the relevant information associated with your project. We'll be using the information in future steps.

2. Working with Jupyter

2.1. DOWNLOAD DEPENDENCIES

To connect bigQuery and Python for the first time we need to download a few dependencies.

```
pip install pandas-gbq
pip install db-dtypes
```

An important aspect to take into consideration is that these commands need to be run in different code boxes, otherwise an error will pop up. Also, note that you don't need to run these commands every time you try to connect to BigQuery. It's only required once.

2.2. DOWNLOAD THE REQUIRED LIBRARIES

```
from google.cloud import bigquery
from google.oauth2 import service_account
import pandas_gbq
```

You should also consider that depending on what you're planning on doing with the data you may need some other libraries.

2.3. CREDENTIALS

There are a few parameters we need to declare before downloading the data.

```
credentials=service_account.Credentials.from_service_account_file('r
oute_to_your_json_file')
project_id = 'your_project_id'
```

This information is crucial and can be found in the .json file that we obtained in step 1.

2.4. DOWNLOAD YOUR DATA

We are going to select the data that we want to download by creating a SQL query. The only requirement is the format shown below.

```
sql = """sql_query"""
```

The next step is to download the data as a dataset.

```
data = pandas_gbq.read_gbq(sql, project_id=project_id,
credentials=credentials)
```

To find more information about pandas_gbq.read_gbq and its possibilities visit: https://pandas.pydata.org/docs/reference/api/pandas.read_gbq.html

2.5. DEVELOP YOUR PYTHON PROJECT

The setup has been completed and the dataset is now available for further processing according to your requirements.

3. Code example

```
from google.cloud import bigquery
from google.oauth2 import service_account
import pandas_gbq

credentials=service_account.Credentials.from_service_account_file('c
redentials.json')
project_id = 'project_id'

sql = """
    SELECT * FROM TABLE

"""
data = pandas_gbq.read_gbq(sql, project_id=project_id,
credentials=credentials)
```

Annex III. Connecting data from BigQuery to R

The code below shows how to connect to a database in BigQuery from a cloud project.

```
install.packages("dplyr")
install.packages("bigrquery")
install.packages("tidyverse")

library(dplyr)
library(bigrquery)
library(tidyverse)

# Select the data from BigQuery
sql <- " select * from table "

# Download the data
data_raw <- bq_project_query("project_id",sql)
df <- bq_table_download(data_raw)

# Examine the data
View(df)
str(df)
```

It is important to highlight that the user trying to access the data must have permission to do so. In addition to this, R will likely ask for extra permission to connect to your google account linked to the project to download the desired data.