NOVA IMS
Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Policy Renewal Optimization Project in Health Insurance

Mohamed Ali Bayoudh

Internship Report presented as the partial requirement for obtaining a Master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

NOVA Information Management School

Instituto Superior de Estatística e Gestão de Informação

Universidade Nova de Lisboa

RENEWAL OPTIMIZATION PROJECT IN HEALTH INSURANCE

by

Mohamed Ali Bayoudh

Internship Report presented as the partial requirement for obtaining a Master's degree in Data Science and Advanced Analytics

Advisor: Flávio Luís Portas Pinheiro

August 2020

DEDICATION

I would love to dedicate this report to my family.

I remember how my parents kept me out of the street at an early age and have always encourage me to learn and have a positive impact in the world.

There is no way I would be here right now without having their support and love.

They have always been and will always be my support system.

Thank you for everything you have done for me.

I could never be grateful enough.

ACKNOWLEDGEMENTS

PRIMARILY, I WOULD LIKE TO THANK ANDRÉ RUFINO, PAULA SANTOS AND DR. FLÁVIO PINHEIRO FOR OFFERING ME THE OPPORTUNITY TO PERFORM AN INTERNSHIP IN AGEAS PORTUGAL GROUP AND SUPERVISING ME DURING THIS PROCESS.

THEN I WOULD LIKE TO THANK ALL THE PRICING AND ADVANCED ANALYTICS TEAM AND ESPECIALLY JORGE ABREU, JÉSSICA ZAQUEU, SUSANA FELIX AND BRUNO VIEIRA FOR ALL THEIR SUPPORT AND SHARING KNOWLEDGE WITH ME.

I ENJOYED EVERY MOMENT WITH YOU AND FEEL REALLY GRATEFUL AND PROUD TO BE PART OF THIS AMAZING TEAM.

MOHAMED ALI BAYOUDH.

ABSTRACT

The Renewal Optimization has been developed over the course of 9 months (September 2018 till June 2019) as part of the thesis for the MSc in Advanced Analytics and Data Science at the Nova Information Management School (Universidade Nova de Lisboa).

The objective was to Predict the Probability of Renewing the policy of our customer. This would allow for more assertive and targeted marketing actions and decision making as well as fine tune the pricing strategy.

The Training Sample is composed of data from the 1st of January 2017 till 30 June 2018 and the results presented reflect a picture of the Médis individual client portfolio from July 1st, 2018 till 31 December 2018 with 68 732 policies tested.

The attributes used in the modelling process cover 6 customer dimensions: demographic, customer profile, product profile, bank variables and usage as well as interaction with the company.

The final model results calculated the Renewal Probabilities of every active policy. These calculations have been divided in deciles where the first group have the lowest Renewal Probability estimated and the last one has the highest Renewal Probability estimated.

To determine the factor that affects the Renewal Rate the most, a comparison has been conducted between the first and the last group (low probability and high probability groups).

Next steps for the project include, but are not limited to, making the results available to all stakeholders and the monitoring plan is also discussed.

KEYWORDS

Renewal, Propensity, Churn, Lift, ROC, Data, Software, SAS, EMBLEM, Tower Watson, Health Insurance

INDEX

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION

1.1. AGEAS PRESENTATION

This report has been conducted after 9-month internship with the International insurance company Group Ageas Portugal more precisely Médis.

Insurance, one of the oldest financial products in history, has seen tremendous progress since its invention in the 18th century. The growth path has accelerated exponentially in the new era of digitalization, where the Internet is disrupting business models and competition is rising leaving businesses under a continuous innovation challenge. In fact, in its July 2019 edition, The Economist published an article with the following title: "the future of insurance is happening without insurance firms."

In this report, we challenge this thought by addressing the case of AGEAS Group, a multinational insurance company with the vision of "becoming the reference partner in insurance, a relevant player in services and the best place to work for entrepreneurs."

Ageas group serves 47 million customers in 14 countries in Europe and Asia thanks to a diversified portfolio. In Portugal the width of the product mix consists of 8 lines which are Home insurance, car insurance, accidents at work insurance, personal accidents insurance, health insurance, life and financial insurance, liability insurance, and boat insurance. The depth and length of each product line vary from one to another with a total of 24 insurance products strengthened by several services for an optimal customer lifetime value. These products are branded under 5 main brand names as follows Ageas Seguros, Medis, Seguro Direto, Ageas Pensões, and Ocidental. Each brand targets a different customer segment and offers tailored products that answer their wants and needs. Our report focuses on the brand Medis, which offers individual health insurance, where customers can choose their own proposals for each stage of their lives. Medis uses an integrated system that keeps track of customer consultation history and various services such as 24-hour nurse advice via the Medis line or application or making available a doctor chosen by the client.

However, the rapid pace of the competitive dynamics of the insurance sector leaves customers with many options and alternatives. The competitive advantage of companies and their value propositions can be easily imitated and even further developed by competitors. It is therefore important not only to allocate resources for innovation to acquire new customers, but also to retain the current ones. Thus, our report will focus on optimizing renewal in the digital age,

using data science, the appropriate modeling methods, and the adequate analytical tools to spot behavioral patterns and suggest a way forward to reflect the findings in the business strategy of Ageas Group Portugal.

1.2. Importance of Renewal

Renewal have always been one of the biggest sources of revenue for any company. They have always been a focus point to increase their profitability and try to do business in a more efficient way. Another way to increase their profitability is by customer acquisition. However, acquiring a new customer is extremely expensive comparing to renewing. In fact, it is said that the cost to acquire a new customer is five to twenty times higher than renewing an existing one because of all the cost of marketing and operations. Moreover, customer who renew are more likely to upsell their options and spend more and this could be related to their Customer Lifetime Value and their relationship with the company. In fact, in Health Insurance we always try to build this relationship of trust it is fundamental for us where Médis will be present in those hard moments and will show why their customer chooses it and this relationship can be seen by the fact that as long as customer stay with the company they become less likely to leave. Another important factor about Renewal is that it could be also an important KPI to understand customer satisfaction as well as the ratio of quality and price. A high Renewal rate generally means that the customer is satisfied with the service provided and would even commit to make it better and give feedbacks however a low Renewal rate means that the customer may have found better options in others company or that he is not satisfied with the service provided.

1.3. IMPORTANCE OF ANALYTICS

In Health insurance Price is one of the main factors to customer renewal it is for sure the most critical factor that make you decides if you should renew or not.

Price is increasing every year because of the increase of risk every year the population is growing up in Portugal life expectancy is increasing as well and according to the world Bank it was 76.31 in 2000 comparing to 81.12 in 2015. Moreover, according to Cancer UK Research half of people born post 1960 will get cancer which also represent an extreme risk to the company. Every year there is a general Inflation in the economy and so the company need to increase their price in order to be updated to the situation we are living in. Finally, all these new technologies, improved medical treatment, new drugs and operations that are introduced every year to show the improvement that is happening in the Health sector are costly and

participate to increase the cost. This is where Analytics enter in action by taking in consideration all these factors that increases the cost then will play an important role to identify the right price to the right people so to define the problem we asked ourselves this business question.

Figure 1.1 Business Question

1.4. The Business Case

We will start by having a look on the renewal rate during the last previous years that have been very stable but slightly decreasing both Renewal rate for Clients which is the number of clients that renewed and Renewal rate for Premium which is the amount of payment that the clients who renewed payed.

Policy renewal rates and Premium renewal rate in 2018 were both extremely high the actual rates could not be shared in this report due to the sensitive nature of this company data.

Médis have a very high Renewal Rate but still need to be improved in order to optimize both renewal rate for clients and premiums by having a data driven approach through modeling to allow the price optimization as well as understand the propensity of renewing a policy. Other things must be taken in consideration such as by doing any price variation Médis will have a strong effect on the market as well as different Médis products will be launched soon that perhaps will give us more flexibility and a better value proposition to increase the renewal rate. After having a deep look on the market and meeting with different stakeholders we defined some objectives for this 9 month Internship.

Start by Analyzing the current client's portfolio and understand the propensity for not renewing. Create a modeling methodology to predict the probability of contract renewal then develop a self-service tool to enable the visualization and use of model so we can gain meaningful customer knowledge to enable refined decision-making concerning pricing.

We have also identified the tools to be used such as SAS Enterprise Guide for data exploration and data preparation, SAS Enterprise Miner for modeling, Towers Watson Emblem for modeling and Microsoft Power BI for visualization.

1.5. DETERMINING THE SIZE OF THE OPPORTUNITY

In other segments of Ageas Group Portugal, the Renewal models coupled with attractive marketing and retention campaigns have provided a huge positive impact and an increase in premium. In Ageas Móbis they implemented the project in 2017 and had a positive impact summarized by a huge increase in premiums acquired.

1.6. STRUCTURE OF THE REPORT

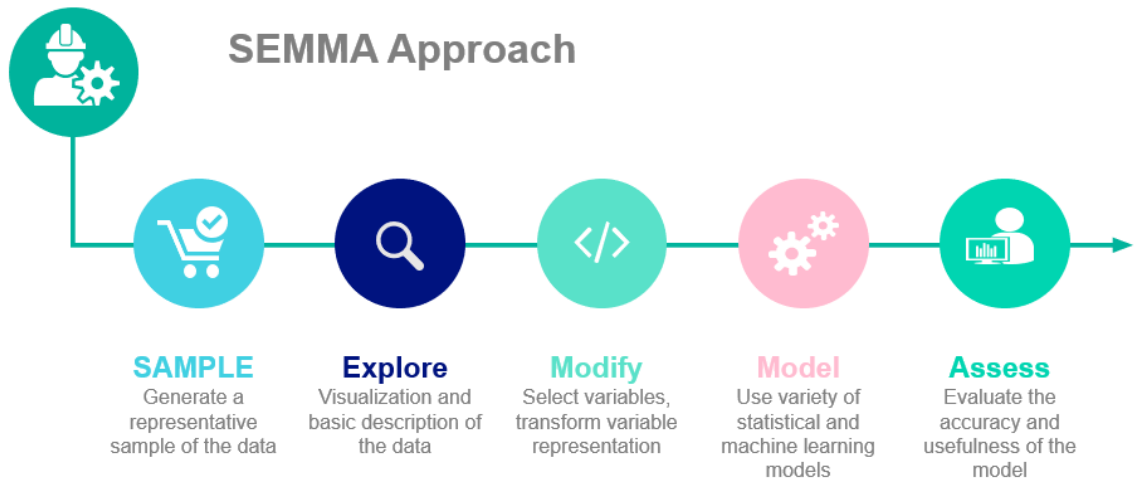
After the presentation, comes the Literature Review in Chapter 2 where we will go through a summary of different models used as well as the method used to assess them. Then in Chapter 3 (Methodology) present the plan of the internship and the steps to follow before variable selection. Chapter 4 (Results of the models) showcase the results of the model tested. Chapter 5 (Conclusions) contains the conclusions and all the deliverables of this project to the stakeholder. Finally, Chapter 6 (Recommendations for future work) includes future recommendations, challenges, and key learning.

2. LITERATURE REVIEW

In the literature review we proceed with the introduction of the analytical approach used providing a good introduction to SEMMA, the modeling techniques implemented as well as defining the right way to assess them.

2.1. SEMMA

The SEMMA Approach is a famous method used by Data Scientist in SAS Enterprise Miner which is the acronyms of the data mining steps of processing the data starts with **S**ampling, **E**xploring, **M**odifying, **M**odeling, and **A**ssessing.



t

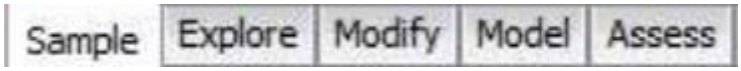


Figure 2.1 SEMMA from SAS Enterprise Miner

The SEMMA Method is perfect to apply in the analytical part of this internship.
According to SAS Enterprise Miner, this data Mining approach will ensure a good preprocessing, Modeling, and the assessment of the model. It is well knowns and used in many industries for diverse problems such as Renewal, Churn, Fraud Detection, Upsell Detection or target a specific proportion of clients.

According to SAS SEMMA, the process flow consists of 5 steps.
In the first phase it consists on creating your Sample universe which are the clients that are the most interesting and fit to our business question with a large number of variables. Then comes the Explore phase which is based on exploring this dataset and understanding how the correlation between the variables and the target is. Followed by Modify phase where some new variables are created and then select the right one for the model. Then, comes the Model phase which consists in applying different algorithm to predict a desired outcome. Finally, Assess the results and see if the model is useful and reliable.

2.2. MODELING

For business purposes the Model must be a regression model in the software Emblem by Towers Watson. It was the perfect way for the company to synchronize the results of the model with Towers Watson Radar and perform simulations and other deep analysis to see the effect of the Price Change.

According to Towers Watson Emblem the software can run Logistic Regression Model with high performance. It can handle a large and complex dataset with a high number of parameters and most importantly it is customized to suits Insurance business requirements. In Ageas Portugal they have been using Emblem Towers Watson for years.

The Logistic Regression is the model that have been implemented in this project it models the probabilities for classification problem with two possible outcomes.

According to Chirstoph, Molnar (2020), It is an extension of a linear regression that can work well for regression but fails in classification. In fact, the Linear model does not score probabilities it simply gives you an output of 0 or 1 and cannot be interpreted as a probability.

However, by using Logistic regression it will use a logistic function to squeeze the output of a linear equation between 0 and 1 and it is defined by:

$$\text{logistic}(\eta) = 1/1 + \exp(-\eta) \qquad (1)$$

Moreover, Logistic Regression provides estimated probabilities in the output which add more significance to the classification for instance it's a huge advantage to know that an observation has 99% renewal probability than 51% renewal probability.

For the Model in Emblem two different ways to balance the data have been implemented (to make the target close to a 50% distribution. First by deleting random observation when the target is equal 1 in order to decrease their proportion. The second method used is called SMOTE which consist on creating virtual observation where the target is equal to 0 to increase their proportion. These observations have the same characteristics with the existing observation that did not renew. However, other Models have been tested for validation and have been used in SAS Enterprise Miner.

According to Chirstoph, Molnar (2020), Decision trees is a Model for classification, it splits the data multiple times creating with more accurate probabilities at every split and different subset of the data are created where every observation belong to these subsets. Decision Trees is a good algorithm that can be efficient for classification and regression. Regression can be effective when the relationship between features and outcome are not linear or where features interact with other. The advantage of the Decision Trees is that it takes less pre-processing to compute a working model, but the disadvantage is the inefficient performance dealing with bigger datasets. The complexity of the calculation of the Decision Trees causes a bad performance either and could result in an overfitting model. Other, easy understandable, Machine Learning algorithms could create useful insights within half of the time frame.

According to Breiman, Leo, (2001), Random Forest is a classifier that evolves from Decision Trees. It consists of many Decision Trees. Each Decision Tree is based on a random selected subset of the complete dataset. To classify a new instance, each decision tree provides a classification for input data; a Random Forest collects the classifications and chooses the most voted prediction as the result. The input of each tree is sampled data from the original dataset. Then, a random selection of a subset of features is done coming from the optional features in order to grow the tree at each node. Each tree is grown without pruning. Essentially, random forest enables many weak or weakly-correlated classifiers to form a strong classifier. Giving the fact that Random Forest uses different Decision Trees, it reduces the variance of each tree. The accuracy per subtree would improve compared with the normal Decision Tree method. Nonetheless, the complex structure and format would make it hard to understand, implement and visualize.

2.3. MODEL ASSESSMENT

After Modelling, one of the most important steps in any analytical project is the assessment of the model and see how it performs. Many methods will be discussed and follow the business needs in order to choose the right model to the problem.

In terms of technical objectives, the goal is to build a model which balances accuracy with number of selected variables, so as to achieve an understanding of their role within the risk of client churn. In every Predictive model there is different kind of prediction that are resumed in the confusion matrix below:

		ACTUAL VALUES	
		Positive	Negative
PREDICTED VALUES	Positive	TP	FP
	Negative	FN	TN

Table 2.3 Confusion Matrix

TP: true Positive, FP: false Positive, FN: false Negative, TN: True Negative

Predictive models aim at maximizing accuracy which are the proportion of observation that are correctly classified within the entire test sample, so the objective is to have as much TP and TN possible and can be calculated as follow:

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN (i.e. entire test sample)}} \quad (2)$$

Other test criteria include Precision, Recall, as well as ROC AUC and Lift curve.

Precision, also known as positive predicted value (PPV), provides a measure for the proportion of correctly classified positive observations, from all observations classified as being positive. In formula:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP (i.e. all classified observations as 1)}} \quad (3)$$

Recall, commonly referred to as true positive rate or sensitivity measures the proportion of correctly identified positive values from all positive values. This can also be thought of as the ability a model has of avoiding false negatives.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN (i.e. all observations which are actually 1)}} \tag{4}$$

ROC Index Measures the model's True Positive and False Positive rates; Higher ROC Index means increased True Positive vs. False Positive outcomes.
Gini Coefficient Indicates the ability (efficiency) of the model to distinguish between those who renewed and those who did not.

Lift score is defined as the ratio between predicted Renewal rate and average Renewal rate.
The underlying idea of lift analysis is to Group observations, typically deciles, based on the predicted Renewal probability the Calculate the true renewal rate per decile as shown in the equation below:

$$\begin{aligned}\text{Lift} &= \frac{\frac{\text{Number of non Renewing in the decile}}{\text{Number of obs. in the decile}}}{\frac{\text{Number of non Renewing in the population}}{\text{Number of obs. in the population}}} \\ &= \frac{\text{non Renewing Rate in Decile}}{\text{non Renewing Rate in Population}}\end{aligned} \tag{5}$$

Lift has been one of the most important metrics not only to assess but also to interpret the model since it has been used to understand customers propensity to renew.
Kolmogorov-Smirnov Statistic measures the maximum distance between the cumulative distributions of those who Renew and those who did not Renew.

$$\text{D} = \max_{1 \le i \le N}\left(\text{F(Yi)} - \frac{\text{i} - 1}{\text{N}}, \frac{\text{i}}{\text{N}} - \text{F(Yi)}\right) \tag{6}$$

Cumulative Percent Captured Response: Cumulative % of non-renewing policies that are captured in the deciles. Non renewing policies are expected to receive higher probability in some specific deciles.

3. METHODOLOGY

In this chapter a detailed explanation about all the process will be provided. First an introduction of the timeline with and explanation of each specific phase. After we proceed by defining the target variable and building the universe table. Then comes the variable collection and preprocessing phase. Finally move to dimension reduction phase using different techniques to select the right variables to the model.

3.1. PROJECT TIMELINE AND STAKEHOLDERS

In total, the project duration was planned to last 40 weeks, and took 55 weeks due to delays, deadline over-estimations and future implementations. Different Phases have been identified as showed in the figure below:

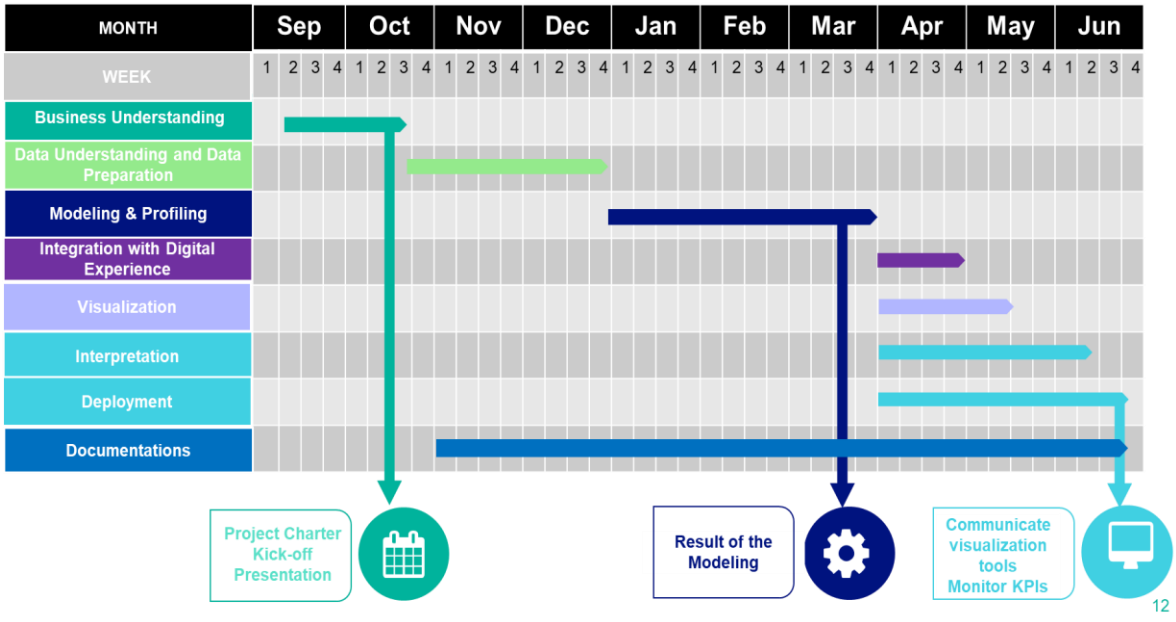


Table 3.1 Internship Timeline per phase per week

In The first phase of Business Understanding. It was primordial to understand the environment you are in. It starts with an overview of the insurance business and its terms as well as understanding the structure of the company, determine your key stakeholders and have personal

meeting in order to know how to answer the business question and build the business case as carried out in the previous section.

In the Data Understanding and Data Preparation Section we started by defining the target variable and build an analytical table universe then build a data dictionary and identify all possible variables that could be related to Renewal. Then, start preprocessing the data, verify the data quality, perform some transformations, and finally select the variable to model.

After Comes the Modeling phase where we tried to build the most accurate model and do the assessment make sure to identify the people that are not Renewing in order to contact them.

In the Next phase of Integration, Visualization and Interpretation we tried to integrate the results of the model with other project such as CLV customer lifetime Value and Digital Experience to see who are the characteristic of people that Renew the most. Moreover, build a Power BI dashboard with all the Interpretation of the model.

Then, we also had the deployment phase where we tried to automatize the model and create a variable called Renewal Probability for every policy in our dataset. This variable is the probability calculated by the model.

Finally the results were communicated to the stakeholders so as to perform different Simulations. Then, a set of actions were defined according to the insights gained and incorporated into a candidate pilot campaign.

Given the nature of the project it had different Stakeholders as shown in the table below:

Project Stakeholders & Position	
Project Sponsor / Head of Médis Direção Técnica	*André Rufino*
Project Sponsor / Head of Pricing Advanced Analytics	*Paula Santos*
Project owner / Data Science Intern	*Mohamed Ali Bayoudh*
Key Stakeholder / Head of Pricing Advanced Analytics	*Paula Santos*
Key Stakeholder / Actuary	*Miguel Nunes*
Key Stakeholder / Head of Data Mining	*Magdalena Neate*
Project Support / Data Science	*Jorge Abreu, Jessica Zaqueu*
Project Support / Data Engineer	*Susana Félix*

Table 3.2 project Stakeholders

3.2. DATA PROCESSING AND VARIABLE SELECTION

3.2.1. Definition of Target Variable

Determining the target variable is a crucial project phase, as it is where the relevant universe is defined, as well as the specific time-reference periods, and most importantly where the target variable is explicitly defined. All possible situations regarding policy renewal are covered to understand which observation we need to train, and which ones not to. This step is a direct and strong interpretation of the business question, and really requires an understanding of what it is the model seeks to answer, as the definition of the target and the modelling universe will define the model outcome.

For this model, renewal target is defined on the policy level. This means, only policy holders who actively renew their policy are taken into consideration. Additionally, there are specific cases of cancelled Médis policies, that are not considered deliberately, with reasons for each case explained below. The goal is to detect – for each observation (policy holder) – behavioral patterns and variable relationships within a specific set of attributes, and specific past time-frame (look back period), with sufficient predictive power to determine the propensity of renewing within a given future time-frame.

3.2.2. Rules & Exclusions

Moreover, a series of rules to determine which policies are to be discarded from the renewal are outlined according to business needs and context:

- Médis employees / Companies /Start-up are removed from data universe, we will only focus on the individual segment
- Policies which have been subjected to re-issuing (including cannibalization) days are also removed from data universe
- Policies which have been issued less than a year ago are not eligible for renewal since they did not close their first year
- Expired policies with the following cancelation reasons are excluded:
    - No longer company worker
    - Subscription error
    - Age limit
    - Death
    - Substitution

- Policy holders who face expiry in the next 12 months due to age limit are excluded
- Policies with option “Vintage” and “Vintage Plus” are also excluded because of the stable increase of premium that these policies have

3.2.3. The Renewal Rate Calculation

Basically, the Customer Renewal Rate, in its simplest form, is calculated as number of customers that renew in a given year +50 days, or over the potential customers 45 days before their renewal date. It is a good indicator of customer satisfaction with the customer service and the efficacy of account management function.

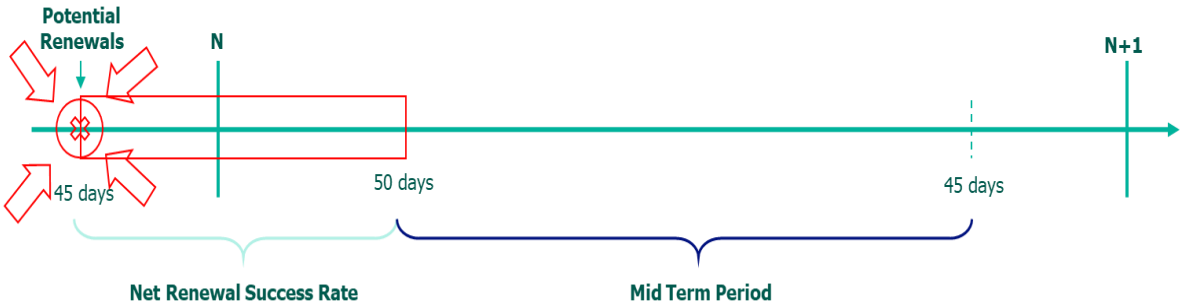


Figure 3.2.1 Renewal Timeline

Potential Renewal:
Policies in force when the renewal is processed in the system and the renewal letter is sent to the Client.

Net Renewal Success Rate (NRSR):

•The first instalment is paid
•There is no cancellation date within the 50 days window.

Net Renewal Success Rate measured in Policies, Annuity N

$$\text{NRSR (Policies)}_{\text{N}} = \frac{\text{R}_{\text{N}}}{\text{PR}_{\text{N}-1|\text{N}}} \qquad (7)$$

$R_N = Renewed\ Policies\ to\ annuity\ N$

$PR_{N-1|N} = Potential\ Renewals\ from\ annuity\ N-1\ to\ annuity\ N$

3.2.4. Data Gathering

Given the current business needs, and operational scenario, the target variable has been built to predict renewal policies in 45 days before the renewing date till 50 days after.
The data collected was from 1 January 2017 till 30June 2018 with 225k observations

- 60% for training
- 20% for validation
- 20% for testing

We also collected data from 1st of July till end of December for back testing.
The structure of the data set starts with a universe table that includes observations of both active and inactive policies at a reference time *t*. The target variable – of binary nature – identifies whether the policy has renewed (according to its policy status, given it is fulfils criteria to be eligible). It is at state 0 if the policy did not renew, or 1 if the policy renewed.
The Universe is composed of policy id, Client id, reference date and the target (policy renewed or not). This table will be the link of any other table to get information about the policies so at the end the data set will have this structure:

ID	Client id	Input var2	…	Input varn	Ref date	Target
1	x_{11}	x_{12}		x_{1n}	*t1*	0
2	x_{21}	x_{22}		X_{2n}	*t2*	1
3	x_{31}	x_{32}		X_{3n}	*t3*	0
…	..	…	…	…	*t4*	…
m	x_{m1}	x_{m2}		x_{mn}	*tm*	

Table 3.2.4 universe table structure

3.2.5. Data Types

To have a better look and a deeper understanding about renewal a wide variety of characteristics and behavior have been collected form Médis data warehouse.
Data gathered within an 18-month period, and that contains information on the policy holder or insured person level, according to the attribute.
Such behaviors can be identified directly or indirectly from information extracted from several data families:

- Policy profile – Premiums, coverages, product
- Non-health policy profile
- Demographic
- Complaints and campaigns
- Claims and Pre-authorization requests
- Interactions with Médis
- Public data

3.3. VARIABLE SELECTION

In order to have a successful predictive model, variable selection is the most critical step and the most important to your analytical project where the objective is to choose wisely the best possible variable for the model. It must go through different steps where every time you decrease the number of variables for analytical reasons as well as business reasons.

3.3.1. Removing outliers

Some observations had an extreme value that are extremely few in the data set and does not help the model to find any correlation to the target. In this phase we proceeded through the nodes filter in SAS Enterprise Miner looking to delete the observation with extreme values. It is important to keep the observation that make sense to the business no matter how high they are such as paying a very high premium are having a huge claim. Moreover, it is preferable that you do not delete more than 5% of your observation to not have a huge effect on your dataset. In this phase 0.7% of records have been removed due to extreme values.

3.3.2. Variables transformation

Some transformations have been performed in order to give a better understanding of the situation. We created several new variables by using some analytical equation and mathematical functions to combine the information provided by different variables. In this phase we created 25 new variables such as the percentage change in premium between 2 years which is one of the most important variables using this formula below:

$$\text{Percentage Change} = \frac{\text{Premium in Year X} - \text{Premium in year X} - 1}{\text{Premium in year X} - 1} \quad (8)$$

3.3.3. Correlation and Descriptive Analysis

Correlation analysis methods are used to identify or study the relationship between numerical or continuous variables in the data set. This statistical method can help to identify features having a strong relationship with the target variable and having a deeper understanding about your dataset.

Descriptive analysis has been performed in Power BI and has been shared with some of the stakeholders in order to provide more insights. This interactive visualization provides important features related to the target value.

For sensitive data, the Power BI results will not be shared however we will provide an example of how a variable can be correlated to the target. Fictive numbers are shared in the graphs because of the importance of the data to the company.

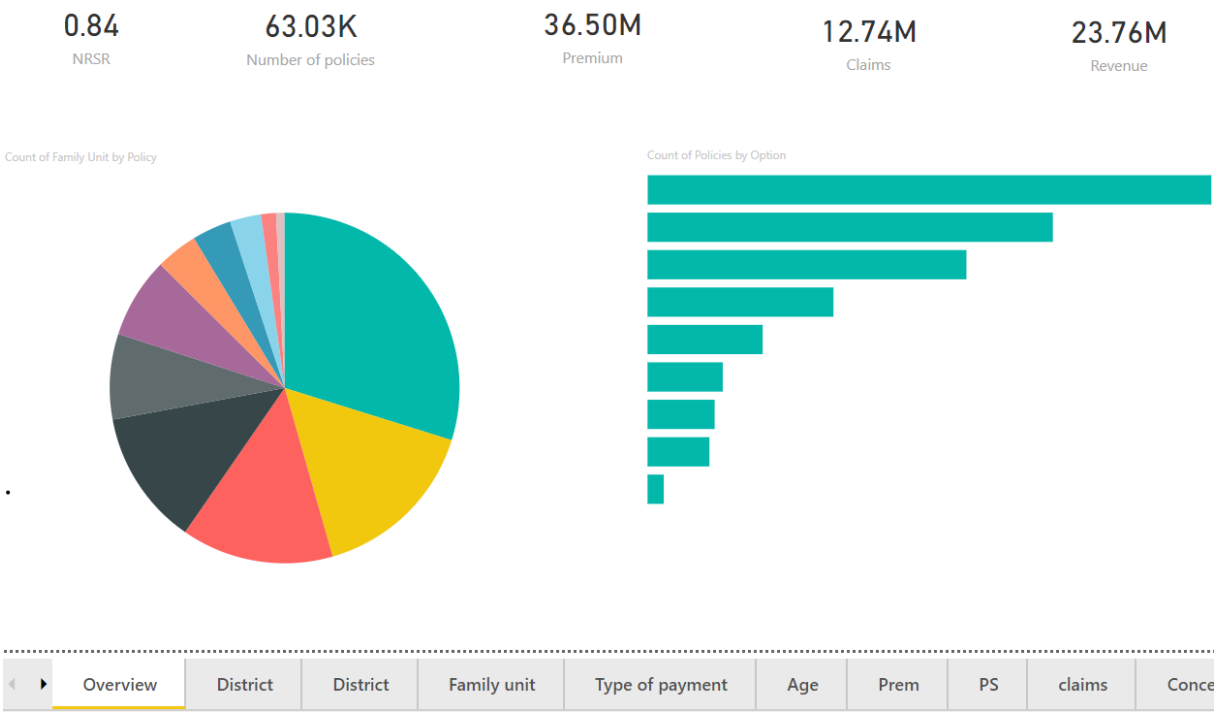


Figure 3.2.2 PowerBI Dashbord. Data Analysis Interface of all variable in the year 2017 data collected from SAS policy tables

In the Power BI dashboard every page was focus on a specific variable and understand if it has a positive or negative impact on the renewal rate as well as showing what happen to the claims and revenue every time a filter have been added.

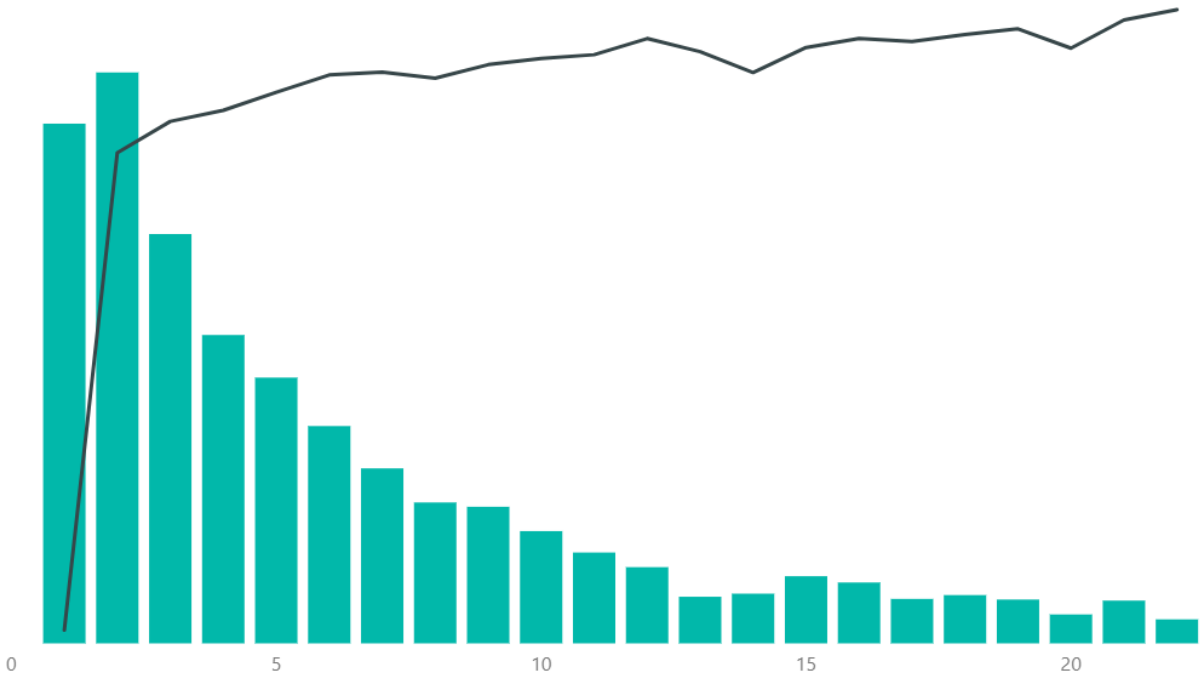


Figure 3.3 Line and stacked column chart. Power bi Var 17 correlation with the target. Green box is the number of policies and the black line is the Renewal Rate in 2017 data collected From SAS Enterprise guide Policy table and displayed in Power BI

As shown in the Line and stacked column Figure 3.3 there is correlation between Var17 and the target where the target is increasing when the Var 17 is increasing as well. Those are the relationship that we were looking at in order to have a better understanding of the dataset.

3.3.4. Missing Values Analysis

Having run the complete analytical base table (ABT) process, the full dataset is composed of 800 variables strong, is submitted to a quick distribution and missing value analysis. All variables with all observations at equal value (range = 0), or with a proportion of missing values of over 60% were removed from the dataset. This process cut the number of attributes from ~800 to ~400.

3.3.5. Chi-Square Tests

A Chi square test is applied to determine the association between the independent variables and the target. The fully functional dataset, rid of missing values and uniform distributions, was then subjected to a Chi-Square test, removing all variables with a 95% probability less than the critical level range (alfa = 5% significance). A Pearson Chi-square test was carried out on binary and categorical variables, considering individual critical values, according to each variable's

number of degrees of freedom (DoF). A Wald Chi-square test was carried out for interval variables. The number of attributes was reduced from ~400 to 70.

3.3.6. PCA

The Principal Component Analysis (PCA) is used for dimensionality reduction: First, to satisfy core necessity of reducing the number of variables used in modelling process, such as replacing a collection of variables within a "family", with one factor which contains information relevant to the attributes it is substituting. Secondly, to capture underlying relationships between attributes which are not necessarily apparent on first analysis. PCAs are mostly used for creating a smaller number of variables or features without ruining the valuable information. This is done by combining correlated variables. The created variables are standardized to a mean of 0 and a std deviation of 1. The PCA method runs on a correlation matrix, given the nature and distribution ranges of the variables being fed into the dimensionality reduction process. The Varimax Orthogonal Rotation method, shown in formula below, is applied in order to initiate a high variation of each axis.

$$R_{VARIMAX} = argmax_R \left(\frac{1}{p} \sum_{j=1}^{k} \sum_{i=1}^{p} (\Lambda R)_{ij}^4 - \sum_{j=1}^{k} \left(\frac{1}{p} \sum_{i=1}^{p} (\Lambda R)_{ij}^2 \right)^2 \right) \quad (9)$$

The formula maximizes the sum of the variances of the squared loadings. The purpose of the PCA is to assign each original variable to exactly one PCA feature. After creating the several different components, the amount of inputs for the model is reduced which results into a faster and more efficient performance.

After Performing PCA we decreased the dataset from 70 variables to 22 factors that were the inputs to the model.

3.3.7. Final Dataset

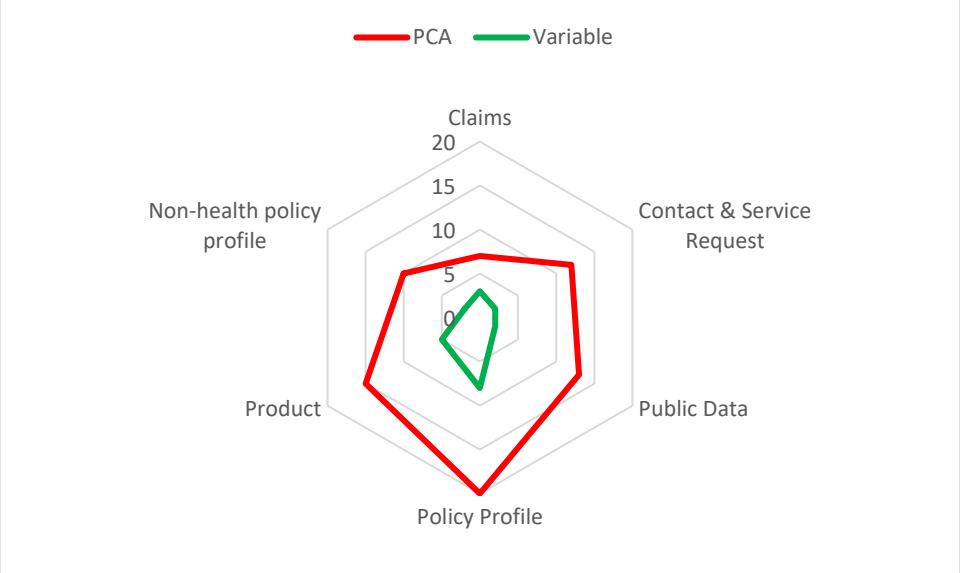


Figure 3.3.7 Factor selection in green and variables in red after running PCA data from the year 2017 collected from SAS Enterprise Miner

As shown in the graph 3.3.7 after Performing PCA the dataset dimension has been reduced from 70 variables to 22 factors.

Here is the family list of the Factors (combine information of similar variables) that have been Selected:

- Claims: Frequency, recency, and cost of claims
- Contact & Service request: Frequency and recency of telephone contact with Médis
- Public Data: Data related to your Concelho
- Policy profile: Includes demographic, tenure, payment frequency, as well as premiums payed
- Product: Indicators of product type and coverages
- Non-health Policy Profile: All non-health profile data

4. RESULTS OF THE MODELS

The result for the Regression Renewal Model performed on Emblem without balancing the data were satisfying and met expectation. For business reasons related to synchronizing all the results with the tariffication software the Emblem Model have been chosen as the wining model. The Regression model was forward with a logit link function and an unlimited number of factors as parameters.

Model Assessments:

	Médis Renewal Model		Médis BENCHMARK			
	Model	**Backtesting**	**Poor**	**Adequate**	**Good**	**Excelent**
Gini Coefficient	0.56	0.54	< 0.20	0.2 to 0.35	0.36 to 0.70	> 0.70
Roc Index	0.78	0.77	< 0.60	0.60 to 0.67	0.68 to 0.85	> 0.85
Cumulative Percent Captured Response	39.4	38.1	<20.5	20.6 to 26.0	26.1 to 33.5	> 33.5

Table 4.1 Model Logistic Regression in Emblem using data from 2017 Assessment

According to the table 4.1 Emblem model was satisfying according to Médis Benchmarks that have been created by Data Scientist in the company. These Benchmarks take into consideration the business as well as the analytical approach. The Emblem model did fit in the good benchmark in Gini Coefficient and Roc Index but in Cumulative Percent Captured Response had an excellent assessment.

As mentioned, before other model has been implemented on SAS Enterprise Miner to validate the results. The results were almost the same slightly better in the Random Forest Model.

	Regression		Random Forest		Decision Tree	
	Model	**Backtesting**	**Model**	**Backtesting**	**Model**	**Backtesting**
Gini Coefficient	0.54	0.52	0.6	0.58	0.58	0.54
Roc Index	0.77	0.76	0.8	0.79	0.79	0.77
Cumulative Percent Captured Response	37.7	36.9	41.2	40.6	39.7	38.3

Table 4.2 Model Assessment of a Logistic Regression, Random Forest ands Decision Tree in SAS Enterprise Miner

As shown in the table 4.2 the Random Forest model was having the best assessment it was in fact the most accurate model. The model has a good assessment in Gini Coefficient and Roc Index but in Cumulative Percent Captured Response had an excellent assessment. However, for business reasons and synchronizing the results with other software as mention before the Emblem model have been chosen.

The Last model that have been implemented was a logistic Regression in SAS Enterprise Miner with a balanced distribution of the target variable due to the high renewal rate the target have. However, the results were not better than the previous model.

	Ramdomly Balanced		SMOTE Balanced	
	Model	**Backtesting**	**Model**	**Backtesting**
Gini Coefficient	0.5	0.46	0.54	0.52
Roc Index	0.75	0.73	0.77	0.76
Cumulative Percent Captured Response	35.8	34.4	39.2	38.5

Table 4.3 Model Assessment of a Logistic Regression balanced data using a random balance and the SMOTE technique

According to the table 4.3 the SMOTE Balanced data was performing better than the random balanced data where it was better in Gini Coefficient, Roc Index and Cumulative Percent Captured Response.

5. CONCLUSIONS

In the Conclusion part an explanation about the use of the model is provided where it helped to understand customer behavior toward renewing, have a specific segmentation of clients and perform a simulation to see the effect of price change in renewal. In this part not all details have been provided due to sensitive information for Ageas Portugal such as specific variables effect on the model so to carry on all the variables that will be shared in this part will be codified.

5.1. SEGMENTATIONS

The estimated probabilities could be also a very good tool to do a segmentation of the Policy that may cancel and try to generate leads which consist on retaining the customer that could cancel due to a specific marketing plan made by the company. After Scoring the test sample we sorted the probabilities estimated and then divided into deciles.

Due to sensitive information the probability of renewing in the deciles have been modified in order to keep the confidentiality of the results however an explanation will be provided by taking in consideration that the renewal rate in the whole universe is 85%. All deciles Renewal Rates have been also changed by fictive numbers.

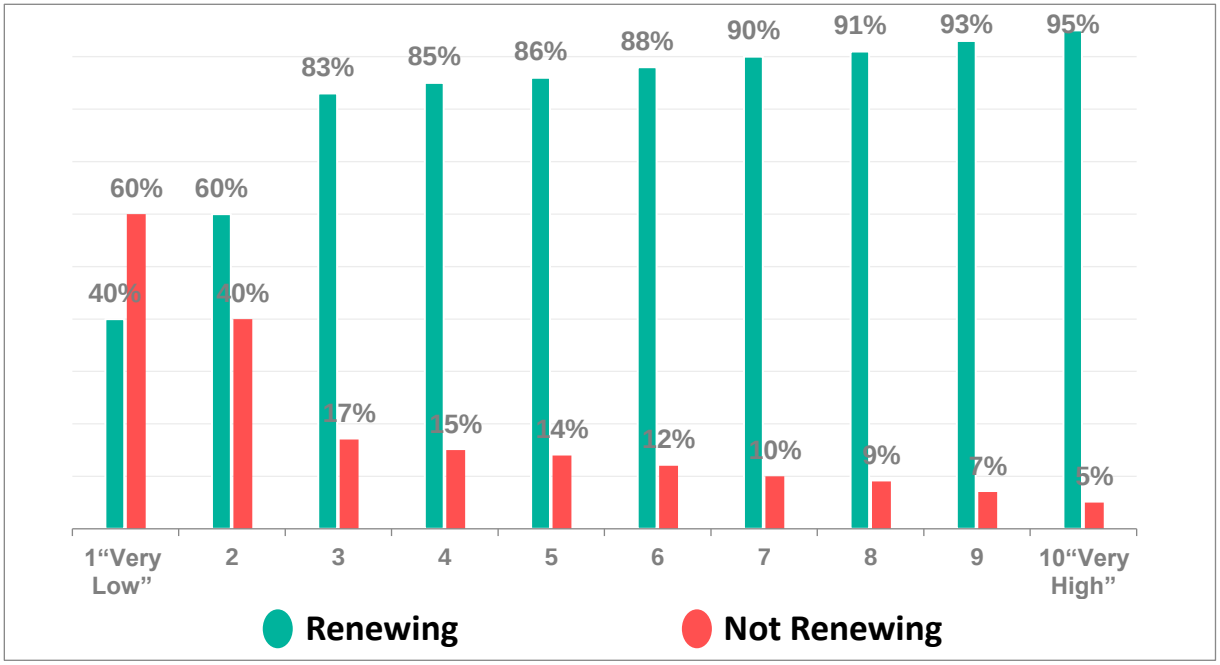


Figure 5.1 Column chart. Segmentation of the client Green box are renewing customer red box are non-renewing customer per decile data from the year 2018 collected from Emblem

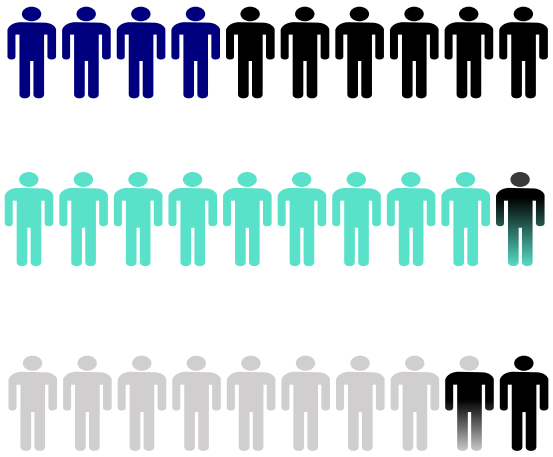
As showed in the graph above the first group recorded the lowest Renewal rate with 40% which is 45p.p. lower than the average and the last group recorded the highest Renewal rate with 95% which is 10p.p. higher than the average.

In fact, if we focus on the Lift, we can determine 4 times better if the policy will leave or not. Then depending to the business decision, we have different sample with different probability to renew that will help you identify the sample of clients you need to approach.

The Segmentation could help identify low renewing probability customer and make a specific marketing plan in order to retain them.

5.2. PROFILING

By comparing the different probabilities estimated we can understand the characteristics of the policy better and what are the variable who are more propense to renew. As mentioned in the previous part, we divided the probability estimated of our testing sample in 10 groups where the first group have the lowest probability estimated and the last group have the highest probability estimated. All deciles Renewal Rates and Universe Renewal Rate are the same fictive numbers introduced in the Segmentation part.



- First GROUP is the lowest group with 40% Renewal
- Last GROUP is the highest group with 95% Renewal
- Average Renewal in the Universe is 85%

Figure 5.2 Deciles Probability of the clients

To have a deeper look on the characteristics propensity to renew we made a comparison between the two extremes the Lowest Renewal group and the Highest Renewal Group

However, for business reasons and sensitive information the variables have been codified.

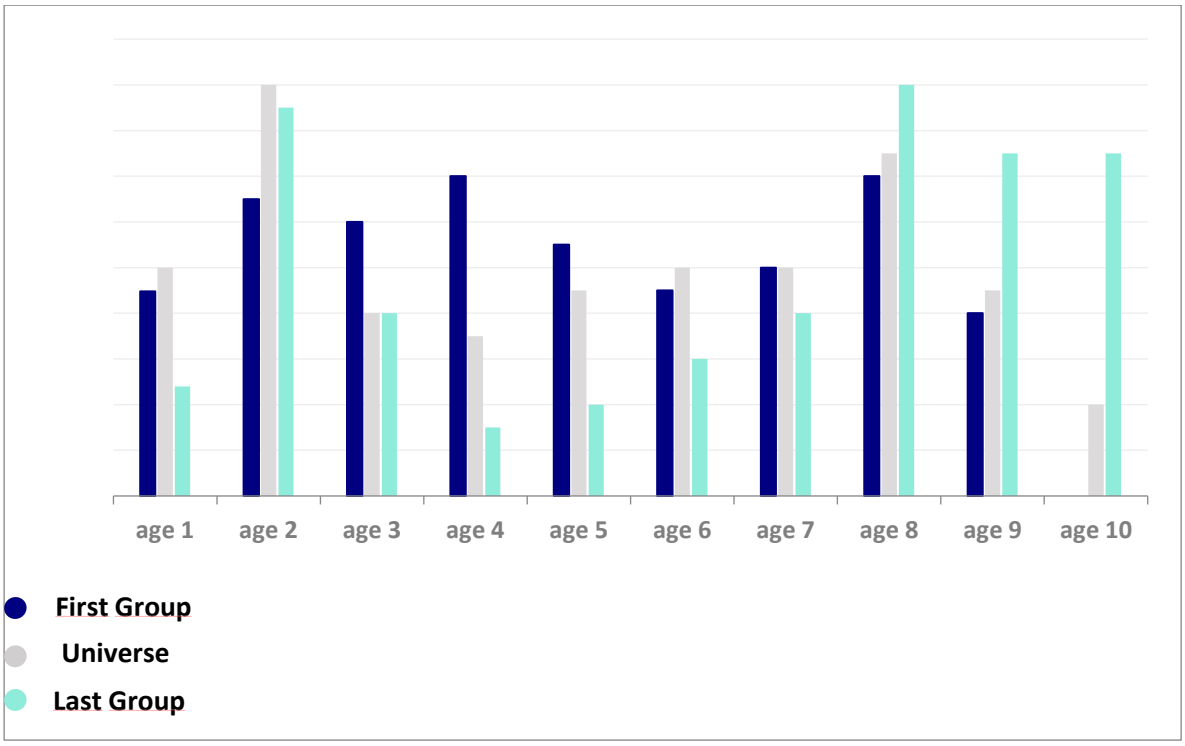


Figure 5.3 Column Chart. Model Variable 5 analysis blue box is the first group grey box is the universe distribution for var 5 group green is last group

According to the Column Chart figure 5.3 Var5 age 8, age 9 and age 10 have a higher proportion in the last group than the first group which make them more propense to renew however age 3, age 4 and age 5 are less propense to renew since they are more concentrated in the first group.

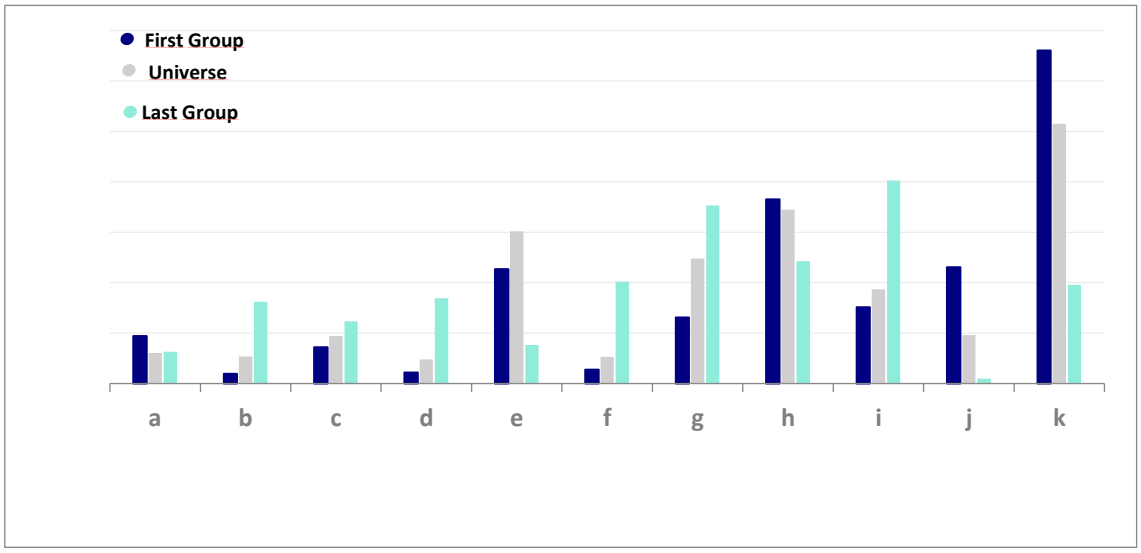


Figure 5.4 Column Chart. Model variable 7 analysis blue box is the first group grey box is the universe distribution for var 7 group green is last group

As shown in the Column Chart figure 5.4 Var 7 j,h and K are less propense to renew with a higher concentration in the first decile and Var 7 b, c, d, f, g and I are more propense to renew with a higher concentration in the last decile

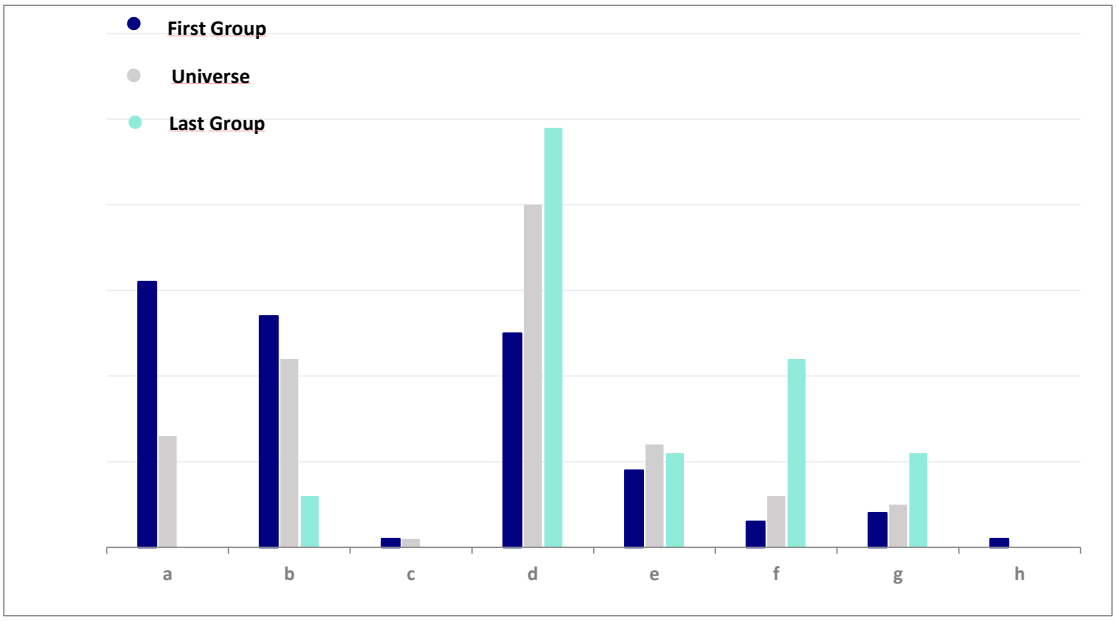


Figure 5.5 Column Chart. Model Variable 3 analysis blue box is the first group grey box is the universe distribution for var 3 group green

As shown in the Column Chart graph 5.5 we can notice that a, b and h from Var 3 are way more concentrated in the first decile therefore they are less propense to renew. Var 3 d, f and g are more propense to renew cause of their higher presence in the last decile.

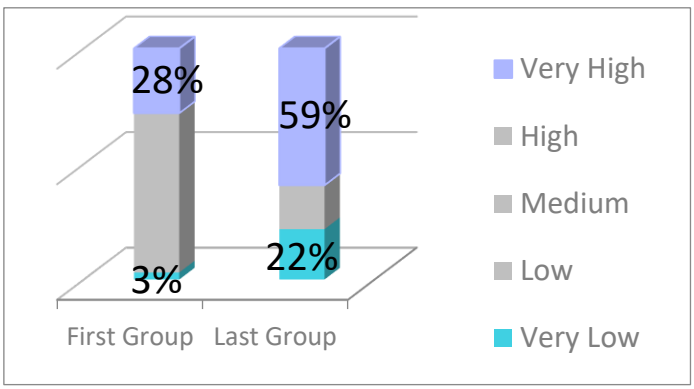


Figure 5.6 Stacked Column Chart. Var 14 distribution comparison between the first group and the last group

According to the Stacked Column Chart graph 5.6 customers that have a very high Var 14 and a very low Var 14 are more propense to renew since they have a higher proportion in the last group.

5.3. SIMULATION

The predictive Renewal model output is an estimated renewal which will give us time to have a deeper study on the Renewal Rate.

Simulation is performed monthly it will help to see in detail and all the change occurring per policy. In fact the pricing team will decide the new premium of the upcoming annuity and by knowing all the other input variable we can predict what will happen to the renewal rate in the future and therefore decide if the price is good or not.
Different test will be done trying to optimize the number of renewal and the premium acquired.

5.4. VISUALISATION

A Power BI visualization have was created and shared with the stakeholders in order to explain the characteristics that influences the Renewal rate.

Also, for further details by inserting the right inputs about a policy such as premium, claims and age, you will have the estimated renewal rate for that policy data.

Figure 5.7 Power BI Model Analysis with filter all important variable and provide an estimation of the renewal rate in 2018 data from the scoring sample in Emblem.

6. RECOMMENDATIONS FOR FUTURE WORKS

6.1. RECOMMENDATIONS FOR FUTURE WORKS

The project was successful and has been implemented directly in Médis. It was also a good case practice that made another partner such as the bank to review their churn model.

However, there is always space for improvement and some next steps works have been identified such as updating the model and using the new variables such as CLV and Web user. Use new technologies such as Data Robot in modeling and comparing the results that we had in Emblem. Finally, the year 2019 was an important year in for the company where changes were made to the tariffication and new products were launched. It is important to monitor the model and see the effect after those changes.

Finally automatize the process in SAS in order to have monthly table that the pricing team can use to perform the simulation.

6.2. Key Learning about the internship

The internship was also very challenging by dealing with a lot of unexpected things and learn a lot of new techniques and software. It helped me to grow and be more efficient at work. I was also agreeably surprised by the work environment where I had space to innovate and bring my own idea and this is not something you can do in every internship. At that moment I remembered what Professor Vanneschi said in our welcome day “After one year in this master you will be able to work and impact international company”. It was true for my case.

The biggest Challenges faced are first the amount of data you have to manage and definitely you need time and experience to get used to that. Another Important point is how to link all those knowledges that you acquire in the master’s degree and relate them to the business reality. Moreover, all the new software that you learn how to use such as Emblem.

Usually Challenges comes with lessons learned and I had the opportunity to interact with a lot of stakeholders and learn so much from them. First always validate your findings, learn how to deal with the current situation and reach your goals and Enhance your soft skills it is such an important skill in the Corporate world.

7. BIBLIOGRAPHY

Breiman, Leo, (2001). *Machine Learning* on Random Forest 45 (1), 5-32.

Chirstoph, Molnar, (2020). *Interpretable Machine Learning* on *Logistic Regression, Chapter4.2.*

Chirstoph, Molnar, (2020). *Interpretable Machine Learning* on *Decision Tree, Chapter4.4.*

Hanley, J., McNeil, B., (1982) on the meaning and use of the area under a receiver operating characteristic (ROC) curve, 29–36.

J., Lescovec, A., Rajaraman, J., D., Ullman, (2010). *Mining of Massive Datasets* on Dimensionality Reduction, Chapeter 11, 405-436.

Kiang, M., (2003), *A comparative assessment of classification methods* on Decision Support Systems, 35, 441- 454.

M., M., Gaber, (2010). *Scientific Data Mining and Knowledge Discovery* on Principles and Foundations.

SAS Enterprise Miner Version 14.3, (2020). On SEMMA.

Towers, Watson, Documentation, (2020). *Emblem Software* on Regression model and usage of the tool