A ESTATÍSTICA COMO FERRAMENTA PREDITIVA – ESTUDO DE CASO DE UMA EMPRESA SEGURADORA

Bernardo Filipe Marques de Almeida[1].

[1]Faculdade de Ciências da Universidade de Lisboa (FCUL), Lisboa, Portugal

RESUMO: A anulação da apólice decorre quando esta é terminada enquanto o risco existe. As duas principais causas são por iniciativa do tomador de seguro ou por falha no pagamento. É do interesse da seguradora proceder a uma tentativa de retenção da mesma apólice, no entanto inviável que este processo seja aplicado a todas as potenciais anulações, pelo que a empresa seguradora deverá concentrar os seus esforços onde conseguirá melhor impacto. É necessário analisar a carteira de clientes e os seus respetivos padrões de anulação para potenciar as acções de retenção da empresa. Para dar resposta a essa necessidade surge o projeto intitulado "Análise e Melhoria do processo de retenção do produto Multirriscos Habitação". O objectivo é descrever os processos de retenção existentes e, como a análise realizada ajudará na tomada de decisão a curto e longo prazo.Esta análise é compartida em três fases: uma análise de contextualização do negócio e descritiva da carteira e das anulações; uma análise clustering que nos permite proceder à segmentação da carteira de clientes que resulta numa campanha de retenção proactiva; a construção de um modelo de regressão logística múltipla que nos assinala os principais fatores de indicação de risco de anulação num cliente.

PALAVRAS-CHAVE: *Clustering*. Regressão Logística Múltipla. Multirriscos Habitação.

ANALYSIS AND IMPROVEMENT OF THE HOUSEHOLD INSURANCE'S RETENTION PROCESS

ABSTRACT: A Household policy cancels when it is terminated whilst its risk is still present. The two main reasons are by the policy holder's initiative or by lack of payment. Whenever this happens it is in the insurance company's best interest to retain the policy. However, it is unfeasible to extend this process to all potential cancellations, therefore the insurance company must focus their efforts to maximize impact. It is necessary to analyse the client portfolio and its respective cancellation patterns to potentialize the company's retention actions. In order to respond to this need, a project entitled "Analysis and improvement of the Household insurance's retention process" appears. With this project the goal is to describe the information relative to the existing retention processes and how the aforementioned analysis will help in short and long term decision making.This analysis is divided into three

stages: a preliminary analysis that contextualizes the existing business; a clustering analysis which allows us to proceed into client segmentation that will result in a proactive retention campaign; the construction of a multiple logistic regression model that specifies the main factors that indicate risk of cancellation in a client.

KEY-WORDS: Clustering. Multiple Logistical Regression. Household insurance policy.

INTRODUÇÃO

Este projecto resulta de uma parceria entre a Faculdade de Ciências da Universidade de Lisboa e a Ocidental Seguros, enquadrada no Mestrado de Matemática Aplicada à Economia e Gestão.

Do ponto de vista profissional, a necessidade deste relatório surgiu de uma vontade de compreender melhor o produto multirriscos, já que nenhuma análise aprofundada tinha sido realizada até ao momento. Este produto, mais conhecido pelo seguro de habitação, é um seguro que assegura o beneficiário do seguro contra riscos referentes ao seu imóvel e/ou conteúdos. O seguro de multirriscos representa quase um quinto das apólices da Ocidental.

A abordagem para compreender melhor os padrões de anulação e, de um ponto de vista mais prático, reduzir as anulações passará por dois pilares: segmentação e modelação. Estes dois estudos ajudar-nos-ão com objectivos a curto e longo prazo. A segmentação será usada para criar uma campanha proativa, procurando o foco num grupo em particular que se adeque a uma campanha específica para que o impacto na anulação seja o mais significativo possível. Existe uma preocupação por parte da empresa em criar uma abordagem personalizada no processo de retenção. O foco final acaba por ser sempre no impacto na taxa de anulação, mas espera-se que, através da análise de segmentação seja possível desenhar campanhas específicas aos perfis identificados de grupos de clientes e trabalhar assim numa vertente de melhoria de qualidade de serviço simultaneamente.

A modelação será usada como apoio para várias ações futuras. Criando um modelo que permita estimar a probabilidade de cada cliente anular (tendo em conta as suas características) pode-se centralizar os esforços onde o risco de anulação é maior, e assim aumentar a taxa de retenção.

OBJETIVO

Foi criado um objetivo geral:

- Análise aprofundada do produto multirriscos de modo a criar uma base para quaisquer análises e ações de marketing futuras;

e cinco objetivos específicos para o estágio:

- Identificação dos principais motivos e meios de anulação do produto;
- Análise dos processos de retenção já existentes e identificação de possíveis melhorias;
- Definição de um grupo de clientes para a campanha de retenção a realizar
- Criação de um modelo que possa estimar a probabilidade do cliente anular.

METODOLOGIA

O projeto decorreu em ambiente empresarial, inserido na equipa de *Data Mining & Retention*, no departamento de Marketing da Ocidental. O período do projecto foi desde 16 de setembro de 2019 a 15 de março de 2020, estando a totalidade dos 6 meses envolvido na resolução das questões discutidas neste relatório. Para além deste período intensivo, a investigação decorreu durante o restante ano de 2020 até à submissão final do relatório.

Este estudo quantitativo, transversal analítico tem objectivos tanto preditivos como explicativos. Para começar, é realizada uma análise preliminar resultante de inquéritos já realizados e de reuniões com as partes envolventes do projecto. Os inquéritos são não probabilísticos, sendo pedidos aos clientes no momento de anulação da sua apólice para um melhor entendimento da sua tomada de decisão. Esta análise inicial é usada para contextualizar e guiar as análises seguintes

Em seguida são usados os dados da base de dados da seguradora, previamente recolhidos e anonimizados. Todo o tratamento e análise dos dados é realizado usando o software *SAS Miner*.

Posteriormente, analisamos o perfil dos clientes a partir de segmentação por clusters, analisando os resultados em termos de relevância estatística (desvio-padrão, distância intra e intercluster) e relevância para o negócio (perfis de clientes distintos e identificáveis). Do resultado escolhido estudamos o cluster com melhor perfil para uma campanha de retenção.

Os métodos usados nesta segmentação serão o algoritmo *k-means* e o método de *Ward*. Foram testados vários algoritmos para compará-los e escolher a segmentação com melhores resultados. Esta comparação era avaliada por três factores: o desvio-padrão, a distância intra-cluster e a distância inter-cluster.

O critério de seleção de variáveis para a segmentação consistiu em escolher as variáveis cujos dados fossem mais fidedignos e que fossem mais relevantes. Este trabalho resultou na inclusão de 5 variáveis categóricas : “Tipo de Venda”, “Fracionamento”, “Segmento de Banco”, “Imóvel e/ou Conteúdo” e “Tipo de Produto” e 3 variáveis contínuas: “Idade”, “Nº de Coberturas”, “Nº de LoB” (Linhas de Negócio).

Seguidamente, combinando o uso da função de ligação logarítmica com os preditores lineares resulta no termo regressão logística (Faraway, 2016). Os métodos usados numa análise que usa regressão logística segue os mesmos princípios usados na regressão linear. O modelo de regressão logística é muito usado para casos onde a variável resposta é binária pela sua simplicidade da sua implementação computacional (Turkman & Silva, 2000).

No tratamento de dados é delineada a escolha de observações, a filtragem e partição dos dados, a imputação e o despiste da multicolinearidade. Daí prossegue-se à dita construção do modelo em si, com a seleção de variáveis, e subsequente avaliação do resultado. Finalmente, o modelo final escolhido é interpretado segundo a estimação dos coeficientes e a informação que daí advém.

A regressão logística é muito sensível a *outliers*, por isso é necessário analisar os dados e ver que observações com comportamentos extraordinários podem deturpar os resultados. É exportada uma lista do programa que nos fornece um sumário estatístico das variáveis, com informações das variáveis. A lista dá-nos a seguinte informação: Média; Desvio-padrão; Mínimo; Máximo; Nº de observações com valor atribuído.

Ademais, os valores máximos destas variáveis são extremamente elevados. Para garantir que *outliers* não deturpa os resultados do modelo, criou-se uma regra que remove todas as observações a mais de 4 desvios-padrão da média.

Outra questão a ter em conta é o número desequilibrado de observações para cada classe da variável resposta. Como apenas 6.3% dos casos anulam, temos uma enorme discrepância de observações entre as apólices anuladas e em vigor. Se considerarmos um modelo que prevê que todos os casos não anulam, teria uma taxa de sucesso de 93.7%, o que sem contextualização induziria em erro.

Assim sendo, é necessário fazer um ajuste a esta proporção entre valores desta variável binária de modo a que o modelo não menospreze o acontecimento que tentamos prever. Procedeu-se a um undersampling aleatório. Esta técnica consiste em remover aleatoriamente observações da classe maioritária de modo a igualar o número na classe minoritária. Com este ajuste garantimos que o modelo é alimentado por dados que não encaminham para um resultado falso por dados desequilibrados.

Resta-nos então definir as proporções das partições da base de dados. Quanto maior for a percentagem de dados utilizados para o treino, melhor será o desempenho geral (Paulino, Guimarães & Shiguemori, 2019). Será usada uma partição de 70% para treino, 30% para validação.

Foi incluído um processo de imputação de valores às observações sem valor. O método de imputação escolhido atribui às variáveis contínuas a média das observações da variável e nas variáveis categóricas imita a distribuição da variável e coloca aleatoriamente valores nas observações de modo a manter as proporções de distribuição da variável entre

as suas categorias.

Este método foi escolhido por duas razões. Em primeiro lugar, a percentagem de valores não atribuídos nas variáveis escolhidas é baixa (a maioria vai até aos 5%), o que reduz o perigo de atenuar quaisquer correlações envolvendo a(s) variável(s) que são imputadas presente neste tipo de imputação. Em segundo lugar, este método é computacionalmente simples, permitindo o seu uso num estudo com uma base de dados complexa.

Após verificar se existe multicolinearidade na lista de variáveis, tem de ser definido quais são usadas no modelo. O ideal é criar um modelo que explique bem a variável resposta da forma mais simples, i.e. usar variáveis suficientes para que a previsão da variável dependente seja adequada mas não usar demasiadas que denigram a simplicidade do modelo e subsequentemente a sua interpretação. Assim sendo, é necessário incluir um método que nos ajude a selecionar as variáveis no modelo final.

Para tal temos três algoritmos:

- Algoritmo de eliminação sequencial (*backward elimination*)
- Algoritmo de inclusão sequencial (*forward selection*)
- Algoritmo de exclusão / inclusão alternativos (*stepwise selection*).

Estes três algoritmos (juntamente com a opção de incluir todas as variáveis) são testados e comparados.

RESULTADOS E DISCUSSÃO

Clustering

À exceção do Cluster 4, cujo perfil já tinha sido de certa forma observado em iterações anteriores e apresenta um conjunto de sub-grupos completamente distintos dos restantes, os desvios-padrão são os mais baixos registados. As distâncias ao centróide são iguais ou menores do que as anteriormente registadas, apresentando valores entre 3.71 e 5.12. As distâncias aos clusters mais próximos revelam valores satisfatórios, com a excepção dos clusters 3 e 5.

Figura 1: Resultados segmentação método de Ward venda activa - 5 clusters

Segment Id	Frequency of Cluster	Root-Mean-Square Standard Deviation	Maximum Distance from Cluster Seed	Nearest Cluster	Distance to Nearest Cluster	Idade_saida
1	906	0.461307	4.486294	3	6.800067	54.674
2	16810	0.494752	5.124511	5	3.386583	58.373
3	76850	0.491443	3.718687	5	1.698522	57.580
4	9293	0.705233	4.551857	3	2.311066	53.893
5	54325	0.489282	4.136398	3	1.698522	61.362

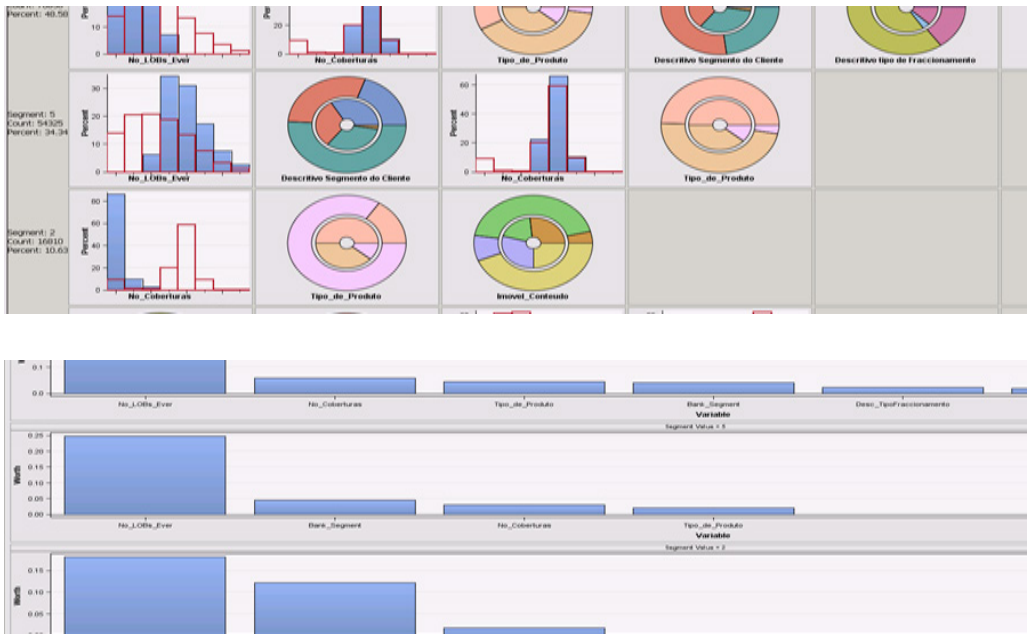
Fonte: Autoria Própria

A distribuição entre os 5 clusters não é uniforme, conforme esperado. O Cluster 3 representa quase metade dos clientes deste estudo, com uma proporção de 49%. O segundo maior cluster, o Cluster 5, detém 34% das observações enquanto os restantes três clusters, Cluster 1, Cluster 2 e Cluster 4 correspondem a menos de 1%, 11% e 6% dos clientes de Multirriscos Habitação de Venda Activa, respectivamente.

Esta observação não é preocupante pois o objetivo não era dividir os clientes em grupos de igual dimensão, mas sim obter grupos de clientes com perfis definidos que nos permitisse criar uma campanha de marketing para um ou mais clusters com risco elevado de anulação.

O Cluster 5 tem o perfil desejado. A sua dimensão é considerável, sendo composto por 58 mil apólices e com uma taxa de anulação de 7.5%, ligeiramente superior à média. O potencial de retenção em prémio anual é muito promissor: a soma de prémios anuais em apólices anuladas é de quase meio milhão de euros.

Figura 2: Variáveis usadas para definir o Cluster 5.



Segmento de Banco	Mass Market	Mass Market Plus	Prestige	Profissionais
Produto	MPC	HomIn	PCT	Outros

Fonte: Autoria Própria

É constituído por clientes com mais linhas de negócio. Relativamente ao segmento de banco, 51% destes clientes são do segmento mais afluente do banco, *Prestige*, com o segundo maior grupo (29%) sendo o *Mass Market Plus*, o grupo de afluência intermédia. Tal como verificado no cluster 3, o Nº de Coberturas é muito semelhante à população, com a exclusão das observações com poucas coberturas alocadas ao Cluster 2. Os tipos de produto dividem-se entre os dois mais recentes, HomIn e Proteção Casa Mais, com 49% e 47% respetivamente.

Começando pelas variáveis discretas, apresentamos abaixo os valores do Cluster 5 comparando com universo de estudo. Olhando para a figura 2, podemos ver que o Cluster 5 é mais afluente que a média. Com 51% dos clientes pertencendo ao segmento mais afluente *Prestige* e com o grupo menos afluente *Mass Market* em menor peso, podemos concluir que este grupo apresenta clientes mais rentáveis do que a média.

Não há grande diferença relativamente ao fracionamento, exceptuando a proporção de apólices com pagamento semestral que é substituída por pagamento anual. Assim sendo, as maiores formas de pagamento deste cluster são anualmente (47%) e mensalmente (41%).

O Cluster 5 apresenta então o melhor balanço entre os três principais fatores:

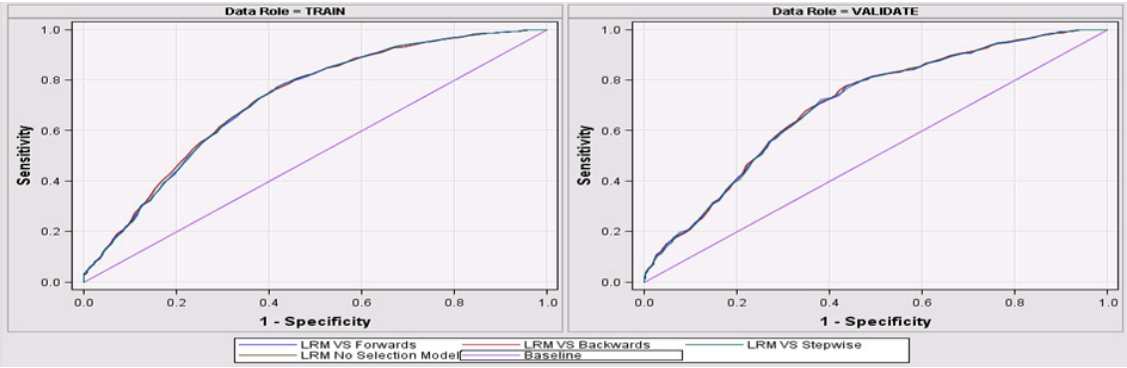
- Taxa de anulação alta – Maior risco de anulação;
- Perfil claramente definido – Maior facilidade em desenhar uma campanha *ad hoc*;
- Volume significante – possibilidade de um impacto significativo.

Modelo de Regressão Logística Múltipla

É necessário usar métricas para definir a qualidade dos diferentes modelos testados de modo a aferir qual o melhor. Para tal, os seguintes 3 métodos de avaliação foram utilizados: Curva ROC e AUR; *Cummulative lift; P-value* dos coeficientes estimados.

A figura 3 mostra-nos a representação das curvas ROC dos 4 modelos testados, tanto para os dados de treino como os de validação. A primeira observação a fazer é que os resultados embora diferentes se apresentam muito semelhantes entre as quatro alternativas, quer em treino que na validação. Como seria de esperar as curvas aproximam-se mais do modelo aleatório (*baseline*) em validação comparativamente ao treino.

Figura 3: Comparação do AUC dos modelos



Fonte: Autoria Própria

Os modelos apresentam todos valores que rondam os 0.72 em treino, enquanto os valores da área abaixo da curva nos dados de validação estão ligeiramente acima de 0.7.

Estes valores tão próximos não nos ajudam a escolher entre as diferentes alternativas. Todavia, os valores de AUC acima de 0.7 indicam-nos que a precisão do modelo atinge uns níveis aceitáveis, ainda que pudessem ser mais altos.

O *cumulative lift* é uma representação visual que nos ajuda a medir a eficácia de um modelo de classificação preditiva. Esta métrica indica-nos o quanto "ganhamos" em usar este modelo para selecionar clientes com maior risco de anulação comparando com escolher aleatoriamente quaisquer clientes para uma campanha de marketing.

As curvas apresentam resultados muito semelhantes. Independentemente da alternativa de seleção de variáveis, a curva começa com valores abaixo de 1.6, baixando de 1.4 a partir do 20º percentil em treino. Em validação, a curva curiosamente começa com valores ainda mais próximos de 1.6, superiores à outra partição de dados. No entanto apresenta uma queda superior, ficando abaixo de 1.4 por volta do 15º percentil.

Os valores confirmam a observação inesperada dos valores na partição de dados para validação serem superiores do que na partição de treino. Os algoritmos *forward selection* e *backward elimination* apresentam um *cumulative lift* de 1.44, tal como a alternativa sem algoritmo, enquanto o algoritmo *stepwise* tem um *cumulative lift* no 1º decil de 1.4.

Esta diferença mínima mantém-se na validação, com *stepwise* a registar valores de 1.46, atrás das três outras opções com 1.49.

O *p-value* corresponde à probabilidade da estatística de teste, sob a validade de , produzir um valor tão ou mais extremo do que os valores observados.

O algoritmo de exclusão/inclusão alternativo é o único que apresenta valores aceitáveis. O *p-value* mais alto é na estimação do coeficiente da variável Sum_PremAnual: 0.01 (figura 4). Todos os restantes *p-values* são inferiores a 0.005.

Figura 4: *P-value* dos coeficientes estimados – algoritmo *stepwise selection*

Likelihood Ratio Test for Global Null Hypothesis: BETA=0

-2 Log Likelihood Intercept Only	-2 Log Likelihood Intercept & Covariates	Likelihood Ratio Chi-Square	DF	Pr > ChiSq
5464.768	4842.692	622.0764	14	<.0001

Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
Desc_TipoFraccionamento	3	25.5587	<.0001
IMP_Bank_Segment	3	27.7857	<.0001
IMP_REP_No_6Mth_Issued_VendaAsso	1	71.8193	<.0001
IMP_REP_Sum_PremAnual	1	6.4033	0.0114
IMP_REP_Sum_PremPaid_VendaAssoc_	1	8.7764	0.0031
IMP_REP_Yrs_Since_Latest_Purchas	1	48.3086	<.0001
Imovel_Conteudo	3	79.5426	<.0001
REP_Months_until_renewal	1	354.4311	<.0001

Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq	Standardized Estimate	Exp(Est)
Intercept		1	1.7521	0.1240	199.56	<.0001		5.766
Desc_TipoFraccionamento	Anual	1	-0.2621	0.0628	17.40	<.0001		0.769
Desc_TipoFraccionamento	Mensal	1	0.1330	0.0665	4.00	0.0455		1.142
Desc_TipoFraccionamento	Semestral	1	0.1217	0.1139	1.14	0.2853		1.129
IMP_Bank_Segment	Mass Market	1	0.1393	0.0836	2.78	0.0956		1.149
IMP_Bank_Segment	Mass Market Plus	1	-0.2458	0.0909	7.31	0.0068		0.782
IMP_Bank_Segment	Prestige	1	-0.2274	0.0949	5.74	0.0165		0.797
IMP_REP_No_6Mth_Issued_VendaAsso		1	-71678.0	8458.0	71.82	<.0001	-48.1478	0.000
IMP_REP_Sum_PremAnual		1	4.844E-6	1.914E-6	6.40	0.0114	0.0532	1.000
IMP_REP_Sum_PremPaid_VendaAssoc_		1	-0.00009	0.000029	8.78	0.0031	-0.0600	1.000
IMP_REP_Yrs_Since_Latest_Purchas		1	-0.1873	0.0270	48.31	<.0001	-0.1418	0.829
Imovel_Conteudo	Conteúdo	1	-0.0905	0.0749	1.46	0.2272		0.913
Imovel_Conteudo	Imóvel	1	-0.3939	0.0807	23.84	<.0001		0.674
Imovel_Conteudo	Imóvel+Conteúdo	1	-0.5348	0.0668	64.02	<.0001		0.586
REP_Months_until_renewal		1	-0.2210	0.0117	354.43	<.0001	-0.3907	0.802

Fonte: Autoria Própria

Tendo em conta que relativamente à área abaixo da curva e ao *cumulative lift* não existem grandes diferenças nos valores registados, a decisão será feita através do último método e o mais importante: o *p-value* dos coeficientes estimados. Conforme já visto, o algoritmo de seleção *stepwise* é o único que satisfaz todas as restrições, pelo que o modelo que daí adveio será o utilizado.

A equação do modelo de regressão logística múltipla, relembrando, é a seguinte:

$$\pi(x) = \frac{e^{g(x)}}{1+e^{g(x)}}$$

em que, aplicando ao caso do modelo construído através do algoritmo *stepwise selection* resulta na equação abaixo:

$$g(x)=\beta_0+\sum_{l=1}^{3}\beta_{1:l}x_{1:l}+\sum_{l=1}^{3}\beta_{2:l}x_{2:l}+\beta_3x_3+\beta_4x_4+\beta_5x_5+\beta_6x_6+\sum_{l=1}^{3}\beta_{7:l}x_{7:l}+\beta_8x_8$$

onde:

- é o *intercept*;
- é o coeficiente associado à variável ;
- é o coeficiente associado à variável *dummy* ;
- representa a variável *dummy*: Fracionamento: Anual;
- representa a variável *dummy*: Fracionamento: Mensal;
- representa a variável *dummy*: Fracionamento: Semestral;
- representa a variável *dummy*: Segmento: *Mass Market*;
- representa a variável *dummy*: Segmento: *Mass Market Plus*;
- representa a variável *dummy*: Segmento: *Prestige*;
- representa a variável: No_6Mth_Issued_VendaAsso;
- representa a variável: Sum_PremAnual;
- representa a variável: Sum_PremPaid_VendaAssoc;
- representa a variável: Yrs_Since_Latest_Purchase;
- representa a variável *dummy*: Objecto segurado: Conteúdo;
- representa a variável *dummy*: Objecto segurado: Imóvel;
- representa a variável *dummy*: Objecto segurado: Imóvel+Conteúdo;
- representa a variável: Months_until_renewal.

Os coeficientes das variáveis do tipo de objeto segurado indicam-nos que os clientes com Imóvel + Conteúdo ou só Conteúdo têm uma probabilidade de anular superior do que os clientes só com o Imóvel coberto, com a opção de objeto segurado combinado a apresentar a maior magnitude.

Relativamente ao fracionamento, o pagamento anual é o que representa menor probabilidade de anulação, enquanto os pagamentos mensais e semestrais assinalam uma maior probabilidade do cliente anular.

A outra variável categórica, o segmento de banco, indica-nos que os clientes mais afluentes presentes nas categorias *Mass Market Plus* e *Prestige* apontam para uma probabilidade menor de anulação enquanto *Mass Market* revela uma inclinação para uma maior probabilidade de anulação.

Olhando agora para as variáveis contínuas, o sinal negativo do coeficiente da variável relativa à proximidade da data de renovação leva-nos a concluir que, quanto maior for o valor desta variável, menor a probabilidade de o cliente anular.

A mesma interpretação pode ser feita no caso da variável do Nº de anos desde a última compra. Curiosamente, e talvez opondo o expectável, quanto maior o intervalo de tempo desde a última compra menor a probabilidade de anulação. Esta descoberta pode dever-se ao facto de clientes que são estáveis, sem necessidade de adquirir novos produtos, apresentarem menor risco de anulação. Por contraste, clientes informados e ativos no mercado segurador podem apresentar maior risco de anular.

As restantes duas variáveis apresentam coeficientes estimados muito baixos. No entanto, tendo em conta que representam variáveis relativas a prémios, o seu valor baixo deve-se não à sua relevância (ou falta dela), mas sim aos valores elevados registados. Multiplicando os valores destas variáveis por este coeficiente pode revelar que esta componente até afete mais do que algumas suprarreferidas.

Estas apresentam sinais opostos, i.e., embora quanto maior o valor de prémio pago em apólices de venda associada menor o risco de anulação, o mesmo não se demonstra no valor de prémio total, o que nos indica que, quanto maior o valor de prémio pago em apólices de venda ativa, maior a probabilidade do cliente anular.

Este modelo poderá ser usado no futuro para campanhas e como apoio à definição de *leads*, todavia o seu principal propósito era ser uma ferramenta na análise do risco de anulação dos clientes Multirriscos, e mais detalhadamente, dos clientes HomIn de venda ativa.

Através desta análise foi possível determinar que factores são mais preeminentes na anulação. Alguns resultados confirmaram a importância de certas variáveis para a probabilidade de anulação de um clientes (como fracionamento e segmento de banco). No entanto, informação valiosa foi recolhida deste processo que até agora não tinha sido notada. A importância da proximidade à renovação é o factor mais importante a explorar no futuro, já que nos permite uma rápida implementação nas ferramentas de retenção já existentes, através da priorização dos clientes cuja data de renovação esteja mais próxima.

Outra observação inesperada terá sido do efeito do Nº de anos desde a última subscrição. Fortalece a noção de divisão entre clientes estáveis e clientes mais voláteis com necessidades de abordagem diferente, sendo que o segundo grupo tem maior risco de anulação.

As variáveis referentes ao prémio apoiam ainda esta linha de pensamento. É do nosso conhecimento que os clientes mais estáveis são clientes nomeadamente de venda associada a crédito, que sentem mais-valias na centralização dos dois serviços, enquanto os clientes de venda ativa estudam mais o mercado e procurar continuamente um melhor serviço, ao detrimento dessa estabilidade. Esta divisão revela-se no efeito das variáveis de prémio na probabilidade de anulação, com o prémio das apólices associadas a crédito a reduzir essa probabilidade enquanto o total (por efeito das apólices de venda ativa) aumenta essa mesma probabilidade.

CONSIDERAÇÕES FINAIS

Revendo os objetivos criados, podemos concluir com satisfação que todos foram cumpridos.

Este relatório começou por referir a análise exploratória que antecedeu a segmentação e modelação ulteriores. Através do estudo de documentos já existentes, reuniões com stakeholders e da análise descritiva efetuada, alguns pontos foram logo registados como sendo de importância para a retenção deste produto:

- Filtragem do fenómeno de canibalização;
- Potencial de venda de clientes cuja apólice é vencida ou crédito é liquidado;
- Sensibilidade ao preço;
- Maior risco de anulação em clientes de venda ativa;
- Preeminência de anulação por cancelamento do débito direto (anulação direta não presencial).

Seguidamente, a análise de clustering resultou na identificação de um grupo de clientes de venda ativa, mais velhos, mais afluentes e com maior associação à empresa. Estes clientes não só se revelam extremamente apelativos pelas suas características que revelam trazer muito valor à empresa, mas também têm a vantagem adicional de terem um perfil claramente distinto o que permite o desenho de uma campanha ad hoc, de modo a ajustá-la melhor às necessidades do cliente.

Por fim, a construção do modelo de regressão logística não só criou uma ferramenta valiosa na seleção de clientes para campanhas de retenção futuras, mas também permitiu aferir quais são de facto as variáveis que mais afectam o risco de anulação do cliente, como a proximidade à data de renovação e o Nº de anos desde a última subscrição. Esta análise cria uma base de conhecimento sobre o produto e a sua anulação que nos permite focar nos fatores que mais afectam a anulação de uma forma mais estruturada, ao invés de nos cingirmos a implementações constantes de campanhas a curto prazo.

É de realçar que todo este projecto foi um trabalho iterativo, repleto de repetições e processos de tentativa e erro até chegar aos resultados obtidos. A necessidade de criar

uma estrutura revelou-se fundamental para usar como apoio durante todo o trabalho, especialmente com todos os ajustes necessários de forma a acomodar as descobertas e os entraves que foram aparecendo ao longo do caminho.

A figura 5 apresenta uma análise SWOT feita ao projecto na sua totalidade. Dividida numa matriz 2∗2, esta análise caracteriza o projecto ao longo de duas vertentes: componentes de origem interna e externa e, se são beneficiais ou prejudiciais ao âmbito do projecto.

Figura 5: Análise SWOT do projecto

Fonte: Autoria Própria

Começando pelas características internas beneficiais, este estudo fornece à empresa um maior conhecimento o perfil do cliente de MRH e dos seus padrões de anulação. Esta base informativa só por si já é proveitosa mas também nos ajuda na tomada de decisão das próximas ações de retenção, enquanto simultaneamente nos mune de informação que nos ajuda a fazer mudanças a um nível estrutural na retenção. A equipa de retenção começará pelo lançamento de uma campanha de retenção proativa segundo a análise feita no capítulo 3. De seguida, podemo-nos focar nos clientes cujas apólices se encontrem perto de renovar, já que se revela um fator preponderante no seu risco de anulação.

Esta informação pode dar seguimento à criação de um método permanente de acompanhamento de clientes com esta característica numa ótica de *customer service*/ retenção criando assim uma mudança estrutural no processo de retenção já instalado. Ademais, o modelo de regressão logística construído servirá para filtrar e ordenar os clientes para futuras campanhas de retenção, permitindo um foco nos clientes cujo risco de anulação seja superior.

Por outro lado, existe uma preocupação relativamente à dificuldade de implementação de campanhas de retenção do produto MRH. Tendo em conta o seu prémio e taxa de anulação baixos, qualquer campanha a considerar não pode incorrer custos iniciais demasiado altos sob o risco de não ser economicamente viável.

Adicionalmente, sendo este um projeto com finalidade tanto profissional como académica, houve sempre um foco no equilíbrio entre o rigor científico necessário e a finalidade de aplicar os conhecimentos adquiridos no negócio. Esta ponte entre a teoria e a prática revelou-se por vezes difícil de percorrer, contudo esta condição foi fundamental no processo e aprendizagem e, juntamente com o ênfase na inovação proporcionou uma base analítica para a retenção dos clientes de Multirriscos Habitação que revelar-se-á muito valiosa.

Uma oportunidade dentro da empresa que certamente beneficiará do trabalho realizado é a preocupação crescente na retenção como componente estratégica de crescimento de carteira. Com um estagnar das vendas a nível transversal devido à pandemia, existe agora uma preocupação rejuvenescida na retenção, o que trará um ímpeto necessário na operacionalização de qualquer ação criada a partir da análise supracitada.

Esta oportunidade vem às custas de uma grande ameaça. O aparecimento do covid-19 trouxe muitas incertezas e, o seu efeito no comportamento dos clientes não é exceção. Toda a análise foi feita com alguns pressupostos, sendo um deles a inexistência futura de uma pandemia global que teria repercussões a níveis económicos, políticos e também de saúde pública. Será certamente difícil aferir o efeito deste “cisne negro” (Taleb, 2007) no padrão de comportamento dos clientes do seguro Multirriscos Habitação e numa perspetiva muito mais macro, a todos nós enquanto sociedade.

REFERÊNCIAS

ALPUIM, T. **Apontamentos para apoio à disciplina de Actividade Seguradora do curso do Mestrado de Matemática Aplicada à Economia e Gestão**. Lisboa: AVR Handouts – FCUL, 2018.

DOBSON, A. J. & BARNETT, A. G. **An Introduction to Generalized Linear Models** (4th ed.). Boca Raton: CRC Press, 2018.

EL HALIMI, Rachel et all. **Optimal tests for random effects in linear mixed models**. Tangier: Laboratoire de Mathématiques et Applications, Faculté des Sciences et Techniques,

Abdelmalek Essaâdi University, 2023.

EVERITT, B. **Cluster Analysis**. West Sussex: Heinemann Educational Books, 1980.

FARAWAY, J. J. **Extending the Linear Model with R** – Generalized Linear, Mixed Effects and Nonparametric Regression Models (2nd ed.). Boca Raton: CRC Press, 2016.

HOSMER, D. W. & LEMESHOW, S. **Applied Logistic Regression** (2nd ed.). Nova Iorque: John Wiley & Sons Inc., 2000.

OLIVEIRA, M. F.. A**pontamentos para apoio à disciplina de Risco em Seguros Vida e não Vida do curso do Mestrado de Matemática Aplicada à Economia e Gestão**. FCUL, Lisboa, 2018.

RASMUSSEN, Louise et all. C**irculating cell-free nucleosomes as biomarkers for early detection of colorectal cancer**. Hvidovre: Oncotarget, 2017.

TALEB, N. N. **The Black Swan.** Random House, 2007.

TURKMAN, M. A. A. & SILVA, G. L. **Modelos Lineares Generalizados** – da teoria à prática. Lisboa: VIII Congresso Anual da Sociedade Portuguesa de Estatística, 2000.

X. Kavelaars et al. **Bayesian multilevel multivariate logistic regression for superiority decision-making under observable treatment heterogeneity**. BMC Medical Research Methodology, 2023.

XU, R. & WUNSCH, D. **Clustering**. Nova Iorque: John Wiley & Sons, 2008.

ZHANG, Shichao & ZHANG, Chengqi & YANG, Qiang. **Data Preparation for Data Mining**. Applied Artificial Intelligence. Sydney: Taylor & Francis, 2003.