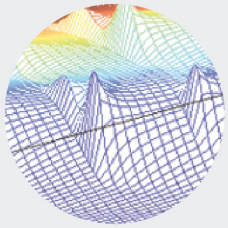
INSTITUTO SUPERIOR DE ENGENHARIA DO PORTO
MESTRADO EM MATEMÁTICA APLICADA À ENGENHARIA E ÀS FINANÇAS






Modelação do cancelamento de apólices de seguro automóvel, por parte do cliente

VASCO ANDRÉ MOTA TORRES CORDEIRO
novembro de 2017

POLITÉCNICO
DO PORTO

Vasco André M. T. Cordeiro

Modelação do cancelamento de apólices de seguro automóvel, por parte do cliente

Instituto Superior de Engenharia do Porto
Outubro de 2017

Vasco André M. T. Cordeiro
vascoandre28@hotmail.com

Modelação do cancelamento de apólices de seguro automóvel, por parte do cliente

Relatório de Estágio
Mestrado em Matemática Aplicada à Engenharia e às Finanças

Instituto Superior de Engenharia do Porto
Outubro de 2017

Abstract

Policy cancellations directly influence daily business operations and have an impact on the risk assumed by insurance companies, which makes clear the need to take an active role in defining strategies that promote customer loyalty and guarantee the means to facilitate monitoring of commercial risks.

It was intended to identify, among the customers with auto insurance policies of an insurance company, characteristics that distinguish those who cancel their insurance policies from other clients, so that, knowing the characteristics of a new individual, one can predict to which group belongs.

Starting point: two data sets containing information on insurance policies and car accidents occurred. The data sets were worked and unified to a final set, composed by 371252 policies active in the year 2015 and, for each of them, data related to twenty variables, one of them being the response variable (policy cancellation, by the customer) to be predicted by the nineteen remaining.

A generalized linear model with logistic regression, constructed with the set of the four most significant variables (*initial date*, *bonus*, *form of payment* and *coverages pack 2*), due its parsimony, was preferred by the insurance company. The corresponding ROC curve has an AUC of approximately 66.6%, with $\text{IC}_{95\%} = [65.8\%, 67.3\%]$. Setting the 80% target for sensitivity, we have a specificity of approximately 42% and an accuracy just over 47%. As an alternative, it is presented a model constructed using five simple variables (*initial date*, *bonus*, *form of payment*, *coverages pack 1*, *policy premium*) and one of the interactions between a pair of these variables (*initial date / form of payment*). This model corresponds to a ROC curve with $AUC = 69.4\%$, $\text{IC}_{95\%} = [68.7\%, 70.1\%]$ and, for the same objective of about 80% sensitivity, achieves 47% of specificity and accuracy just over 51%.

Having as goal for the model the objective of the balance between good adjustment, parsimony and interpretation, it was noticed that only the quality of adjustment is not what it is desired.

Key-Words: customer loyalty, generalized linear models, logistic regression, ROC curve

Resumo

Os cancelamentos de apólices influenciam diretamente as operações comerciais diárias e têm impacto no risco assumido pelas companhias de seguros, o que torna evidente a necessidade de ter um papel ativo na definição de estratégias que levem à promoção da fidelidade do cliente e na garantia de meios que facilitem a monitorização dos riscos comerciais.

Quis-se identificar, entre os clientes com seguro automóvel de uma companhia de seguros, características que distinguem aqueles que cancelam as suas apólices de seguro dos outros clientes, de modo que, conhecidas as características de um novo indivíduo, se possa prever a que grupo pertence.

Ponto de partida: dois conjuntos de dados contendo informações relativas a apólices de seguro e a sinistros ocorridos. Os conjuntos de dados foram trabalhados e unificados até um conjunto final, composto por 371252 apólices ativas no ano de 2015 e, para cada uma delas, dados relativos a vinte variáveis, sendo uma delas a variável resposta (anulação da apólice, por parte do cliente) a ser predita pelas dezanove restantes.

Um modelo linear generalizado, em particular uma regressão logística, construído com o conjunto das quatro variáveis mais significativas (*data inicial*, *bonificação*, *forma de pagamento* e *pack de coberturas 2*), dada a sua parcimónia, foi preferido pela companhia de seguros. A curva ROC correspondente tem uma AUC de aproximadamente 66.6%, com $\text{IC}_{95\%} = [65.8\%, 67.3\%]$. Fixando o objetivo de 80% para a sensibilidade, tem-se uma especificidade de aproximadamente 42% e uma exatidão pouco superior a 47%. Como alternativa, apresenta-se um modelo, construído com recurso a 5 variáveis simples (*data inicial*, *bonificação*, *forma de pagamento*, *pack de coberturas 1*, *prémio apólice*) e a uma das interações entre um par dessas variáveis (*data inicial / forma de pagamento*). A este modelo corresponde uma curva ROC com $AUC = 69.4\%$, $\text{IC}_{95\%} = [68.7\%, 70.1\%]$ e, para o mesmo objetivo de cerca de 80% de sensibilidade, consegue 47% de especificidade e exatidão pouco acima de 51%.

Tendo como meta para o modelo o objetivo do equilíbrio entre bom ajustamento, parcimónia e interpretação, verificou-se que apenas a qualidade de ajustamento deixa a desejar.

Palavras-Chave: fidelidade do cliente, modelos lineares generalizados, regressão logística, curva ROC

Dedicatória

Exemplo, Amor, Futuro

Um avô, Mota,
Que me mostrou como crescer

A mulher, Sofia
Clara realização do meu sonho
O meu Para Sempre

Quem ela nos traz
Com Ele, duas dádivas de três letras

Agradecimentos

Aos professores do ISEP,
pela sabedoria transmitida nesta nova etapa de minha formação

Entre estes, em particular, à diretora de mestrado, Dra. Stella Abreu,
pois, desde o primeiro dia, vestiu a camisola pelo curso

À Ageas Portugal, Companhia de Seguros, S.A.,
pela oportunidade de realização deste estágio

Ao Dr. Luís Maranhão, elo com a Ageas,
pelos primeiros passos, pela disponibilidade e prestabilidade que sempre evidenciou

À minha orientadora, Dra. Sandra Ramos,
por toda a ajuda, prontidão, simpatia e compreensão demonstradas

Conteúdo

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

1.1 Fidelidade do cliente na área dos seguros

Para uma companhia de seguros, um cliente pode ser definido como uma pessoa, empresa ou qualquer outra organização com uma ou mais apólices de seguro, de um ou mais tipos [21]. Adquirida uma primeira apólice, começa a ser recolhida, ao longo do tempo, informação estatística de maior ou menor relevância, como coberturas abrangidas, renovações, novas apólices, reivindicações, cancelamentos ou queixas.

As companhias de seguros têm de gerir uma grande quantidade de riscos, que podem ser classificados de diferentes maneiras. Uma possibilidade é distinguir entre risco financeiro e risco operacional. O risco financeiro é classificado como risco de responsabilidade, o risco que a companhia de seguros está a assumir ao vender contratos de seguro, ou como risco de ativos, associado à gestão de ativos de uma seguradora. Estes dois tipos de riscos estão diretamente relacionados com a atividade comercial da companhia de seguros e são, portanto, bem conhecidos e facilmente geridos por meio de várias técnicas quantitativas. O risco que não pode ser classificado como risco de ativos ou de responsabilidade é chamado de risco operacional, e é subdividido em risco comercial, impulsionado pelo ambiente competitivo, e risco de evento, como, por exemplo, um sistema informático com erros ou avarias. O risco comercial é definido como a variabilidade do valor comercial intrínseco devido ao volume de negócios e às flutuações das margens desencadeadas pelo ambiente competitivo.

Os cancelamentos de apólices influenciam diretamente as operações comerciais diárias e têm impacto no risco assumido pelas companhias de seguros. Por si só, isto torna evidente a necessidade, por parte das companhias de seguros, de ter estratégias que levem à promoção da fidelidade do cliente e meios que facilitem a monitorização dos riscos comerciais.

As companhias de seguros precisam de monitorizar a fidelidade do cliente e o risco comercial pelas seguintes razões [20]:

- *Recolher informações sobre a qualidade do portfólio.* A qualidade do portfólio depende do nível de rentabilidade de um cliente, sendo maior quando o nível de rentabilidade é maior. Clientes de alta rentabilidade são clientes cujo comportamento de sinistros observado está

abaixo das expectativas (menor custo e/ou menor variabilidade). Uma vez que novas apólices vão sendo constantemente subscritas ou canceladas, a qualidade do portfólio também está em constante alteração. A monitorização da fidelidade de clientes permite que as seguradoras identifiquem essas mudanças e, assim, possam evitar possíveis perdas de lucros da empresa.

- *Implementar com sucesso as estratégias de captação e retenção de clientes.* Diferentes estratégias de marketing são geralmente aplicadas a diferentes categorias de clientes. Portanto, é necessário gerar dados específicos sobre a fidelidade e rentabilidade de cada cliente. Classificando os clientes por seus níveis de fidelidade e rentabilidade permite que a empresa crie o procedimento de marketing mais eficiente para cada cliente individual e, portanto, maximize o impacto positivo nas margens globais do negócio. Para maximizar os benefícios, o curso de ação mais comum é manter clientes bons e de alta rentabilidade e identificar clientes maus e de baixa rentabilidade, persuadindo os últimos a deixar a empresa.

- *Avaliar a competitividade do mercado de seguros.* Mercados de seguros altamente competitivos causam altas flutuações na composição das carteiras. A empresa é capaz de avaliar o impacto da competitividade do mercado no risco de negócios de seguros, analisando essas flutuações e a evolução da rentabilidade do cliente. Flutuações desfavoráveis da qualidade da carteira e da rentabilidade dos clientes podem afetar a estabilidade e a solvência da companhia de seguros. Portanto, ao monitorizar essas flutuações e o risco comercial, a empresa protege-se contra perdas potenciais.

- *Recolher informações sobre a posição da empresa no mercado.* Quando uma apólice é cancelada, a companhia de seguros costuma saber por qual concorrente o cliente trocou a apólice dado que, quando o cliente decide transferir uma determinada apólice, é muitas vezes a nova companhia de seguros que comunica o cancelamento à seguradora anterior em nome do cliente. Assim, ao monitorizar os seus próprios registos de novos contratos e cancelamentos, a empresa aprende sobre como se compara aos seus concorrentes em termos de eficácia do recrutamento de clientes, bem como sobre a sua posição no mercado de seguros. Nesse sentido, o papel do agente é muito importante. A maioria das pessoas não tem relação direta com a companhia de seguros, estabelecendo toda a relação por intermédio de um agente (se o agente mudar de companhia de seguros, muitos dos seus clientes o seguirão, mesmo que a antiga companhia de seguros tenha bons produtos).

Na abordagem do estudo da fidelidade do cliente dentro do setor de seguros, é obrigatório considerar a perspetiva econométrica. Os modelos econométricos são amplamente utilizados no campo atuarial, especialmente no seguro automóvel. Um dos antecedentes mais importantes nesse sentido é o contributo de Dionne e Vanasse [14] em que eles predizem o prémio puro do titular da apólice em função de suas características de risco. Mais tarde, Dionne, Gourieroux e Vanasse [12] usaram um modelo probit ordenado para prever o número de reivindicações em apólices individuais. As decisões de compra também foram estudadas e um modelo probit dicotómico foi usado para analisar o que faz o cliente escolher uma apólice com franquia.

Pinquet [29, 30] usou modelos de regressão de Poisson e modelos log-lineares para explicar a frequência e a magnitude do custo das reivindicações. Abrahamse e Carroll [1] usaram um modelo de regressão linear múltipla para analisar a proporção de reivindicações graves em várias

zonas nos EUA. Dionne, Laberge-Nadeau, Desjardins, Messier e Maag [13] usaram um modelo de regressão logística para analisar a probabilidade de haver pelo menos uma reivindicação no período de um mês, para avaliar o impacto de uma nova legislação para obter a carta de condução no Canadá em 1991.

Todas as contribuições anteriores foram destinadas a avaliar o risco em termos de número de reivindicações e/ou a sua gravidade. O principal objetivo é fornecer uma estimativa apropriada para preços e seleção de riscos. Nos últimos anos, outras questões interessantes chegaram à cena do seguro automóvel. Um exemplo vem de Artís, Ayuso e Guillen [7, 8] e é a aplicação de modelos com variáveis dependentes qualitativas na deteção de fraude no seguro automóvel.

1.2 O estágio: descrição, objetivos e metodologia

A presente dissertação surge no âmbito do estágio a decorrer em parceria com a Ageas Portugal, Companhia de Seguros, S.A., parte integrante do plano de estudo do Mestrado em Matemática Aplicada à Engenharia e às Finanças, na disciplina de Dissertação/Projeto/Estágio, lecionada pelo Instituto Superior de Engenharia do Porto (ISEP).

A Ageas é um grupo segurador internacional, sediado em Bruxelas, com 190 anos de experiência e de conhecimento. Presente em 16 países da Europa e da Ásia, a empresa propõe soluções Vida e Não Vida a milhões de Clientes individuais e empresariais. A Ageas é um dos maiores grupos seguradores europeus, é líder na Bélgica e encontra-se entre os principais líderes de mercado na maioria dos países em que está presente. Tem mais de 40 000 colaboradores (incluindo parcerias não consolidadas) e está presente na Bélgica, Reino Unido, Luxemburgo, França, Itália, Portugal, Turquia, China, Malásia, Índia, Tailândia, Vietname, Laos, Cambodja, Singapura e Filipinas. Está presente em Portugal desde 2005, operando já através de marcas reconhecidas. Primeiro, através da Médis e da Ocidental e, mais recentemente, com a aquisição da AXA Portugal. Com esta aquisição, consegue complementar os seus canais de distribuição já existentes com uma rede de mediadores profissionais e um canal de venda direta.

O tema sobre o qual incidiu o estudo foi proposto pela entidade que acolheu o estágio, a qual pretendia que o mesmo servisse como comparação com estudos realizados anteriormente.

Este estágio pretendeu responder ao seguinte objetivo principal: identificar características que distinguem os clientes que cancelam apólices de seguro dos outros clientes, de modo que, conhecidas as características de um novo indivíduo, se possa prever a que grupo pertence.

O estágio não decorreu em ambiente empresarial, embora tenham sido realizadas reuniões periódicas nas instalações da empresa que proporciona o estágio, a Ageas Portugal, Companhia de Seguros, S.A..

O estágio dividiu-se nas seguintes fases:

- Pesquisa bibliográfica;
- Limpeza e organização da base de dados;
- Identificação de variáveis de interesse;
- Análise descritiva das variáveis mais relevantes;
- Criação de um modelo de regressão;
- Análise crítica dos resultados e escrita do relatório.

A orientação do estágio esteve ao cargo da orientadora responsável do ISEP, a Dra. Sandra Ramos, e do orientador da companhia de seguros, o Dr. Luís Maranhão.

Toda a análise dos dados foi realizada com o recurso à linguagem *R* (versão 3.4.2), por meio do software *RStudio* (versão 1.0.153).

R é uma linguagem e ambiente para computação estatística e gráficos, semelhante à linguagem e ambiente S que foi desenvolvido na Bell Laboratories (anteriormente AT&T, agora Lucent Technologies) por John Chambers e outros colegas. Pode considerar-se R como uma diferente implementação de S. Embora existam algumas diferenças importantes, grande parte do código escrito para S é executado inalterado sob R. R fornece uma grande variedade de técnicas estatísticas (modelação linear e não-linear, testes estatísticos clássicos, análise de séries temporais, classificação, *clustering*, ...) e gráficas e é altamente extensível. A linguagem S é muitas vezes a primeira escolha para pesquisa em metodologia estatística e R fornece, neste sentido, uma via de código aberto. Um dos pontos fortes do R é a facilidade com que se podem produzir representações gráficas de grande qualidade, incluindo símbolos matemáticos e fórmulas, sempre que necessário.

RStudio é um ambiente de desenvolvimento integrado (IDE) para R. Inclui uma consola, um editor de texto com realce de sintaxe que suporta a execução direta de código, bem como ferramentas para representação gráfica, histórico, depuração e gestão do ambiente de trabalho.

1.3 Organização deste relatório

Neste capítulo, começa por fazer-se uma abordagem geral ao tema sobre o qual incide o trabalho, a fidelidade do cliente, e à forma como tem sido trabalhado ao longo do tempo na perspetiva das companhias de seguros. Segue-se uma apresentação do estágio, descrevendo as suas características, objetivos delineados e metodologia implementada. A introdução termina com a descrição detalhada da forma como está organizado este relatório.

No segundo capítulo, dá-se a conhecer toda a metodologia estatística que é utilizada ao longo deste trabalho. É descrito o tipo de modelação a efetuar e o método de seleção de variáveis empregue. São também apresentados variados conceitos, essenciais para boa compreensão, bem como os testes de hipóteses a que se recorreu durante cada procedimento.

Pretende-se, no terceiro capítulo, que o leitor compreenda, em linhas gerais, a forma como se trabalhou a base de dados recebida da companhia de seguros. Descrevem-se, pormenorizadamente, as alterações dimensionais que a base de dados sofreu, até se obter o conjunto de dados final. Procede-se à apresentação das variáveis preditivas e da variável resposta, todas elas presentes no conjunto de dados final.

O capítulo 4 contém uma análise a cada uma das variáveis preditivas mais significativas, com a indicação, sempre que necessária, da categorização que foi efetuada. Antes da apresentação das variáveis, para melhor entendimento das tabelas e gráficos aí exibidos, são descritos, em pormenor, todos os dados referenciados.

No capítulo 5, é justificada a opção pela utilização de um Modelo Linear Generalizado (GLM), mais especificamente um modelo de regressão logística. Segue-se a apresentação dos resultados dos testes de independência efetuados entre as variáveis preditivas e variável resposta. Descreve-se o algoritmo usado na seleção de variáveis a incluir no modelo, bem como o critério de comparação necessário à sua implementação. É apresentado o modelo de quatro variáveis obtido, interpretam-se os coeficientes do modelo e procede-se à sua avaliação. São ainda apresentados modelos alternativos, juntamente com algumas das medidas que permitem a sua avaliação.

No último capítulo apresentam-se as conclusões gerais a partir dos resultados obtidos, referem-se as limitações encontradas na realização do mesmo e discriminam-se possibilidades para trabalho futuro relativo ao tema.

Para o fim, fica a revelação, em apêndice, do código em linguagem R construído e utilizado em todas as fases de manuseamento e apresentação dos dados, bem como a inevitável indicação da bibliografia.

Capítulo 2

Fundamentação Teórica

No capítulo que aqui se inicia, dá-se a conhecer toda a metodologia estatística que é utilizada ao longo deste trabalho. É descrito o tipo de modelação a efetuar e o método de seleção de variáveis empregue. São também apresentados variados conceitos, essenciais para boa compreensão, bem como os testes de hipóteses a que se recorreu durante cada procedimento.

2.1 Teste de independência do χ^2

Quando estamos na presença de dados que resultam de contagens, é recorrente a utilização de tabelas de frequências, em que estão discriminadas todas as classificações dos dados, segundo as suas várias características. Estas classificações são exaustivas e mutuamente exclusivas. A estas tabelas, no caso de termos duas ou mais variáveis, dá-se o nome de tabelas de contingência.

Numa primeira fase, interessa averiguar se as variáveis aleatórias segundo as quais foi feita a classificação cruzada são ou não independentes, ou seja, se é possível que exista alguma associação entre elas. As hipóteses em teste são:

$$H_0 : \text{as variáveis são independentes} \quad \text{vs} \quad H_1 : \text{as variáveis não são independentes}$$

No caso de duas variáveis com r e c categorias, a estatística de teste terá distribuição χ^2 com $(r-1)(c-1)$ graus de liberdade. O valor observado da estatística de teste é dada por

$$\chi^2_{obs} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(o_{ij} - e_{ij})2}{e_{ij}},$$

onde o_{ij} e e_{ij} representam, respetivamente, as frequências observada e esperada na célula correspondente ao cruzamento na tabela das categorias i e j.

Sob a validade da hipótese nula, rejeita-se H_0 se $\chi^2_{obs} \geq \chi^2_{1-\alpha;(r-1)(c-1)}$, ou seja, quando a estatística de teste é maior ou igual que o quantil de probabilidade $1-\alpha$ de uma distribuição χ^2 com $(r-1)(c-1)$ graus de liberdade.

Note-se que o teste não deve ser utilizado se mais de 20% das frequências esperadas, sob hipótese de independência, forem inferiores a 5 ou se alguma delas for mesmo igual a 0.

2.2 Modelos Lineares Generalizados

Os Modelos Lineares Generalizados (GLM) consistem numa generalização do modelo linear, em que se pretende que apresentem a estrutura de uma regressão linear e que partilhem como característica o facto de a variável resposta seguir uma distribuição pertencente a uma família de distribuições com propriedades específicas, a família exponencial. Nelder e Wedderburn [27] propuseram, como uma extensão dos modelos lineares clássicos, uma síntese de vários modelos estatísticos, a que deram o nome de GLM.

Considere-se uma variável aleatória Y, variável resposta, e um vetor $\boldsymbol{x} = (x_1, \ldots, x_k)^\mathsf{T}$ constituído por k variáveis preditivas, também designadas de covariáveis, que se acredita explicar parte da variabilidade inerente a Y. Tanto a variável resposta como as covariáveis consideradas na construção do modelo podem ser de qualquer natureza: contínua, discreta ou dicotómica.

O modelo linear clássico é definido por Turkman e Silva [34] da forma

$$Y = Z\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

em que Z é uma matriz de dimensão $n \times (k+1)$, tendo um vetor unitário na primeira coluna e a transposição da matriz X das covariáveis nas restantes, associada a um vetor $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\mathsf{T}$ de parâmetros e $\boldsymbol{\varepsilon}$ é um vetor de erros aleatórios com distribuição que se supõe ser normal de média 0 e variância dada por $\sigma^2 I$.

$$Z = \begin{pmatrix} 1 & x_{11} & x_{21} & \cdots & x_{k1} \\ 1 & x_{12} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{kn} \end{pmatrix}$$

Sob esta hipótese, $\mu = E(Y|Z) = Z\boldsymbol{\beta}$. Significa isto que o valor esperado da variável resposta é uma função linear das variáveis preditivas.

Os GLM estendem o conceito do modelo linear clássico na medida em que:

1. a variável resposta não tem de seguir distribuição normal, podendo estar distribuída segundo qualquer das distribuições da família exponencial
2. apesar de se manter uma estrutura linear, pode ter-se qualquer função a relacionar o valor esperado com o vetor das covariáveis, desde que seja diferenciável

Definição 2.1 (Família Exponencial) *Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial de dispersão (ou simplesmente família exponencial) se a sua função densidade de probabilidade (f.d.p.) ou função massa de probabilidade (f.m.p.) se puder escrever na forma*

$$f(y|\theta, \phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\},$$

onde θ e ϕ são parâmetros escalares, $a(\cdot)$, $b(\cdot)$ e $c(\cdot, \cdot)$ são funções reais conhecidas.

Um Modelo Linear Generalizado (GLM) é composto por três componentes:

- um **componente aleatório** que especifica a distribuição condicional da variável resposta, Y_i (para a i-ésima de n observações independentes), dados os valores das variáveis explicativas do modelo
- um **componente estrutural ou sistemático** (também conhecido como preditor linear), combinação linear de preditores x_{ij}

$$\eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{2.1}$$

- uma **função de ligação** $g(\cdot)$, monótona e diferenciável, que é aplicada a cada componente de $\mu = E(Y)$ e que o relaciona com o preditor linear

$$g(\mu_i) = \eta_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{2.2}$$

Dado que a função de ligação é invertível, temos ainda que

$$\mu_i = g^{-1}(\eta_i) = g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik}) \tag{2.3}$$

As funções de ligação mais usuais, bem como as suas inversas, são apresentadas na Tabela 2.1.

Tabela 2.1: Funções de ligação

Ligação	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
identidade	μ_i	η_i
logarítmica	$\ln \mu_i$	e^{η_i}
inversa	μ_i^{-1}	η_i^{-1}
quadrática inversa	μ_i^{-2}	$\eta_i^{-1/2}$
raiz quadrada	$\sqrt{\mu_i}$	η_i^2
logit	$\ln \dfrac{\mu_i}{1-\mu_i}$	$\dfrac{1}{1+e^{-\eta_i}}$
probit	$\Phi^{-1}(\mu_i)$	$\Phi(\eta_i)$
log-log	$-\ln[-\ln(\mu_i)]$	$e^{-e^{-\eta_i}}$
complementar log-log	$\ln[-\ln(1-\mu_i)]$	$1-e^{-e^{\eta_i}}$

NOTA: μ_i é o valor esperado da resposta; η_i é o preditor linear; $\Phi(\cdot)$ é a função de distribuição da distribuição normal *standard*.

A existência de variáveis qualitativas leva à necessidade da sua codificação à custa de variáveis *dummy*; para uma variável qualitativa com q categorias, necessitamos de $q-1$ variáveis binárias para a representar, sendo que todas estas variáveis devem estar presentes em Z.

2.3 Regressão Logística

Uma das distribuições de probabilidade que pertence à classe de distribuições da família exponencial é a distribuição binomial. Neste trabalho, dar-se-á especial importância ao caso em que o componente aleatório do GLM segue distribuição binomial e o conjunto de variáveis explicativas podem ser de qualquer natureza. Quando a variável resposta é do tipo dicotómico, deve utilizar-se a regressão logística para modelar a probabilidade de ocorrência de uma das duas realizações das classes desta variável.

Considere-se um vetor de k variáveis explicativas $\boldsymbol{x_i}$ $(i = 1, \ldots, n)$ e as variáveis resposta Y_i que assumem o valor 1 (sucesso - presença de determinada característica) com probabilidade π_i e o valor 0 (insucesso - ausência da característica) com probabilidade $1 - \pi_i$. Assim, Y_i $(i = 1, \cdots, n)$ são variáveis aleatórias com distribuição de Bernoulli com $E(Y_i) = \mu_i = \pi_i$.

A sua função de probabilidade pode ser escrita na forma

$$f(y_i|\pi_i) = \pi_i^{y_i}(1-\pi_i)^{1-y_i}, \quad y_i = 0, 1, \quad i = 1, \ldots, n \tag{2.4}$$

$$= \exp\left\{ y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right) - (-\ln(1-\pi_i)) \right\} \tag{2.5}$$

o que prova a sua pertença à família exponencial, definida anteriormente.

Neste caso, a função de ligação é a função *logit*

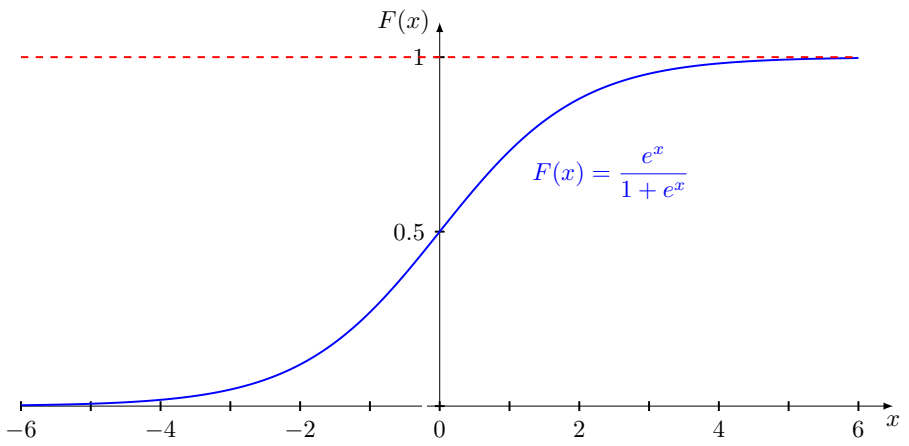
$$\theta_i = logit(\pi_i) = \ln\left(\frac{\pi_i}{1-\pi_i}\right) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} \tag{2.6}$$

em que a probabilidade do sucesso é dada por

$$\pi_i = \frac{1}{1+e^{-\eta_i}} = \frac{e^{\eta_i}}{1+e^{\eta_i}} \tag{2.7}$$

Facilmente se verifica que a função $F : \mathbb{R} \to [0, 1]$, definida por $F(x) = \dfrac{e^x}{1+e^x}$ e representada na Figura 2.1 é uma função de distribuição.

Figura 2.1: Função de distribuição logística



A esta função chama-se função de distribuição logística e, por isso mesmo, o modelo binomial com função de ligação *logit* é conhecido por modelo de regressão logística.

Para obter estimativas para o vetor $\boldsymbol{\beta}$ dos parâmetros do modelo pode ser utilizado o método de máxima verosimilhança. Pretende-se encontrar valores para $(\beta_0, \beta_1, \ldots, \beta_k)$ que maximizem a log-verosimilhança.

A função de verosimilhança é dada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi_i^{y_i} (1 - \pi_i)^{1-y_i} \tag{2.8}$$

A log-verosimilhança é obtida aplicando o logaritmo a (2.8)

$$\ln\left(L(\boldsymbol{\beta})\right) = \sum_{i=1}^{n} y_i \ln \pi_i + (1 - y_i) \ln(1 - \pi_i) \tag{2.9}$$

e usando depois a igualdade (2.7) para efetuar a substituição

$$\ln\left(L(\boldsymbol{\beta})\right) = \sum_{i=1}^{n} \left(y_i \left(\beta_0 + \sum_{i=1}^{n} \beta_k x_{ik} \right) + \ln \left(1 + \exp \left\{ \beta_0 + \sum_{i=1}^{n} \beta_k x_{ik} \right\} \right) \right) \tag{2.10}$$

Há que derivar parcialmente em ordem a cada parâmetro a expressão em (2.10).

Igualando cada uma das equações a 0 são obtidas as estimativas de máxima verosimilhança, $\hat{\boldsymbol{\beta}}$, dos parâmetros do modelo. Na verdade, a obtenção destas estimativas implica o recurso a métodos numéricos já que, geralmente, as equações não têm solução analítica. Nelder e Wedderburn propuseram um algoritmo que opera segundo uma sequência de problemas de mínimos quadrados ponderados (para maior compreensão, ler descrição do algoritmo em [34]).

Obtidas estas estimativas, podemos então calcular estimativas para as probabilidades ajustadas, $\hat{y}_i$.

2.4 Seleção de variáveis preditivas e AIC

Tal como em qualquer modelo linear, também um GLM enfrenta o problema da seleção das variáveis. É um processo que cresce de complexidade com o aumento do número de variáveis, dado o rápido crescimento dos possíveis efeitos e interações. Podem identificar-se dois objetivos concorrentes: o modelo deve ser complexo o suficiente para se ajustar aos dados; por outro lado, o ajuste deve ser suave, evitando um sobre-ajuste dos dados e, preferencialmente, ser de relativamente simples interpretação. Assim, o interesse consiste em reconhecer o modelo mais parcimonioso, isto é, com o menor número de variáveis explicativas, que ofereça uma boa capacidade de interpretação do problema em análise e ainda se ajuste bem aos dados. Um equilíbrio entre três fatores: bom ajustamento, parcimónia e interpretação.

Um algoritmo *stepwise* (passo a passo) de **seleção progressiva** adiciona sequencialmente variáveis. A cada passo, seleciona a que proporciona maior melhoria no ajuste. O processo

termina quando nenhuma outra adição de variáveis se mostra capaz de melhorar o ajuste, de acordo com a significância estatística ou um critério de avaliação da qualidade do ajuste do modelo. Uma variante deste procedimento verifica, a cada passo, se alguma das variáveis previamente adicionadas ainda é necessária. Já um algoritmo de **eliminação regressiva** parte de um modelo complexo e remove variáveis sequencialmente. A cada passo, seleciona aquela cuja sua remoção provoca menor efeito, através do maior *valor-p* num teste de significância ou a menor deterioração do ajuste, segundo determinado critério. O processo termina quando mais nenhuma eliminação origina piores ajustes.

O **Critério de Informação de Akaike** (AIC) é baseado na função log-verosimilhança, com a introdução de um fator de correção como modo de penalização da complexidade do modelo. Um modelo é avaliado pela comparação do ajuste amostral esperado com o verdadeiro ajuste do modelo. Apesar de um modelo mais simples estar mais longe da realidade que um modelo mais complexo, para uma dada amostra o modelo mais simples pode levar a um melhor ajuste, precisamente pelo princípio de parcimónia. Dado um conjunto de potenciais modelos, o melhor é aquele que tende a ter um ajuste amostral mais próximo do verdadeiro ajuste do modelo.

O AIC é calculado da seguinte forma:

$$AIC_q = -2 \ln L(\hat{\boldsymbol{\beta}}) + 2q$$

em que $\ln L(\hat{\boldsymbol{\beta}})$ representa a função log-verosimilhança para um modelo com q variáveis explicativas (logo $q+1$ parâmetros) e $\hat{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança de $\boldsymbol{\beta}$. Quanto menor o AIC, melhor a qualidade de ajuste do modelo.

2.5 Teste da Razão de Verosimilhanças

A estatística de razão de verosimilhanças é mais utilizada para comparar modelos encaixados, ou seja, modelos em que um é submodelo do outro.

Considere-se

$$H_0 : \boldsymbol{\beta_r} = 0 \qquad vs \qquad H_1 : \boldsymbol{\beta_r} \neq 0,$$

sendo $\boldsymbol{\beta_r}$ um subvetor de $\boldsymbol{\beta}$, de dimensão r.

Trata-se de testar um modelo sem as r covariáveis relativas aos parâmetros supostos nulos sob a hipótese H_0. É o que acontece, por exemplo, para cada variável policotómica que, tomando $r+1$ valores distintos, gera a necessidade de construção de r novas variáveis dicotómicas para a sua correta representação. Interessa testar globalmente se os r parâmetros são significativamente diferentes de zero.

A estatística de razão de verosimilhanças, ou estatística de Wilks, é definida por

$$\Lambda = -2 \ln \frac{\max_{H_0} L(\boldsymbol{\beta})}{\max_{H_0 \cup H_1} L(\boldsymbol{\beta})}$$

O teorema de Wilks estabelece que, sob certas condições de regularidade, a estatística Λ tem, sob H_0, uma distribuição assintótica χ^2 sendo o número de graus de liberdade igual à diferença

entre o número de parâmetros a estimar sob $H_0 \cup H_1$ (neste caso p) e o número de parâmetros a estimar sob H_0 (neste caso $p-q$) [34].

A hipótese nula é rejeitada a favor da hipótese alternativa, a um nível de significância α, se o valor observado da estatística Λ for superior ao quantil de probabilidade $1-\alpha$ de um χ^2_q.

2.6 *Odds Ratio*

A *odds ratio* consiste numa medida de associação utilizada em regressão logística para completar o teste à significância das covariáveis. As *odds* de sucesso são definidas pelo quociente entre as probabilidades de sucesso e insucesso, sendo que, para uma probabilidade de sucesso π,

$$odds = \frac{\pi}{1-\pi}.$$

A própria probabilidade de sucesso é função das *odds*, já que se tem $\pi = \dfrac{odds}{odds+1}$.

A *odds ratio*, como o próprio nome indica, consiste na razão entre duas *odds*.

$$\theta = \frac{odds_1}{odds_2} = \frac{\pi_1/(1-\pi_1)}{\pi_2/(1-\pi_2)} \tag{2.11}$$

A *odds ratio* pode tomar qualquer valor não negativo. Para dois eventos independentes com $\pi_1 = \pi_2$, $odds_1 = odds_2$ e $\theta = 1$. Este valor para θ serve assim como base para a comparação. Quando $\theta > 1$, as *odds* de sucesso são maiores para o evento 1. Desta forma, os integrantes deste evento são mais propícios a sucessos que os do evento 2, isto é, $\pi_1 > \pi_2$. Quando $\theta < 1$, acontece exatamente o caso contrário, ou seja, $\pi_1 < \pi_2$.

Substituindo as expressões na Equação 2.11 pelas probabilidades obtidas a partir do modelo de regressão logística (consultar Equação 2.7), pode estimar-se o valor de θ a partir dos coeficientes do modelo de regressão. Para o caso em que se tem uma variável independente dicotómica que assume valores 0 ou 1,

$$\hat{\theta} = \frac{\left(\frac{e^{\hat{\beta_0}+\hat{\beta_1}}}{1+e^{\hat{\beta_0}+\hat{\beta_1}}}\right) / \left(\frac{1}{1+e^{\hat{\beta_0}+\hat{\beta_1}}}\right)}{\left(\frac{e^{\hat{\beta_0}}}{1+e^{\hat{\beta_0}}}\right) / \left(\frac{1}{1+e^{\hat{\beta_0}}}\right)} = \frac{e^{\hat{\beta}_0+\hat{\beta}_1}}{e^{\hat{\beta}_0}} = e^{\hat{\beta}_1} \tag{2.12}$$

Um intervalo de confiança para a *odds ratio* pode ser obtido calculando os extremos de um intervalo de confiança para β_1 e procedendo à exponenciação dos seus extremos. Os extremos deste intervalo terão assim a forma

$$\exp\left[\hat{\beta}_1 \pm z_{1-\alpha/2} \times \widehat{SE}\left(\hat{\beta}_1\right)\right]$$

em que $\widehat{SE}\left(\hat{\beta}_1\right) = \left[\widehat{Var}\left(\hat{\beta}_1\right)\right]^{1/2}$, relação entre erro padrão e variância da estimativa para o parâmetro β_1.

Para maior detalhe pode consultar-se Hosmer e Lemeshow [23].

2.7 Teste de Hosmer e Lemeshow

O **teste de Hosmer-Lemeshow** [23] é muito utilizado em regressão logística com a finalidade de testar a qualidade do ajuste, ou seja, o teste comprova se o modelo proposto pode explicar bem o que se observa. O teste avalia o modelo ajustado através das distâncias entre as probabilidades ajustadas e as probabilidades observadas.

A qualidade do teste é baseada na divisão da amostra segundo as probabilidades ajustadas com base nos valores dos parâmetros estimados pela regressão logística. Os valores ajustados são dispostos de forma crescente e, de seguida, separados em g grupos de tamanho aproximadamente igual. Hosmer e Lemeshow propõe que seja utilizado $g = 10$.

Na literatura há pouca orientação sobre como escolher o número de grupos. As simulações mostradas por Hosmer e Lemeshow foram baseadas no uso de $g > p + 1$, em que p é o número de covariáveis do modelo ajustado. Se as frequências esperadas em alguns dos grupos forem muito pequenas, a estatística do teste de Hosmer-Lemeshow calculada entretanto pode não ser confiável. Neste caso, devemos especificar um número menor de grupos, não se podendo utilizar menos de 3 grupos, pois com $g < 3$ é impossível calcular a estatística do teste.

Antes do cálculo da estatística do teste, é necessário estimar a frequência esperada dentro de cada grupo. Para isso dividimos a variável resposta, que é dicotómica. Para $Y = 1$, a frequência esperada estimada é dada pela soma das probabilidades estimadas de todos os indivíduos dentro daquele grupo. Para $Y = 0$, a frequência esperada estimada é dada pela soma dos complementares das probabilidades estimadas de todos os indivíduos dentro daquele grupo.

Tendo as frequências esperadas, calculamos a estatística de Hosmer e Lemeshow, $\hat{C}$, que é obtida da seguinte forma:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n_k \bar{\pi}_k)^2}{n_k \bar{\pi}_k (1 - \bar{\pi}_k)},$$

em que:

- n_k é o número de indivíduos no k-ésimo grupo
- $\bar{\pi}_k = \sum_{j=1}^{c_k} \frac{m_j \bar{\pi}_j}{n_k}$
- c_k é o número total de combinações de níveis dentro do k-ésimo decil
- $o_k = \sum_{j=1}^{c_k} y_j$ é o número total de respostas dentro do grupo k

Hosmer e Lemeshow mostrou por simulação que a estatística do teste segue, aproximadamente, uma distribuição χ^2 com $g - 2$ graus de liberdade, quando o modelo está especificado corretamente.

2.8 Matriz de Confusão e Curva ROC

Quando se desenvolvem modelos de previsão de resultados, é importante validar os resultados de forma a quantificar o seu poder discriminativo e identificar um procedimento ou método

como bom ou não para determinada análise. No entanto, deve ter-se presente que a simples quantificação de acertos num conjunto de teste não reflete necessariamente o quão eficiente um modelo é, pois essa quantificação dependerá fundamentalmente da qualidade e distribuição dos dados no conjunto de teste.

A chamada **matriz de confusão**, como se pode verificar na Tabela 2.2, trata-se de uma tabela de contingência em que se apresentam, por linhas, os valores previstos e, por colunas, os valores verdadeiros. Consideram-se valores positivos que o modelo julgou positivos como verdadeiros positivos (acerto), valores positivos que o modelo julgou negativos como falsos negativos (erro), valores negativos que o modelo julgou como negativos como verdadeiros negativos (acerto) e valores negativos que o modelo julgou positivos como falsos positivos (erro).

Tabela 2.2: Matriz de confusão

		Valor Verdadeiro (confirmado por análise)	
		positivos	negativos
Valor Previsto (predito pelo teste)	positivos	**VP** Verdadeiro Positivo	**FP** Falso Positivo
	negativos	**FN** Falso Negativo	**VN** Verdadeiro Negativo

Esta matriz serve de base para as medidas seguidamente apresentadas.

Sensibilidade. A proporção de verdadeiros positivos: a capacidade do modelo para predizer corretamente a condição para casos que realmente a têm.

$$P(\hat{Y} = 1|Y = 1) = \frac{\text{ACERTOS POSITIVOS}}{\text{TOTAL DE POSITIVOS}} = \frac{VP}{VP + FN}$$

Especificidade. A proporção de verdadeiros negativos: a capacidade do modelo para predizer corretamente a ausência da condição para casos que realmente não a têm.

$$P(\hat{Y} = 0|Y = 0) = \frac{\text{ACERTOS NEGATIVOS}}{\text{TOTAL DE NEGATIVOS}} = \frac{VN}{VN + FP}$$

Exatidão. A proporção de predições corretas, sem considerar o que é positivo e o que negativo, mas sim o acerto global. É altamente suscetível a desequilíbrios no conjunto de dados, pelo que pode facilmente induzir a uma conclusão errada sobre o desempenho do modelo.

$$P\left(\left[\hat{Y} = 1|Y = 1\right] \cup \left[\hat{Y} = 0|Y = 0\right]\right) = (VP + VN)/(P + N)$$

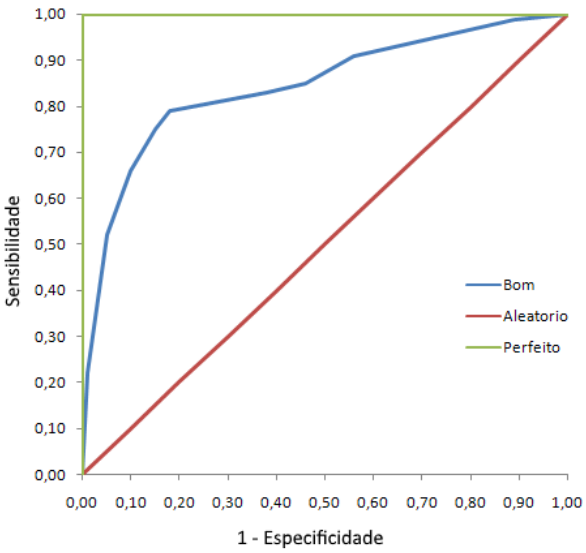
Quando estamos na presença de uma variável resposta binária (1 se o evento se verifica e 0 caso contrário) é necessário escolher uma regra de predição ($\hat{Y} = 0$ *ou* 1), já que $\hat{\pi}$ está entre 0 e 1.

É intuitivo pensar que se o valor de $\hat{\pi}_i$ for grande, $\hat{Y}_i = 1$ e se $\hat{\pi}_i$ for pequeno, $\hat{Y}_i = 0$. Torna-se assim evidente a necessidade de definir um ponto de corte, ou um limiar de decisão, para se classificar e contabilizar o número de predições positivas e negativas. Dada a arbitrariedade que existe para a sua seleção, uma boa prática consiste na comparação do desempenho dos modelos, sob o efeito de diferentes pontos de corte.

Uma forma bastante utilizada para determinar o ponto de corte é através da **Curva ROC** (*Receiver Operating Characteristic Curve*). Esta foi desenvolvida por engenheiros elétricos e engenheiros de sistemas de radar durante a Segunda Guerra Mundial para detetar objetos inimigos em campos de batalha, também conhecida como teoria de deteção de sinais. Há várias décadas que a análise ROC tem sido utilizada em medicina, radiologia, psicologia e outras áreas. Mais recentemente, foi introduzida em áreas como a aprendizagem automática e a prospeção de dados.

Para cada ponto de corte são calculados valores de sensibilidade e especificidade, que podem então ser dispostos num gráfico denominado curva ROC, que apresenta no eixo das ordenadas os valores de sensibilidade e nas abcissas o complementar da especificidade, ou seja, o valor (1-especificidade). Acaba por se tratar de encontrar uma combinação ótima entre sensibilidade e especificidade.

Figura 2.2: Exemplos de curvas ROC



Um classificador perfeito corresponderia a uma linha horizontal no topo do gráfico, o que dificilmente é alcançado. Na prática, curvas consideradas boas estarão entre a linha diagonal e a linha perfeita e quanto maior a distância à linha diagonal, melhor o modelo. Exemplos destes casos são apresentados na Figura 2.2. A linha diagonal indica uma classificação aleatória, um modelo que seleciona aleatoriamente *outputs* positivos ou negativos. Mesmo que a curva esteja localizada abaixo da diagonal, pode ser convertida, bastando para isso inverter os seus *outputs*, o que faz com que a curva também seja invertida.

Tabela 2.3: AUC e poder discriminante do modelo

Área sob a curva ROC (AUC)	Poder discriminante do modelo
0.5	Sem poder discriminativo
$]0.5; 0.7[$	Discriminação fraca
$[0.7; 0.8[$	Discriminação aceitável
$[0.8; 0.9[$	Discriminação boa
≥ 0.9	Discriminação excecional

Uma medida padrão para a comparação de modelos é a área sob a curva (AUC), que pode ser obtida por métodos de integração numérica, como por exemplo, o método dos trapézios.

Teoricamente, quanto maior a AUC, melhor o modelo. Hosmer e Lemeshow [23] apresentam valores indicativos da AUC que podem ser utilizados para classificar o poder discriminante do modelo de regressão e que são dados pela Tabela 2.3.

Podemos dizer que as curvas ROC descrevem a capacidade discriminativa de um teste diagnóstico para um determinado número de pontos de corte. Isto permite evidenciar os valores para os quais existe maior otimização da sensibilidade em função da especificidade.

Capítulo 3

Base de Dados

Pretende-se, com este capítulo, que o leitor compreenda, em linhas gerais, a forma como se trabalhou a base de dados recebida da companhia de seguros, desde a limpeza dos dados à criação de variáveis a partir das já existentes. É possível aqui perceber a forma como, em termos dimensionais, a base de dados foi sofrendo alterações para dela ser possível obter resultados. É feita a apresentação das variáveis presentes no conjunto de dados final, sobre o qual incidirá a análise e modelação, apresentadas em capítulos posteriores.

3.1 Operações na base de dados

Trabalhou-se com base em dois conjuntos de dados: Auto2015 e Auto2015sin.
A necessidade de cruzar informações das apólices com informações dos sinistros originou que se trabalhassem ambos os conjuntos no sentido de acoplar a informação e ter em auto451, para cada apólice, o número de sinistros observados.

Auto2015 é a base de dados principal.
Cada observação contém todas as informações relativas a uma apólice de seguro automóvel.

Tabela 3.1: Operações na base de dados - fase um

Operações (consultar script em $A.1$ e $A.3$)	Dimensão dos dados $[linhas \times colunas]$
Auto2015	$[598793 \times 273]$
Eliminar colunas vazias	$[598793 \times 174]$
Identificar e eliminar colunas com informação desprezável	$[598793 \times 109]$
Eliminar colunas com dados redundantes	$[598793 \times 103]$
Formatação de campos e criação de novos (auto21)	$[598793 \times 116]$
Retirar apólices criadas em 2016 ou canceladas antes de 2015 (auto3)	$[598233 \times 116]$
Retirar apólices anuladas em 2016 (auto4)	$[597800 \times 116]$
Criação de novo campo: *nap*	$[597800 \times 117]$
Filtrar por apólices de ligeiros de passageiros (auto451)	$[402257 \times 117]$

Estão descritas na Tabela 3.1 as primeiras operações realizadas na base de dados.

A formatação de campos refere-se, por exemplo, a agregação de campos, redefinição de formatos de dados (factor, data), categorização de campos, etc.

Foram retiradas da análise as apólices anuladas em 2016 pois, mesmo tendo estado em vigor em 2015, o interesse é apenas avaliar apólices de 2015 *em vigor* vs. *anuladas*. O facto de serem em número reduzido, apenas 433, também faz com que esta eliminação não cause algum dano à análise.

O campo *dur* mede a exposição ao risco, em dias. Consiste no número de dias de 2015 em que cada apólice está exposta ao risco de anulação. Dividindo este valor por 365, é obtido um valor em $[0, 1]$, denominado por *nap*.

Auto2015sin contém dados relativos a sinistros.
Cada observação apresenta as informações referentes a cada sinistro.

Tabela 3.2: Operações na base de dados - fase dois

Operações (consultar script em $A.2$)	Dimensão dos dados $[linhas \times colunas]$
Auto2015sin	$[94725 \times 6]$
Formatação de campos como data e criação de campo *anosin*	$[94725 \times 7]$
Filtrar por apólices com sinistros em 2015 e $reserve \geq 50$	$[47369 \times 7]$
Criação de nova tabela: nrsin	$[43222 \times 2]$

A nova tabela nrsin, decorrente das operações descritas na Tabela 3.2 e contendo apenas 2 campos, associa a cada apólice o número de sinistros observados.

auto451 resulta do cruzamento dos dois conjuntos de dados trabalhados e anteriormente descritos, de forma a termos, para cada apólice de ligeiros de passageiros, o número de sinistros observados.

Tabela 3.3: Operações na base de dados - fase três

Operações (consultar script em $A.1$ e $A.5$)	Dimensão dos dados $[linhas \times colunas]$
Cruzamento das tabelas, fazendo com que auto451 inclua o número de sinistros por apólice	$[402257 \times 118]$
Criação de novo campo: *DL*	$[402257 \times 119]$
Filtrar por campos de interesse (mini)	$[402257 \times 25]$
Filtrar por observações sem *missings* (dados.plus)	$[371252 \times 25]$
Eliminação de campos que não serão variáveis do modelo (dados)	$[371252 \times 20]$

A Tabela 3.3 descreve a fase final das várias operações realizadas na base de dados, até se chegar ao conjunto de dados final.

O novo campo criado - *DL* - acabará por constituir a variável resposta para o modelo a criar e identifica, de uma forma dicotómica, se cada uma das apólices foi ou não anulada no ano de 2015. Note-se que, apólices não renovadas pelo cliente, para efeitos deste trabalho, são também classificadas em *DL* como anuladas.

Após reduzir o conjunto de dados às variáveis com maior potencial significativo, recorrendo à experiência no setor, optou-se por eliminar as observações com dados incompletos, bem como se retirou do conjunto de dados final variáveis que haviam sido utilizadas para a definição de outros campos. Chega-se, deste modo, a um conjunto de dados composto por 371252 observações (apólices), com 19 variáveis preditivas e 1 variável resposta.

3.2 Apresentação das variáveis

Campo identificador:

NPOLIZA	identificador da apólice

Variáveis preditivas:

grp_codbonus	bonificação do contrato
formapa	forma de pagamento
grp_anoini	data inicial
grp_cotpot	prémio apólice
codrisc	classificador de risco de responsabilidade civil
grp_codcomb	combustível
codgara	garagem
grp_jovem	campanha jovem
jubilado	jubilado
zona_pr	zona tarifária
codutil	tipo de utilização
class_dp	classificador de risco de danos próprios
grp_pot	potência
grp_modapac1	pack de coberturas 1
grp_modapac2	pack de coberturas 2
grp_anonas	data de nascimento
grp_anocar	data da carta
grp_estcivi	estado civil
sexo	sexo do cliente

Variável resposta:

DL	anula/não anula

Capítulo 4

Análise dos Dados por Campos

No processo de definição das variáveis preditivas para o modelo, recorreu-se, frequentemente, à categorização de campos da base de dados.

A análise realizada a todos os campos constantes da base de dados foi resumida de acordo com a descrição que se segue.

Para cada campo, a primeira tabela apresenta os seguintes dados:

Campo - categorias do campo em questão

Em vigor - número de apólices em vigor

Anuladas - número de apólices anuladas

Taxa Anulação - peso relativo das apólices anuladas

$$\frac{\text{Anuladas}}{\text{Em vigor} + \text{Anuladas}}$$

Tax100 - Taxa de anulação, em percentagem, relativa à média

NAP - para cada apólice é calculado o nº de dias do ano de 2015 em que esta teve exposição ao risco. Este nº de dias é armazenado na variável 'dur' e está compreendido entre 0 e 365. Nesta coluna, apresenta-se, por categoria, o somatório do quociente entre 'dur' e o total de dias de um ano não bissexto

$$\sum \frac{\text{dur}}{365}$$

Sinistros - número de sinistros ocorridos

Frequência - quociente entre o número de sinistros e o NAP. Traduz uma relação entre os sinistros ocorridos e a exposição da apólice ao risco

$$\frac{\text{Sinistros}}{\text{NAP}}$$

Freq100 - Frequência de sinistralidade, em percentagem, relativa à média

Para as colunas *Em vigor*, *Anuladas*, *NAP*, *Sinistros*, é apresentada a soma na última linha da tabela.

Para as colunas *Taxa anulação*, *Frequência*, são apresentadas as correspondentes médias ponderadas.

A segunda tabela contém, para cada categoria do campo, a taxa de anulação já apresentada na tabela anterior bem como a sua distribuição pelas diferentes causas de anulação: DL, Seg, Comp, Neu (consultar script em A.1). Estas quatro causas de anulação são as categorias resultantes do campo *codcaus*, como descrito de seguida.

Campo: *codcaus*
Novo campo: *grp_ codcaus*

codcaus	grp_codcaus	info
21, 301	DL	Decreto-Lei
11, 12, 13, 14, 16, 17, 31, 32, 33, 37	Seg	Segurado
23, 24, 905, 909	Comp	Companhia
15, 26, 302, 901, 903, 999	Neu	Neutra

A única das causas para anulação que advém da vontade do cliente é a anulação por Decreto-Lei. Por isso mesmo, é esta a causa que define a variável resposta do modelo (DL).

De seguida, apresenta-se um resumo dos resultados da análise a cada uma das variáveis preditivas, com a indicação, sempre que necessária, da categorização que foi efetuada. Resultam desta análise as tabelas e gráficos apresentados (consultar script em $A.3$ e $A.4$). Devido ao elevado número de variáveis preditivas, optou-se por incluir neste resumo apenas as oito covariáveis que viriam a se demonstrar estatisticamente mais significativas no sentido de virem a ser incluídas em algum dos modelos a apresentar.

4.1 Bonificação

Campo: *codbonus* - código identificador da bonificação do contrato
Novo campo: *grp_ codbonus*

codbonus	grp_codbonus	info
$0-6$	M	malus
7	N	null
$8-16$	B	bonus $0-40$
$17-...$	B40	bonus >40

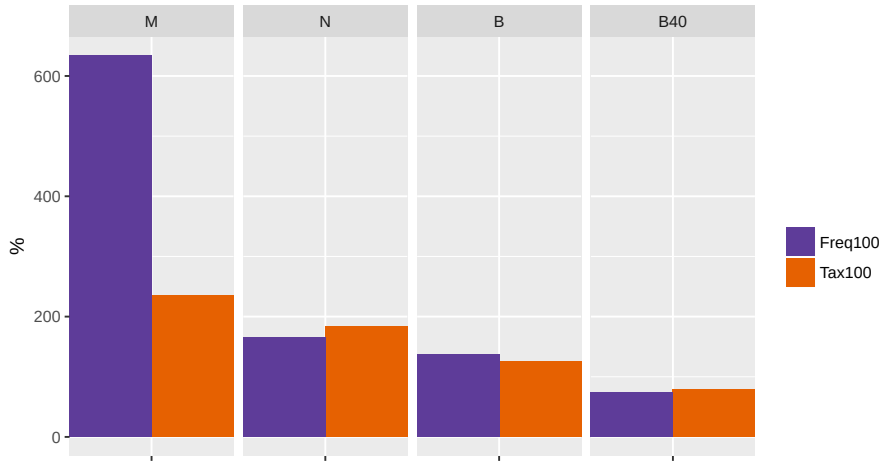
Cada código de bonificação foi agrupado numa de quatro categorias: M, N, B ou B40. Estas distinguem entre apólices com penalização, posição neutra e dois níveis de bonificação.

Tabela 4.1: Bonificação

Bonificação	Em vigor	Anuladas	Taxa Anulação	Tax100	NAP	Sinistros	Frequência	Freq100
M	1 013	732	41.9	235.4	1 318.01	818	62.1	633.7
N	7 432	3 605	32.7	183.7	5 317.86	864	16.2	165.3
B	108 566	31 311	22.4	125.8	110 587.2	14 907	13.5	137.8
B40	188 022	30 571	14	78.7	192 745.8	13 809	7.2	73.5
	305 033	66 219	17.8	100	309 968.87	30 398	9.8	100

É possível notar que tanto Tax100 como Freq100 diminuem em sentido inverso ao da bonificação: quanto maior a bonificação, menores os valores destes indicadores (ver Tabela 4.1 e Figura 4.1). Salienta-se o facto de a frequência de sinistralidade, no caso das apólices com penalização, ser mais de seis vezes superior à média.

Figura 4.1: Bonificação - Tax100 vs Freq100



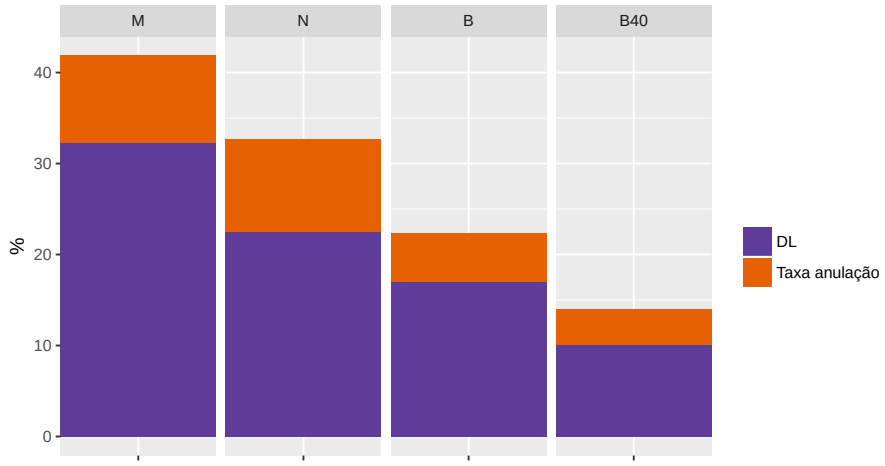
Em qualquer das categorias de bonificação, a anulação por DL é a que tem maior relevo.

Tabela 4.2: Bonificação - anulação e distribuição por causas

Bonificação	Taxa anulação	DL	Seg	Comp	Neu
M	41.9	32.3	4.8	4.4	0.5
N	32.7	22.5	7.2	1.2	1.8
B	22.4	17	3.7	0.9	0.8
B40	14	10.1	3	0.4	0.6
	17.8	13.2	3.4	0.6	0.7

Pode verificar-se que a anulação por DL é tanto maior quanto menor o nível de bonificação (ver Tabela 4.2 e Figura 4.2). As apólices com penalização chegam a ter mais de o triplo de anulação por DL (32.3%) que aquelas em que existe maior bonificação (10.1%).

Figura 4.2: Bonificação - Anulação por DL



4.2 Forma de Pagamento

formapa	info
1	A - anual
2	S - semestral
4	T - trimestral
5	M - mensal

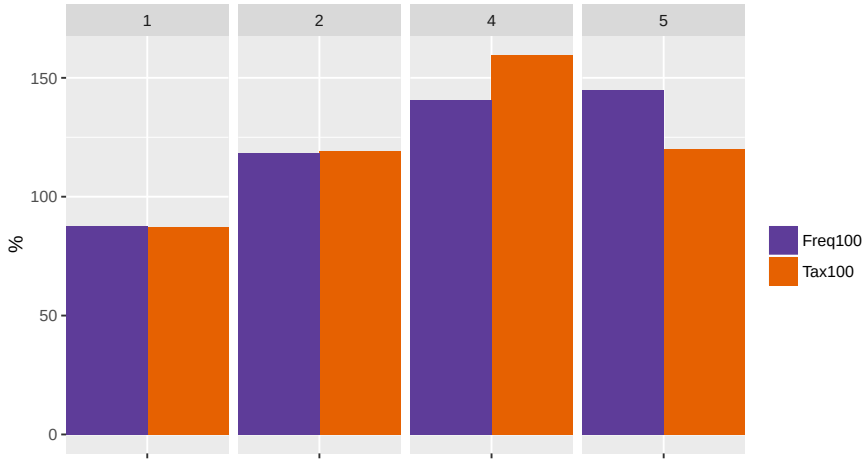
Para a forma de pagamento, não houve necessidade de nenhuma nova codificação da variável.

Tabela 4.3: Forma de Pagamento

Forma de Pagamento	Em vigor	Anuladas	Taxa Anulação	Tax100	NAP	Sinistros	Frequência	Freq100
A	207 498	37 982	15.5	87.1	210 961.75	18 150	8.6	87.8
S	65 921	17 715	21.2	119.1	67 515.72	7 844	11.6	118.4
T	15 541	6 150	28.4	159.6	15 859.37	2 182	13.8	140.8
M	16 073	4 372	21.4	120.2	15 632.02	2 222	14.2	144.9
	305 033	66 219	17.8	100	309 968.86	30 398	9.8	100

Verifica-se uma tendência para Tax100 e Freq100 serem mais elevadas quando a forma de pagamento é mais frequente (ver Tabela 4.3 e Figura 4.3). No caso da taxa de anulação, o maior valor ocorre nos pagamentos trimestrais, enquanto que, para a frequência de sinistralidade, o valor mais elevado é mesmo o dos pagamentos mensais.

Figura 4.3: Forma de Pagamento - Tax100 vs Freq100



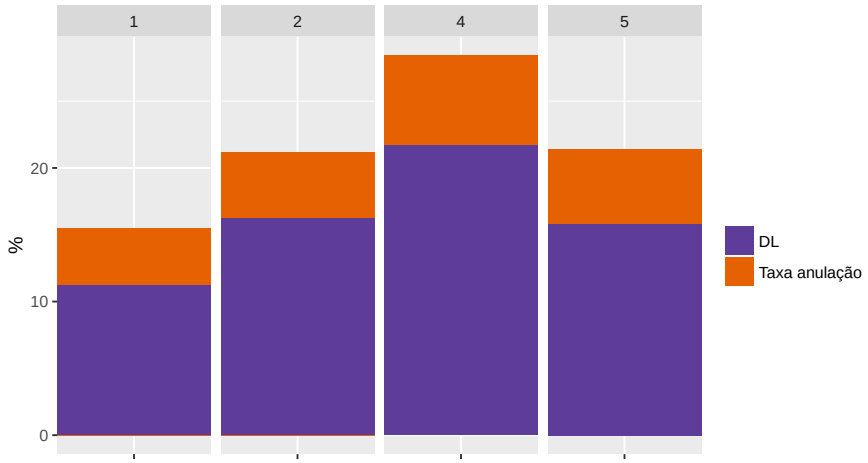
Em qualquer das categorias de forma de pagamento, a anulação por DL é a que tem maior relevo.

Tabela 4.4: Forma de Pagamento - anulação e distribuição por causas

Forma Pagamento	Taxa anulação	DL	Seg	Comp	Neu
A	15.5	11.2	3.2	0.5	0.6
S	21.2	16.2	3.5	0.8	0.7
T	28.4	21.7	5	0.9	0.9
M	21.4	15.8	3.9	0.9	0.8
	17.8	13.2	3.4	0.6	0.7

O valor mais baixo de anulação por DL acontece nas apólices com pagamento anual (11.2%) (ver Tabela 4.4 e Figura 4.4). Em todos os outros casos, pagamentos mais faseados, a anulação é superior, tomando o valor mais elevado no caso dos pagamentos trimestrais (21.7%).

Figura 4.4: Forma de Pagamento - Anulação por DL



4.3 Data Inicial

Campo: *dataini* → *anoini*
Novo campo: *grp_anoini*

anoini	grp_anoini	info
1900 – 2009	A	Antes de 2010
2010 – 2012	B	2010 – 2012
2013 – 2014	C	2013 – 2014
2015	D	2015

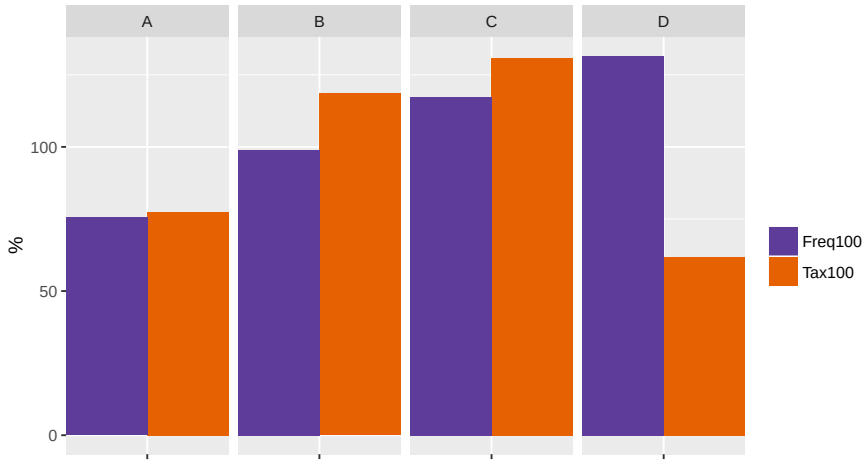
Este novo campo resultou da extração e posterior categorização do ano de início da apólice. Foram agrupadas todas a apólices anteriores a 2010, 2010 a 2012, 2013 a 2014 e aquelas que só foram criadas em 2015.

Tabela 4.5: Data Inicial

Data Inicial	Em vigor	Anuladas	Taxa Anulação	Tax100	NAP	Sinistros	Frequência	Freq100
Antes de 2010	104 714	16 798	13.8	77.5	112 880.69	8 352	7.4	75.5
2010 – 2012	48 664	13 026	21.1	118.5	55 067.28	5 336	9.7	99
2013 – 2014	97 890	29 727	23.3	130.9	111 603.33	12 789	11.5	117.3
2015	53 765	6 668	11	61.8	30 417.56	3 921	12.9	131.6
	305 033	66 219	17.8	100	309 968.86	30 398	9.8	100

Percebe-se que as apólices com maior período de vigência são as que apresentam menores valores nos itens Tax100 e Freq100 (ver Tabela 4.5 e Figura 4.5). O menor valor de Tax100 nas apólices com data inicial em 2015 poder-se-á explicar pelo facto de para estas ainda nem ter decorrido um ano desde que se encontram em vigor.

Figura 4.5: Data Inicial - Tax100 vs Freq100



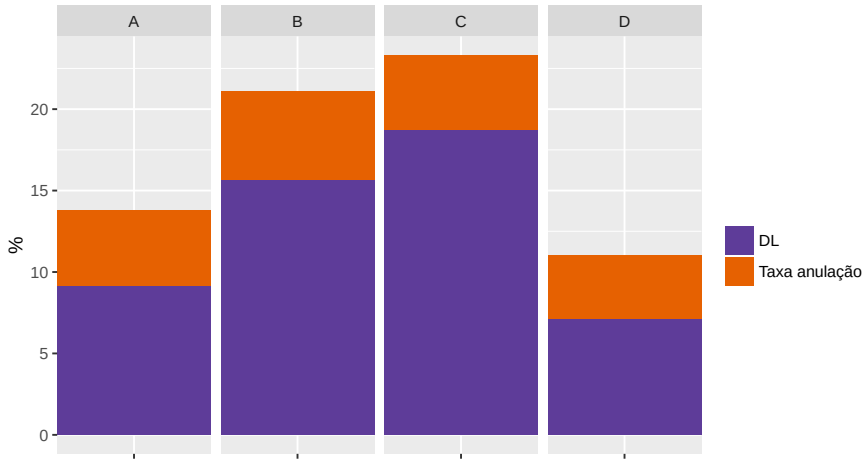
Em qualquer das categorias de data inicial, a anulação por DL é a que tem maior relevo.

Tabela 4.6: Data Inicial - anulação e distribuição por causas

Data Inicial	Taxa anulação	DL	Seg	Comp	Neu
Antes de 2010	13.8	9.1	3.5	0.6	0.7
2010 – 2012	21.1	15.6	3.8	1	0.7
2013 – 2014	23.3	18.7	3.5	0.6	0.4
2015	11	7.1	2.5	0.1	1.3
	17.8	13.2	3.4	0.6	0.7

Os menores valores de anulação por DL verificam-se nas apólices mais recentemente criadas (7.1%) e nas com mais de 5 anos de vigência (9.1%) (ver Tabela 4.6 e Figura 4.6). As que foram criadas entre 2010 e 2014 apresentam todas valores de anulação por DL acima de 15%.

Figura 4.6: Data Inicial - Anulação por DL



4.4 Prémio Apólice

Campo: *cotpot*
Novo campo: *grp_ cotpot*

cotpot	grp_cotpot	info
$[0, 150[$	A	< 150
$[150, 250[$	B	$150 - 250$
$[250, 500[$	C	$250 - 500$
$[500, 10850]$	D	≥ 500

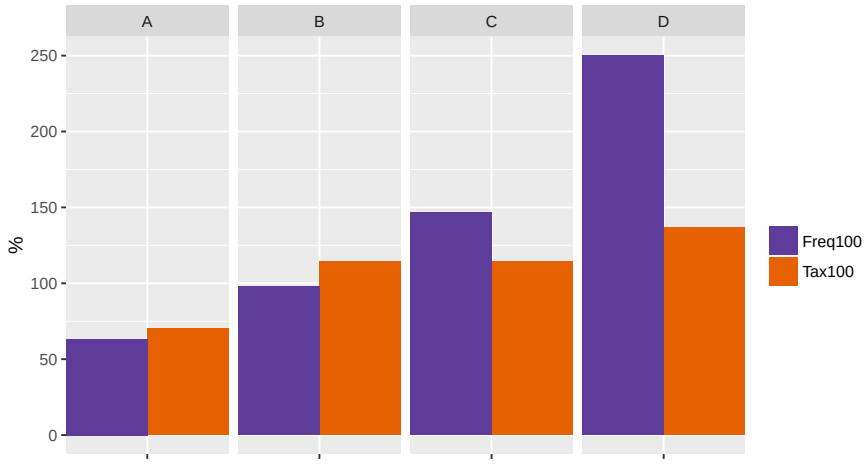
A partir do prémio anual referente a cada apólice existente, foram criadas quatro categorias, de acordo com o valor em questão. Os valores que separam as categorias são 150, 250 e 500 euros.

Tabela 4.7: Prémio Apólice

Prémio Apólice	Em vigor	Anuladas	Taxa Anulação	Tax100	NAP	Sinistros	Frequência	Freq100
< 150	112 716	16 036	12.5	70.2	106 864.82	6 582	6.2	63.3
$150 - 250$	131 701	33 844	20.4	114.6	140 079.68	13 460	9.6	98
$250 - 500$	48 549	12 456	20.4	114.6	50 232.59	7 224	14.4	146.9
≥ 500	12 067	3 883	24.3	136.5	12 791.78	3 132	24.5	250
	305 033	66 219	17.8	100	309 968.87	30 398	9.8	100

Tanto Tax100 como Freq100 crescem com o aumento do prémio da apólice, ou seja, quanto maiores se apresentam os prémios, maiores são também os valores destes indicadores (ver Tabela 4.7 e Figura 4.7). As apólices com prémios anuais superiores ou iguais a 500 euros têm uma taxa de sinistralidade duas vezes e meia superior à média.

Figura 4.7: Prémio Apólice - Tax100 vs Freq100



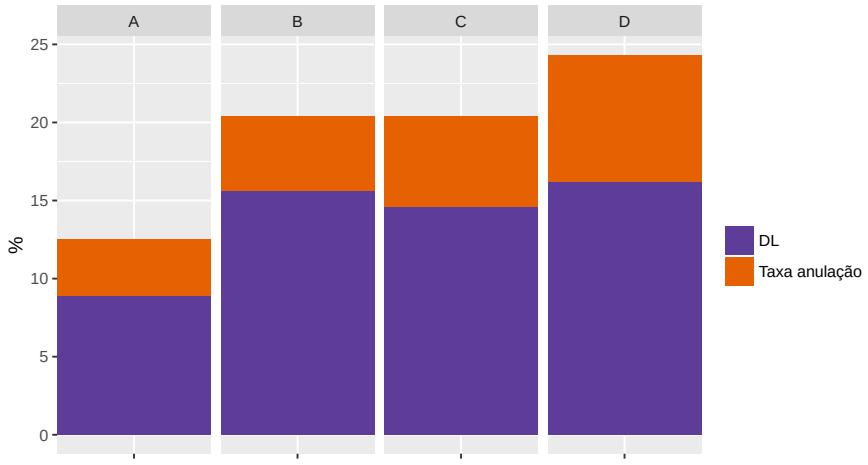
Em qualquer das categorias de prémio da apólice, a anulação por DL é a que tem maior relevo.

Tabela 4.8: Prémio Apólice - anulação e distribuição por causas

Prémio Apólice	Taxa anulação	DL	Seg	Comp	Neu
< 150	12.5	8.9	2.7	0.3	0.5
$150 - 250$	20.4	15.6	3.5	0.6	0.7
$250 - 500$	20.4	14.6	4	0.8	1
≥ 500	24.3	16.2	5.1	1.5	1.5
	17.8	13.2	3.4	0.6	0.7

Pode salientar-se o valor mais baixo de anulação por DL nas apólices com prémios abaixo de 150 euros (8.9%) (ver Tabela 4.8 e Figura 4.8). Todas as outras categorias de prémios têm um valor de anulação por DL muito semelhante (entre 14.6% e 16.2%).

Figura 4.8: Prémio Apólice - Anulação por DL



4.5 Combustível

Campo: *codcomb*
Novo campo: *grp_ codcomb*

codcomb	grp_codcomb	info
	D	incluído em diesel
D	D	diesel
E	E	elétrico
G	G	gasolina
M	M	misto

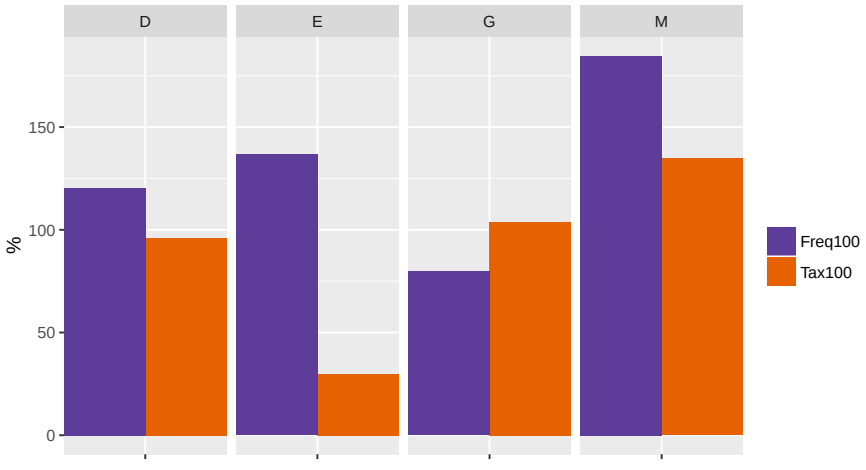
Dado serem em número reduzido, optou-se por incluir as apólices em que não se tem a indicação do tipo de combustível utilizado na categoria com maior número de apólices (diesel).

Tabela 4.9: Combustível

Combustível	Em vigor	Anuladas	Taxa Anulação	Tax100	NAP	Sinistros	Frequência	Freq100
D	155 876	32 260	17.1	96.1	157 198.61	18 513	11.8	120.4
E	18	1	5.3	29.8	14.91	2	13.4	136.7
G	149 120	33 952	18.5	103.9	152 733.26	11 879	7.8	79.6
M	19	6	24	134.8	22.09	4	18.1	184.7
	305 033	66 219	17.8	100	309 968.87	30 398	9.8	100

As categorias relevantes para análise são apenas D (diesel) e G (gasolina), já que as outras dizem respeito a menos de 20 apólices cada (ver Tabela 4.9 e Figura 4.9). Nestas categorias, os valores de Tax100 estão próximos da média, ligeiramente superiores na gasolina. Já quanto a Freq 100, o diesel é superior à média em 20% e a gasolina inferior à média nos mesmos 20%.

Figura 4.9: Combustível - Tax100 vs Freq100



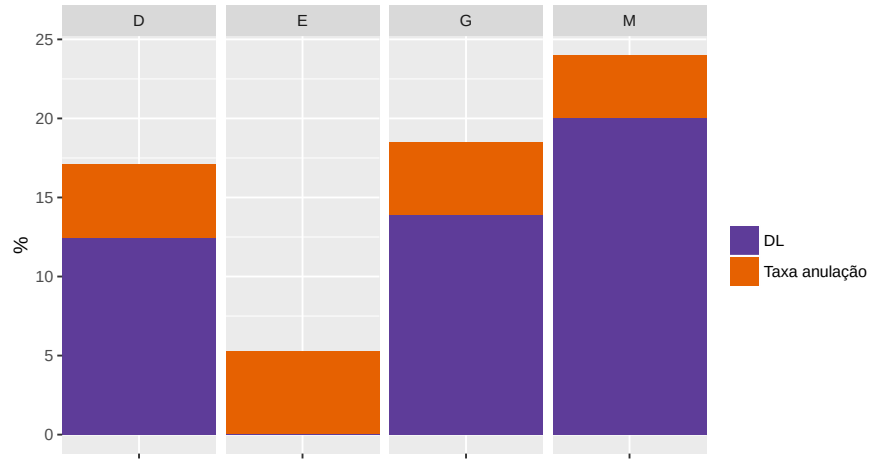
Em qualquer das categorias de combustível consideradas em análise (apenas D e G), a anulação por DL é a que tem maior relevo.

Tabela 4.10: Combustível - anulação e distribuição por causas

Combustível	Taxa anulação	DL	Seg	Comp	Neu
D	17.1	12.4	3.2	0.7	0.8
E	5.3	0	0	5.3	0
G	18.5	13.9	3.6	0.5	0.6
M	24	20	4	0	0
	17.8	13.2	3.4	0.6	0.7

Os valores de anulação por DL nas categorias respeitantes a viaturas movidas a diesel e gasolina são muito semelhantes (12.4% e 13.9%, respetivamente) (ver Tabela 4.10 e Figura 4.10).

Figura 4.10: Combustível - Anulação por DL



4.6 Pack de Coberturas 1

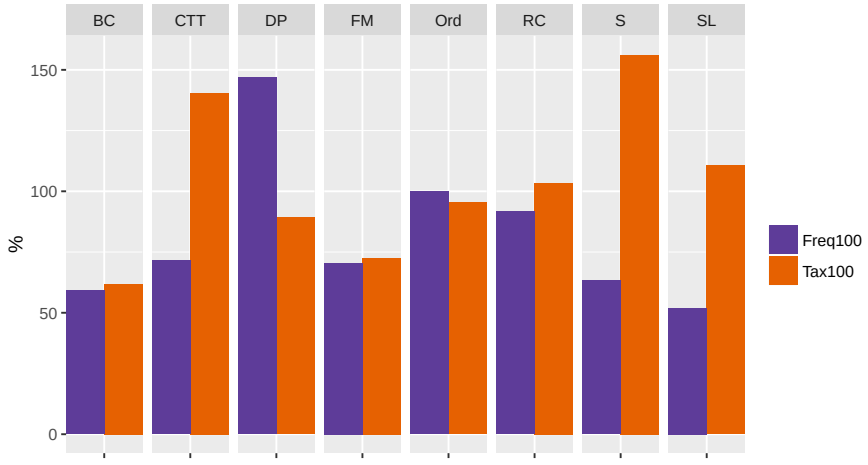
O campo tem como origem variados códigos, respeitantes à lista de coberturas associadas à apólice. Os diversos códigos foram agrupados em oito diferentes categorias: Bons Condutores (BC), CTT, Danos Próprios (DP), Forças Militares (FM), Ordens Profissionais (Ord), Responsabilidade Civil (RC), Simply (S) e Serviço à Lista (SL).

Tabela 4.11: Pack Coberturas 1

Pack de Coberturas 1	Em vigor	Anuladas	Taxa Anulação	Tax100	NAP	Sinistros	Frequência	Freq100
BC	4896	603	11	61.8	5208.95	301	5.8	59.2
CTT	72	24	25	140.4	85.72	6	7	71.4
DP	56011	10587	15.9	89.3	56759.11	8172	14.4	146.9
FM	3458	513	12.9	72.5	3719.01	256	6.9	70.4
Ord	2346	481	17	95.5	2576.17	253	9.8	100
RC	232120	52193	18.4	103.4	235043.12	21050	9	91.8
S	2250	865	27.8	156.2	2313.64	143	6.2	63.3
SL	3880	953	19.7	110.7	4263.15	217	5.1	52
	305033	66219	17.8	100	309968.87	30398	9.8	100

Os maiores valores de Tax100 verificam-se nas categorias S (56.2% acima da média) e CTT (40.4% acima da média), enquanto que encontramos os menores nas categorias BC (38.2% abaixo da média) e FM (27.5% abaixo da média) (ver Tabela 4.11 e Figura 4.11). A única categoria que tem Freq100 acima da média é a que diz respeito às apólices com cobertura de danos próprios (46.9% acima da média). SL, BC e S são as 3 categorias com o valor mais baixo na frequência de sinistralidade (entre 36.7% e 48% abaixo da média).

Figura 4.11: Pack Coberturas 1 - Tax100 vs Freq100



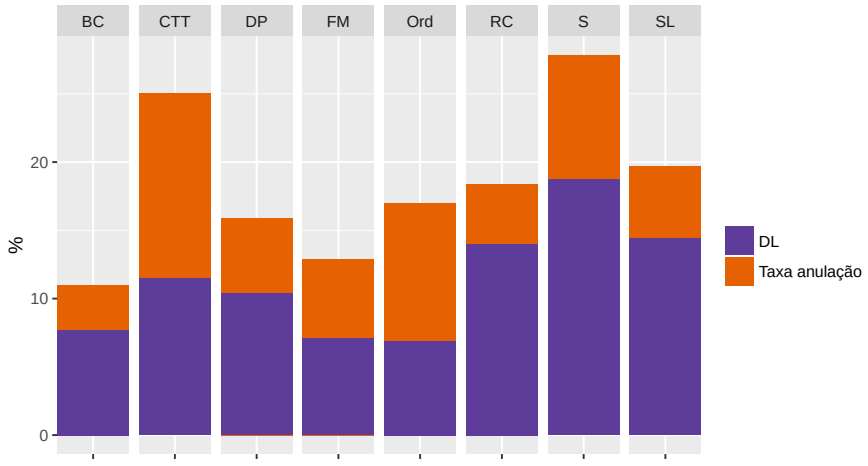
Com a exceção da categoria das ordens profissionais, em que há equilíbrio entre a anulação por DL e uma das outras causas de anulação, a anulação por DL é a predominante ao longo das diferentes categorias de pack de coberturas 1.

Tabela 4.12: Pack Coberturas 1 - anulação e distribuição por causas

Pack de Coberturas 1	Taxa anulação	DL	Seg	Comp	Neu
BC	11	7.7	2.3	0.5	0.5
CTT	25	11.5	6.2	0	7.3
DP	15.9	10.4	3.7	0.8	1
FM	12.9	7.1	4	0.5	1.3
Ord	17	6.9	8	0.4	1.6
RC	18.4	14	3.2	0.6	0.6
S	27.8	18.7	6.6	0.8	1.7
SL	19.7	14.4	3.7	0.6	1
	17.8	13.2	3.4	0.6	0.7

A anulação por DL é mais elevada nas categorias S (18.7%), SL (14.4%) e RC (14%) (ver Tabela 4.12 e Figura 4.12). As categorias Ord (6.9%), FM (7.1%) e BC (7.7%) são as que têm menores valores neste tipo de anulação de apólices.

Figura 4.12: Pack Coberturas 1 - Anulação por DL



4.7 Pack de Coberturas 2

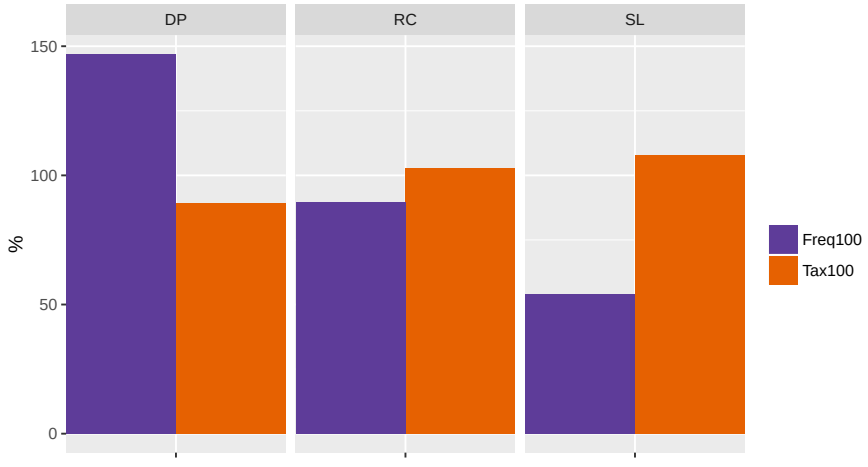
A origem deste campo é exatamente a mesma que a de Pack de Coberturas 1. Neste caso, os diversos códigos foram agrupados apenas em três diferentes categorias: Danos Próprios (DP), Responsabilidade Civil (RC) e Serviço à Lista (SL).

Tabela 4.13: Pack Coberturas 2

Pack de Coberturas 2	Em vigor	Anuladas	Taxa Anulação	Tax100	NAP	Sinistros	Frequência	Freq100
DP	56833	10734	15.9	89.3	57649.31	8280	14.4	146.9
RC	243964	54477	18.3	102.8	247672.62	21873	8.8	89.8
SL	4236	1008	19.2	107.9	4646.93	245	5.3	54.1
	305033	66219	17.8	100	309968.86	30398	9.8	100

Embora não se notem diferenças significativas de Tax100 entre as categorias, a categoria que engloba as apólices com cobertura de danos próprios apresenta um elevado valor de Freq100 (ver Tabela 4.13 e Figura 4.13). Neste aspeto, as apólices com SL apresentam o valor menor, cerca de 50% em relação à média.

Figura 4.13: Pack Coberturas 2 - Tax100 vs Freq100



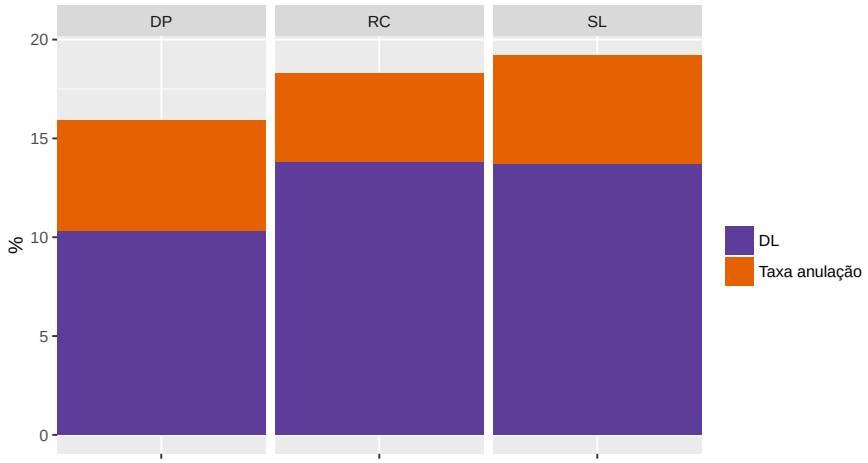
Em qualquer das categorias de pack de coberturas 2, a anulação por DL é a que tem maior relevo.

Tabela 4.14: Pack Coberturas 2 - anulação e distribuição por causas

Pack de Coberturas 2	Taxa anulação	DL	Seg	Comp	Neu
DP	15.9	10.3	3.7	0.8	1
RC	18.3	13.8	3.3	0.6	0.6
SL	19.2	13.7	3.8	0.6	1.1
	17.8	13.2	3.4	0.6	0.7

Os valores de anulação por DL são bastante próximos entre as três categorias (ver Tabela 4.14 e Figura 4.14). Enquanto RC e SL têm perto de 14% das apólices anuladas por DL, a categoria RC tem pouco mais de 10%.

Figura 4.14: Pack Coberturas 2 - Anulação por DL



4.8 Estado Civil

Campo: *estcivi* **Novo campo:** *grp_estcivi*

estcivi	grp_estcivi	info
0	0	sem info
1	1	solteiro
2	2	casado
3	3	divorciado
4	4	separado
5	5	viúvo
6	6	união de facto
7	0	incluído em 0

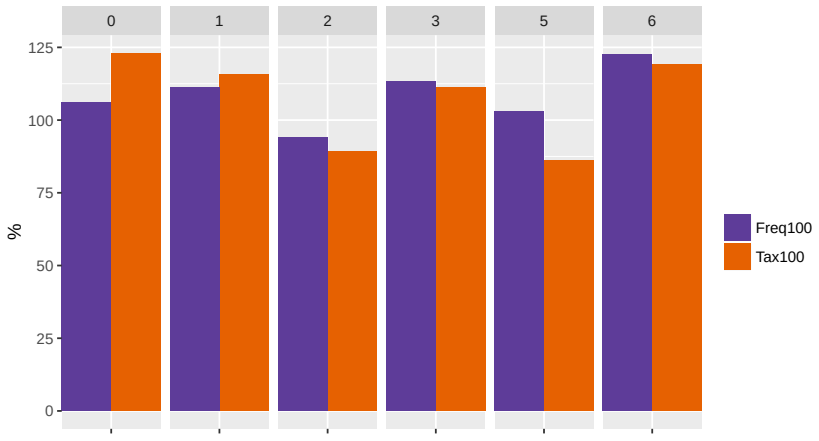
Uma vez que se observou o código 7 de estado civil em apenas uma das apólices, optou-se por juntar esta apólice às que não apresentam nenhuma informação sobre o estado civil. Todas as outras categorias se mantiveram como originalmente.

Tabela 4.15: Estado Civil

Estado Civil	Em vigor	Anuladas	Taxa Anulação	Tax100	NAP	Sinistros	Frequência	Freq100
0	45090	12645	21.9	123	46769.9	4850	10.4	106.1
1	52365	13624	20.6	115.7	53450.39	5818	10.9	111.2
2	188052	35446	15.9	89.3	190057.44	17574	9.2	93.9
3	10840	2669	19.8	111.2	10983.25	1220	11.1	113.3
5	5673	1024	15.3	86	5738.77	581	10.1	103.1
6	3013	811	21.2	119.1	2969.12	355	12	122.4
	305033	66219	17.8	100	309968.87	30398	9.8	100

Excluiu-se da análise as apólices para as quais não há informação sobre o estado civil. Os valores mais elevados de Tax100 ocorrem nas apólices de clientes unidos de facto, solteiros e divorciados (respetivamente 19.1%, 15.7% e 11.2% acima da média) e os valores mais baixos para os clientes viúvos e casados (respetivamente 14% e 10.7% abaixo da média) (ver Tabela 4.15 e Figura 4.15). O mais alto valor de Freq100 ocorre também nos clientes unidos de facto (22.4% acima da média). A única categoria que apresenta um valor abaixo da média, neste item, é a dos casados (6.1%).

Figura 4.15: Estado Civil - Tax100 vs Freq100



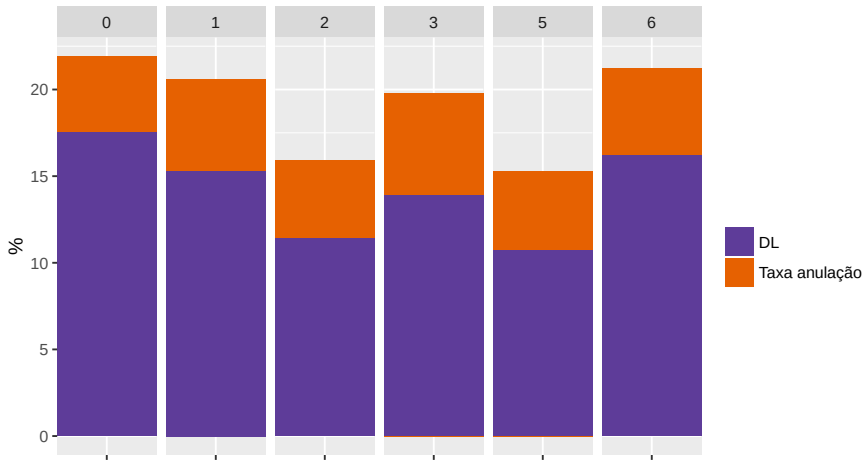
Em qualquer das categorias de estado civil, a anulação por DL é a que tem maior relevo.

Tabela 4.16: Estado Civil - anulação e distribuição por causas

Estado Civil	Taxa anulação	DL	Seg	Comp	Neu
0	21.9	17.5	3.2	0.6	0.5
1	20.6	15.3	3.8	0.8	0.8
2	15.9	11.4	3.3	0.5	0.7
3	19.8	13.9	4.2	0.9	0.8
5	15.3	10.7	3.3	0.7	0.6
6	21.2	16.2	3.7	0.5	0.7
	17.8	13.2	3.4	0.6	0.7

As categorias que incluem apólices de clientes viúvos e casados são as que apresentam menores valores de anulação por DL, ambas perto de 11% (ver Tabela 4.16 e Figura 4.16). O oposto acontece nas categorias dos unidos de facto e solteiros, em que os valores se situam pouco acima de 15%.

Figura 4.16: Estado Civil - Anulação por DL



Capítulo 5

Modelação

Neste capítulo, começa-se por justificar a opção por um Modelo Linear Generalizado (GLM), mais especificamente um modelo de regressão logística. Segue-se a apresentação dos resultados dos testes de independência efetuados entre as variáveis preditivas e variável resposta. Descreve-se a utilização de um algoritmo *stepwise* (passo-a-passo) para seleção de variáveis a incluir no modelo, com base num critério de comparação, o Critério de Informação de Akaike (AIC). É apresentado o modelo de quatro variáveis obtido, faz-se a interpretação dos coeficientes do modelo e procede-se à sua avaliação. Por fim, são ainda apresentados modelos alternativos, juntamente com algumas das medidas que permitem a sua avaliação.

5.1 Tipo de modelo

Desde o primeiro contato com a companhia de seguros que ficou estabelecido que a modelação seria feita por meio de um GLM. Esta opção explica-se pelo à vontade que a empresa apresenta com este tipo de modelação, dado ter sido usado no apoio a outras situações. Quer-se que uma nova ferramenta de apoio possa ser uma ajuda no dia-a-dia da companhia e a simplicidade de um GLM e da regressão logística são uma vantagem, pois mesmo quem não tenha estado na génese da sua construção consegue usar o modelo nas suas análises. Essenciais também, obviamente, as características do problema apresentado: pretende-se prever uma variável resposta dicotómica (anula *vs* não anula) a partir de um conjunto de variáveis preditivas.

5.2 Testes de independência entre variáveis preditivas e a variável resposta

Foram realizados testes de independência entre cada uma das variáveis preditivas e a variável resposta, de forma a perceber se existe uma relação de dependência estatisticamente significativa entre as variáveis.

$$H_0 : \text{a variável preditiva e a variável resposta são independentes} \quad \text{vs} \quad H_1 : \text{as variáveis não são independentes}$$

Tabela 5.1: Resultados dos testes de independência entre as variáveis preditivas e a variável resposta

Teste χ^2 de Pearson: DL vs	χ^2	graus de liberdade	*valor-p*
sexo	1.1309	2	$= 0.5681$
codutil	16.013	3	$= 0.001127$
codgara	21.692	3	$= 7.56e-05^{**}$
zona_pr	28.999	3	$= 2.24e-06$
grp_pot	111.78	3	$< 2.2e-16$
class_dp	143.11	10	$< 2.2e-16$
grp_codcomb	181.97	3	$< 2.2e-16^{**}$
codrisc	321.27	12	$< 2.2e-16^{**}$
grp_modapac2	583.16	2	$< 2.2e-16$
grp_modapac1	1 074	7	$< 2.2e-16$
jubilado	1 112.7	1	$< 2.2e-16^{*}$
grp_jovem	1 178.6	3	$< 2.2e-16$
grp_estcivi	1 913.3	5	$< 2.2e-16$
grp_anocar	2 024.5	3	$< 2.2e-16$
grp_anonas	2 759.1	4	$< 2.2e-16$
formapa	3 029.6	3	$< 2.2e-16$
grp_cotpot	3 121	3	$< 2.2e-16$
grp_codbonus	5 072	3	$< 2.2e-16$
grp_anoini	7 479	3	$< 2.2e-16$

* teste χ^2 de Pearson com correção de continuidade de Yates
** a aproximação χ^2 pode estar incorreta

Nestes testes, a hipótese nula (H_0) consiste na independência entre cada par de variáveis, enquanto a hipótese alternativa (H_1) rejeita essa independência. Para um nível de significância de 0.05, verifica-se na Tabela 5.1 que estes testes permitiram rejeitar a independência entre a variável resposta e todas as variáveis preditivas, com excepção da variável *sexo* ($p = 0.5681 > \alpha = 0.05$). Assim sendo, ao contrário do que acontece com as outras variáveis preditivas, não se pode rejeitar a possibilidade de independência entre *sexo* e *anulação por DL*.

5.3 Seleção do modelo logístico

Tendo em consideração as 19 variáveis preditivas e as 371252 apólices da base de dados, pretende-se ajustar um modelo de regressão logística aos dados, ou seja, estimar π_i, probabilidade da i-ésima apólice de seguro ser anulada pelo cliente, dado o vetor de covariáveis $\boldsymbol{z_i} = (1, x_{i1}, x_{i2}, \ldots, x_{i19})$. Como foi visto em (2.7), π_i pode definir-se como

$$\pi_i = \frac{e^{\boldsymbol{z_i}\boldsymbol{\beta}}}{1 + e^{\boldsymbol{z_i}\boldsymbol{\beta}}} \tag{5.1}$$

com $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_{19})^\top$ e $i = 1, 2, \ldots, 371252$.

Procura-se um modelo simplificado que inclua apenas as variáveis mais importantes para explicar a probabilidade de sucesso π_i.

Para proceder à seleção de variáveis a incluir no modelo final, optou-se por partir do modelo nulo, adicionando uma variável a cada passo. A variável que é adicionada ao modelo é aquela para a qual o Critério de Informação de Akaike (AIC) apresenta o menor valor.

Apresentam-se na Tabela 5.2 as avaliações comparativas dos modelos de regressão logística por meio do AIC, utilizando um algoritmo *stepwise* de seleção progressiva. Note-se que AIC_{i+1} ($i \geq 0$) denota os resultados dos modelos com recurso a $i+1$ variáveis preditivas, após fixar i variáveis no modelo. O * indica o campo para o qual foi obtido o menor valor de AIC.

O menor AIC apresentado é obtido em AIC_{17}, ou seja, por este critério comparativo, o melhor modelo tem 17 variáveis preditivas, excluindo apenas *codgara* e *codutil*.

Efetuando um teste da razão de verosimilhança para modelos encaixados,

$$H_0 : \text{o modelo reduzido ajusta-se melhor que o modelo saturado} \quad \text{vs} \quad H_1 : \text{o modelo reduzido não se ajusta melhor que o modelo saturado}$$

temos um comparativo de ajustamento entre o modelo saturado (construído com todas as variáveis preditivas) e o modelo reduzido a 17 variáveis. Percebe-se que o *valor-p* observado ($p = 0.9214 > \alpha = 0.05$) não nos permite excluir H_0, ou seja, a possibilidade de o modelo reduzido se ajustar melhor que o modelo saturado.

Tabela 5.2: Comparação entre modelos, por meio do AIC, utilizando um algoritmo *stepwise* de seleção progressiva

DL ∼	AIC_1	AIC_2	AIC_3	AIC_4	AIC_5	AIC_6	AIC_7	AIC_8	AIC_9	AIC_{10}
grp_anoini	253550*	–	–	–	–	–	–	–	–	–
grp_codbonus	255913	249763*	–	–	–	–	–	–	–	–
formapa	257732	251179	248219*	–	–	–	–	–	–	–
grp_modapac1	259235	252392	248439	246720*	–	–	–	–	–	–
grp_cotpot	257403	251345	248253	246901	243887*	–	–	–	–	–
grp_anonas	257932	251857	248983	247763	246300	243516*	–	–	–	–
grp_estcivi	258645	252696	249263	247831	246356	243570	243241*	–	–	–
grp_codcomb	260140	253173	249447	247933	246640	243680	243242	242969*	–	–
zona_pr	260283	253553	249764	248210	246715	243675	243312	243033	242726*	–
codrisc	260036	253155	249443	247897	246524	243666	243259	242985	242783	242550*
sexo	260306	253549	249726	248152	246666	243841	243434	243149	242840	242599
grp_jovem	259425	252590	249302	247923	246524	243872	243432	243151	242881	242643
grp_modapac2	259730	252461	248838	247024	246708	243821	243451	243176	242905	242655
grp_anocar	258728	252271	249272	247957	246591	243851	243510	243236	242957	242706
jubilado	259222	252903	249483	248079	246548	243694	243488	243213	242936	242691
grp_pot	260210	253335	249617	248078	246660	243736	243340	243067	242912	242664
class_dp	260192	253367	249641	248107	246684	243776	243388	243113	242946	242707
codgara	260284	253555	249768	248225	246725	243892	243522	243247	242975	242731
codutil	260290	253553	249767	248224	246719	243892	243521	243246	242974	242731

DL ∼	AIC_{11}	AIC_{12}	AIC_{13}	AIC_{14}	AIC_{15}	AIC_{16}	AIC_{17}	AIC_{18}	AIC_{19}
sexo	242418*	–	–	–	–	–	–	–	–
grp_jovem	242471	242339*	–	–	–	–	–	–	–
grp_modapac2	242481	242350	242272*	–	–	–	–	–	–
grp_anocar	242527	242394	242291	242222*	–	–	–	–	–
jubilado	242513	242375	242296	242230	242179*	–	–	–	–
grp_pot	242533	242400	242321	242254	242203	242161*	–	–	–
class_dp	242543	242414	242335	242269	242219	242176	242153*	–	–
codgara	242555	242423	242344	242277	242227	242184	242166	242158*	–
codutil	242556	242423	242344	242277	242227	242184	242166	242159	242163

5.4 Modelo 4 variáveis

Por conveniência e dado que o ganho não era substancial, optou-se por trabalhar com um modelo apenas com 4 variáveis preditivas, sendo elas: *data inicial*, *bonificação*, *forma de pagamento*, *pack de coberturas 2*.

A escolha das variáveis seguiu os resultados do algoritmo *stepwise* de seleção progressiva, apresentado na Tabela 5.2. Foram selecionadas as 4 variáveis indicadas em AIC_4, mas optou-se por trocar *pack de coberturas 1* (variável adicionada na 4ª iteração do algoritmo) por *pack de coberturas 2*, já que tem por base o mesmo campo da base de dados mas apresenta menos categorias, o que facilita o seu seguimento. Apresenta-se, na Tabela 5.3, a descrição do modelo.

Tabela 5.3: *Output* modelo 4 variáveis

	Variável dependente: DL			
	Estimativa	Erro padrão	z	*valor-p*
(Intercept)	-1.95148	0.05811	-33.584	$< 2e-16^{***}$
grp_anoiniB	0.48430	0.01633	29.659	$< 2e-16^{***}$
grp_anoiniC	0.65426	0.01371	47.737	$< 2e-16^{***}$
grp_anoiniD	-0.63493	0.02123	-29.910	$< 2e-16^{***}$
grp_codbonusN	0.07629	0.06069	1.257	$= 0.209$
grp_codbonusB	-0.56094	0.05520	-10.162	$< 2e-16^{***}$
grp_codbonusB40	-1.02321	0.05542	-18.461	$< 2e-16^{***}$
formapa2	0.35090	0.01236	28.381	$< 2e-16^{***}$
formapa4	0.72347	0.01949	37.113	$< 2e-16^{***}$
formapa5	0.39506	0.02210	17.874	$< 2e-16^{***}$
grp_modapac2RC	0.48547	0.01498	32.411	$< 2e-16^{***}$
grp_modapac2SL	0.78574	0.04564	17.217	$< 2e-16^{***}$
		Observações	334 127	
		Log Verosimilhança	$-123\,499.800$	
		AIC	247 023.500	

Signif. códigos: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Na última coluna da Tabela 5.3 pode consultar-se o *valor-p* correspondente a cada coeficiente da sub-classe. Os valores obtidos permitem-nos concluir, para um nível de significância de 0.05, que apenas o coeficiente referente à sub-classe da ausência de bonificação/penalização é que não tem um valor estatisticamente significativo ($p = 0.209 > \alpha = 0.05$). Desta forma, não se pode rejeitar a possibilidade de independência entre esta sub-classe e a *anulação por DL*.

5.5 Interpretação dos coeficientes do modelo

Tomando em consideração cada um dos campos utilizados na definição do modelo de 4 variáveis, são calculados e sujeitos a interpretação as *odds ratio* de cada classe, relativamente a uma classe de referência. É ainda apresentado um intervalo de confiança a 95% para cada um dos valores.

Tabela 5.4: *Odds Ratio* e Intervalos de Confiança para a variável *Data Inicial*

Data Inicial	*Classe de referência*: Antes de 2010		
	Odds Ratio	IC 95%	Variação
grp_anoiniB: 2010 − 2012	1.623	[1.572, 1.676]	↗ 62.3%
grp_anoiniC: 2013 − 2014	1.924	[1.873, 1.976]	↗ 92.4%
grp_anoiniD: 2015	0.530	[0.508, 0.552]	↘ 47.0%

Tomando como classe de referência a das apólices com data inicial anterior a 2010 (Tabela 5.4), verifica-se que a maior probabilidade de anulação ocorre naquelas com data inicial de 2013-2014. Nestas apólices, é 92.4% mais provável a anulação. Também as apólices com data inicial de 2010-2012 apresentam maior probabilidade de anulação (mais 62.3%) que as que temos como referência, ao passo que as apólices de 2015, em virtude de terem sido criadas no ano em análise, são as que apresentam menor probabilidade de anulação.

Tabela 5.5: *Odds Ratio* e Intervalos de Confiança para a variável *Bonificação*

Bonificação	*Classe de referência*: malus		
	Odds Ratio	IC 95%	Variação
grp_codbonusN: 0	1.079	[0.959, 1.216]	↗ 7.9%
grp_codbonusB: bónus 0 − 40	0.571	[0.512, 0.636]	↘ 42.9%
grp_codbonusB40: bónus > 40	0.359	[0.323, 0.401]	↘ 64.1%

Tendo como referência a classe das apólices com penalização (malus) (Tabela 5.5), verifica-se que as apólices que apresentam algum tipo de bonificação têm muito menor probabilidade de anulação, sendo que as que apresentam menor probabilidade são exatamente as que têm maior bonificação (menos 64.1%). As apólices sem penalização ou bonificação revelam uma maior probabilidade de anulação (mais 7.9%), mas é um valor não significativo, já que o próprio IC inclui o valor 1.

Tabela 5.6: *Odds Ratio* e Intervalos de Confiança para a variável *Forma de Pagamento*

Forma de Pagamento		*Classe de referência*: anual	
	Odds Ratio	IC 95%	Variação
formapa2: semestral	1.420	[1.386, 1.455]	↗ 42.0%
formapa4: trimestral	2.062	[1.984, 2.142]	↗ 106.2%
formapa5: mensal	1.484	[1.421, 1.550]	↗ 48.4%

Considerando como classe de referência a das apólices com pagamento anual (Tabela 5.6), deduz-se que estas são as com menor probabilidade de anulação. A maior probabilidade de anulação acontece nas que têm pagamento trimestral (mais 106.2%), enquanto que as que são pagas semestralmente e mensalmente têm, respetivamente, 42% e 48.4% mais probabilidade de anulação que as da classe de referência.

Tabela 5.7: *Odds Ratio* e Intervalos de Confiança para a variável *Pack de Coberturas 2*

Pack de Coberturas 2		*Classe de referência*: DP	
	Odds Ratio	IC 95%	Variação
grp_modapac2RC: RC	1.625	[1.578, 1.673]	↗ 62.5%
grp_modapac2SL: SL	2.194	[2.005, 2.398]	↗ 119.4%

Quanto ao campo *Pack de Coberturas 2* e tendo como referência a classe das apólices com seguro de danos próprios (Tabela 5.7), percebe-se que estas são as que têm menor probabilidade de anulação. Pelo contrário, as apólices com cobertura SL são as que têm maior risco de anulação (mais 119.4%) e, para as com cobertura RC, a sua probabilidade de anulação é superior em 62.5%.

5.6 Avaliação do modelo

5.6.1 Qualidade do ajuste

Efetuou-se um teste χ^2 de Hosmer e Lemeshow,

$$H_0 : \text{o modelo ajusta-se bem aos dados} \quad \text{vs} \quad H_1 : \text{o modelo não se ajusta bem aos dados}$$

obtendo-se resultados que fazem com que se rejeite o bom ajustamento do modelo aos dados ($p < 2.2 \times 10^{-16} < \alpha = 0.05$).

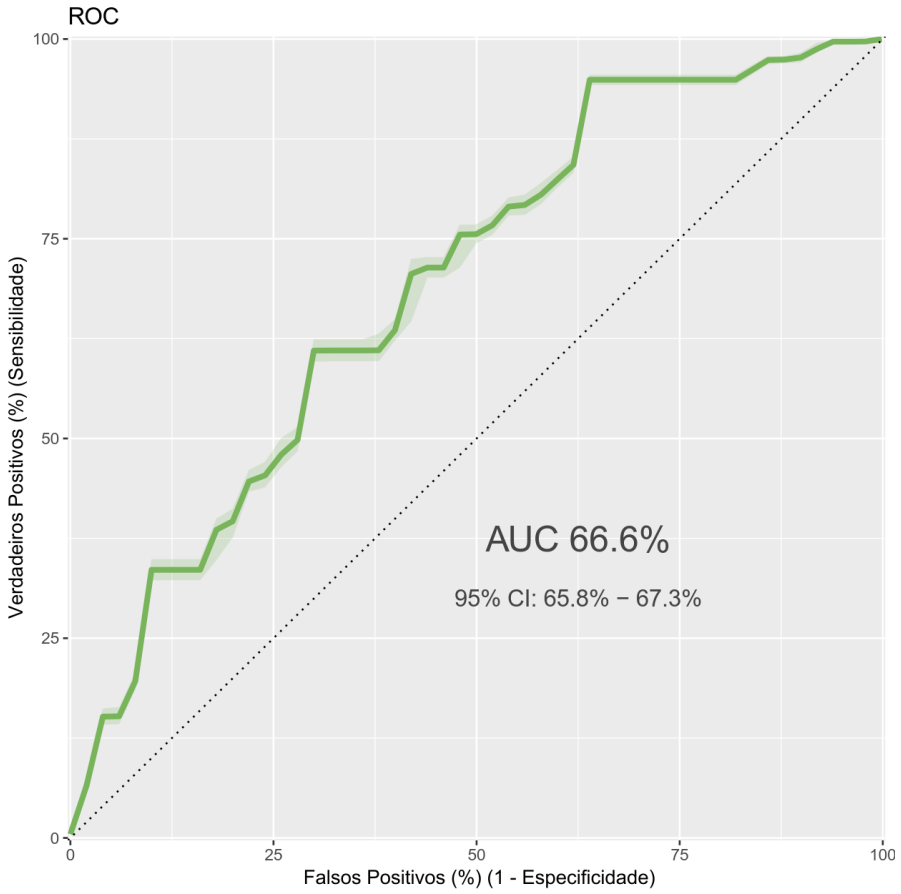
Note-se que, neste teste, os valores ajustados são ordenados por ordem crescente e, em seguida, agrupados em g grupos de dimensões aproximadamente iguais. Na literatura há pouca orientação sobre como escolher o número de grupos, sendo o desempenho do teste sensível a esta escolha. Neste trabalho, considerou-se $g = 10$ [23], o que implicou que, em alguns grupos,

as frequências esperadas tomassem valores pequenos, conduzindo a um valor da estatística de teste pouco confiável.

5.6.2 Desempenho preditivo

No sentido de podermos avaliar a capacidade preditiva do modelo, apresentamos na Figura 5.1 a **curva ROC** correspondente.

Figura 5.1: Curva ROC



A esta curva ROC, corresponde uma $AUC \approx 66.6\%$, com $\text{IC}_{95\%} = [65.8\%, 67.3\%]$, pelo que se tem de classificar o poder discriminativo deste modelo como fraco a aceitável, conforme descrito na Tabela 2.3.

Quis a companhia de seguros definir o **ponto de corte** de forma que lhes permita a identificação de 80% dos clientes que vão anular a apólice, ou seja, estabelecer a sensibilidade como 80%. Para este valor, obtém-se a seguinte matriz de confusão

	$Obs_i = 0$	$Obs_i = 1$
$Pred_i = 0$	13573 (42.1%)	961 (19.7%)
$Pred_i = 1$	18665 (57.9%)	3926 (80.3%)

o que torna evidente a grande percentagem de falsos positivos (57.9%).

Este facto acaba por ter um especial relevo no cálculo da **exatidão** deste modelo, com este ponto de corte. Nestas condições, obtemos uma exatidão de 47.1%.

5.7 Modelos alternativos

Apesar de a seguradora ter escolhido explicitamente o modelo apresentado na Seção 5.4, achou-se por bem investigar a possível existência de outros modelos satisfatórios e efetuar a sua comparação.

Os modelos comparados na Tabela 5.8 são o modelo saturado, o que resulta do algoritmo *stepwise* de seleção progressiva, o que foi escolhido pela empresa (Base), bem como outros modelos com duas a seis variáveis simples. Em relação a estes últimos, com base na Tabela 5.2, partiu-se do modelo com as variáveis de maior significância (B), procurou-se inserir no modelo a melhor interação entre um par de variáveis simples (BI) e, numa terceira abordagem, excluir do modelo a variável simples com menor significância (BID), desde que se obtivesse um modelo diferente de todos os anteriores.

Manteve-se o critério dos 80% de sensibilidade para definição do ponto de corte, pelo que os valores de especificidade e exatidão apresentados são os resultantes do estabelecimento deste critério.

Tabela 5.8: Alternativas ao modelo base

Modelo	AUC	Ponto de corte	Sensibilidade	Especificidade	Exatidão
Saturado	69.1%	0.0998	80.0%	46.8%	51.1%
Stepwise	69.1%	0.0998	80.1%	46.7%	51.1%
Base	66.6%	0.1007	80.3%	42.1%	47.1%
Base + I	67.6%	0.1014	80.0%	44.6%	49.2%
2B	64.3%	0.0835	81.1%	37.9%	43.6%
2BI	65.2%	0.0987	82.2%	37.8%	43.7%
3B	66.1%	0.1080	81.4%	41.7%	46.9%
3BI	67.0%	0.1221	81.9%	42.9%	48.1%
4B	66.5%	0.0970	80.0%	42.8%	47.7%
4BI	67.7%	0.0991	80.1%	44.6%	49.3%
4BID	67.0%	0.1221	81.9%	42.9%	48.1%
5B	68.5%	0.0963	80.0%	45.4%	50.0%
5BI	69.4%	0.1011	80.1%	47.0%	51.3%
5BID	68.7%	0.1071	87.9%	35.0%	42.0%
		0.1072	78.9%	45.8%	50.2%
6B	68.5%	0.1011	80.4%	45.3%	49.9%
6BI	69.5%	0.1055	80.6%	46.3%	50.8%

Nos vários modelos indicados na Tabela 5.8, sempre que foi usada uma interação entre variáveis simples, a que se constituiu como melhor alternativa foi a interação entre a data inicial da apólice e a sua forma de pagamento.

Os resultados obtidos nas várias experiências, levam a que se considere que o modelo 5BI poderia constituir um modelo alternativo ao originalmente selecionado. Este modelo, descrito na Tabela 5.9, tem como variáveis *data inicial*, *bonificação*, *forma de pagamento*, *pack de coberturas 1*, *prémio apólice* e *interação data inicial / forma de pagamento*, sendo que à sua curva ROC corresponde uma $AUC \approx 69.4\%$, com $\text{IC}_{95\%} = [68.7\%, 70.1\%]$.

Tabela 5.9: *Output* modelo 5BI

	Variável dependente: DL			
	Estimativa	Erro padrão	z	*valor-p*
(Intercept)	-2.345356	0.079842	-29.375	$< 2e-16^{***}$
grp_anoiniB	0.512894	0.020269	25.304	$< 2e-16^{***}$
grp_anoiniC	0.653631	0.016923	38.623	$< 2e-16^{***}$
grp_anoiniD	-2.183069	0.056502	-38.637	$< 2e-16^{***}$
grp_codbonusN	0.179442	0.063073	2.845	0.004441^{**}
grp_codbonusB	-0.244503	0.056406	-4.335	$1.46e-05^{***}$
grp_codbonusB40	-0.580040	0.057022	-10.172	$< 2e-16^{***}$
formapa2	0.123415	0.026899	4.588	$4.47e-06^{***}$
formapa4	0.028758	0.053746	0.535	0.592608
formapa5	-0.262945	0.068415	-3.843	0.000121^{***}
grp_modapac1CTT	0.469660	0.342964	1.369	0.170870
grp_modapac1DP	-0.917437	0.059471	-15.427	$< 2e-16^{***}$
grp_modapac1FM	-0.042921	0.084698	-0.507	0.612326
grp_modapac1Ord	-0.594047	0.096000	-6.188	$6.09e-10^{***}$
grp_modapac1RC	0.224565	0.055282	4.062	$4.86e-05^{***}$
grp_modapac1S	0.095489	0.081022	1.179	0.238577
grp_modapac1SL	0.508198	0.069908	7.270	$3.61e-13^{***}$
grp_cotpotB	0.487979	0.013234	36.872	$< 2e-16^{***}$
grp_cotpotC	0.958304	0.021837	43.884	$< 2e-16^{***}$
grp_cotpotD	1.474614	0.034723	42.468	$< 2e-16^{***}$
grp_anoiniB:formapa2	-0.006424	0.038443	-0.167	0.867293
grp_anoiniC:formapa2	0.200974	0.032085	6.264	$3.76e-10^{***}$
grp_anoiniD:formapa2	2.024143	0.067279	30.086	$< 2e-16^{***}$
grp_anoiniB:formapa4	0.102028	0.070225	1.453	0.146259
grp_anoiniC:formapa4	0.411702	0.061336	6.712	$1.92e-11^{***}$
grp_anoiniD:formapa4	2.945705	0.084043	35.050	$< 2e-16^{***}$
grp_anoiniB:formapa5	-0.045170	0.088171	-0.512	0.608439
grp_anoiniC:formapa5	0.347828	0.075672	4.596	$4.30e-06^{***}$
grp_anoiniD:formapa5	3.108215	0.095208	32.647	$< 2e-16^{***}$
		Observações	334 127	
		Log Verosimilhança	$-120\,398.600$	
		AIC	240 855.100	

Signif. códigos: 0 *** 0.001 ** 0.01 * 0.05 . 0.1 1

Capítulo 6

Conclusão

6.1 Conclusões gerais

Partindo de um conjunto de dezanove variáveis preditivas, com o objetivo de previsão da variável resposta identificadora da anulação da apólice por parte do cliente, utilizou-se um algoritmo *stepwise* de seleção de variáveis, conjugado com o critério de informação de Akaike, no sentido de encontrar o conjunto de variáveis a incluir no modelo final. Este procedimento permitiu, por um lado, excluir duas das variáveis preditivas (*garagem* e *tipo de utilização*) e, por outro, ordenar as variáveis em termos da sua contribuição para o modelo.

Apesar de se ter chegado a um modelo de dezassete variáveis preditivas, a companhia de seguros optou pela adoção de um modelo mais parcimonioso, apenas com quatro das dezassete variáveis. São elas: *data inicial*, *bonificação*, *forma de pagamento* e *pack de coberturas 2*. Para o modelo linear generalizado, com regressão logística, construído com estas quatro variáveis preditivas, a curva ROC correspondente tem uma AUC de aproximadamente 66.6%, com $\text{IC}_{95\%} = [65.8\%, 67.3\%]$, pelo que se pode classificar como fraco a aceitável o seu poder discriminativo. Fixando o objetivo de 80% para a sensibilidade, tem-se uma especificidade de aproximadamente 42% e uma exatidão pouco superior a 47%.

A busca por um modelo alternativo, leva a considerar como hipótese viável um modelo construído com recurso a cinco variáveis simples (*data inicial*, *bonificação*, *forma de pagamento*, *pack de coberturas 1*, *prémio apólice*) e a uma das interações entre um par dessas variáveis (*data inicial / forma de pagamento*). A este modelo corresponde uma curva ROC com $AUC = 69.4\%$, $\text{IC}_{95\%} = [68.7\%, 70.1\%]$ e, para o mesmo objetivo de cerca de 80% de sensibilidade, consegue 47% de especificidade e exatidão pouco acima de 51%.

Tendo como meta para o modelo o objetivo do equilíbrio entre bom ajustamento, parcimónia e interpretação, verificou-se que apenas a qualidade de ajustamento deixa a desejar.

6.2 Limitações

Apesar de a base de dados conter grande quantidade de informação relativa a cada apólice, há uma clara consciência de que existem variados fatores não constantes da base de dados que podem ser determinantes para as escolhas que os clientes fazem neste âmbito. De entre estes possíveis fatores, salientam-se:

- o *papel dos mediadores de seguros.* A fidelidade dos clientes está, em grande medida, dependente da relação que estabelecem com os mediadores de seguros, que são, variadas vezes, o único elo entre cliente e companhia de seguros. Está muitas vezes nas mãos destes mediadores, pela confiança estabelecida entre si e o cliente, a decisão sobre a escolha da companhia onde se subscreve o seguro ou, por exemplo, as coberturas abrangidas pelo mesmo. Há assim, por parte da companhia de seguros, que ter noção da importância que podem ter as suas remunerações e/ou comissões nas ações que desenvolvem junto de atuais e possíveis futuros clientes.
- a influência do *enquadramento macroeconómico* na tomada de decisão pela subscrição de seguros e na dimensão da sua carteira
- as *ações de marketing* desenvolvidas pela própria companhia e pelas suas concorrentes. Durante e imediatamente após o desenvolvimento de ações de marketing, percebem-se reações do mercado a estas iniciativas.

6.3 Trabalho futuro

Não se pretende que este trabalho constitua um ponto final na abordagem ao problema colocado. É mais um passo no sentido da compreensão de um fenómeno e espera-se que seja uma ajuda para quem o decida abraçar noutro momento.

Tendo este trabalho sido realizado a partir de dados de apólices em vigor em 2015, seria interessante efetuar previsões para a anulação em 2017, com os dados mais recentes de 2016, a partir do modelo escolhido e talvez de um dos modelos alternativos. Mais tarde, verificar e comparar o desempenho preditivo de cada um.

Dado que ficou definida à partida a modelação por meio de um GLM, seria uma hipótese para trabalho futuro a abordagem ao mesmo problema com outros tipos de métodos.

Apêndice A

Código R

A.1 Primeiros passos

Importação de dados e criação de 'Auto2015.RData'

```
Auto2015 <- read.csv("C:/#DISPE_AXA/dados origem/Auto2015.csv")
save.image("Auto2015.RData")
load("Auto2015.RData") #usar para chamar os dados
```

Limpeza de campos

```
## Eliminar colunas vazias
vazias <- c('datausin','grupoeco','codbon','codcart','codigor','codprem','
    codris0','codris1','codris2','numpres','franq_rc','act_cap0','act_cap_','
    agr_ida0','agr_idad','ancons2','ano_tari','cap2','capdan2','capfar1','
    capfar2','captsf1','captsf2','cartve0','cat1','cil2','codava','codcob0','
    codcob1','codcobt','codmoda','codsitu1','cod_cob2','cod_mod_','cod_pa_0',
    'cod_pa_m','cod_pre0','data_ma0','desc_rc','despt','flag_bon','freqdan2',
    'lincapt','lotaca2','matr2','minvpta','pb2','pdmp02','pre_sim2','p_pa0','
    p_pas','p_ri0','p_ri1','p_ri2','p_ri3','p_ri4','p_ris','p_sim2','txagr0',
    'txagrr','txagrs','txgera','tx_ag0','tx_ag1','tx_ag2','tx_agr','tx_sin','
    NGRUPO','codcond','codgeno','codposi','codsusp','cod_enca','datafim','
    deduca','dtfimc','dtinici','duraca','ind_reca','matricu1','modalid','
    promocao','px_prest','tipmovi','tipovei','totsino','tx_cor_s','x','cod_
    com2','cod_com3','cod_comi','data_in2','data_ter','funca','modali','
    numsegu','num_ints','txdesc','tx_cor_v')
auto1 <- Auto2015[,! names(Auto2015) %in% vazias]

## Eliminar colunas desprezadas
despr <- c('cap_ext','datapro','dataemi','comp','cdcamp','dtjovem','d_propos
    ','d_proaxa','d_tripro','agent','cap1','capita0','caprc','cartver','
    freqdan1','matr1','pdmp01','tari','znactu','informac','freqrc','areains',
    'balca','lotaca1','premcob','prsimpl2','mod_auto','codveic','codcircu','
    coduso','cve_veic','protocol','nacional','codext','excl_bon','excl_mal','
    ind_cose','modelo','pass_car','unrisc','cod_fra2','cve_tran','veicprin','
    datater','datate0','DNI','dini_anu','dfim_anu','d_grava','anuidade','cap_
    dp','fcobra','escrit','cod_p_e','mat_peg','apoli1','codinfo','ligeiro','
    sverde','matricul','side_car','numluga','capita','capitac','ind_ints')
auto2 <- auto1[,! names(auto1) %in% despr]
```

```
## Eliminar colunas redundantes
redund <- c('anocons','ancons1','cat2','formaco','codcate','causa_anu')
auto21 <- auto2[,! names(auto2) %in% redund]
```

Agregação de campos

```
### Completar colunas nr sinistros
k0 <- as.numeric(levels(auto21$ano0))[auto21$ano0] # ?factor ou help(factor)
k1 <- as.numeric(levels(auto21$ano1))[auto21$ano1]
k2 <- as.numeric(levels(auto21$ano2))[auto21$ano2]
kt <- as.numeric(levels(auto21$anot))[auto21$anot]
auto21[is.na(auto21$nsin0),'nsin0'] <- k0[is.na(auto21$nsin0)]
auto21[is.na(auto21$nsin1),'nsin1'] <- k1[is.na(auto21$nsin1)]
auto21[is.na(auto21$nsin2),'nsin2'] <- k2[is.na(auto21$nsin2)]
auto21[is.na(auto21$nsint),'nsint'] <- kt[is.na(auto21$nsint)]
rm(k0,k1,k2,kt)

### completar coluna zona tarifária
auto21[is.na(auto21$zona_pr),'zona_pr'] <- auto21[is.na(auto21$zona_pr),'
   zonatar']

#eliminar colunas absorvidas por outras
absor <- c('ano0','ano1','ano2','anot','zonatar')
auto21 <- auto21[,! names(auto21) %in% absor]

rm(vazias,despr,redund,absor) #remove as listas de campos apagados
```

Formatação de campos como factor

```
auto21$formapa <- as.factor(auto21$formapa) # forma pagamento
auto21$codsitu <- as.factor(auto21$codsitu) # situação apólice
auto21$codrisc <- as.factor(auto21$codrisc) # codrisc
auto21$codgara <- as.factor(auto21$codgara) # codgara
auto21$zona_pr <- as.factor(auto21$zona_pr) # zona_pr
auto21$codutil <- as.factor(auto21$codutil) # codutil
auto21$cod_cate <- as.factor(auto21$cod_cate) # categoria
auto21$estcivi <- as.factor(auto21$estcivi) # estado civil
auto21$sexo <- as.factor(auto21$sexo) # sexo
auto21$sexo_ch <- as.factor(auto21$sexo_ch) # sexocondutor habitual
```

Formatação de campos como Date

```
library(lubridate)
auto21$datamatr <- as.Date(ymd(auto21$datamatr))
auto21$dataini <- as.Date(ymd(auto21$dataini))
auto21$dataven <- as.Date(ymd(auto21$dataven))
auto21$dataumov <- as.Date(ymd(auto21$dataumov))
auto21$datault <- as.Date(ymd(auto21$datault))
auto21$datanas <- as.Date(ymd(auto21$datanas))
auto21$datacar <- as.Date(ymd(auto21$datacar))
```

Criação de novos campos

1. anos

```
auto21$anoini <- year(ymd(auto21$dataini)) # anoini - ano de
auto21$anoult <- year(ymd(auto21$datault)) # anoult - ano de anulação
auto21$anonas <- year(ymd(auto21$datanas)) # anonas - ano de nascimento
auto21$anocar <- year(ymd(auto21$datacar)) # anocar - ano de
```

2. status - situação da apólice

```
#apólice anulada 0, apólice em vigor 1
auto21$status <- ifelse(auto21$codsitu == 5,0,1)
```

3. grp_codcomb - combustível

```
#valores em falta passam a Diesel (maior freq)
auto21$grp_codcomb <- auto21$codcomb
levels(auto21$grp_codcomb)[levels(auto21$grp_codcomb) == ''] <- 'D'

# ----------------------------------TESTE-----------------------------------
# table(auto21$codcomb)
# table(auto21$grp_codcomb)
```

4. grp_jovem - jovem

```
#valores em falta passam a N (maior freq)
auto21$grp_jovem <- auto21$jovem
levels(auto21$grp_jovem) <- c('N','I','J','N','S')

# ----------------------------------TESTE-----------------------------------
# table(auto21$jovem)
# table(auto21$grp_jovem)
```

5. grp_estcivi - estado civil

```
#valores 7 passam a 0
auto21$grp_estcivi <- auto21$estcivi
levels(auto21$grp_estcivi)[levels(auto21$grp_estcivi) == '7'] <- '0'

# ----------------------------------TESTE-----------------------------------
# table(auto21$estcivi)
# table(auto21$grp_estcivi)
```

6. dur - exposição ao risco, em dias

```
source("nap.R")
auto21$dur <- nap(auto21$codsitu,auto21$dataini,auto21$datault)

# ----------------------------------TESTE-----------------------------------
# abc <- auto21[! is.na(auto21$dur), #extrai ini=2016
#               c('codsitu','status','dataini','anoini','datault',
#               'anoult','datanas','anonas','datacar','anocar','dur')]
# neg <- abc[abc$dur < 0,] #anuladas antes de 2015
# plus <- abc[abc$dur > 365,] #anuladas em 2016
```

Agrupar por categorias

```
auto21$grp_codbonus <- cut(auto21$codbonus, breaks = c(0,7,8,17,27), labels = c('M','N','B','B40'), right = FALSE)
auto21$grp_anoini <- cut(auto21$anoini, breaks = c(1900,2010,2013,2015,2017), labels = c('A','B','C','D'), right = FALSE)
auto21$grp_cotpot <- cut(auto21$cotpot, breaks = c(0,150,250,500,10850), labels = c('A','B','C','D'), right = FALSE)
auto21$grp_pot <- cut(auto21$pot, breaks = c(0,80,100,150,1500), labels = c('A','B','C','D'), right = FALSE)
auto21$grp_anonas <- cut(auto21$anonas, breaks = c(1799,1960,1970,1980,1990,2016), labels = c('A','B','C','D','E'), right = FALSE)
auto21$grp_anocar <- cut(auto21$anocar, breaks = c(1899,2007,2011,2013,3000), labels = c('A','B','C','D'), right = FALSE)

# Categorias 'codcaus'
# Decreto de Lei [DL] : 21, 301
# Segurado [Seg] : 11, 12, 13, 14, 16, 17, 31, 32, 33, 37
# Companhia [Comp] : 23, 24, 905, 909
# Neutra [Neu] : 15, 26, 302, 901, 903, 999
# 302 sem enquadramento!!
declei <- c(21, 301)
segu <- c(11, 12, 13, 14, 16, 17, 31, 32, 33, 37)
compa <- c(23, 24, 905, 909)
neut <- c(15, 26, 302, 901, 903, 999)
auto21$grp_codcaus[auto21$codcaus %in% declei] <- 'DL'
auto21$grp_codcaus[auto21$codcaus %in% segu] <- 'Seg'
auto21$grp_codcaus[auto21$codcaus %in% compa] <- 'Comp'
auto21$grp_codcaus[auto21$codcaus %in% neut] <- 'Neu'
auto21$grp_codcaus <- as.factor(auto21$grp_codcaus)
rm(declei, segu, compa, neut)

# Categorias 'modapac1'
# Simply [S] : 0
# Bons condutores [BC] : 14, 32
# Forças militares [FM] : 15, 16, 40, 91
# Ordens profissionais [Ord] : 21, 22, 23, 24, 92
# Danos Próprios [DP] : 3, 4, 6, 58, 59, DD, DF, DG, ED, EF, EG
# Responsabilidade Civil [RC] : 5, 25, DE, N9, EE, U1
# CTT [CTT] : 7, 8, 9, 90
# SL [SL] : 96, 99
# 25 sem enquadramento!!

s1 <- 0
bc1 <- c(14, 32)
fm1 <- c(15, 16, 40, 91)
ord1 <- c(21, 22, 23, 24, 92)
dp1 <- c(3, 4, 6, 58, 59, 'DD', 'DF', 'DG', 'ED', 'EF', 'EG')
rc1 <- c(5, 25, 'DE', 'N9', 'EE', 'U1')
ctt1 <- c(7, 8, 9, 90)
sl1 <- c(96, 99)
auto21$grp_modapac1[auto21$modapac %in% s1] <- 'S'
auto21$grp_modapac1[auto21$modapac %in% bc1] <- 'BC'
auto21$grp_modapac1[auto21$modapac %in% fm1] <- 'FM'
auto21$grp_modapac1[auto21$modapac %in% ord1] <- 'Ord'
auto21$grp_modapac1[auto21$modapac %in% dp1] <- 'DP'
```

```
auto21$grp_modapac1[auto21$modapac %in% rc1] <- 'RC'
auto21$grp_modapac1[auto21$modapac %in% ctt1] <- 'CTT'
auto21$grp_modapac1[auto21$modapac %in% sl1] <- 'SL'
auto21$grp_modapac1 <- as.factor(auto21$grp_modapac1)
rm(s1, bc1, fm1, ord1, dp1, rc1, ctt1, sl1)

# Categorias 'modapac2'
# Danos Próprios [DP] : 3, 4, 6, 8, 9, 16, 23, 24, 58, 59, DD, DF, DG, ED, EF, EG
# Responsabilidade Civil [RC] : 0, 5, 7, 14, 15, 21, 22, 25, 32, 40, DE, N9, EE, U1
# SL [SL] : 90, 91, 92, 96, 99
# 25 sem enquadramento!!

dp2 <- c(3, 4, 6, 8, 9, 16, 23, 24, 58, 59, 'DD', 'DF', 'DG', 'ED', 'EF', 'EG')
rc2 <- c(0, 5, 7, 14, 15, 21, 22, 25, 32, 40, 'DE', 'N9', 'EE', 'U1')
sl2 <- c(90, 91, 92, 96, 99)
auto21$grp_modapac2[auto21$modapac %in% dp2] <- 'DP'
auto21$grp_modapac2[auto21$modapac %in% rc2] <- 'RC'
auto21$grp_modapac2[auto21$modapac %in% sl2] <- 'SL'
auto21$grp_modapac2 <- as.factor(auto21$grp_modapac2)
rm(dp2, rc2, sl2)

# ----------------------------------TESTE-----------------------------------
# a <- auto21[,c('grp_codbonus','grp_anoini','grp_cotpot','grp_pot',
#                'grp_anonas','grp_anocar','codcaus','grp_codcaus')]
```

Aplicação de Filtros

```
#retirar apólices com início em 2016 ou canceladas antes de 2015
auto3 <- auto21[! is.na(auto21$dur) & auto21$dur >= 0,]
#retirar apólices anuladas em 2016
auto4 <- auto3[auto3$dur <= 365,]
#Novo campo: nap - dur em [0,1]
auto4$nap <- auto4$dur / 365

# ----------------------------------TESTE-----------------------------------
# def <- auto4[, c('codsitu','status','dataini','datault','anoult','dur','nap')]
```

```
auto451 <- auto4[auto4$cod_cate == 451,] # ligeiros passageiros
auto451 <- droplevels(auto451) # drop unused factors after filtering
# apólices de ligeiros passageiros anuladas em 2015
auto451.2015 <- auto4[auto4$cod_cate == 451 & !is.na(auto4$anoult) & auto4$anoult == 2015,]
auto451.2015 <- droplevels(auto451.2015)
```

Criação de 'Auto451.RData'

```
rm(Auto2015,auto1,auto2,auto3)
save.image("Auto451.RData")
```

Cruzamento de tabelas

```
# Load dos dados gerados em 'axa_script.R' e 'axa_sin.R'
load("Auto451.RData")
load("Nr_sin2015.RData")
library(dplyr)
library(ggplot2)

# Cruzamento das tabelas 'auto451' e 'nrsin', fazendo com que 'auto451'
# inclua o nr de sinistros por apólice
auto451 <- left_join(auto451,nrsin) #dplyr
# variável resposta: binomial anulação por DL
declei <- c(21, 301)
auto451$DL <- ifelse(auto451$codcaus %in% declei,1,0)
rm(declei)

# ----------------------------------TESTE---------------------------------
#a <- auto451[,c('NPOLIZA','status','anoult','dur','nr_sin','DL')]
```

Dados para a regressão

```
vars <- c('NPOLIZA','grp_codbonus','formapa','grp_anoini','grp_cotpot','
    codrisc','grp_codcomb','codgara','grp_jovem','jubilado','zona_pr','
    codutil','class_dp','grp_pot','grp_modapac1','grp_modapac2','grp_anonas',
    'grp_anocar','grp_estcivi','sexo','DL', 'status', 'nap', 'nr_sin', 'grp_
    codcaus')
mini <- auto451[, names(auto451) %in% vars]
rm(vars)
```

Criação de 'mini.RData'

```
rm(autosin, nrsin, sin2015,auto21, auto4, auto451, auto451.2015, nap)
save.image("mini.RData")
```

A.2 Sinistros

Importação de dados e criação de 'Auto2015sin.RData'

```
Auto2015sin <- read.csv("C:/#DISPE_AXA/dados␣origem/Auto2015sin.csv")
load("Auto2015sin.RData") #usar para chamar os dados
```

```
# Datas - formatação dos campos como data
library(lubridate)
Auto2015sin$DATAPAR <- as.Date(ymd(Auto2015sin$DATAPAR))
Auto2015sin$DATASIN <- as.Date(ymd(Auto2015sin$DATASIN))
```

Criação de novos campos
anosin - ano de anulação

```
Auto2015sin$anosin <- year(ymd(Auto2015sin$DATASIN))

# passar dados de formatos data.frame para formato tbl e usar package dplyr
library(dplyr)
sin2015 <- tbl_df(Auto2015sin) # data.frame -> tbl
```

Aplicação de Filtros

```
# apólices com ano de sinistro 2015 e reserve >= 50
autosin <- filter(sin2015, anosin == 2015 & reserve >= 50)
```

Nova tabela - nr sinistros por apólice

```
nrsin <- count(autosin, NPOLIZA) #2 campos
nrsin <- rename(nrsin, nr_sin = n) #renomeação do campo nr sinistros
rm(Auto2015sin)

# ----------------------------------TESTE---------------------------------
# totalsin <- summarise(nrsin, sum(nr_sin)) # total obs
```

Criação de 'Nr_sin2015.RData'

```
save.image("Nr_sin2015.RData")
```

A.3 Funções

tabela.R

```
# Esta função tem como input:
#   - tabl          -> data.frame a consultar
#   - campo         -> seleção do campo           [campo 'codsitu']
#   - nome_campo    -> nome a atribuir ao campo   [campo 'dataini']
#
# e o seu output consiste numa tabela com os seguintes campos:
#
#  nome_campo | 'Em vigor' | 'Anuladas' | 'Taxa anulação' | 'Tax100' |
#  NAP |'Sinistros' | 'Frequência' | 'Freq100'
# A tabela a consultar deverá conter os campos: status, nap, nr_sin.
# ------------------------------------------------------------------------

tabela <- function(tabl, campo, nome_campo = 'Campo')
{
levs <- levels(campo) #requer campo formatado como factor
nr_sin <- nap <- freq <- vig <- anu <- tx <- rep(0,length(levs))

for (i in 1:length(levs))
{
vig[i] <- sum(tabl$status[campo == levs[i]
& tabl$status == 1], na.rm = TRUE)
anu[i] <- length(campo[campo == levs[i]]) - vig[i]
tx[i] <- round(anu[i] / (vig[i] + anu[i]),3) * 100
nap[i] <- round(sum(tabl$nap[campo == levs[i]], na.rm = TRUE),2)
nr_sin[i] <- sum(tabl$nr_sin[campo == levs[i]], na.rm = TRUE)
freq[i] <- round(nr_sin[i] / nap[i],3) * 100
}

#totais
levs[i+1] <- NA
vig[i+1] <- sum(vig)
```

```
anu[i+1] <- sum(anu)
tx[i+1] <- round(anu[i+1] / (vig[i+1] + anu[i+1]),3) * 100
nap[i+1] <- sum(nap)
nr_sin[i+1] <- sum(nr_sin)
freq[i+1] <- round(nr_sin[i+1] / nap[i+1],3) * 100

freq100 <- tx100 <- rep(0,length(levs))
for (i in 1:length(levs))
{
tx100[i] <- round(tx[i] / tx[length(levs)],3) * 100
freq100[i] <- round(freq[i] / freq[length(levs)],3) * 100
}

campotab <- matrix(c(levs,vig,anu,tx,tx100,nap,nr_sin,freq,freq100), ncol=9)
colnames(campotab) <- c(nome_campo,'Em␣vigor','Anuladas','Taxa␣anulação','
    Tax100','NAP','Sinistros','Frequência','Freq100')
campotab.table <- as.table(campotab)
rm(i)
return(as.data.frame(campotab))
}

# ----------------------------------TESTE-----------------------------------
# a <- tabela(tabl, tabl$estcivi)
# b <- tabela(tabl, tabl$sexo,'Sexo')
```

tabela2.R

```
# Esta função tem como input:
#   - tabl          -> data.frame a consultar
#   - campo         -> seleção do campo          [campo 'codsitu']
#   - nome_campo    -> nome a atribuir ao campo  [campo 'dataini']
#
# e o seu output consiste numa tabela com os seguintes campos:
#
#  nome_campo | 'Taxa anulação' | 'DL' | 'Seg' | 'Comp' | 'Neu'
# A tabela a consultar deverá conter os campos: status, nap, nr_sin e grp_
    codcaus.
# --------------------------------------------------------------------------

tabela2 <- function(tabl, campo, nome_campo = 'Campo')
{
levs <- levels(campo) #requer campo formatado como factor
vig <- anu <- tx <- rep(0,length(levs))
neu <- comp <- seg <- dl <- rep(0,length(levs))

for (i in 1:length(levs))
{
vig[i] <- sum(tabl$status[campo == levs[i]
& tabl$status == 1], na.rm = TRUE)
anu[i] <- length(campo[campo == levs[i]]) - vig[i]
tx[i] <- round(anu[i] / (vig[i] + anu[i]),3) * 100
dl[i] <- round(sum(!is.na(tabl$grp_codcaus[campo == levs[i]
& tabl$grp_codcaus == 'DL']), na.rm = TRUE)
/ (vig[i] + anu[i]),3) * 100
seg[i] <- round(sum(!is.na(tabl$grp_codcaus[campo == levs[i]
& tabl$grp_codcaus == 'Seg']), na.rm = TRUE)
```

```
/ (vig[i] + anu[i]),3) * 100
comp[i] <- round(sum(!is.na(tabl$grp_codcaus[campo == levs[i]
& tabl$grp_codcaus == 'Comp']), na.rm = TRUE)
/ (vig[i] + anu[i]),3) * 100
neu[i] <- round(sum(!is.na(tabl$grp_codcaus[campo == levs[i]
& tabl$grp_codcaus == 'Neu']), na.rm = TRUE)
/ (vig[i] + anu[i]),3) * 100
}

#totais
levs[i+1] <- NA
vig[i+1] <- sum(vig)
anu[i+1] <- sum(anu)
tx[i+1] <- round(anu[i+1] / (vig[i+1] + anu[i+1]),3) * 100
dl[i+1] <- round(sum(!is.na(tabl$grp_codcaus[tabl$grp_codcaus == 'DL']),
na.rm = TRUE) / (vig[i+1] + anu[i+1]),3) * 100
seg[i+1] <- round(sum(!is.na(tabl$grp_codcaus[tabl$grp_codcaus == 'Seg']),
na.rm = TRUE) / (vig[i+1] + anu[i+1]),3) * 100
comp[i+1] <- round(sum(!is.na(tabl$grp_codcaus[tabl$grp_codcaus == 'Comp']),
na.rm = TRUE) / (vig[i+1] + anu[i+1]),3) * 100
neu[i+1] <- round(sum(!is.na(tabl$grp_codcaus[tabl$grp_codcaus == 'Neu']),
na.rm = TRUE) / (vig[i+1] + anu[i+1]),3) * 100

campotab <- matrix(c(levs,tx,dl,seg,comp,neu), ncol=6)
colnames(campotab) <- c(nome_campo,'Taxa␣anulação','DL','Seg','Comp','Neu')
campotab.table <- as.table(campotab)
rm(i)
return(as.data.frame(campotab))
}

# ----------------------------------TESTE-----------------------------------
# a <- tabela2(tabl, tabl$estcivi)
# b <- tabela2(tabl, tabl$sexo,'Sexo')
```

nap.R

```
# Esta função tem como input:
#   - cod       -> situação da apólice  [campo 'codsitu']
#   - dataini   -> início da apólice    [campo 'dataini']
#   - datault   -> anulação da apólice  [campo 'datault']
# e devolve o nr de dias do ano de 2015 em que a apólice teve exposição ao
   risco.
# Este nr de dias é armazenado na variável 'dur' e está compreendido entre 0
    e 365.
# ------------------------------------------------------------------------

nap <- function(cod,dataini,datault=19000101)
{
library(lubridate)
Li <- as.Date(ymd(20150101)); Ls <- as.Date(ymd(20151231))
Ini <- dataini; Ult <- datault

ifelse(Ini > Ls,
dur <- NA, # --> delete
ifelse(cod == 5,
ifelse(Ini < Li, dur <- Ult - Li + 1, dur <- Ult - Ini + 1),
```

```
ifelse(Ini < Li, dur <- Ls - Li + 1, dur <- Ls - Ini + 1)))
}

# ----------------------------------TESTE-----------------------------------
# a<-nap(5,as.Date(ymd(20160101)),as.Date(ymd(20150205)))
# b<-nap(5,as.Date(ymd(19921003)),as.Date(ymd(20150928)))
# c<-nap(5,as.Date(ymd(20150106)),as.Date(ymd(20150110)))
# d<-nap(0,as.Date(ymd(20140506))) #usa default para datault
# e<-nap(0,as.Date(ymd(20150506))) #usa default para datault
```

aval.R

```
# Esta função tem como input:
#   - mdl           -> modelo a avaliar
#   - coff          -> ponto de corte
#
# e o seu output consiste numa tabela com os seguintes campos:
#
#     AUC / Cut-off / Accuracy / Sensibility / Specificity
# ------------------------------------------------------------------------

aval <- function(mdl, coff)
{
fitted.results <- predict(mdl, newdata = teste, type = 'response')

library(ROCR)
pr <- prediction(fitted.results, teste$DL)
auc <- performance(pr, measure = "auc") # area parcial fpr.stop=0.5
auc <- auc@y.values[[1]]

# accuracy ---------------------------------------------------------------
pred_ac <- ifelse(fitted.results > coff, 1, 0)
misClassifError <- mean(pred_ac != teste$DL) #média dos valores lógicos
acc <- 1 - misClassifError

tab <- xtabs(~ pred_ac + teste$DL)
prp <- prop.table(tab,2)

result <- matrix(c(auc, coff, prp[2,2], prp[1,1], acc), ncol = 5)
colnames(result) <- c('AUC', 'Cut-off', 'Sensibility', 'Specificity', '
   Accuracy')
return(as.table(result))
}

# ----------------------------------TESTE-----------------------------------
# aval(model5BI, 0.15)
```

A.4 Tabelas e gráficos

Tabelas resumo

```
load("mini_fix.RData")
```

```
tabl <- dados.plus # seleccionar aqui a data.frame a consultar

source("tabela.R") # call function
tb_codbonus    <- tabela(tabl, tabl$grp_codbonus,'Bonificação')
tb_formapa     <- tabela(tabl, tabl$formapa,'Forma Pagamento')
tb_dataini     <- tabela(tabl, tabl$grp_anoini,'dataini')
tb_cotpot      <- tabela(tabl, tabl$grp_cotpot,'cotpot')
tb_codrisc     <- tabela(tabl, tabl$codrisc,'Codrisc')
tb_codcomb     <- tabela(tabl, tabl$grp_codcomb,'Combustível')
tb_codgara     <- tabela(tabl, tabl$codgara,'Garagem')
tb_jovem       <- tabela(tabl, tabl$grp_jovem,'Jovem')
tb_jubilado    <- tabela(tabl, tabl$jubilado,'Jubilado')
tb_zona_pr     <- tabela(tabl, tabl$zona_pr,'Zona Tarifária')
tb_codutil     <- tabela(tabl, tabl$codutil,'Codutil')
tb_class_dp    <- tabela(tabl, tabl$class_dp,'Class_dp')
tb_pot         <- tabela(tabl, tabl$grp_pot,'Potência')
tb_modapac_1   <- tabela(tabl, tabl$grp_modapac1,'Pack de Coberturas 1')
tb_modapac_2   <- tabela(tabl, tabl$grp_modapac2,'Pack de Coberturas 2')
tb_datanas     <- tabela(tabl, tabl$grp_anonas,'Data de Nascimento')
tb_datacar     <- tabela(tabl, tabl$grp_anocar,'Data da Carta')
tb_estcivi     <- tabela(tabl, tabl$grp_estcivi,'Estado Civil')
tb_sexo        <- tabela(tabl, tabl$sexo,'Sexo')

source("tabela2.R") # call function
tb_codbonus2   <- tabela2(tabl, tabl$grp_codbonus,'Bonificação')
tb_formapa2    <- tabela2(tabl, tabl$formapa,'Forma Pagamento')
tb_dataini2    <- tabela2(tabl, tabl$grp_anoini,'dataini')
tb_cotpot2     <- tabela2(tabl, tabl$grp_cotpot,'cotpot')
tb_codrisc2    <- tabela2(tabl, tabl$codrisc,'Codrisc')
tb_codcomb2    <- tabela2(tabl, tabl$grp_codcomb,'Combustível')
tb_codgara2    <- tabela2(tabl, tabl$codgara,'Garagem')
tb_jovem2      <- tabela2(tabl, tabl$grp_jovem,'Jovem')
tb_jubilado2   <- tabela2(tabl, tabl$jubilado,'Jubilado')
tb_zona_pr2    <- tabela2(tabl, tabl$zona_pr,'Zona Tarifária')
tb_codutil2    <- tabela2(tabl, tabl$codutil,'Codutil')
tb_class_dp2   <- tabela2(tabl, tabl$class_dp,'Class_dp')
tb_pot2        <- tabela2(tabl, tabl$grp_pot,'Potência')
tb_modapac2_1  <- tabela2(tabl, tabl$grp_modapac1,'Pack de Coberturas 1')
tb_modapac2_2  <- tabela2(tabl, tabl$grp_modapac2,'Pack de Coberturas 2')
tb_datanas2    <- tabela2(tabl, tabl$grp_anonas,'Data de Nascimento')
tb_datacar2    <- tabela2(tabl, tabl$grp_anocar,'Data da Carta')
tb_estcivi2    <- tabela2(tabl, tabl$grp_estcivi,'Estado Civil')
tb_sexo2       <- tabela2(tabl, tabl$sexo,'Sexo')
```

Conjuntos de tabelas

```
t1 <- list(tb_codbonus, tb_formapa, tb_dataini, tb_cotpot, tb_codrisc, tb_
   codcomb, tb_codgara,
tb_jovem, tb_jubilado, tb_zona_pr, tb_codutil, tb_class_dp, tb_pot, tb_
   modapac_1,
tb_modapac_2, tb_datanas, tb_datacar, tb_estcivi, tb_sexo)

t2 <- list(tb_codbonus2, tb_formapa2, tb_dataini2, tb_cotpot2, tb_codrisc2,
   tb_codcomb2,
tb_codgara2, tb_jovem2, tb_jubilado2, tb_zona_pr2, tb_codutil2, tb_class_dp2
   ,
```

```
tb_pot2, tb_modapac2_1, tb_modapac2_2, tb_datanas2, tb_datacar2, tb_estcivi2
    , tb_sexo2)
```

Criação de 'tabelas.RData'

```
rm(dados, dados.plus, tabl, teste, treino, ind.teste, tabela, tabela2)
save.image("tabelas.RData")
```

Apresentação LaTeX

```
load("tabelas.RData")
library(xtable)

cap <- c('Bonificação','Forma␣de␣Pagamento','dataini','cotpot','codrisc',
'Combustível','Garagem','Jovem','Jubilado','Zona␣Tarifária','codutil',
'class_dp','Potência','Pack␣Coberturas␣1','Pack␣Coberturas␣2','Data␣de␣
    Nascimento',
'Data␣da␣Carta','Estado␣Civil','Sexo')
ali <- c('rccccccccc', 'rcccccc')

t <- t2 #t1 ou t2
i <- 19 #selecionar campo (1 a 19)
j <- ifelse(ncol(t[[i]]) == 9, 1, 2) #definição do align
print(xtable(t[[i]], caption = cap[i], align = ali[j]),
booktabs = TRUE,
size = "footnotesize",
include.rownames = FALSE,
caption.placement = "top")
rm(cap, ali, t, i, j)
```

Gráficos após cada tabela

```
load('tabelas.RData')

library(ggplot2)

names <- c('Bonif', 'ForPag', 'DataIni', 'PrApo', 'Codrisc', 'Combu', 'Gara'
    , 'Jovem',
'Jubil', 'ZonTarif', 'Codutil', 'Clas_dp', 'Pot', 'Cobert1', 'Cobert2',
'DatNas', 'DataCar', 'EstCivi', 'Sexo')

# Rearranjar dados das tabelas para gráficos de Tax100 e Freq100 (após 1as
    tabelas)

i <- 1 #usar sem ciclo para reordenar facets
#fct <- list(c('M','N','B','B40'), )

for (i in 1:length(t1)){
filename1 <- paste('TaxFreq_', names[i], '.pdf', sep='')
mypath1 = file.path('report/imagens', filename1)

pdf(file = mypath1, height = 4, width = 7)
tabela <- t1[[i]] # escolha da tabela por campo
a <- tabela[-nrow(tabela),c(1,5,9)] # retirar última linha e selecionar
    colunas
vals <- c(as.numeric(levels(a$Tax100))[a$Tax100],
```

```
as.numeric(levels(a$Freq100))[a$Freq100])
f100 <- c(rep('Tax100', nrow(a)), rep('Freq100',nrow(a)))
fcol1 <- c(rep(as.vector(a[,1]),2))
ab <- data.frame(vals,f100,fcol1)
# ab$fcol1 <- factor(ab$fcol1, levels = c('M','N','B','B40')) #reorder
    facets Bonif (i=1)
rm(tabela,a,vals,f100,fcol1)
print(ggplot(ab, aes(x = '', y = vals, fill = f100)) +
geom_bar(width = 1, position = 'dodge', stat = 'identity')  +
# ggtitle('Tax100 vs Freq100') +
xlab('') + ylab('%') +
facet_grid(facets =. ~ fcol1) +
scale_fill_manual(values = c('#5e3c99','#e66101'),
name = ''))
dev.off()
}
rm(ab, filename1, i, mypath1)

# Rearranjar dados das tabelas para gráficos de anulação por DL (após 2as
    tabelas)

for (i in 1:length(t2)){
filename2 <- paste('anulDL_', names[i], '.pdf', sep='')
mypath2 = file.path('report/imagens', filename2)

pdf(file = mypath2, height = 4, width = 7)
tabela <- t2[[i]] # escolha da tabela por campo
a <- tabela[-nrow(tabela),c(1,2,3)] # retirar última linha e selecionar
    colunas
vals <- c(as.numeric(levels(a$`Taxa anulação`))[a$`Taxa anulação`],
as.numeric(levels(a$DL))[a$DL])
taxadl <- c(rep('Taxa␣anulação', nrow(a)), rep('DL',nrow(a)))
fcol1 <- c(rep(as.vector(a[,1]),2))
ab <- data.frame(vals,taxadl,fcol1)
# ab$fcol1 <- factor(ab$fcol1, levels = c('M','N','B','B40')) #reorder
    facets Bonif (i=1)
rm(tabela,a,vals,taxadl,fcol1)
print(ggplot(ab, aes(x = '', y = vals, fill = taxadl)) +
geom_bar(width = 1, position = 'identity', stat = 'identity')  +
# ggtitle('Anulação por DL') +
xlab('') + ylab('%') +
facet_grid(facets =. ~ fcol1) +
scale_fill_manual(values = c('#5e3c99','#e66101'),
name = ''))
dev.off()
}
rm(ab, filename2, i, mypath2)
```

A.5 Modelos

Tratamento de missings

```
load("mini.RData")
```

```
cl <- c('NPOLIZA', 'status', 'nap', 'nr_sin', 'grp_codcaus')
# exclusão temporária de colunas em 'cl' para identificar missings
tmp.dados <- mini[, !(names(mini) %in% cl)]
ids <- complete.cases(tmp.dados) #indices das linhas sem missings
dados.plus <- mini[ids, ] # apenas linhas de mini sem missings
dados <- mini[ids, !(names(mini) %in% cl) ] #retira colunas para stepwise
    backward
rm(cl, tmp.dados, ids)
```

Subconjunto de dados equilibrados

```
# library(dplyr)
# a <- dados[dados$DL == 1,]
# b <- dados[dados$DL == 0,]
# ind.DL0 <- sort(sample(1:nrow(b), size = nrow(a)))
# b.filt <- b[ind.DL0,]
# dados.filt <- bind_rows(a, b.filt)
# rm(a,b,ind.DL0,b.filt)
```

Conjuntos de treino / teste

```
##balanced
# ind.teste <- sort(sample(1:nrow(dados.filt),
#                          size = round(0.1 * nrow(dados.filt))))
# treino <- dados.filt[-ind.teste, ]
# teste <- dados.filt[ind.teste, ]
##imbalanced
ind.teste <- sort(sample(1:nrow(dados), size = round(0.1 * nrow(dados))))
treino <- dados[-ind.teste, ]
teste <- dados[ind.teste, ]
```

Criação de 'mini_fix.RData'

```
rm(autosin, nrsin, sin2015,auto21, auto4, auto451, auto451.2015, nap, mini)
save.image("mini_fix.RData")
#write.csv(treino, "C:/#DISPE_AXA/treino.csv")

# começar aqui para trabalhar sempre com os mesmos conjuntos de treino e
    teste
load("mini_fix.RData")
```

Testes Qui-quadrado de independência de variáveis

```
chisq.test(dados$formapa,dados$DL)
chisq.test(dados$codrisc,dados$DL)
chisq.test(dados$codgara,dados$DL)
chisq.test(dados$jubilado,dados$DL)
chisq.test(dados$zona_pr,dados$DL)
chisq.test(dados$codutil,dados$DL)
chisq.test(dados$class_dp,dados$DL)
chisq.test(dados$sexo,dados$DL)
chisq.test(dados$grp_codcomb,dados$DL)
chisq.test(dados$grp_jovem,dados$DL)
chisq.test(dados$grp_estcivi,dados$DL)
chisq.test(dados$grp_codbonus,dados$DL)
```

```
chisq.test(dados$grp_anoini,dados$DL)
chisq.test(dados$grp_cotpot,dados$DL)
chisq.test(dados$grp_pot,dados$DL)
chisq.test(dados$grp_anonas,dados$DL)
chisq.test(dados$grp_anocar,dados$DL)
chisq.test(dados$grp_modapac1,dados$DL)
chisq.test(dados$grp_modapac2,dados$DL)
```

Definição do modelo

```
#modelo completo
full.model <- glm(DL ~ grp_anocar + grp_anoini + grp_codbonus + grp_anonas +
    grp_cotpot + grp_modapac1 + formapa + grp_estcivi + grp_codcomb + zona_
   pr + codrisc + sexo + grp_jovem + grp_modapac2 + jubilado + grp_pot +
   class_dp + codgara + codutil,
        family = binomial(), data = treino)

# Stepwise Regression
# mini[,-1] ; dados ; treino
min.mod <- glm(DL ~ 1, family = binomial(), treino)
max.mod <- glm(DL ~ ., family = binomial(), treino)
res <- step(min.mod, direction = 'forward', scope = formula(max.mod)) #
   forward
# res <- step(max.mod, direction = 'backward') #backward
summary(res)

# modelo com melhor resultado após stepwise
model <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_modapac1 + grp_
   cotpot + grp_anonas + grp_estcivi + grp_codcomb + zona_pr + codrisc +
   sexo + grp_jovem + grp_modapac2 + grp_anocar + jubilado + grp_pot + class
   _dp,
        family = binomial(), data = treino)

# modelo Maranhão v1 (ignorado após analisar Odds Ratios de zona_pr)
model1 <- glm(DL ~ grp_anoini + grp_codbonus + formapa + zona_pr + grp_
   modapac2,
        family = binomial(), data = treino)

# modelo Maranhão v2 (modelo final / interação)
modelBase <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_modapac2,
        family = binomial(), data = treino)
modelBaseI <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_modapac2 +
   grp_anoini:formapa,
        family = binomial(), data = treino)

# procura de novos
add1(model3B, ~.^2) # procura melhor interação entre 2 campos
drop1(model3BI)

# modelo com 2 a 6 variáveis
model2B <- glm(DL ~ grp_anoini + formapa,
        family = binomial(), data = treino)
model2BI <- glm(DL ~ grp_anoini + formapa + grp_anoini:formapa,
        family = binomial(), data = treino)
model3B <- glm(DL ~ grp_anoini + grp_codbonus + formapa,
        family = binomial(), data = treino)
```

```
model3BI <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_anoini:formapa,
      family = binomial(), data = treino)
#model3BID = model2BI
model4B <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_modapac1,
      family = binomial(), data = treino)
model4BI <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_modapac1 + grp_anoini:formapa,
      family = binomial(), data = treino)
model4BID <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_anoini:formapa,
      family = binomial(), data = treino)
model5B <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_modapac1 + grp_cotpot,
      family = binomial(), data = treino)
model5BI <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_modapac1 + grp_cotpot + grp_anoini:formapa,
      family = binomial(), data = treino)
model5BID <- glm(DL ~ grp_anoini + formapa + grp_modapac1 + grp_cotpot + grp_anoini:formapa,
      family = binomial(), data = treino)
model6B <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_modapac1 + grp_cotpot + grp_anonas,
      family = binomial(), data = treino)
model6BI <- glm(DL ~ grp_anoini + grp_codbonus + formapa + grp_modapac1 + grp_cotpot + grp_anonas + grp_anoini:formapa,
      family = binomial(), data = treino)
#model6BID = model5BI
```

Tabela para LaTeX

```
library(stargazer)
stargazer(model2, label = 'mM', digit.separator = '␣', single.row = TRUE)
```

Avaliação do modelo

```
# Antes
# 1) importar dados
load("mini_fix.RData")
# 2) correr modelos em 'modelos.R'

# Seleção do modelo a avaliar
mdl <- modelBase

## Odds-Ratio e Intervalos de Confiança
#sem package
# OR <- exp(cbind(OddsRatio <- coef(mdl), confint(mdl)))
#com package
# library(oddsratio)
# ORm2 <- calc.oddsratio.glm(data = treino, model = mdl, CI = 0.95)
# save.image("modOR.RData")
load("modOR.RData")

## Avaliação do modelo (ajustamento)

# Goodness of Fit: Likelihood Ratio Test
```

```
library(lmtest)
lrtest(full.model, model)
lrtest(model, model5BI)

# Goodness of Fit: Pseudo R^2
library(pscl)
pR2(mdl) # look for 'McFadden'

# Goodness of Fit: Hosmer-Lemeshow Test
a <- mdl$y
b <- fitted(mdl) # usa dados de treino
library(ResourceSelection)
hos.test <- hoslem.test(a, b, g = 10)
hos.test
cbind(hos.test$expected,hos.test$observed)
rm(a,b,hos.test)

# Tests of Individual Predictors: Wald Test
library(survey)
regTermTest(mdl, "grp_anoini")
regTermTest(model, "grp_pot")
regTermTest(full.model, "grp_pot")

# Tests of Individual Predictors: Variable Importance
library(caret)
varImp(model)

## Avaliação do modelo (desempenho preditivo)

fitted.results <- predict(mdl, newdata = teste, type = 'response')

library(classifierplots)
memory.limit(size = 8068) #increase memory limit
#classifierplots(teste$DL,fitted.results) #cria 1 plot
classifierplots_folder(teste$DL,fitted.results,"D:/OneDrive/#AXA/pasta")
#cria imagens separadamente numa pasta

# function aval
source("aval.R")
aval(model3BI, 0.1221)
```

Bibliografia

[1] Abrahamse, A. F. e S. J. Carroll: *The frequency of excess claims for automobile personal injuries*. Em *Automobile insurance: Road safety, new drivers, risks, insurance fraud and regulation*, págs. 131–149. Springer, 1999.

[2] Agresti, A.: *Categorical Data Analysis*. Wiley Series in Probability and Statistics. Wiley-Interscience, 2nd ed., 2002, ISBN 0471360937.

[3] Agresti, A.: *An Introduction to Categorical Data Analysis*. John Wiley & Sons, Inc., 2nd ed., 2007, ISBN 0471226181.

[4] Agresti, A.: *Categorical Data Analysis*. John Wiley & Sons. Inc., 2013, ISBN 0470463635.

[5] Agresti, A.: *Foundations of Linear and Generalized Linear Models*. Wiley series in probability and statistics. John Wiley & Sons, Inc., 2015, ISBN 1118730034.

[6] Alice, M.: *How to perform a Logistic Regression in R*. https://www.r-bloggers.com/how-to-perform-a-logistic-regression-in-r/, set. 2015.

[7] Artís, M., M. Ayuso e M. Guillen: *Modeling Different Types of Automobile Insurance Fraud Behavior in the Spanish Market*. Insurance: Mathematics and Economics, 24:67–81, fev. 1999.

[8] Artís, M., M. Ayuso e M. Guillen: *Detection of Automobile Insurance Fraud With Discrete Choice Models and Misclassified Claims*. Journal of Risk and Insurance, 69:325 – 340, set. 2002.

[9] Cookson, T.: *Automatically Save Your Plots to a Folder*. https://www.r-bloggers.com/automatically-save-your-plots-to-a-folder/, abr. 2011.

[10] Dahl, D. B.: *xtable: Export Tables to LaTeX or HTML*, 2016. https://CRAN.R-project.org/package=xtable, R package version 1.8-2.

[11] Defazio, A.: *classifierplots: Generates a Visualization of Classifier Performance as a Grid of Diagnostic Plots*, 2017. https://CRAN.R-project.org/package=classifierplots, R package version 1.3.3.

[12] Dionne, G., C. Gouriéroux e C. Vanasse: *Evidence of adverse selection in automobile insurance markets*. Em *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*, págs. 13–46. Springer, 1999.

[13] Dionne, G., C. Laberge-Nadeau, D. Desjardins, S. Messier e U. Maag: *Analysis of the economic impact of medical and optometric driving standards on costs incurred by trucking*

firms and on the social costs of traffic accidents. Em *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*, págs. 323–351. Springer, 1999.

[14] Dionne, G. e C. Vanasse: *Automobile insurance ratemaking in the presence of asymmetrical information*. Journal of Applied Econometrics, 7(2):149–165, 1992.

[15] Dobson, A. J.: *An introduction to generalized linear models*. Chapman & Hall/CRC, 2002.

[16] Faraway, J. J.: *Practical Regression and ANOVA using R*. University of Bath, 2002.

[17] Fischetti, T.: *Data Analysis with R*. PACKT PUB, 2015, ISBN 1785288148.

[18] Grolemund, G. e H. Wickham: *Dates and Times Made Easy with lubridate*. Journal of Statistical Software, 40(3):1–25, 2011. http://www.jstatsoft.org/v40/i03/.

[19] Grolemund, G. e H. Wickham: *R for Data Science*. http://r4ds.had.co.nz/, 2016.

[20] Guillen, M., J. P. Nielsen e A. M. Pérez-Marín: *The Need to Monitor Customer Loyalty and Business Risk in the European Insurance Industry*. The Geneva Papers on Risk and Insurance - Issues and Practice, 33(2):207, 2008. https://doi.org/10.1057/gpp.2008.1.

[21] Guillen, M., J. Parner, C. Densgsoe e A. M. Pérez-Marín: *Customer loyalty in the insurance industry: a logistic regression approach*. Em *II Conference in Actuarial Science and Finance on Samos, Karlovasi-Samos, Greece*, 2002.

[22] Hlavac, M.: *stargazer: Well-Formatted Regression and Summary Statistics Tables*. Harvard University, Cambridge, USA, 2015. http://CRAN.R-project.org/package=stargazer, R package version 5.2.

[23] Hosmer, D. W., S. Lemeshow e R. X. Sturdivant: *Applied Logistic Regression*. John Wiley & Sons, Inc., 3rd ed., 2013, ISBN 0470582472.

[24] Khakbaz, S. B., N. Hajiheydari e M. Pourestarabadi: *Car Insurance Risk Assessment with Data Mining for an Iranian Leading Insurance Company*. International Journal of Business and Economics Research, 3(3):128, 2014. http://dx.doi.org/10.11648/j.ijber.20140303.12.

[25] Kleinbaum, D. G. e M. Klein: *Logistic Regression*. Springer-Verlag GmbH, 3rd ed., 2010, ISBN 1441917411.

[26] Marôco, J.: *Análise Estatística com o SPSS Statistics. 6ª Ed.* ReportNumber, Lda, 2014, ISBN 9899676343. http://www.reportnumber.pt/ae/.

[27] Nelder, J. A. e R. W. M. Wedderburn: *Generalized Linear Models*. Journal of the Royal Statistical Society. Series A (General), 135(3):370, 1972. http://dx.doi.org/10.2307/2344614.

[28] Paulino, C., D. Pestana, J. Branco, J. Singer, L. Barroso e W. Bussab: *Glossário Inglês-Português de Estatística*. http://glossario.spestatistica.pt/, 2011.

[29] Pinquet, J.: *Allowance for cost of claims in bonus-malus systems*. ASTIN Bulletin: The Journal of the IAA, 27(1):33–57, 1997.

[30] Pinquet, J.: *Allowance for hidden information by heterogeneous models and applications to insurance rating.* Em *Automobile Insurance: Road Safety, New Drivers, Risks, Insurance Fraud and Regulation*, págs. 47–78. Springer US, 1999.

[31] R Core Team: *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2017. https://www.R-project.org/.

[32] Sing, T., O. Sander, N. Beerenwinkel e T. Lengauer: *ROCR: Visualizing classifier performance in R.* Bioinformatics, 21(20):7881, 2005. http://rocr.bioinf.mpi-sb.mpg.de.

[33] Stulp, G.: *ggplotgui: Create Ggplots via a Graphical User Interface*, 2017. https://github.com/gertstulp/ggplotgui/, R package version 1.0.0.

[34] Turkman, M. A. A. e G. L. Silva: *Modelos Lineares Generalizados - da teoria à prática.* Em *VIII Congresso Anual da Sociedade Portuguesa de Estatística, Lisboa*, 2000.

[35] Wickham, H.: *ggplot2: Elegant Graphics for Data Analysis.* Springer-Verlag New York, 2009, ISBN 978-0-387-98140-6. http://ggplot2.org.

[36] Wickham, H., R. Francois, L. Henry e K. Müller: *dplyr: A Grammar of Data Manipulation*, 2017. https://CRAN.R-project.org/package=dplyr, R package version 0.7.2.