LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO

CIÊNCIAS ATUARIAIS

TRABALHO FINAL DE MESTRADO

TRABALHO DE PROJETO

MODELOS PARA ESTIMAR TAXAS DE RETENÇÃO DE CLIENTES APLICAÇÃO A UMA CARTEIRA DE SEGURO AUTOMÓVEL

JOANA MENDONÇA VASCONCELOS ROMÃO

OUTUBRO - 2019

LISBON SCHOOL OF ECONOMICS & MANAGEMENT
UNIVERSIDADE DE LISBOA

MESTRADO

CIÊNCIAS ATUARIAIS

TRABALHO FINAL DE MESTRADO

TRABALHO DE PROJETO

MODELOS PARA ESTIMAR TAXAS DE RETENÇÃO DE CLIENTES APLICAÇÃO A UMA CARTEIRA DE SEGURO AUTOMÓVEL

JOANA MENDONÇA VASCONCELOS ROMÃO

ORIENTAÇÃO:
PROF. DR. JOÃO MANUEL DE SOUSA ANDRADE E SILVA

OUTUBRO - 2019

Agradecimentos

Gostaria de agradecer ao meu orientador Professor Doutor João Andrade e Silva o apoio e total disponibilidade durante o desenvolvimento deste projeto, à entidade seguradora que me possibilitou a recolha dos dados e à minha família e amigos, cujo apoio foi imensurável.

Apresentação

Modelos para Estimar Taxas de Retenção de Clientes

Aplicação a uma Carteira de Seguro Automóvel

Joana Mendonça Vasconcelos Romão

O acesso à informação tem-se tornado cada vez mais fácil. A comparação entre condições tarifárias de diferentes seguradoras é hoje mais frequente, com efeito nas taxas de retenção de clientes e respetivos contratos de seguro. A importância que é dada a este tema é cada vez maior e a construção de ferramentas para estimar as referidas taxas permite tomar medidas para a retenção de negócio rentável e o agravamento dos prémios de contratos menos rentáveis.

Este trabalho teve como objetivo estimar a probabilidade de retenção à data de vencimento de uma apólice de seguro, numa carteira do ramo automóvel.

Verificado o problema de desequilíbrio entre as classes da variável resposta, a escolha das metodologias a usar baseou-se essencialmente na procura de aumentar a exatidão do modelo final e contornar esse problema.

Palavras-chave: Taxas de Retenção; Variação Anual no Prémio; Modelos Lineares Generalizados; Desequilíbrio entre classes; *Machine Learning*; *Gradient Boosted Trees*; *Balanced Random Forest*.

Presentation

Models to Predict the Customers Retention Rates

An Application to a Motor Insurance Portfolio

Joana Mendonça Vasconcelos Romão

With an increasingly easy accessibility to information, there is a growing concern about customer retention rates. Insurers are giving more importance on having accurate tools to monitor the policies renewal process, making them allowed to keep with the profitable business and increase premiums on the less profitable one.

The objective of this study was to estimate the probability of renewing a policy in a motor insurance portfolio.

To be working with an imbalance data set made us try different modelling methodologies, where all of them were chosen based on the need to increase the predictive performance of the model.

Keywords: Retention Rates; Annual Change in Premium; Generalized Linear Models; Imbalanced Data; Machine Learning; Gradient Boosted Trees; Balanced Random Forest.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Acrónimos

AIC - *Akaike Information Criterion*.

BRF – *Balanced Random Forest.*

GBM – *Gradient Boosting Model.*

MLG – Modelo Linear Generalizado.

WRF – *Weighted Random Forest*.

OOB – Erro *out-of-bag.*

CV – *Cross Validation.*

ROC - Receiver Operating Characteristic Curve.

AUC – *Area Under the Curve*.

1 Introdução

A competitividade no mercado segurador tem aumentado nos últimos anos. Atualmente, encontrar o melhor preço para um contrato de seguro, aquele que para além de garantir os níveis de rentabilidade se assume competitivo quando comparado com os valores de mercado, é um trabalho considerado de grande importância. Desta forma, o estudo do comportamento dos clientes à data da renovação do seu contrato, nomeadamente face a diferentes variações no prémio, tem-se verificado um tema relevante no âmbito da gestão de uma carteira.

Motivado pelo exposto, este trabalho trata a modelação atuarial de taxas de retenção de clientes com diferentes perfis e face a diferentes variações de prémio. Mais especificamente, estima a probabilidade de renovação de uma apólice de seguro com vista à obtenção, numa fase posterior, de uma ferramenta que permita otimizar os níveis de variação de prémio.

O trabalho baseia-se em dados reais do mercado segurador português, pertencentes a uma linha de negócio destinada a clientes particulares do ramo automóvel.

Uma vez que para clientes particulares as datas de vencimento se distribuem uniformemente ao longo do ano, a companhia em análise pretende que a utilização do modelo seja numa base mensal. Uma monitorização com essa frequência permitirá a otimização das taxas de agravamento de prémio de acordo com os valores de retenção, tanto atuais como estimados.

Para a escolha da metodologia a utilizar, além do tipo de uso a dar ao modelo, outros fatores devem ser tidos em conta. O problema em estudo, o tipo de variável a estimar, as características da amostra e os destinatários da informação são alguns exemplos, sendo que na maioria dos casos diversas metodologias são aplicáveis e o processo de escolha passa por testá-las no nosso conjunto de dados.

Neste projeto foram abordados diferentes modelos preditivos de variáveis binárias, mas cujos níveis de precisão e desempenho variam.

O primeiro a ser apresentado é uma regressão logística. Os Modelos Lineares Generalizados mostram-se à partida adequados pois é com base num conjunto de

variáveis de classificação (perfil do cliente) que é estimado o resultado da prova de Bernoulli, cujo sucesso foi por nós considerado a manutenção do contrato de seguro. No entanto, os MLG tendem a simplificar interações complexas entre fatores e dificilmente atingem a granularidade pretendida em certos problemas de classificação. Os métodos *Tree-based* contornam essas fraquezas e resultam na grande maioria dos casos em modelos com melhor desempenho preditivo. Inserem-se nos métodos de *Machine Learning* para tarefas de classificação que, de forma bastante eficiente constroem um conjunto de condições de fácil interpretação e implementação.

O desequilíbrio na distribuição das classes é um problema comum em análises de taxas de retenção/abandono e acontece quando uma das classes abrange apenas uma pequena minoria dos dados. Esta particularidade verificou-se na nossa amostra e dificultou a procura de um modelo de exatidão satisfatória. De acordo com anteriores trabalhos de Burez e Van den Poel (2009) as técnicas de *Gradient Boosting* possibilitam a modelação de dados desequilibrados com melhoria significativa das métricas de desempenho preditivo.

Boosting consiste na construção de algoritmos com o objetivo específico de complementar um outro de exatidão superior, mas que sozinho apresenta ainda lacunas de *performance*. É uma técnica que a cada iteração dá uma maior importância aos casos que foram anteriormente mal classificados e, por isso, é habitualmente bem-sucedido em exemplos de distribuição desequilibrada entre classes. *Gradient Boosting* é um tipo de *boosting* com particularidades na seleção dos sucessivos conjuntos de dados de treino e também no cálculo da contribuição de cada estimativa para o algoritmo final. O segundo modelo que apresentamos é resultado da aplicação desta técnica a árvores de decisão.

Outro meio de contornar o desequilíbrio entre classes é a aplicação de *down-sampling* à classe com a maioria das observações (Kubat e Matwin (1997)). O terceiro e último modelo obteve-se por aplicação do método *Random Forest* induzido por *down-sampling*.

Random Forest é, tal como as *Gradient Boosted Trees*, um modelo que junta várias árvores de decisão para efeitos de melhoria do desempenho preditivo e redução de instabilidade (propensão a grandes alterações face a pequenas mudanças na amostra de dados). As árvores são obtidas por reamostragem do conjunto de dados de treino e posteriormente combinadas por maioria de voto entre as estimativas obtidas. *Random*

Forest trabalha com uma seleção aleatória de características classificativas, acrescentando assim diversidade ao modelo e diminuindo a probabilidade de sobre ajuste aos dados de treino. A utilização de apenas parte do conjunto de características recolhidas permite atingir bons resultados em amostras com elevado número de observações e variáveis de classificação. Sendo esse o nosso caso, esta foi mais uma das razões que nos levou a testar este método.

Por fim, apresentamos uma comparação entre os resultados obtidos recorrendo a métricas de desempenho, concluindo com uma apreciação sobre qual o método que constitui a melhor escolha para este problema.

2 Amostra de Dados

A fase de recolha de dados é decisiva no que se refere à fiabilidade do modelo final. É importante garantir que o conjunto de dados tem uma dimensão suficiente e que reflete de melhor forma possível a realidade que se pretende modelar.

Cada apólice proposta a renovação constitui uma observação e a amostra inclui 302152 observações. Aplicou-se o método *Holdout* (Kohavi, 1995) tendo resultado dois conjuntos de dados, um de treino e um de teste, com 80% e 20% do total, respetivamente.

O período temporal analisado foi de 12 meses pelo facto de se tratar de uma linha de negócio para clientes individuais do ramo automóvel. Cada contrato tem a duração de um ano e as renovações distribuem-se de forma igualitária ao longo desse período. Outra importante particularidade tida em conta foi a existência de um período de trinta dias para pagamento do prémio devido e que, portanto, impossibilitou a recolha da decisão do cliente antes de findo esse período.

O âmbito acordado foi o conjunto das renovações de 2018 por ser o período de um ano completo mais atual.

2.1Variável Resposta

A variável resposta é binária e representa a decisão de renovação ou abandono do contrato de seguro. A informação foi obtida a partir dos recibos de prémios à data da recolha da amostra. Se pago significa que a apólice renovou e a variável resposta toma o valor 1.

A medida de exposição define-se pelo número total de contratos propostos a renovar. Na linha de negócio em análise uma apólice corresponde a apenas um objeto seguro, o veículo ligeiro de uso habitual do cliente. Há, no entanto, casos em que é contratado um seguro para o reboque e é prática comum da companhia inclui-lo na mesma apólice. Apesar disso, é irrelevante a análise da renovação ao nível do objeto, pois é improvável que a renovação ou cancelamento não inclua ambos.

Justificado pela procura de melhores condições tarifárias, assiste-se por vezes à anulação da apólice e emissão imediata de uma nova. Este fenómeno tem o nome de Migração e é habitualmente considerado em ajuste de modelos de retenção. No nosso caso as migrações foram encaradas como cancelamentos, já que pretendemos uma ferramenta para a análise de níveis de elasticidade de prémio, e uma migração é de facto uma resposta a alterações tarifárias. É importante referir que a recolha da causa de anulação do contrato permitiu perceber que uma migração pode ter outras causas, como a alteração do agente de mediação. Esses casos originam uma anulação sem que a causa esteja relacionada com alterações tarifárias. Por essa razão, fomos contabilizá-los e por constituírem uma minoria da amostra optou-se por exclui-los em detrimento da consideração do motivo de anulação como fator explicativo.

Pela mesma razão, cancelamentos por perda total ou venda do veículo foram também excluídos.

Alterações às condições do contrato (como a contratação de novas coberturas ou a alteração de capitais seguros) implicam na maioria dos casos variações no prémio tendo sido, por isso, tema discutido durante este projeto. No entanto, a importância destas situações é maior em apólices de frota, onde as alterações são mais frequentes e com maior impacto nas somas seguras. Por essa razão, e pela complexidade que traria ao processo de recolha dos dados, optou-se por não se considerar a causa inerente à variação do prémio.

2.2 Variáveis Explicativas

Segue-se uma breve descrição das variáveis de classificação incluídas na amostra.

2.2.1 Tomador de Seguro

No que diz respeito ao tomador, é fácil prever que, como autor da tomada de decisão, seja relevante para a construção de um modelo deste tipo. Espera-se que aqueles que

demonstram maiores taxas de retenção sejam os de mais **idade**, maior **antiguidade** e possuindo **outros contratos** na companhia. Trata-se de variáveis que se pensa estarem intimamente ligadas à fidelidade do cliente.

Habilitações literárias, **profissão** e **estado civil** são também relevantes, mas raramente fiáveis. É informação que muda no tempo e muito raramente atualizada pelo cliente, pelo que não foi considerada.

O **distrito** também se espera que tenha influência pelas disparidades a nível económico e social que se verificam entre zonas do país.

2.2.2 Condutor Habitual

Na maioria dos casos o tomador é o condutor habitual da viatura. Quando isso não acontece está-se, muitas vezes, a falar de condutores mais jovens com apólices em nome de um parente. Essas apólices são tendencialmente de rentabilidade inferior e por isso com provável influência nas taxas de retenção. A **idade** e a **antiguidade da carta de condução** foram os fatores recolhidos neste âmbito.

2.2.3 Apólice

Prémio, indicadores de rentabilidade e formas de pagamento constituem os três grandes grupos de fatores de importância no que se refere ao contrato de seguro. Foram considerados a **variação no prémio** (valores absolutos e relativos), **prémios propostos** (anterior e a vigorar em caso de renovação**)**, **contagem de sinistros** (recolha do número de sinistros reportados em dois períodos distintos, ano anterior e três anos anteriores), **frequência de pagamento** e **forma de pagamento**, com especial foco em pagamentos mensais por débito em conta como meio de promoção da retenção (os agravamentos de prémio são cobrados em montantes menores e sem qualquer ação por parte do cliente).

É importante referir que foram incluídas as anulações ocorridas durante o termo da apólice pelas elevadas taxas de abandono verificadas em contratos com fracionamento

trimestral, antes de paga a totalidade das prestações. Considerar apenas as anulações à data do aniversário seria sobrestimar a retenção desses contratos.

A importância do histórico de sinistralidade para a modelação das taxas de retenção reside no facto de o mesmo estar diretamente relacionado com o prémio devido. Por outro lado, condutores com mais sinistros tendem a não abandonar a companhia pela dificuldade que terão em encontrar melhores propostas no mercado.

A informação que nos dá a contagem de sinistros no ano anterior difere da que podemos retirar da sua contagem nos três anos anteriores. O número de sinistros reportados durante o ano anterior explica com maior exatidão a propensão a renovar o contrato. Isto porque são, em todos os casos, sinistros com proximidade temporal à data da decisão. Por outro lado, a volatilidade que resulta da análise dos sinistros nos últimos três anos é visivelmente menor pois permite distinguir entre sinistros que são casos pontuais ou que acontecem com frequência. Assim sendo, são variáveis que se complementam, mas que prevemos virem a apresentar uma elevada correlação. Esse facto poderá constituir um fator de exclusão de uma delas no processo de obtenção do modelo linear generalizado (para mais informações sobre este tema considere a secção 6.1.1).

Ainda sobre os prémios propostos ao cliente, é esperado que apresentem correlação não nula com os fatores tarifários presentes nesta seleção, casos esses que virão a ser alvo de especial análise.

O **mês da renovação** faz também parte do conjunto de variáveis independentes pois não excluímos que possa existir uma tendência sazonal.

2.2.4 Canais de Distribuição

O **agente mediador** é o principal canal de distribuição nesta linha. Assim sendo, várias medidas de classificação do mesmo foram selecionadas, nomeadamente níveis de **rentabilidade de negócio**, **volume de carteira** e **experiência**. Lidando diretamente com o cliente, acreditamos que o agente terá um papel importante no nosso modelo.

A **Delegação**, agrupamento de zonas do país em que cada mediador pertence a uma única delegação consoante a zona onde desenvolve o seu negócio, foi também um fator incluído na nossa amostra.

As **comissões** de angariação e cobrança (em valor relativo) foram igualmente incluídas.

2.2.5 Veículo

Relativamente ao objeto seguro foram recolhidos, **antiguidade, valor em novo, marca, cilindrada, combustível, peso bruto**, presença de **dissuasores de furto** e informação diversa sobre as coberturas contratadas. Variáveis relativas ao veículo são importantes indicadores do tipo de risco e também do poder de compra do cliente. Por outro lado, as **coberturas**, **somas seguras** e **franquias** informam sobre o grau de tolerância ao risco do cliente. Aparentemente, ambos com influência nos níveis de elasticidade do prémio.

2.2.6 Dados Externos

A fonte de dados externos foi o Instituto Nacional de Estatística (INE). A informação recolhida consiste em **taxas de desemprego** e valores de **densidade populacional** por distrito.

3 Metodologia

3.1 Modelos Lineares Generalizados

Os modelos lineares generalizados (McCullagh e Nelder, 1989) resultam, tal como o nome indica, da generalização dos modelos lineares. Permitem quantificar a relação entre a variável resposta (ou dependente) e as variáveis explicativas (ou independentes) e cumprem dois aspetos essenciais:

I. A distribuição da variável resposta pertence à família de dispersão exponencial.

II. Uma transformação da média da variável resposta (necessariamente monotónica e diferenciável) escreve-se como combinação linear dos fatores explicativos.

3.1.1 A Família Exponencial de Distribuições

Uma variável aleatória Y pertence à família de dispersão exponencial se a sua função densidade puder ser escrita da seguinte forma:

$$f_y(y;\theta,\varphi) = \exp\left(\frac{\big(y\theta - b(\theta)\big)}{a(\varphi)} + c(y,\varphi)\right),$$

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

θ é denominado por parâmetro natural e φ por parâmetro de dispersão. A sua especificação, bem como das funções $a(.), b(.)\ e\ c(.)$ depende da distribuição de Y.

Dada uma amostra aleatória de Y, $(y_1, \dots, y_n)$, em que a distribuição de Y pertence à família de dispersão exponencial, a função logarítmica de verosimilhança de $(y_1, \dots, y_n)$ é dada por:

$$l(\theta, \varphi; y_1, \dots, y_n) = \sum_{i=1}^{n} ln(f_y(y_i; \theta_i, \varphi) = \sum_{i=1}^{n} \left[\frac{y_i\theta_i - b(\theta_i)}{a(\varphi)} + c(y_i, \varphi)\right],$$

onde $f_y(y; \theta, \varphi)$ denota a função densidade de Y.

Sabendo que,

$$E\left[\frac{\partial l}{\partial \theta}\right] = 0 \quad \text{e} \quad E\left[\frac{\partial^2 l}{\partial \theta}\right] + E\left[\left(\frac{\partial l}{\partial \theta}\right)^2\right] = 0,$$

é possível provar as seguintes expressões relativas à média e variância de Y:

$$E\lfloor Y \rfloor = b'(\theta) \quad \text{e} \quad var(Y) = a(\varphi)b''(\theta),$$

onde b' e b'' denotam a primeira e segunda derivadas de b, respetivamente (Nelder e Wedderburn, 1972).

Para explicitar a dependência da média, a variância de Y é por vezes escrita da forma:

$$var(Y) = a(\varphi)V(\mu),$$

onde V (μ) denota a função variância e é definida por $V(\mu) = b''(\theta)$.

3.1.2 Preditores Lineares e Funções de Ligação

O preditor linear η define-se por:

$$\eta = \sum_{j=1}^{p} x_j\beta_j$$

onde β_j são os parâmetros escalares e ($x_1, \dots, x_n$) as variáveis explicativas.

Adicionalmente, dada uma função $g(.)$ monotónica e diferenciável, temos que:

$$g(\mu) = \eta = \sum_{j=1}^{p} x_j \beta_j.$$

Portanto, $g(.)$ permite a ligação entre a média e o preditor linear e denomina-se por função de ligação. Define, portanto, a relação entre a média da variável Y e os fatores explicativos ($x_1, \dots, x_n$).

Se $g(\mu) = \theta$ então $g(.)$ é chamada de ligação canónica.

O processo de modelação em MLG implica uma definição prévia da distribuição da variável resposta Y assim com da função de ligação. Para além disso, as observações de Y ($y_1, \dots, y_n$, onde n é a dimensão da amostra) devem ser independentes entre si.

3.1.3 Estimação dos Parâmetros

Nos MLG a estimação dos parâmetros é feita por uso do método da Máxima Verosimilhança. Os coeficientes do preditor linear são obtidos por maximização da função log de verosimilhança. Assim sendo, é necessário recorrer à função inversa de $g(.)$ para expressar a média da variável resposta em função do preditor linear. Obtém-se, dessa forma, que:

$$\mu = g^{-1}(\eta).$$

Derivando $l(\theta, \varphi; y_1, \dots, y_n)$ em ordem a cada parâmetro a estimar e igualando a zero obtêm-se as equações de primeira ordem para a maximização da função. As soluções constituem as estimativas para os parâmetros.

O sistema de equações a resolver é o seguinte:

$$\sum_{i=1}^{n} \left[\frac{y_i - b'(\theta_i)}{a(\varphi)} \right] \frac{\partial \theta_i}{\partial \beta_j} = 0, j = 1, \dots, p,$$

onde p é o número total de parâmetros β_j.

3.1.4 Distribuição da Varável Resposta

Considere-se Y uma variável aleatória. Se Y ~ *Bernoulli* (μ) então,

$$f_Y(y) = (\mu^y)(1-\mu)^{1-y},\ \forall\, y \in \{0,1\}.$$

É possível demonstrar que a distribuição de *Bernoulli* pertence à família exponencial fazendo $\theta = ln\left(\frac{\mu}{1-\mu}\right), \varphi = 1, a(\varphi) = \varphi = 1, b(\theta) = ln\left(1+e^{\theta}\right)$ e $c(y,\varphi) = 0$.

A nossa variável resposta tem a distribuição de *Bernoulli* (μ). A função de ligação escolhida foi a canónica, ou seja, a transformação logística. Ela transforma o preditor linear numa quantidade compreendida entre zero e um que é, de facto, o requerido no nosso caso.

Suponhamos que $g(\mu) = ln\left(\frac{\mu}{1-\mu}\right)$ é função de ligação. Então,

$$\eta = ln\left(\frac{\mu}{1-\mu}\right) \Leftrightarrow \mu = \frac{1}{1+e^{\eta}}.$$

MLG nestas condições denomina-se por Regressão Logística.

3.1.5 Medidas da Qualidade do Ajuste do Modelo

Modelos *Nested*:

A utilização dos procedimentos de teste seguintes requer que os modelos tenham a mesma função de ligação e que a variável resposta tenha a mesma distribuição, isto é, que os modelos sejam *nested*.

Modelo Saturado define-se pelo modelo que contém tantos parâmetros quantas observações. Os valores observados coincidem com os valores ajustados e o ajuste à amostra é perfeito. O sobre ajuste aos dados de treino torna o modelo pouco útil pois virá a desempenhar de forma fraca em dados futuros. No entanto, o modelo saturado é o ideal

para fornecer uma medida de comparação no momento de avaliar a qualidade do nosso modelo.

Considere-se a fórmula da *deviance* (Δ):

$$\Delta = 2\,(\check{l} - \hat{l}),$$

onde $\check{l}$ e $\hat{l}$ denotam as funções log de verosimilhança dos modelos saturado e ajustado, respetivamente.

Analisando a fórmula da *deviance* facilmente se conclui que o objetivo é minimizar o seu valor.

Para distribuições da variável resposta cuja *deviance* é assintoticamente qui-quadrado distribuída, um possível procedimento de teste à qualidade de ajuste do modelo consiste na comparação do seu valor com a distribuição $\chi^2_{(n-p)}$, onde n é a dimensão da amostra e p o número de parâmetros do modelo a testar.

Considere-se a fórmula da *deviance* (Δ) para a distribuição de *Bernoulli*:

$$\Delta = -2\sum_{i=1}^{n}\left\{\widehat{\mu_i} ln\left(\frac{\widehat{\mu_i}}{1-\widehat{\mu_i}}\right) + \ln(1-\widehat{\mu_i})\right\}.$$

A *deviance* é inapropriada para analisar a qualidade do ajuste de uma regressão logística. Os valores observados y_i não surgem na fórmula senão por meio de $\widehat{\mu_i}$. Para além disso, Δ não é assintoticamente qui-quadrado distribuído (Collett 2003).

Considere-se a estatística qui-quadrado de Pearson:

$$\sum_{i=1}^{n}\left[\frac{y_i - \widehat{\mu_i}}{\widehat{\mu_i}(1-\widehat{\mu_i})}\right].$$

Apesar de mais informativa que a *deviance*, a medida também não é fiável pelo facto da aproximação ao qui-quadrado não ser satisfatória.

Para testar restrições aos parâmetros o *Likelihood Ratio Test* é aplicável.

Os resultados que optámos por apresentar referem-se ao teste de Hosmer-Lemeshow. O teste é feito por análise das diferenças entre as probabilidades estimadas e as observadas. Baseia-se na divisão da amostra em g grupos segundo as suas probabilidades ajustadas, devidamente ordenadas da menor para a maior. Hosmer e Lemeshow propõem a utilização de $g = 10$, e portanto um teste aos decis de risco. A hipótese nula (H_0) é que o modelo se ajusta à amostra e a estatística de teste é definida por:

$$\hat{C} = \sum_{k=1}^{g} \frac{(o_k - n'_k \overline{\pi_k})^2}{n'_k \pi_k (1 - \pi_k)},$$

onde:

n'_k é o número de observações no k-ésimo grupo;

$\overline{\pi_k} = \sum_{j=1}^{n'_k} \frac{\widehat{\pi_j}}{n'_k}$, sendo $\widehat{\pi_j}$ a probabilidade estimada para a observação j;

$o_k = \sum_{j=1}^{n'_k} y_j$, sendo y_j o valor da j-ésima observação.

Assumindo que o modelo está corretamente especificado, $\hat{C}$ segue assintoticamente uma distribuição qui-quadrado com $g - 2$ graus de liberdade (Hosmer e Lemeshow, 2000).

Por outro lado, o processo de escolha do modelo final passa por testes à nulidade dos parâmetros escalares $\beta_j, \forall\, j \in \{1, \dots, p\}$, onde p é o número total de parâmetros. O Teste de Wald é um dos utilizados para esse efeito. Sob a hipótese nula H_0: $\beta_i = 0$ (com $i \in \{1, \dots, p\}$ e p número total de parâmetros escalares) a estatística de teste segue assintoticamente uma distribuição $Z \sim N(0,1)$, e define-se por:

$$W_i = \frac{\widehat{\beta_i}}{se(\widehat{\beta_i})},$$

onde $\hat{\beta}_i$ é a estimativa de máxima verosimilhança para o parâmetro, e $se(\hat{\beta}_i)$ o seu desvio padrão.

Modelos *Non-Nested*:

Para comparar modelos que não verificam as condições de modelos *nested*, ou como indicador complementar, é frequentemente utilizado o *Akaike Information Criterion* (Akaike, 1974). O critério define-se por:

$$AIC = -2l + 2p,$$

onde l denota a função log de verosimilhança do modelo e p o número total de parâmetros estimados.

O termo que depende do número de parâmetros estimados representa a inclusão de uma penalidade por cada um introduzido no modelo. Um novo parâmetro é considerado significativo se originar o aumento de l em mais do que 1.

Entre dois modelos prevalece o que tiver um menor valor de AIC.

3.2 Métodos *Tree-based*

Os métodos *tree-based* (Lantz, 2013) são métodos de *Machine Learning* supervisionados e não-paramétricos e apresentam-se graficamente como árvores de decisão.

Árvores de decisão são estruturas de dados formadas por um conjunto de elementos que armazena regras, elementos esses que têm o nome de nós. O nó raiz é o ponto de partida e liga-se a outros, chamados nós filhos. Esses, por sua vez, podem ligar-se a outros nós filhos.

Os nós que não possuem filhos são denominados por nós terminais e representam a decisão a ser tomada.

Estes métodos consistem, essencialmente, na obtenção recursiva de partições de um espaço de observações nas quais um modelo de classificação simples é aprendido.

Iremos considerar apenas cortes ortogonais, isto é, em cada corte será tida em consideração apenas uma característica classificativa. Este facto previne a construção de

decisões de complexidade mais elevada, constituindo uma limitação a nível preditivo, mas também um ganho em eficiência computacional.

O critério de paragem não é inevitavelmente a atribuição da mesma classe a todos os exemplos, ou a inexistência de características de classificação capazes de distinguir os restantes exemplos. Aliás, permitir que uma árvore atinja o seu comprimento máximo pode levar ao sobre ajuste do modelo aos dados e, consequentemente, à obtenção de um modelo fraco quando a ser utilizado noutros grupos de dados. Assim sendo, dois métodos são possíveis de ser aplicados.

Pre-pruning permite o crescimento da árvore até um ponto previamente definido, que pode ter como base um número limite de decisões ou de observações no nó terminal.

Post-pruning permite o crescimento da árvore até ao seu ponto máximo. A redução do tamanho acontece depois e resulta da avaliação das taxas de erro das estimativas.

Apesar do método de *pré-pruning* evitar partições desnecessárias, *post-pruning* apresenta vantagens a nível preditivo por permitir a avaliação de diferentes níveis de crescimento, cujos índices de *performance* são difíceis de prever.

Ponto ótimo de corte define-se pela regra classificativa que maximiza a pureza das sub-regiões de dados obtidas. A pureza determina a homogeneidade na classificação dos exemplos, ou seja, uma sub-região é tanto mais homogénea quanto menor for a diversidade de exemplos nela contida. Existem vários critérios de impureza, mas todos se baseiam no cálculo do ganho de informação (isto é, índice de pureza do espaço) associado à regra classificativa que origina o corte.

$$InfoGain(R, R_e, R_d) = H(R) - \frac{\#R_e * H(R_e) + \#R_d * H(R_d)}{\#R},$$

onde H é a impureza da região, R a região atual, R_e a sub-região da esquerda, R_d a sub-região da direita e # a quantidade de observações numa dada região.

O índice Gini é um dos critérios de impureza mais utilizado em problemas de classificação.

$$gini(R) = \sum_{i=1}^{c} p_i\,(1 - p_i),$$

onde c representa o número total de classes presente na sub-região em questão e p a proporção de observações pertencente à classe i.

As árvores de decisão são bastante utilizadas em problemas de classificação por serem de fácil interpretação e também por permitirem a inclusão de todo o tipo de variáveis, isto é, numéricas, nominais e até variáveis em que parte dos valores é desconhecido.

No entanto, uma só árvore é, por vezes, insuficiente para atingir o nível de desempenho preditivo desejado. Também a instabilidade[1] associada à utilização de uma só árvore constitui uma preocupação. Os métodos *Ensemble* são métodos baseados na ideia de que a combinação de diversos algoritmos mais fracos fará resultar um mais forte (onde fraco e forte se refere à exatidão das estimativas obtidas).

Esses métodos são bastante utilizados em problemas de classificação. Um modelo constituído por várias árvores de decisão permitirá não só aumentar o desempenho preditivo como também contornar os problemas de instabilidade referidos.

3.3 Métodos *Ensemble* para Problemas de Classificação

Existem vários métodos *Ensemble* e a escolha de qual se adequa ao nosso problema depende, uma vez mais, das características da amostra de dados. A nossa análise vai basear-se nos métodos *Boosting* e *Random Forest*.

[1] Instabilidade refere-se ao facto de pequenas alterações na amostra de treino originarem grandes alterações na lógica inerente às decisões implementadas pelo algoritmo.

3.3.1 *Boosting*

Boosting (Schapire e Freund, 2012) usa combinações de modelos de desempenho preditivo insuficiente, treinados em conjuntos de dados obtidos por reamostragem da amostra de treino. O processo de reamostragem é gerado especificamente para construção de algoritmos complementares, baseando-se nos erros da iteração anterior. Assim sendo, para além de importantes ferramentas para a diminuição das taxas de classificação errada, estes métodos são considerados de elevada eficiência, na medida em que vários modelos com níveis de desempenho insuficiente são o que basta para obter o modelo que se ajusta à amostra da forma pretendida.

No que diz respeito à combinação dos vários modelos obtidos, *Boosting* fá-lo por ponderação dos níveis de *performance*, ou seja, por forma a minimizar uma determinada função de custo.

O processo iterativo inicia-se com as estimativas $\hat{f}(x_i) = 0$ e os resíduos $r_i = y_i, \forall\, i \in \{1, \dots, n\}$ e n dimensão da amostra. O número total de iterações B é fixado. Posteriormente, é obtido o número de iterações ótimo, que será aquele a partir do qual o erro começa a aumentar para a amostra de teste devido ao sobre ajuste aos dados de treino. Esse é o número de iterações de deverá ser utilizado para obter as estimativas e avaliar a exatidão do modelo.

Para b = 1, …, B o algoritmo consiste no seguinte:

1. Ajuste de uma árvore $\hat{f}^b$, de comprimento pré-definido, aos resíduos r_i. O parâmetro ν $(0 < \nu \leq 1)$ representa a taxa de aprendizagem.

2. Atualização das estimativas:

$\hat{f}(x_i) \leftarrow \hat{f}(x_i) + \nu\lambda_b \hat{f}^b(x_i)$, $\forall\, i \in \{1, \dots, n\}$ e para λ_b escolhido de forma a minimizar a função de custo.

3. Atualização dos resíduos:

$$r_i \leftarrow r_i + \nu\lambda_b \hat{f}^b(x_i)\,, \qquad \forall\, i \in \{1, \dots, n\}$$

O modelo final é dado por:

$$\hat{f}(x_i) = \sum_{b=1}^{B} \nu\lambda_b \hat{f}^b(x_i) \ , \forall\, i \in \{1, \dots, n\}$$

Gradient Boosting (Breiman, 1997) generaliza a técnica de *boosting* admitindo a otimização de uma função de custo diferenciável. É, portanto, um processo de otimização por gradiente descendente. É habitualmente utilizado com árvores de decisão, principalmente as CART, árvores de classificação e regressão.

Friedman altera, posteriormente, o algoritmo, considerando que λ_b varia para cada nó terminal da árvore (ou seja, λ_{bj} em que j varia entre 1 e o número total de nós terminais), dando-lhe o nome de TreeBoost (Friedman, 2001).

3.3.2 *Random Forest*

Random Forest (Breiman, 2001) é também um método *Ensemble*. Neste caso, as árvores de decisão são *unpruned* e existe uma seleção aleatória de características classificativas que farão parte da construção do algoritmo. Esse facto dá diversidade ao modelo e evita também o sobre ajuste aos dados de treino. Também permite aumentar a eficiência computacional, facto especialmente relevante para amostras de grandes dimensões e elevado número de variáveis de classificação. O número de características a selecionar é definido previamente e não deve ser elevado.

O processo de reamostragem é feito por *bootstrap* e as estimativas são, por fim, combinadas por maioria de voto.

4 Especificidades da Amostra

Desequilíbrio entre Classes

Problemas de desequilíbrio entre classes acontecem quando um dos níveis da variável resposta constitui apenas uma pequena minoria das observações presentes na amostra. Um menor volume de dados dificulta a aprendizagem da máquina e as taxas de erro elevam-se para a classe com menor representação.

Algoritmos objetivando uma redução da taxa de erro de classificação não permitem solucionar o problema e dá-se lugar à aplicação de técnicas com especial foco nos erros da classe em minoria. Para exemplificar, um modelo em que 99% das observações pertencem à classe 0 e cuja previsão seja sempre 0 terá uma taxa de erro de apenas 1% mas não terá, efetivamente, utilidade nenhuma.

As abordagens possíveis para minorar o problema são diversas. *Cost Sensitive Learning* consiste na definição de custos a aplicar a erros de classes distintas. Em amostras desequilibradas irá interessar aplicar um maior custo a erros da classe em minoria.

Como exemplo dessa técnica pode referir-se a *Weighted Random Forest*, em que os pesos pré-definidos são utilizados tanto no processo de crescimento da árvore (para determinação dos pontos de corte) como nos nós terminais, aquando da obtenção da estimativa para a classificação dos exemplos em causa.

O uso de técnicas de amostragem é outra forma de contornar os problemas de desequilíbrio entre classes. *Down-sampling* é um exemplo, e consiste na redução da classe em maioria. De forma semelhante, *Over-sampling* baseia-se no aumento da classe em minoria. Neste âmbito, Chen et al., 2004 testaram a aplicação das *Balanced Random Forests*, técnica definida pela combinação entre métodos *Ensemble* e técnicas de *down-sampling*.

Mais especificamente, eles mostraram que, tanto as BRF como as WRF apresentam resultados bastante satisfatórios quando comparados com técnicas como o C4.5 (Quinlan,

1993), 1-NN (Bremner et al., 2005), Standard RIPPER (Cohen, 1996), *One-sided sampling* (Kubat e Matwin, 1997), SHRINK (Kubat et al., 1998), SMOTE (Chawla et al., 2002) e SMOTEboost (Chawla et al., 2003). Entre os dois (BRF e WRF) não foi encontrado um que prevalecesse ao nível do desempenho preditivo. Foi, no entanto, apontada uma desvantagem ao método WRF. Um erro na classe em maioria terá menor impacto no modelo do que aquele que deveria ter. Esta foi a razão que nos levou a decidir ajustar uma BRF, já que não foi possível testar ambas.

5 Desempenho Preditivo

5.1 *K-fold cross validation* e Estimativa de Erro *out-of-bag*

Um teste eficaz ao modelo implica o uso de uma amostra representativa do universo em estudo e diferente da de treino. A técnica mais comum é conhecido como método *Holdout* e consiste em dividir aleatoriamente a amostra total em duas partes, uma para treino e outra para teste. São habitualmente utilizados 80% e 20% dos dados para os conjuntos de treino e de teste, respetivamente.

Existem, no entanto, técnicas baseadas em processos de reamostragem que visam melhorar a *performance* preditiva do modelo final: OOB (*out of bag*) e CV (*cross-validation*) são exemplos dessas técnicas.

Ambas visam estimar o número de iterações a partir do qual o erro de validação começa a aumentar devido ao sobre ajuste do modelo aos dados de treino. O processo de validação por utilização destas técnicas ocorre durante o ajuste do modelo, o que significa que, posteriormente, o modelo é ainda testado na amostra de teste, isto é, na partição de 20% obtida pelo método *Holdout*.

O que diferencia estas duas técnicas é a forma como é criada a subamostra de validação. A estimativa de erro *out-of-bag* (OOB) é obtida a partir de uma amostra criada por *bootstrap aggregating*, ou *bagging* (Breiman, 1994). Para cada observação x_i ($\forall\, i \in \{1, \dots, n\}$ e n dimensão da amostra) da amostra de treino o erro OOB é a média dos erros estimados para cada árvore cuja amostra obtida por *bootstrap* não contém x_i.

Cross Validation tem por base a técnica *Repeated Holdout* (Lantz, 2013), diferindo no facto de incluir uma divisão da amostra em k partições que substitui a produção consecutiva de amostras. A ideia de particionar a amostra produz resultados vantajosos no processo de construção e validação do modelo, pois evita obter amostras de treino e de teste semelhantes a cada iteração.

Por exemplo, se k = 10, a amostra terá 10 partições. Em cada uma das dez iterações o modelo é ajustado em 90% da amostra e validado nos restantes 10%. As amostras de validação são precisamente as partições obtidas inicialmente.

Ao contrário do que acontece na técnica CV, no cálculo da estimativa de erro *out-of-bag* a amostra de validação não é independente da de treino, possibilitando a utilização da totalidade das observações para treinar o modelo. No entanto, OOB subestima frequentemente a performance do modelo e consequentemente o número ótimo de iterações.

5.2 Medidas de Desempenho para Métodos de Classificação

As matrizes de confusão são uma ferramenta útil na análise de desempenho de um modelo de classificação. A partir de um limiar pré-definido, cada estimativa de probabilidade é classificada como "positiva" ou "negativa" consoante a probabilidade se encontre acima ou abaixo desse limite. Esse valor é conhecido como ponto de corte.

A dimensão da tabela depende do número de níveis do fator a ser explicado. No nosso caso a tabela apresenta dimensão 2x2 pelo facto de existirem apenas duas classes, e o evento "positivo" será a renovação do contrato. A sua estrutura é a seguinte:

		Estimado	
Observado		0	1
	0	VN	FP
	1	FN	VP

Tabela 5.1: Matriz de Confusão.

Como legenda da tabela 3.1 temos:

- VN (verdadeiros negativos): corretamente classificados como "negativo";

- FN (falsos negativos): incorretamente classificados como "negativo";
- FP (falsos positivos): incorretamente classificados como "positivo";
- VP (verdadeiros positivos): corretamente classificados como "positivo".

Dado P número total de eventos positivos (VP + FP) e N número total de eventos negativos (VN + FN), *accuracy* define a taxa de acerto e é dada por:

$$Accuracy = (VP + VN)/(P + N),$$

sendo, naturalmente, a taxa de erro dada por $1 - Accuracy$.

A estatística $Kappa$ indica o quão legítimo é o valor da *accuracy* pois mede o afastamento entre observações atuais e esperadas, fruto do acaso.

$$Kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

onde $Pr(a)$ é a proporção atual de concordância e $Pr(e)$ a esperada. Mais especificamente:

$$Pr(a) = ACC$$

$$Pr(e) = \frac{VN + FP}{P + N} * \frac{VN + FN}{P + N} + \frac{FN + VP}{P + N} * \frac{FP + VP}{P + N}.$$

Pretende-se que $Kappa$ se aproxime o máximo possível de 1.

Outras métricas incluem:

$$Sensibilidade = \frac{VP}{(VP + FN)},$$

$$Especificidade = \frac{VN}{(VN + FP)},$$

$$Weighted\ Accuracy = \beta * Sensibilidade + (1 - \beta) * Especificidade,$$

em que $\beta \in]0,1[$.

A utilização das medidas mencionadas anteriormente não é possível sem antes ser selecionado o ponto de corte. Para isso, é habitualmente usada a curva ROC (*Receiver*

Operating Characteristic Curve). Para cada ponto de corte entre 0 e 1, a curva faz corresponder valores de Sensibilidade e Especificidade. A figura 5.1 representa um exemplo da curva ROC.

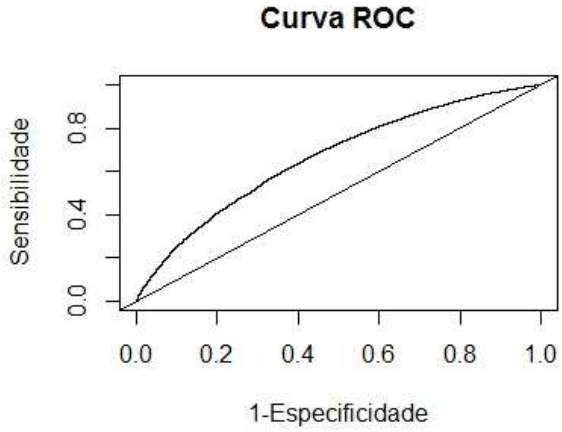

Figura 5.1: Curva ROC (exemplo).

A sensibilidade e a especificidade medem, respetivamente, a proporção de exemplos positivos e negativos que foram bem classificados. Ambas variam entre 0 e 1 com valores próximos de 1 a representar, logicamente, modelos de melhor *performance*. No entanto, a otimização conjunta destas duas métricas depende de um inevitável *tradeoff*. Por exemplo, um maior número de casos verdadeiros positivos significa que, perante uma determinada taxa de erro, o número de verdadeiros negativos será menor. O equilíbrio ótimo baseia-se nos objetivos do problema em mãos, e o custo do erro associado a cada nível da variável resposta é o ponto de partida para a definição desse mesmo equilíbrio.

Portanto, em qualquer curva ROC, o ponto de corte ótimo é aquele que se aproxima mais do par ordenado (0,1), o ponto de sensibilidade e especificidade iguais a 1. AUC (*Area Under the Curve*) permite quantificar a rapidez com que é atingido esse ponto. Um modelo cuja habilidade preditiva é equivalente ao acaso tem AUC igual a 0,5 e a curva ROC coincide com a reta que na figura 5.1 faz um ângulo de 45º com a origem. AUC tem como máximo 1 e o objetivo é maximizar o seu valor.

6 Aplicação Prática

Nesta secção serão apresentados os três modelos obtidos e especificadas algumas etapas da sua construção. O *software* utilizado foi o R. Recorreu-se também ao *software* Emblem, nomeadamente para suporte ao ajuste da regressão logística.

O detalhe ao nível dos *outputs* apresentados é reduzido por questões de confidencialidade.

6.1 Regressão Logística

6.1.1 Fatores Correlacionados

A inclusão de variáveis com elevada correlação entre si leva ao aumento da volatilidade da estimativa do vetor de parâmetros. Para além disso, dadas duas variáveis colineares, ou quase colineares, testes de hipótese sobre a significância de uma delas pressupondo a presença da outra, evidenciam que a variável não é efetivamente significativa para o modelo, quando a realidade pode ser diversa. Assim sendo, o uso de uma medida de correlação, como por exemplo *Cramér's V*, permite eliminar *a priori* os fatores que seriam prejudiciais ao ajuste correto do modelo.

O primeiro passo no ajuste da regressão consistiu em eliminar variáveis com o indicador *Cramér's V* acima de 0,5 (a matriz de correlações foi obtida por utilização do Emblem, o qual procede à transformação de todas as variáveis em fatores). As escolhas apresentadas são as que foram feitas com base numa análise marginal. Um teste à inclusão das variáveis excluídas nesta fase é obviamente necessário para obter a garantia de que estamos a escolher as que realmente têm uma maior significância para o modelo.

Os valores de prémio, anterior e a vigorar em caso de renovação, e o seu diferencial são obviamente bastante correlacionados. Optámos por testar o prémio proposto para renovação e a variação, tanto percentual como em valor absoluto.

O cálculo de *Cramér's V* para as variáveis idade do tomador e idade do condutor habitual também revelou um valor elevado (0,864). A razão é o facto de serem poucos os casos em que o condutor habitual do veículo difere do tomador. Foi, portanto, selecionado para teste a idade do tomador.

A mesma análise para as variáveis Distrito, Delegação, taxa de desemprego e densidade populacional devolveu valores também acima de 0,5. Neste caso foi escolhido o Distrito pois considerámos ser de todas a mais adequada. Por exemplo, a granularidade associada às taxas de desemprego quando a explicar a retenção poderia não coincidir com a granularidade necessária quando considerada a influência da densidade populacional nessas mesmas taxas. Quanto à Delegação a justificação foi precisamente a inversa, a granularidade é maior do que a que existe ao nível do Distrito, significando um aumento da complexidade do modelo com possibilidade de sobre ajuste aos dados de treino. Tal como referido anteriormente, não se dispensa o teste à inclusão das variáveis excluídas nesta primeira fase.

Quanto ao histórico de sinistralidade, o valor de *Cramér's V* obtido para as variáveis número de sinistros reportados no ano anterior e número de sinistros reportados nos três anos anteriores foi de 0,439. Testes à significância da inclusão de cada uma delas, bem como de ambas acompanhada de uma interação entre si, permitiram excluir o número de sinistros reportados no ano anterior.

A leitura das próximas subsecções relativas à regressão (subsecções 6.1.2, 6.1.3 e 6.1.4) deverá ser acompanhada por uma consulta do Anexo A o qual inclui, entre outros, informação relativa às variáveis incluídas e o *output* de R relativo ao sumário do modelo.

6.1.2 Ajuste de Polinómios Ortogonais

De seguida serão apresentados gráficos de taxas de retenção observadas por fator, visando ilustrar decisões tomadas relativamente ao tratamento das variáveis contínuas.

Considere-se a figura 6.1 que mostra a evolução da taxa de retenção como função da variação no prémio.

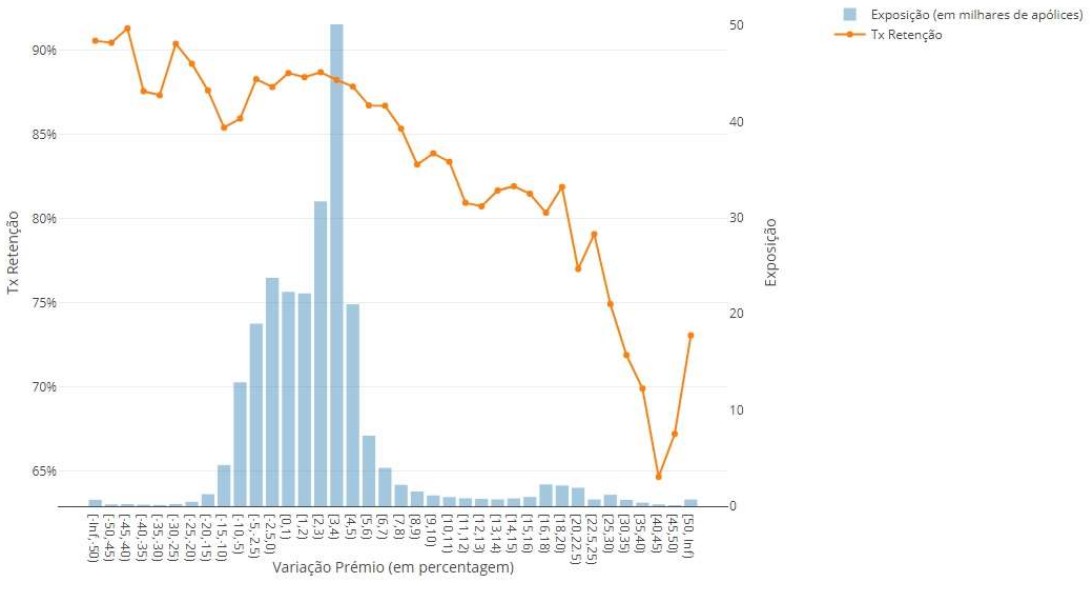

Figura 6.1: Taxas de retenção observadas por nível de variação de prémio.

As taxas de retenção apresentaram, tal como esperado, uma tendência oposta à variação percentual no prémio. No entanto, o seu decréscimo não é uniforme. O declive é visivelmente maior para variações acima de 18%. Essa diferença deve-se ao facto de apólices com prémios mais elevados assumirem, nesses casos, um importante papel na diminuição das taxas de retenção. Recorrendo ao software Emblem, foram ajustados dois polinómios de primeiro grau, com declives distintos. Foi também incluída uma interação entre um indicador de presença de pelo menos uma cobertura de danos próprios e a variação percentual no prémio, já que a inclusão de danos próprios eleva o prémio.

A figura 6.2 apresenta o mesmo tipo de análise, mas para o valor de prémio da apólice.

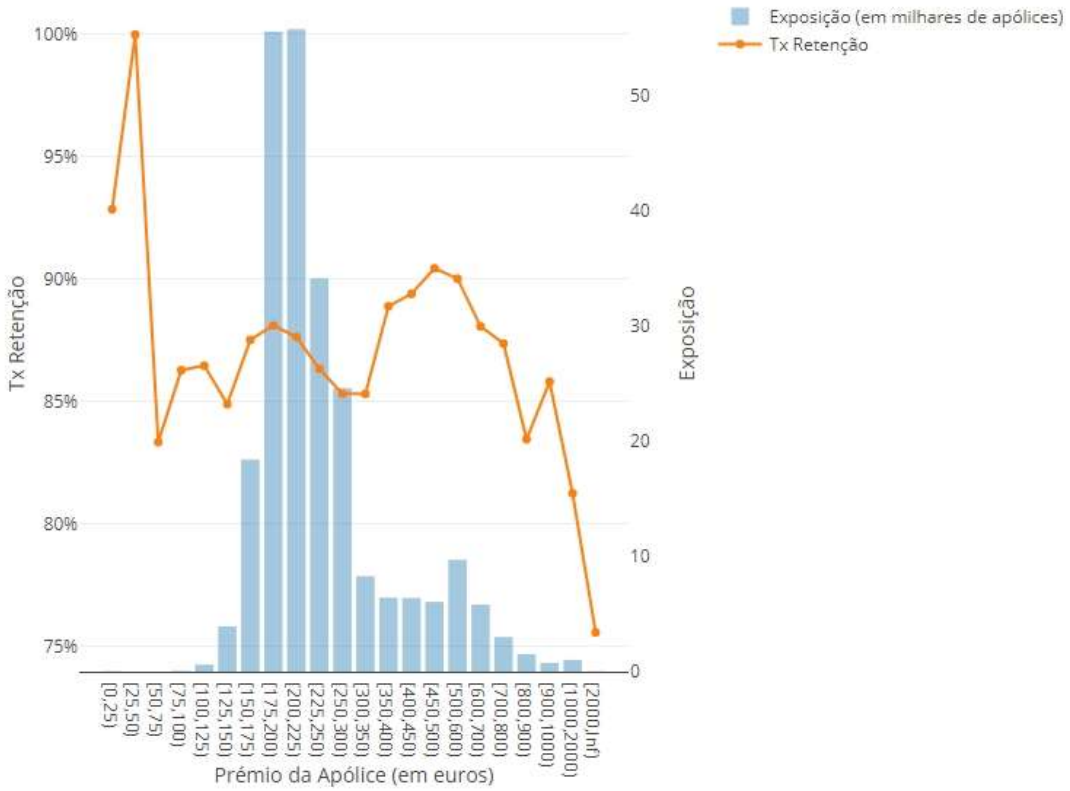

Figura 6.2: Taxas de retenção observadas por nível de valor do prémio da apólice.

A relação que existe entre o prémio da apólice e as taxas de retenção não é, tal como esperado, muito explícita. Com o auxílio do software Emblem, foi possível ajustar polinómios ortogonais estatisticamente significativos para o modelo.

Foi também testada uma interação entre o prémio da apólice e um fator de categorias de desconto, já que sabíamos haver bandas de valor de prémio onde a aplicação de desconto é mais comum e que por isso se associam a taxas de retenção mais elevadas.

O diferencial no prémio que resulta da aplicação de agravamentos e/ou descontos é parte da variação anual. Os descontos advêm, por exemplo, de campanhas promocionais ou de *cross-selling*, e os agravamentos relacionam-se principalmente com a ocorrência de sinistro.

O gráfico seguinte sugere a sua inclusão no modelo:

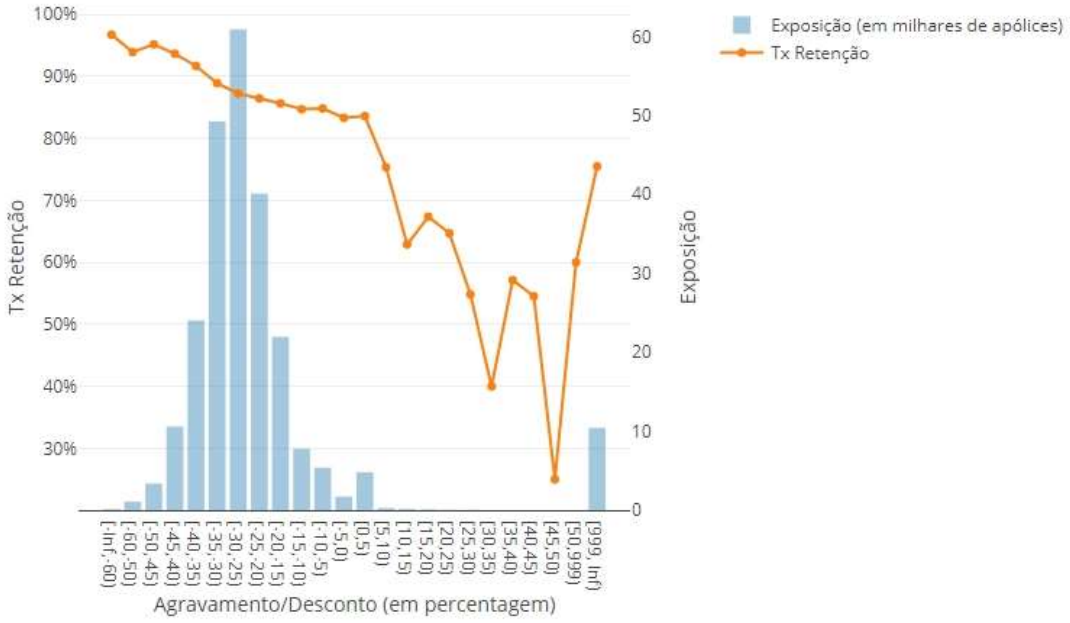

Figura 6.3: Taxas de retenção observadas por nível de variação do prémio por aplicação de agravamentos e/ou descontos.[2]

Uma vez que os níveis de agravamento não apresentam exposição suficiente decidiu-se agrupá-los e construir uma variável com bandas de desconto a variar entre 20 e 45 %. Foi ajustado um polinómio ortogonal de segundo grau e obteve-se significância estatística. Essa variável foi também utilizada para explicar a interação do prémio com o nível de desconto.

A diferença na taxa de retenção que se verifica na última banda corresponde às observações com valor desconhecido. Para modelar esta questão foi criada uma variável binária que diferencia esses casos. É possível verificar (por observação do sumário do modelo contido na secção A.2 do Anexo A) que o fator é estatisticamente significativo.

Para as restantes variáveis contínuas recorreu-se, tal como anteriormente, ao software Emblem, o qual permite a análise das taxas de retenção ao longo da reta real dado que as variáveis já selecionadas para inclusão no modelo tomam os valores correspondentes aos seus níveis base.

[2] A última banda de valores inclui apenas os casos de valor desconhecido.

6.1.3 Interações entre Fatores

O gráfico 6.4 apresenta a taxa de retenção observada por número de anos de antiguidade do tomador na companhia.

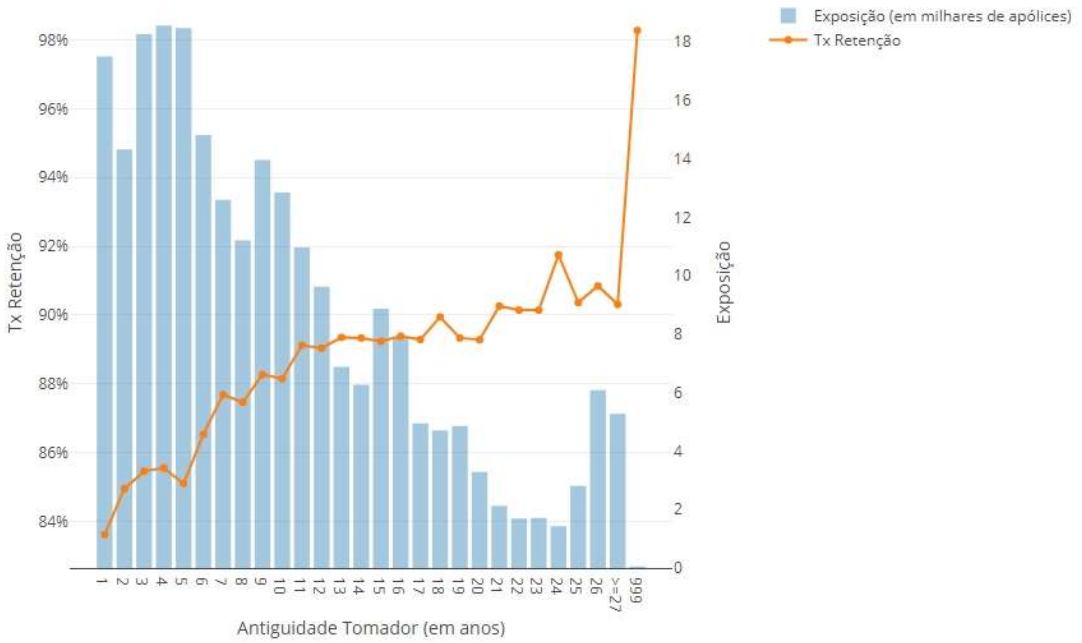

Figura 6.4: Taxas de retenção observadas por nível de antiguidade do tomador (em anos).[3]

Tal com previsto, a tendência é crescente. O gráfico apresenta, no entanto, um abrandamento na velocidade de crescimento.

Considere-se agora o mesmo tipo de análise para a contagem de sinistros nos três anos anteriores.

[3] O valor 999 refere-se aos casos de valor desconhecido.

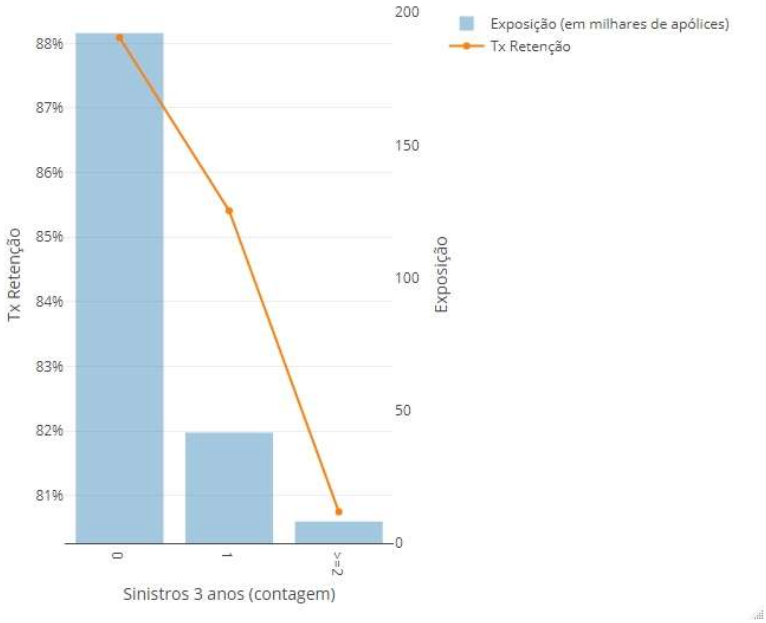

Figura 6.5: Taxas de retenção observadas por nível relativo ao número de sinistros reportados nos três anos anteriores.

É possível verificar que a retenção diminui com o aumento do número de sinistros, provavelmente devido aos agravamentos no prémio. Sabendo que um maior período de exposição ao risco aumenta a probabilidade de ocorrência de sinistro, testámos uma interação entre a antiguidade do tomador e o histórico de sinistros. Concluímos, por fim que, a taxa de sinistralidade contribui para o aumento do número de anulações entre os 5 e os 11 anos de antiguidade.

Para além disso, a diferença no valor da taxa de retenção que se observa para clientes com antiguidade 1 (correspondente à segunda renovação do seu contrato) levou-nos a incluir no modelo uma variável binária que toma o valor "Sim" se o cliente vai renovar o seu contrato pela segunda vez e "Não" caso contrário. A variável é estatisticamente significativa e a estimativa do parâmetro é positiva, sendo o nível base o "Não". Tendo em conta que numa segunda renovação o número de sinistros reportados é na grande maioria dos casos pequeno, ou mesmo nulo, torna-se fácil interpretar o resultado obtido. É importante referir que foi decidido excluir as apólices em renovação pela primeira vez (o habitualmente denominado Novo Negócio) por serem casos com pouco tempo de exposição ao risco e cujas métricas de rentabilidade são frequentemente analisadas à parte da restante carteira.

Outra importante interação incluída foi entre a antiguidade do veículo e um indicador de presença da cobertura de colisão, uma vez que a contratação da mesma está associada a veículos mais novos.

6.1.4 Avaliação do Modelo

Para avaliar o modelo, do ponto de vista global, foi utilizado o teste de Hosmer-Lemeshow (ver secção 3.1.5).

Teste de Hosmer-Lemeshow

O valor obtido para a estatística de teste foi de 13,123 que compara com um qui-quadrado com 8 graus de liberdade devolvendo um *p-value* igual a 0,1077. Para um nível de significância de 5% os dados mostram evidência que o modelo se ajusta. O *output* dos testes obtidos por utilização do R encontra-se disponível na secção A.3 do Anexo A.

O modelo apresenta uma exatidão aceitável, mas não a desejada. Há um grande desequilíbrio na distribuição das classes. As observações pertencentes à classe negativa (se negativa denotar as anulações) constituem uma minoria da total dimensão da amostra (a taxa de abandono observada na amostra de treino é de aproximadamente 12,6%). Este facto é responsável pelo aumento da taxa de erro na classe com menor exposição.

Continuando numa avaliação global do modelo, as tendências esperadas foram efetivamente as verificadas. Analisando algumas das que considerámos menos óbvias, podemos começar por referir o fato dos efeitos do fracionamento mensal e trimestral serem de sinal contrário. Isso explica-se pelas anulações verificadas antes de pagas as quatro prestações, tal como referido anteriormente na secção 2.2.3. A escolha de fracionamento mensal do prémio obriga ao pagamento por débito em conta e por isso nos surgem tendências de sinal contrário.

Outra situação é a existência de estimativas de sinal igual para os indicadores de presença de coberturas de danos próprios e da cobertura de colisão (note-se que o nível base é para uma delas "presente" e para outra "não presente"). Seria de esperar que a cobertura de colisão estivesse presente na grande maioria dos casos em que se contrata danos próprios. No entanto, a figura seguinte mostra que a exposição é efetivamente diferente.

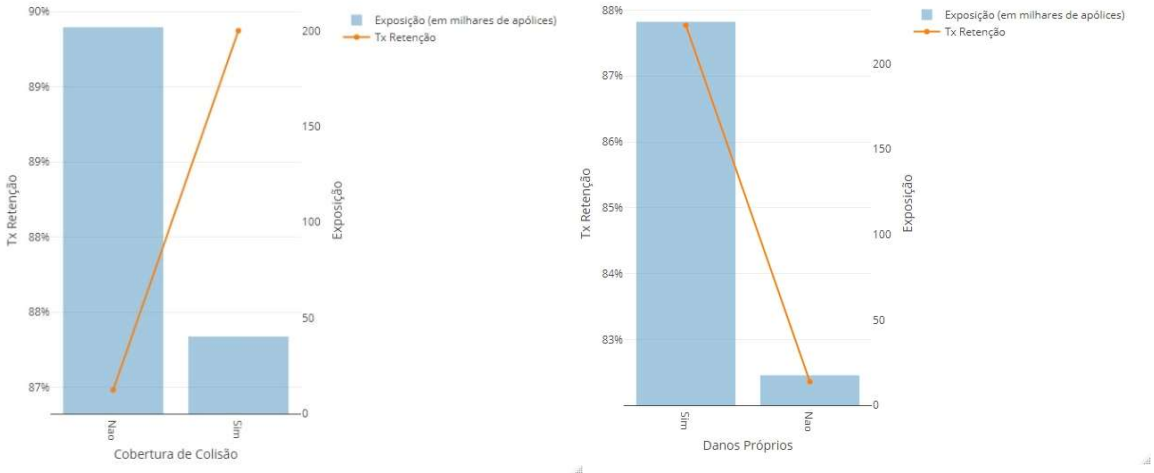

Figura 6.6: Taxas de retenção observadas por nível dos indicadores de Danos Próprios e Cobertura de Colisão.

A razão é o facto da cobertura Quebra de Vidros estar incluída no grupo dos danos próprios e ser muitas vezes contratada com apenas a responsabilidade civil obrigatória e a assistência em viagem.

O passo seguinte poderia passar por uma melhoria ao modelo logit. No entanto, a utilização das metodologias que apresentámos na secção 3.3 permite minimizar o problema do desequilíbrio entre classes sendo, portanto, a melhor escolha nesta fase do projeto.

6.2 Modelos de *Machine Learning*

Para a modelação das taxas de retenção utilizando *Machine Learning* foi incluída a totalidade das variáveis recolhidas inicialmente. A elevada correlação verificada entre algumas delas não constituiu problema. Na secção B.1 do Anexo B é possível consultar a listagem de variáveis consideradas, tanto para as *Gradient Boosted Trees* como para a *Balanced Random Forest*.

6.2.1 *Gradient Boosted Trees*

Para a construção de um algoritmo de *Gradient Boosting* aplicado a árvores de decisão foi utilizado o package "Generalized Boosted Regression Models" (gbm) *do software* R, com base no qual se pode aplicar o algoritmo *TreeBoost* (Friedman, 2001).

Ao contrário do que acontece com a utilização de uma única árvore de decisão, o método GBM é resistente ao sobre ajuste. Neste âmbito, as duas diferentes técnicas descritas na secção 5.1 são aplicáveis, OOB (*out of bag*) e CV (*cross-validation*).

Para a obtenção das *Gradient Boosted Trees* utilizámos a técnica CV. O modelo foi treinado com um máximo de 1000 árvores e depois obtido o número ótimo de iterações a utilizar no cálculo dos valores estimados. Para obter mais esclarecimentos sobre os restantes parâmetros usados poderá consultar a secção B.2 do Anexo B onde encontrará o *script* de R respetivo.

Apenas duas das quarenta e quatro variáveis incluídas não tiveram qualquer influência no treino do modelo (o valor da comissão de cobrança do mediador e o indicador de contratação de uma soma segura para responsabilidade civil acima do valor obrigatório por lei).

6.2.2 *Balanced Random Forest*

Tal como introduzido anteriormente, o terceiro e último modelo que apresentamos é uma *Random Forest* induzida por *down-sampling* (ver secção B.3 do Anexo B onde se apresenta o *script* de R e os respetivos *outputs*). Isso foi conseguido por inclusão do argumento "sample.fraction" na função "ranger" do R. Uma vez que o objetivo era equilibrar o número de observação de cada classe da variável resposta, de tal forma que não se perdesse demasiada informação relativa aos preditores da classe originalmente em maioria, optou-se por aplicar a proporção de 0,5 a cada. Assim sendo, o processo de reamostragem passou a garantir a existência de tantos casos de anulação como de renovação do contrato.

Utilizou-se um máximo de 15 variáveis na classificação de cada nó e de cada árvore (a escolha delas é aleatória tal como descrito anteriormente).

A totalidade das variáveis incluídas foi utilizada na classificação dos exemplos durante o treino do modelo, sendo que a importância do valor da comissão de cobrança do mediador é quase nula. A razão de isso acontecer em ambos os modelos é o facto do valor da comissão ser praticamente constante entre as observações da nossa amostra.

7 Comparação de Resultados

Nesta secção serão utilizadas as métricas de desempenho descritas na secção 5.2 para comparar os três modelos obtidos. As figuras seguintes representam as curvas ROC relativas a cada um deles.

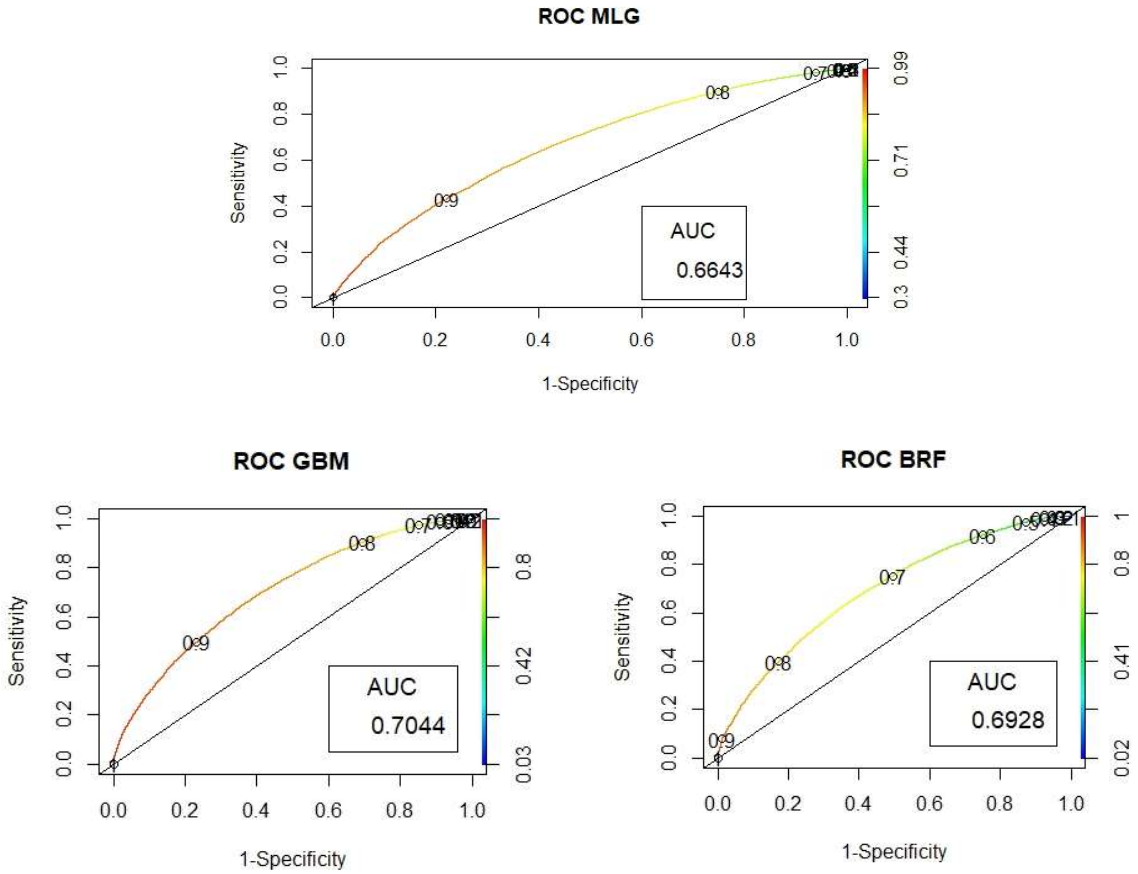

Figura 7.1 Curvas ROC e valores de AUC para os modelos MLG, GBM e BRF.

Pelo valor da AUC são as *Gradient Boosted Trees* que apresentam o melhor desempenho preditivo, sendo o MLG o modelo mais fraco.

Por observação dos gráficos anteriores é possível concluir que o ponto de corte que minimiza a distância ao ponto ótimo (0,1) varia entre os três modelos e que em todos eles se afasta de 0,5. A razão é o desequilíbrio verificado entre as classes da variável resposta. As tabelas seguintes apresentam medidas de desempenho para cada um dos modelos e para diversos pontos de corte.

MLG						
Medida/Ponto de Corte	**0,5**	**0,75**	**0,8**	**0,85**	**0,87**	**0,89**
Weighted Accuracy (β=0,5)	0.5021	0.5426	0.5754	0.6101	0.6182	0.6120
Accuracy	0.876	0.8561	0.8203	0.7183	0.6391	0.5349
Kappa	0.0072	0.1141	0.1542	0.1452	0.1247	0.0947

Gradient Boosted Trees						
Medida/Ponto de Corte	**0,5**	**0,75**	**0,8**	**0,85**	**0,87**	**0,89**
Weighted Accuracy (β=0,5)	0.53065	0.5780	0.6056	0.6354	0.6444	0.6398
Accuracy	0.8802	0.8603	0.831	0.7469	0.6818	0.5873
Kappa	0.1001	0.1954	0.2136	0.1888	0.1648	0.1284

Balanced Random Forest				
Medida/Ponto de Corte	**0,5**	**0,72**	**0,73**	**0,75**
Weighted Accuracy (β=0,5)	0.5519	0.6352	0.6364	0.6355
Accuracy	0.8718	0.6798	0.6563	0.6057
Kappa	0.1497	0.1551	0.1471	0.13

Tabela 7.1: Avaliação do desempenho preditivo dos modelos MLG, GBM e BRF.

Dos pontos de corte testados os que permitem obter melhores resultados foram 0,87 para o MLG e as *Gradient Boosted Trees* e 0,73 para a *Balanced Random Forest*. O ponto de corte é mais baixo para a BRF pois houve um processo de *down-sampling* que permitiu que a cada iteração a amostra tivesse tantas observações positivas como negativas e que, portanto, a probabilidade de retenção fosse de 50% (que compara com os 87,4% observados na amostra com desequilíbrio entre classes).

O GBM é o que apresenta as três métricas mais elevadas. Juntando ao fato de ser também aquele cujo valor da AUC é mais elevado fica claro que é o melhor método a utilizar neste problema.

8 Conclusão

Numa ótica de otimização do valor a cobrar à data da renovação de um contrato de seguro, o estudo das taxas de retenção face a variações no prémio é atualmente considerado de uma importância comparável ao estudo do risco.

Durante este projeto analisaram-se diferentes metodologias de modelação para a taxa de retenção e considerou-se, por fim, as *Gradient Boosted Trees* como sendo o modelo com melhor desempenho preditivo relativamente à nossa amostra de dados.

Sugere-se que, na sequência deste trabalho, se construa um processo de otimização do prémio baseado numa medida rentabilidade.

Para estudar a elasticidade do prémio existem os *Price Tests,* que consistem na aplicação de um aumento no prémio de apólices escolhidas aleatoriamente. O teste é considerado bastante útil já que não existe correlação com nenhum outro fator explicativo das taxas de retenção. No entanto, são testes difíceis de aplicar.

O processo de otimização que propomos pressupõe a impossibilidade de aplicar um *Price Test*. Sugerimos a definição de taxas de variação de prémio, a aplicar à data da renovação, com base nos rácios de perda esperados para cada contrato (em que rácio de perda esperado se define pela razão entre o prémio de risco e o prémio efetivamente cobrado).

Um modelo como o que se obteve neste projeto será então útil para obter as taxas de variação de prémio que permitem à companhia reter o negócio que objetiva, tanto em termos de volume como de rentabilidade.

No cálculo da variação de prémio deve ser também tido em conta o rácio de perda atual (isto é, a razão entre os custos com sinistros e o prémio cobrado) bem como o número de sinistros reportados, uma vez que os agravamentos no prémio deverão concentrar-se nos contratos menos rentáveis. Por outro lado, clientes com sinistros tendem a aceitar melhor os aumentos no prémio.

Esse processo deve ser mensal e incluir apenas as apólices cujo vencimento seja no mês seguinte uma vez que é importante considerar o histórico de sinistralidade mais atual possível.

Ao longo deste texto foram ainda referidos alguns tópicos que, pela complexidade que trariam ao projeto, ou mesmo pelo tempo limitado disponível, não foram totalmente explorados. Um deles foi o facto de não ter sido incluída a causa da variação no prémio. Um aumento devido a alterações nas condições do contrato, como inclusão de novas coberturas, deveria ser diferenciado de variações no prémio relativas a taxas de sinistralidade. Por não ser comum acontecer aumentos de prémio significativos por alterações ao contrato, esta questão foi desconsiderada. No entanto, é algo que deveria ser estudado de forma mais profunda e, adicionalmente, medido o impacto da sua inclusão/exclusão.

Relativamente ao desempenho preditivo dos modelos obtidos, é considerado por nós que há lugar para melhorias. Sugere-se o teste a outros métodos de tratamento de amostras desequilibradas, como as *Weighted Random Forests*.

Anexo A - Regressão Logística

A.1 Detalhes sobre os fatores incluídos na regressão

Distrito 1: Braga e Lisboa.
Distrito 2: Angra do Heroísmo, Beja, Castelo Branco, Évora, Funchal, Horta e Ponta Delgada.
Distrito 3: Aveiro, Bragança, Guarda, Leiria, Setúbal, Viana do Castelo, Vila Real e Viseu.
Distrito 4: Coimbra, Faro, Portalegre e Santarém.
Distrito 5: Porto.
Os casos de valor desconhecido foram incluídos no grupo Distrito 3 pela semelhança ao nível das taxas de retenção observadas.

Tipo Mediador 1: Agentes com nível de experiência médio e alto.
Tipo Mediador 2: Agentes com nível de experiência baixo, Corretores e Negócio Direto.

Marca 2: Opel, Volkswagen, Nissan, BMW, Audi.
Marca 3: Mercedes-Benz, Mitsubishi, Toyota.
Marca 1: Restantes em carteira.

Fatores Explicativos				
Nome	**Descrição**	**Número de Níveis**	**Nível base**	**Intervalo da reta real onde se aplica o polinómio ortogonal (se aplicável)**
Fracionamento	Frequência de pagamento do prémio	3	Anual e Semestral	-
Tipo_Mediador	Tipologia relativa à experiência	2	Tipo_Mediador1	-
Marca	Marca do veículo	3	Marca1	-
Distrito	Distrito do Tomador	5	Distrito1	-
Desc_Retencao[4]	Indicador de aplicação de desconto de retenção	2	Não	-
Colisao	Indicador de presença da cobertura de colisão	2	Não	-

[4] Desconto especialmente aplicado a contratos cujo objetivo é reter.

Debito_Direto	Indicador de pagamento por débito direto	2	Não	-
Idade_Tomador	Idade do tomador (em anos)	-	-	-
Anos_Carta	Antiguidade da carta de condução (em anos)	-	-	-
Mediador_Rentabilidade	Tipologia relativa à rentabilidade da carteira do mediador	3	Média e Baixa	-
Apolices_Outros_Ramos	Número de apólices de outros ramos	-	-	-
Danos_Próprios	Indicador de presença de pelo menos uma cobertura de danos próprios	2	Sim	-
Sinistros_3anos	Número de sinistros nos três anos anteriores	3	0	-
AntigT_1ano	Antiguidade do tomador na companhia (em anos) - indicador de antiguidade igual a 1 ano	2	Não	-
AntigT_polOrt_5-11	Antiguidade do tomador na companhia (em anos)	-	-	[5,11]
AntigT_polOrt_16-21		-	-	[16,21]
AgravDesc_desconhecido	Diferença no prémio por aplicação de agravamentos e/ou descontos (indicador de valor desconhecido)	2	Não	-
AgravDesc_polOrt_-45a-20	Diferença no prémio por aplicação de agravamentos e/ou descontos (%)	-	-	]-45,-20[
Premio_polOrt_175-225	Prémio proposto para renovação (em euros)	-	-	]175,225[
Premio_polOrt_600-900		-	-	]600,900[
DifPerc_polOrt_18-35	Diferença no prémio (%)	-	-	]18,35[
DifPerc_polOrt_2-8		-	-	]2,8[
Vol_Carteira_Med_polOrt_10000-70000	Volume da carteira do mediador (prémio total em euros)	-	-	]10000,70000[

Vol_Carteira_Med_< 200000	Volume da carteira do mediador (prémio total em euros) - indicador de valor inferior a 200.000 euros	2	Não	-
AntigVeic_polOrt_1-4	Antiguidade do veículo (em anos)	-	-	[1,4]
AntigVeic_polOrt_4-10		-	-	[4,10]
AntigVeic_polOrt_17-24		-	-	[17,24]

Tabela A.1: Fatores Explicativos.

A.2 Sumário do Modelo

Call:
glm(formula = Renew ~ . + poly(AgravDesc_polOrt_-45a-20, degree = 2, raw = F) +
poly(Anos_Carta, degree = 2, raw = F) +
poly(Apolices_Outros_Ramos, degree = 2, raw = F) +
Danos_Proprios(Nao):DifPerc_polOrt_2-8 +
Colisao (Sim):AntigVeic_polOrt_0-4 +
Colisao(Sim):AntigVeic_polOrt_17-24 +
Desc_Retencao (Sim):AgravDesc_desconhecido (Sim) + AgravDesc_polOrt_-45a-
20:Premio_polOrt_175-225+
TipoMediador2:AgravDesc_polOrt_-45a-20 +
Sinistros_3anos:AntigT_polOrt_5-11,
family = binomial, data = RET_train)

Deviance Residuals:

Min	1Q	Median	3Q	Max
-2.8926	0.3636	0.4587	0.5538	1.6109

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	5.140e+00	1.899e-01	27.073	< 2e-16	***
Apólice:					
Fraccionamento (Mensal)	6.570e-02	2.866e-02	2.293	0.021862	*
Fraccionamento (Trimestral)	-3.653e-01	2.267e-02	-16.109	< 2e-16	***
Desc_Retencao (Sim)	2.730e-01	1.753e-02	15.575	< 2e-16	***
Debito_Direto (Sim)	4.420e-01	2.432e-02	18.177	< 2e-16	***
Sinistros_3anos (1)	1.693e-02	4.986e-02	0.340	0.734190	
Sinistros_3anos (>=2)	1.200e-01	9.536e-02	1.258	0.208326	

AgravDesc_desconhecido (Sim)	-5.980e-01	3.040e-02	-19.673	< 2e-16	***
Premio_polOrt_175-225	-2.527e-01	2.244e-02	-11.262	< 2e-16	***
Premio_polOrt_600-900	-9.126e-02	1.397e-02	-6.533	6.44e-11	***
DifPerc_polOrt_18-35	-3.559e-02	3.187e-03	-11.170	< 2e-16	***
DifPerc_polOrt_2-8	-6.021e-02	3.596e-03	-16.741	< 2e-16	***
poly(AgravDesc_polOrt_-45a-20, degree = 2, raw = F)1	-9.848e+01	2.927e+01	-3.365	0.000766	***
poly(AgravDesc_polOrt_-45a-20, degree = 2, raw = F)2	3.517e+01	3.873e+00	9.080	< 2e-16	***
Tomador:					
Distrito2	-3.696e-01	2.405e-02	-15.368	< 2e-16	***
Distrito3	-1.187e-01	1.780e-02	-6.669	2.58e-11	***
Distrito4	-2.607e-01	2.183e-02	-11.939	< 2e-16	***
Distrito5	1.477e-01	2.294e-02	6.439	1.20e-10	***
Idade_Tomador	9.231e-03	9.453e-04	9.765	< 2e-16	***
poly(Apolices_Outros_Ramos, degree = 2, raw = F)1	5.276e+01	3.624e+00	14.557	< 2e-16	***
poly(Apolices_Outros_Ramos, degree = 2, raw = F)2	-1.665e+01	3.527e+00	-4.722	2.33e-06	***
AntigT_1ano (Sim)	9.023e-02	2.363e-02	3.818	0.000135	***
AntigT_polOrt_5-11	2.606e-02	3.303e-03	7.891	3.00e-15	***
AntigT_polOrt_16-21	1.599e-02	5.161e-03	3.098	0.001947	**
Condutor Habitual:					
poly(Anos_Carta, degree = 2, raw = F)1	1.939e+01	2.967e+00	6.534	6.40e-11	***
poly(Anos_Carta, degree = 2, raw = F)2	-1.446e+01	2.705e+00	-5.346	8.97e-08	***
Canal de Distribuição:					
TipoMediador2	-1.077e-01	6.261e-02	-1.720	0.085444	.
Mediador_Rentabilidade (Alta)	8.180e-02	1.767e-02	4.630	3.66e-06	***
Vol_Carteira_Med_polOrt_10000-70000	4.152e-02	3.402e-03	12.205	< 2e-16	***
Vol_Carteira_Med_<200000 (Sim)	-8.706e-02	1.450e-02	-6.004	1.92e-09	***
Veículo:					
Marca2	9.105e-02	1.416e-02	6.430	1.28e-10	***
Marca3	2.218e-01	1.889e-02	11.742	< 2e-16	***
AntigVeic_polOrt_0-4	-1.141e-01	2.096e-02	-5.444	5.22e-08	***
AntigVeic_polOrt_4-10	-6.823e-02	5.191e-03	-13.143	< 2e-16	***
AntigVeic_polOrt_17-24	-4.721e-02	2.541e-03	-18.579	< 2e-16	***
Coberturas:					
Colisao (Sim)	-3.022e+00	6.222e-01	-4.857	1.19e-06	***
Danos_Proprios (Nao)	-3.558e-01	4.051e-02	-8.781	< 2e-16	***
Interações:					
Danos_Proprios (Nao):DifPerc_polOrt_2-8	3.923e-02	9.934e-03	3.949	7.86e-05	***
Colisao (Sim):AntigVeic_polOrt_0-4	7.140e-02	8.269e-03	8.635	< 2e-16	***
Colisao (Sim):AntigVeic_polOrt_17-24	1.457e-01	3.664e-02	3.977	6.97e-05	***
Desc_Retencao (Sim):AgravDesc_desconhecido (Sim)	-3.850e-01	5.594e-02	-6.883	5.87e-12	***
AgravDesc_polOrt_-45a-20:Premio_polOrt_175-225	2.542e-02	4.318e-03	5.887	3.93e-09	***

TipoMediador2:AgravDesc_polOrt_-45a-20	-2.961e-02	1.301e-02	-2.275	0.022894	*
Sinistros_3anos (>=2):AntigT_polOrt_5-11	-4.609e-02	1.131e-02	-4.076	4.57e-05	***
Sinistros_3anos (1):AntigT_polOrt_5-11	-1.424e-02	6.038e-03	-2.358	0.018371	*

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1					
(Dispersion parameter for binomial family taken to be 1)					
Null deviance: 183502 on 242028 degrees of freedom Residual deviance: 174828 on 241984 degrees of freedom					
AIC: 174918					
Number of Fisher Scoring iterations: 5					

Tabela A.2: Sumário da regressão logística.

A.3 Teste de Hosmer e Lemeshow

```
>hl.train=hoslem.test(RET_train$Renew,fitted(glmTrain))
> hl.train

        Hosmer and Lemeshow goodness of fit (GOF) test

        data:  RET_train$Renew, fitted(glmTrain)
        X-squared = 13.123, df = 8, p-value = 0.1077
```

Anexo B – Modelos de *Machine Learning*

B.1 Variáveis incluídas

Apólice:
Fracionamento
Indicador de presença do desconto de retenção
Indicador de pagamento por débito direto em conta
Número de sinistros nos três anos anteriores
Número de sinistros no ano anterior
Prémio proposto para renovação
Prémio da anuidade anterior
Diferença percentual no prémio
Diferença no prémio em valor absoluto
Diferença percentual no prémio por desconto e/ou agravamento
Mês da renovação
Antiguidade da apólice (em anos)
Tomador:
Distrito
Concelho
Antiguidade do tomador na companhia (em anos)
Número de apólices de outros ramos
Idade
Prémio Total em apólices do ramo automóvel (permite identificar tomadores com mais do que uma apólice nesse ramo)
Condutor Habitual:
Antiguidade da Carta de Condução (em anos)
Canal de Distribuição (Mediador):
Tipologia relativa à experiência
Classificação quanto à rentabilidade da carteira
Classificação quanto à rentabilidade da carteira Automóvel
Volume de prémios da carteira total
Comissão de Angariação
Comissão de Cobrança
Número do Mediador (identifica de forma única cada Mediador)

Canal de Distribuição (outros):
Delegação
Veículo:
Marca
Antiguidade
Peso Bruto
Cilindrada
Dissuasores de Furto (tipo)
Potência
Combustível
Valor em Novo
Coberturas:
Indicador de presença de pelo menos uma cobertura de Danos Próprios
Indicador de presença da cobertura de Colisão
Franquia em valor
Franquia em percentagem
Soma segura na cobertura de Colisão (anuidade anterior)
Soma segura na cobertura de Colisão (à data da renovação)
Indicador de presença de uma soma segura em responsabilidade civil acima de 7.290.000 euros (valor obrigatório por lei)
Fatores Externos:
Densidade Populacional
Taxa de Desemprego

Tabela B.1: Variáveis incluídas em ambos os modelos de *Machine Learning*.

B.2 *Gradient Boosted Trees*

Modelo:

```
> system.time(
      tree_gbm <- gbm(
        formula = Renew_num~.,
        data = train_ML[-1],
        distribution = "bernoulli",
        n.trees=1000,
        interaction.depth = 4,
        shrinkage=0.1,
        cv.folds=3,
        class.stratify.cv=TRUE))
> tree_gbm
gbm(formula = Renew_num ~ ., distribution = "bernoulli",
   data = train_ML[-1], n.trees = 1000, interaction.depth = 4,
   shrinkage = 0.1, cv.folds = 3, class.stratify.cv = TRUE)
A gradient boosted model with bernoulli loss function.
1000 iterations were performed.
The best cross-validation iteration was 625.
There were 44 predictors of which 42 had non-zero influence.
```

Número ótimo de iterações:

```
> best_iteration <- gbm.perf( object = tree_gbm,
                              method = "cv")
```

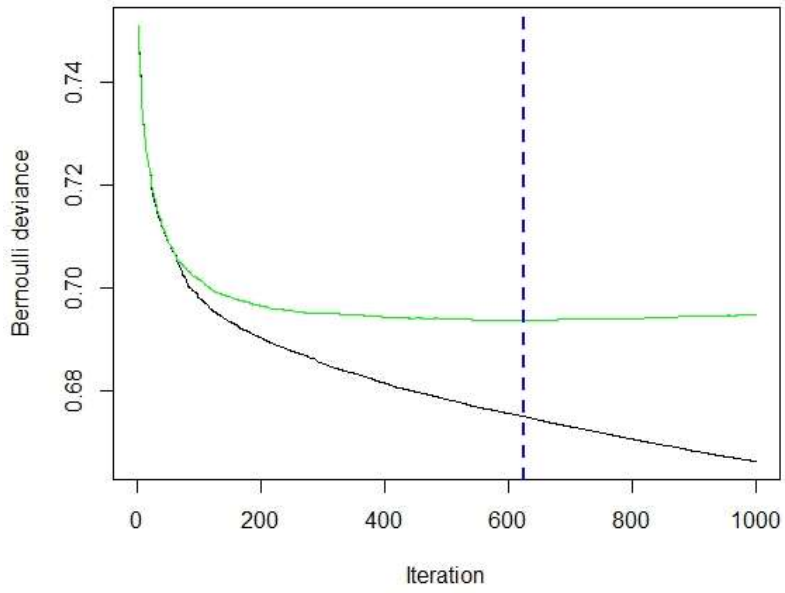

Figura B.1: Estimativa do erro de validação pelo método *cross validation* (apresentação da iteração ótima para o modelo GBM).

B.2 *Balanced Random Forest*

Modelo:

```
> brf <- ranger(Renew ~ ., train_ML,
              mtry=15,
              importance="impurity",
              probability = TRUE,
              replace = TRUE,
              sample.fraction = c(0.5, 0.5))

> brf
Ranger result
Call:
 ranger(Renew ~ ., train_ML, mtry = 15, importance = "impurity",     probability = TRUE,
replace = TRUE, sample.fraction = c(0.5,          0.5))
```

```
Type:                         Probability estimation
Number of trees:              500
Sample size:                  242029
Number of independent variables:  44
Mtry:                         15
Target node size:             10
Variable importance mode:     impurity
Splitrule:                    gini
OOB prediction error (Brier s.):  0.1226921
```

Referências

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control 19 (6), 716–723.

Breiman, L. (1994). Bagging Predictors. Technical Report 421. Retrieved 2019-07-28. Statistics Department University of California, Berkeley.

Breiman, L. (1997). Arcing The Edge. Technical Report 486. Statistics Department University of California, Berkeley.

Breiman, L. (2001). Random forest. Machine Learning, 45, 5–32.

Bremner D., Demaine E., Erickson J., Iacono J., Langerman S., Morin P., and Toussaint G. (2005). Output-sensitive algorithms for computing nearest-neighbor decision boundaries. Discrete and Computational Geometry, 33 (4): 593-604.

Burez, J. and Van den Poel, D. (2009). Handling class imbalance in customer churn prediction. Expert Systems with Applications 36 (3), 4626–4636.

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). Smote: Synthetic minority oversampling technique. Journal of Artificial Intelligence Research, 16, 321–357.

Chawla, N. V., Lazarevic, A., Hall, L. O., & Bowyer, K. W. (2003). SMOTEboost: Improving prediction of the minority class in boosting. In 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, (pp. 107–119).

Chen, C., Liaw, A., Breiman, L. (2004). Using Random Forest to Learn Imbalanced Data.

Cohen, W. (1996). Learning Trees and Rules with Set-Valued Features. American Association for Artificial Intelligence (AAAI).

Collett, D. (2003). Modelling Binary Data (Second ed). Boca Raton, FL: Chapman and Hall/CRC.

Friedman, J. H. (2001). Greedy Function Approximation: A Gradient Boosting Machine.

Hosmer, D. W. and Lemeshow, S. (2000). Applied Logistic Regression, 2nd ed. New York: Wiley.

Instituto Nacional de Estatística (2019). Desemprego registado por 100 habitantes com 15 ou mais anos de idade (%) por Local de residência (NUTS - 2013); Anual. Período de referência dos dados: 2017. Disponível em: https://www.ine.pt.

Instituto Nacional de Estatística (2019). Densidade populacional (N.º/ km²) por Local de residência (NUTS - 2013); Anual. Período de referência dos dados: 2017. Disponível em: https://www.ine.pt.

Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In: International joint Conference on artificial intelligence. [S.l.: s.n.], 1995. v. 14, p. 1137–1145.

Kubat, M. & Matwin, S. (1997). Addressing the curse of imbalanced data sets: One-sided sampling. In Proceedings of the 14th International conference on Machine Learning, (pp. 179–186). Morgan Kaufmann.

Kubat, M., Holte, R., & Matwin, S. (1998). Machine learning for the detection of oil spills in satellite radar images. Machine Learning, 30, 195–215.

Lantz, Brett (2013). Machine Learning with R, 1st ed. Birmingham: Packt Publishing.

McCullagh, P. and Nelder, J. A. (1989). Generalized Linear Models, 2nd ed. London: Chapman and Hall.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. Journal of the Royal Statistical Society. Series A 135 (3), 370–384.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers.

Schapire, R., and Freund, Y. (2012). : Boosting – Foundations and Algorithms Understanding Rule Learners. The MIT Press.