

FUNDAÇÃO GETÚLIO VARGAS
ESCOLA DE MATEMÁTICA APLICADA
MESTRADO EM MODELAGEM MATEMÁTICA DA INFORMAÇÃO

GABRIEL CARNEIRO LENTO

***Random forest* em dados desbalanceados: uma aplicação na modelagem de *churn* em
seguro saúde**

RIO DE JANEIRO
2017

GABRIEL CARNEIRO LENTO

***Random forest* em dados desbalanceados: uma aplicação na modelagem de *churn* em seguro saúde**

Dissertação apresentada à Escola de Matemática Aplicada da Fundação Getúlio Vargas para a obtenção de Título de Mestre em Modelagem Matemática da Informação.

Orientador: Eduardo Fonseca Mendes

RIO DE JANEIRO
2017

Lento, Gabriel Carneiro

Random forest em dados desbalanceados: uma aplicação na modelagem de churn em seguro saúde / Gabriel Carneiro Lento. - 2017.

41 f.

Dissertação (mestrado) – Fundação Getulio Vargas, Escola de Matemática Aplicada.

Orientador: Eduardo Fonseca Mendes.

Inclui bibliografia.

1. Aprendizado do computador. 2. Mineração de dados (Computação).
3. Seguro-saúde. I. Mendes, Eduardo Fonseca. II. Fundação Getulio Vargas.
Escola de Matemática Aplicada. III. Título.

CDD – 006.3


GABRIEL CARNEIRO LENTO

**RANDOM FOREST EM DADOS DESBALANCEADOS: UMA APLICAÇÃO NA
MODELAGEM DE *CHURN* EM SEGURO SAÚDE.**

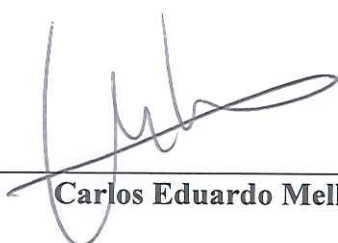
Dissertação apresentada ao Curso de Mestrado em Modelagem Matemática da Informação da Escola de Matemática Aplicada da Fundação Getúlio Vargas para obtenção do grau de Mestre em Modelagem Matemática da Informação.

Data da defesa: 27/03/2017.

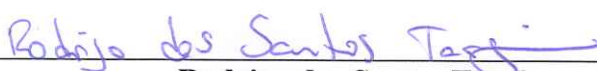
ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA



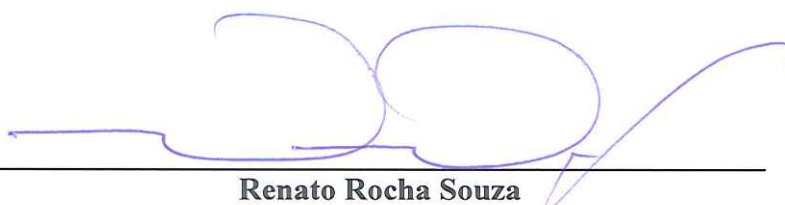
Eduardo Fonseca Mendes
Orientador (a)



Carlos Eduardo Mello



Rodrigo dos Santos Targino



Renato Rocha Souza

À minha esposa Julia, família e amigos

AGRADECIMENTOS

À minha esposa Julia, pelos momentos em que se privou de suas vontades para me apoiar na dissertação. Seu apoio incondicional foi muito importante para que esse trabalho se concretizasse.

Ao meu orientador Eduardo Fonseca Mendes, pelos ensinamentos, por ter acreditado em mim e por ter dedicado horas de lazer em compartilhar seus conhecimentos.

À minha família, em especial à minha mãe Heloisa, minha irmã Ana, minha sogra Thaiza, meu sogro Araujo e minha cunhada Natalia, pelo apoio e torcida para que desse tudo certo.

Aos meus amigos fiéis que sempre se preocuparam em me apoiar, mesmo com a falta da minha presença em eventos importantes.

RESUMO

Neste trabalho estudamos o problema de *churn* em seguro saúde, isto é, a previsão se o cliente irá cancelar o produto ou serviço em até um período de tempo pré-estipulado. Tradicionalmente, regressão logística é utilizada para modelar a probabilidade de cancelamento do serviço. Atualmente, técnicas modernas de *machine learning* vêm se tornando cada vez mais populares para esse tipo de problema, com exemplos nas áreas de telecomunicação, bancos, e seguros de carro, dentre outras. Uma das grandes dificuldades nesta modelagem é que apenas uma pequena fração dos clientes de fato cancela o serviço, o que significa que a base de dados tratada é altamente desbalanceada. Técnicas de *under-sampling* e *over-sampling* são utilizadas para contornar esse problema. Neste trabalho, aplicamos *random forests*, que são combinações de árvores de decisão ajustadas em subamostras dos dados, construídas utilizando *under-sampling* e *over-sampling*. Ao fim do trabalho comparamos métricas de ajustes obtidas nas diversas especificações dos modelos testados e avaliamos seus resultados dentro e fora da amostra. Observamos que técnicas de *random forest* utilizando sub-amostras não balanceadas com o tamanho menor do que a amostra original apresenta a melhor performance dentre as *random forests* utilizadas e uma melhora com relação ao praticado no mercado de seguro saúde. Palavras-chave: *Random Forest*, *churn*, dados desbalanceados, seguro saúde, *under-sampling*, *over-sampling*.

ABSTRACT

In this work we study churn in health insurance, that is predicting which clients will cancel the product or service within a preset time-frame. Traditionally, the probability whether a client will cancel the service is modeled using logistic regression. Recently, modern machine learning techniques are becoming popular in churn modeling, having been applied in the areas of telecommunications, banking, and car insurance, among others. One of the big challenges in this problem is that only a fraction of all customers cancel the service, meaning that we have to deal with highly imbalanced class probabilities. Under-sampling and over-sampling techniques have been used to overcome this issue. We use random forests, that are ensembles of decision trees, where each of the trees fits a subsample of the data constructed using either under-sampling or over-sampling. We compare the distinct specifications of random forests using various metrics that are robust to imbalanced classes, both in-sample and out-of-sample. We observe that random forests using imbalanced random samples with fewer observations than the original series present a better overall performance. Random forests also present a better performance than the classical logistic regression, often used in health insurance companies to model churn. Keywords: Random Forest; imbalanced class; churn; health insurance; under-sampling; over-sampling

SUMÁRIO

1	Introdução	12
2	Metodologia	15
2.1	Regressão Logística	15
2.2	Critério de informação de Akaike	15
2.3	Método <i>stepwise</i> para seleção de variáveis	16
2.4	Árvore de Decisão	16
2.5	<i>Bootstrap</i> e <i>Bagging</i>	17
2.6	Técnicas de balanceamento de amostras	18
2.7	<i>Random Forest</i>	19
2.8	<i>Balanced Random Forest</i>	20
3	Fonte de Dados	22
4	Resultados	24
4.1	Especificação dos modelos	24
4.2	Métricas de avaliação da performance	24
4.3	Comparação da performance	26
5	Conclusão	30
	Referências	31
A	Evolução do setor	33
B	Estratégia para manutenção dos clientes	34
C	Revisão da literatura para outros segmentos	35
D	Análise Exploratória	36
E	Tabelas completas <i>random forest</i>	38

LISTA DE FIGURAS

FIGURA 1 – Exemplo de estrutura de Árvore de Decisão	17
FIGURA 2 – Importância das variáveis	28
FIGURA 3 – Variável Exploratória: Idade do Cliente	36
FIGURA 4 – Variável Exploratória: Receita paga pelo cliente	37
FIGURA 5 – Variável Exploratória: Custo de Utilização (12 meses)	37

LISTA DE TABELAS

TABELA 1 – Algoritmo <i>random forest</i> para regressão e classificação	19
TABELA 2 – Lista de variáveis	22
TABELA 3 – Matriz de Confusão	26
TABELA 4 – Resultados <i>random forest</i> considerando a métrica <i>g-mean</i>	26
TABELA 5 – Resultados <i>random forest</i> considerando a métrica <i>weighted accuracy</i> . . .	27
TABELA 6 – Resultados <i>random forest</i> considerando a métrica <i>f-measure</i>	27
TABELA 7 – Métricas Modelo de Regressão Logística	28
TABELA 8 – Resultado Modelo de Regressão Logística	29
TABELA 9 – Comparativo <i>out-of-sample random forest</i> e <i>logit</i>	29
TABELA 10 –Evolução do total de operadoras ativas	33
TABELA 11 –Resultados Grupo 1	38
TABELA 12 –Resultados Grupo 2	39
TABELA 13 –Resultados Grupo 3 Número de Sub-Amostras: 200, 500 e 1.000	40
TABELA 14 –Resultados Grupo 3 Número de Sub-Amostras: 5.000, 10.000 e 15.000 . .	41
TABELA 15 –Resultados Grupo 4 Número de Sub-Amostras: 200, 500 e 1.000	42
TABELA 16 –Resultados Grupo 4 Número de Sub-Amostras: 5.000, 10.000 e 15.000 . .	43

1 Introdução

O cenário da saúde pública no Brasil encontra-se precário em quase todo o território nacional. Por isso, os consumidores que desejam um serviço com maior qualidade precisam contratar seguros saúde de empresas privadas (ROCHA; CHRISTENSEN, 2000). Com o aumento de consumidores neste mercado, a regulação estabelecida pelo governo faz com que o mercado privado se torne altamente fiscalizado e por isso o número de empresas operando no setor vêm diminuindo com o tempo (Apêndice A). Como marco do setor, destacamos a Lei 9.656/98 e o Estatuto do Idoso (Lei 10.741/03). A partir dessas leis, os consumidores com mais de 59 anos ganharam o direito da isenção de reajuste em contratos já existentes por aumento da idade (ANS, 2017), além de outros benefícios. Caso o cliente com esse perfil mude de empresa, ele estará sujeito a novos preços que, em geral, são mais altos.

Por outro lado, as empresas investem cada vez mais na melhoria no relacionamento com seus clientes, para que consigam oferecer um serviço mais adequado às necessidades de seus clientes e devido ao aumento da criticidade dos consumidores ao escolherem um prestador de serviços. Por esses fatores, o CRM (*customer relationship management*) torna-se uma ferramenta essencial para que as empresas estabeleçam práticas, estratégias e tecnologias que, ao longo do ciclo de vida dos clientes, serão essenciais para uma relação saudável. A aquisição, manutenção e retenção de clientes representam, de forma macro, o papel do CRM.

O termo *churn* é utilizado quando se trata da terceira etapa macro do CRM: a retenção de clientes. Segundo Kotler (1998) a definição geral de *churn* é "o ato do cliente escolher uma companhia em detrimento de outra". Definições específicas são encontradas em alguns casos, como por exemplo em Owczarczuk (2010), onde o cliente *churn* é "aquele que não recebeu nem realizou ligações durante 6 meses". Portanto, é extremamente importante que não haja ambiguidade na definição de *churn*. Neste estudo, definimos *churn* como o cancelamento do serviço, o seguro saúde, por parte do cliente.

Tendo em vista que o custo de aquisição de um novo cliente pode ser de 5 até 7 vezes maior do que o de manutenção (KOTLER, 1998), grandes empresas vêm investindo em modelos para previsão de *churn* e definindo estratégias para que isso aconteça (Apêndice B). Mais que isso, de acordo com Reichheld e Sasser (1990), as empresas podem aumentar seus lucros entre 25% e 85% ao reduzir a taxa de cancelamento em 5%. Logo, uma boa modelagem de *churn* permite a identificação de clientes propensos a cancelar e com foco nesses clientes os programas de retenção podem ser mais concentrados, reduzindo seus custos. Dentre iniciativas de retenção de clientes podemos destacar desconto em serviços, oferecimento de cobertura nova ou extensão das coberturas já existentes.

Na literatura de seguro saúde, podemos destacar dois estudos. Su et al. (2011) propôs um modelo multivariado para segmentação da base de clientes em dois *clusters* utilizando técnicas não-supervisionadas (*Hierarchical clustering*, *K-means clustering* e *SOM/Kohonen technique*) e ajustando uma regressão logística a cada grupo. O resultado encontrado na modelagem foi positivo, mostrando que, combinados com relatórios de perfil de cliente, as técnicas de mineração de dados podem ajudar as empresas a alocar recursos limitados com base em fatos, em vez de intuição. Morik e Köpcke (2004) apresentaram uma aplicação de *Support Vector Machines* considerando ou não *timestamp* em uma base com contratos de seguros saúde, além de outros tipos de seguros. Trataram problemas como: classe desbalanceada, onde apenas 7,7% dos clientes foram classificados como *churners*, base de dados com alta dimensão, dados esparsos e classes de variáveis que pouco distinguiam *churners* de *não churners*. O resultado do estudo mostrou que os modelos de previsão que consideram *timestamp* funcionaram melhor. Além

desses estudos, outras aplicações são feitas para esse tipo de problema em outros segmentos (Apêndice C).

Este estudo emprega técnica de *random forest* e *balanced random forests*, utilizando tanto *under-sampling* quanto *over-sampling* nas construções de suas sub-amostras, para previsão do cancelamento do seguro saúde pelo cliente. A comparação dessas duas técnicas de balanceamento da amostra são importantes pois a base de dados utilizada neste estudo é extremamente desbalanceada, com apenas 6% dos clientes classificados como *churners*. Além disso, comparamos os resultados com a regressão logística, técnica muito utilizada na literatura para a previsão de *churn* (KIM; JUN; LEE, 2014; SU et al., 2011; OWCZARCZUK, 2010; BUCKINX; POEL, 2005; MOZER et al., 2000; EIBEN; KOUDIJS; SLISSER, 1998). Neste estudo utilizaremos a regressão logística com a seleção de variáveis pelo método de *stepwise* da mesma forma como aplicado pelo mercado profissional.

Random forests são *ensembles* de árvores de decisão construídas utilizando sub-amostras dos dados e das variáveis explicativas. Sua utilização no estudo teve como motivação estudos com aplicação em modelagem de *churn* encontrados na literatura (XIE et al., 2009; BENOIT; POEL, 2012; CHEN; LIAW; BREIMAN, 2004), além disso, *random forest* está implementado em diversos pacotes e é de fácil interpretação. *Under-sampling* corresponde a construir sub-amostras balanceadas dos dados de tal forma que a classe majoritária é sub-representada. Por outro lado, *over-sampling* corresponde a construção de amostras de tal forma que a classe minoritária é amostrada com mais frequência, sendo esta super-representada. Esta última técnica não aumenta a quantidade de informação, porém faz com que a classe menor possua mais peso na função de perda. Chen, Liaw e Breiman (2004) propuseram duas maneiras de lidar com classes desbalanceadas em *random forests*: as *weighted random forests* atribuem pesos diferentes para as classes, e as *balanced random forests* utilizam-se de *under-sampling* para construir as sub-amostras. Não é claro qual dos dois métodos possui uma performance melhor nas bases aplicadas. Neste estudo utilizamos uma pequena modificação do algoritmo de *balanced random forests*.

Dividimos as técnicas utilizadas em quatro grupos com diferentes configurações, levando em consideração se a amostra é balanceada, o número de árvores e o tamanho de sub-amostras aplicadas. Determinamos configurações e métricas para avaliação da performance dos modelos. Avaliamos as configurações vencedoras para cada grupo levando em consideração as métricas estipuladas. Ao final, comparamos as performances dos grupos com relação as métricas *g-mean*, *weighted accuracy* e *f-measure* e o melhor método, segundo nossos critérios, utiliza-se de sub-amostras não-balanceadas com o tamanho menor do que amostra original. O número de árvores possui pouca influência quando comparado com outras variáveis, desde que o número seja grande o suficiente.

O estudo contribui para a literatura acadêmica sobre *churn* em seguro saúde, pouco desenvolvida na direção adotada na dissertação. Além disso, não encontramos trabalhos no Brasil tratando deste tema em seguro saúde, fato ainda mais relevante dado a instabilidade nas regras do mercado brasileiro, impostas pela Agência Nacional de Saúde. Neste trabalho é desmostrada a eficácia na aplicação de *random forests* quando observamos amostras desbalanceadas. Apesar do foco do trabalho ser em *random forests* por se adequarem naturalmente a este tipo de problema, não é descartada a aplicação de outras técnicas modernas de *machine learning* utilizadas na modelagem preditiva de *churn*.

Este trabalho está organizado da seguinte forma: no Capítulo 2 apresentamos a metodologia utilizada. Apresentamos o algoritmo utilizado para *random forest* e sua modificação *balanced random forest*. Capítulo 3 descreve a fonte de dados, apresentando as variáveis e filtros utilizados no estudo. No capítulo 4 encontram-se as especificações dos modelos, métricas

utilizadas e comparação da performance. O capítulo final conclui o trabalho e apresenta direções para estudo futuro. Os apêndices agrupam informações mais detalhadas sobre a evolução do setor de saúde suplementar, algumas estratégias aplicadas pelas empresas para manutenção de seus clientes, revisão da literatura de *churn* considerando outros segmentos, uma análise exploratória extensa com a análise de todas as variáveis utilizadas no estudo e as tabelas completas com os resultados de *random forest*.

2 Metodologia

Neste capítulo faremos uma abordagem das metodologias utilizadas no estudo. Além de definir e detalhar cada técnica, é importante que haja compreensão para posterior interpretação dos resultados obtidos. A seção se inicia abordando a técnica mais popular para tratamento de modelagem de *churn*: a regressão logística, em seguida apresenta a medida AIC, de qualidade de ajuste do modelo estatístico e o método de seleção de variáveis *stepwise*. Em seguida é apresentado o método *bootstrap* e a definição de *bagging*. Logo após, são detalhadas as técnicas de amostragem utilizadas neste estudo e *random forest*. Por fim, é apresentada a técnica de *balanced random forest* e a modificação realizada para este estudo a partir de seu algoritmo.

2.1 Regressão Logística

A regressão logística é um dos principais modelos aplicados na modelagem de dados binários (PAULA, 2004; FRIEDMAN; HASTIE; TIBSHIRANI, 2001; JR; LEMESHOW; STURDIVANT, 2013). Isso deve-se a sua simplicidade, facilidade na interpretação dos parâmetros, e ampla aplicabilidade. Neste estudo, utilizamos o modelo logístico onde

$$\Pr(\text{churn}|\mathbf{x}) = \pi(\mathbf{x}; \beta)$$

é a probabilidade de “churn” dado o vetor de variáveis explicativas \mathbf{x} e um vetor de parâmetros β . O modelo logístico é tal que a razão de chances é uma função linear das variáveis explicativas, isto é

$$\ln \left\{ \frac{\pi(\mathbf{x}; \beta)}{1 - \pi(\mathbf{x}; \beta)} \right\} = \beta_1 + \beta_2 x_2 + \dots + \beta_p x_p, \quad (2.1)$$

onde $\mathbf{x} = (1, x_2, \dots, x_p)^T$ contém valores observados de variáveis explicativas e $\beta = (\beta_1, \beta_2, \dots, \beta_p)$ contém valores dos parâmetros desconhecidos.

Os vetor de parâmetros $\beta \in \mathbb{R}^p$ é estimado maximizando-e a verossimilhança do modelo ou, equivalentemente, minimizando-se a sua entropia cruzada. De forma geral, o parâmetro estimado maximiza a probabilidade que este modelo tenha gerado os dados. Para mais informações sobre os algoritmos de estimação veja Paula (2004, Cap. 2) e Friedman, Hastie e Tibshirani (2001, Cap. 4).

2.2 Critério de informação de Akaike

O critério de informação Akaike (AIC) é uma medida relativa da qualidade do ajuste de um modelo estatístico. Segundo Friedman, Hastie e Tibshirani (2001, Cap. 7), a ideia é penalizar o modelo pelo número de parâmetros. O critério AIC é

$$AIC = -\frac{2}{n} \cdot \text{loglik} + 2 \cdot \frac{d}{n}, \quad (2.2)$$

onde p é o número de regressores, n é o tamanho da amostra e loglik é o log da função de verossimilhança avaliada no seu máximo. Ao comparar dois ou mais modelos, o preferido será aquele que apresentar o menor valor do AIC. Uma vantagem deste critério é que ele é útil não apenas no caso do desempenho do modelo em termos de previsões dentro da amostra quanto fora da amostra (GUJARATI, 2009).

2.3 Método *stepwise* para seleção de variáveis

O método *stepwise* para seleção de variáveis é um procedimento que consiste em otimizar o processo de seleção de variáveis a partir de um conjunto inicial de variáveis explicativas. O procedimento é útil nos estágios iniciais de análise do modelo, especialmente quando existe um número muito grande de possíveis variáveis explicativas. As variáveis são testadas no modelo uma a uma e a escolha é feita de acordo com a estatística AIC de cada modelo univariado (JR; LEMESHOW; STURDIVANT, 2013).

No primeiro passo do algoritmo, são computados os valores de AIC parciais para testar se a variável será incluída ou não no modelo. No segundo passo, os AICs parciais são calculados, onde leva-se em conta apenas as variáveis que já foram consideradas nos passos anteriores. A partir daí, os valores calculados de AIC parciais para as variáveis candidatas são comparados com um valor de AIC de referência a priori, que determina a entrada ou não da variável. A cada passo do algoritmo, é feito um teste para a remoção de cada variável individualmente. Neste estágio, a variável analisada é tratada como resposta e as outras que já foram consideradas no modelo são tratadas como explicativas. O procedimento termina no momento em que não houver mais variáveis significativas para serem incluídas no modelo (JR; LEMESHOW; STURDIVANT, 2013).

2.4 Árvore de Decisão

As árvores de decisão são modelos estatísticos representadas por um conjunto de condições que divide a amostra de aprendizagem em partes cada vez menores, como por exemplo na figura 1. para maiores informações veja Friedman, Hastie e Tibshirani (2001, Cap. 9). Ou seja, em um problema de classificação onde temos as classes $\{0 = \text{não churn}, 1 = \text{churn}\}$, a única mudança necessária no algoritmo de árvore correspondem aos critérios de classificação dos nós e de "poda" da árvore. Portanto, para um nó m representando a região R_m com N_m observações,

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{x_i \in R_m} I(y_i = k), \quad (2.3)$$

representa a proporção de observações da classe k no nó m . Sendo assim, são classificadas as observações no nó m para a classe $k(m) = \operatorname{argmax}_m \hat{p}_{mk}$, $k \in \{0, 1\}$, ou seja, para a classe majoritária referente ao nó m .

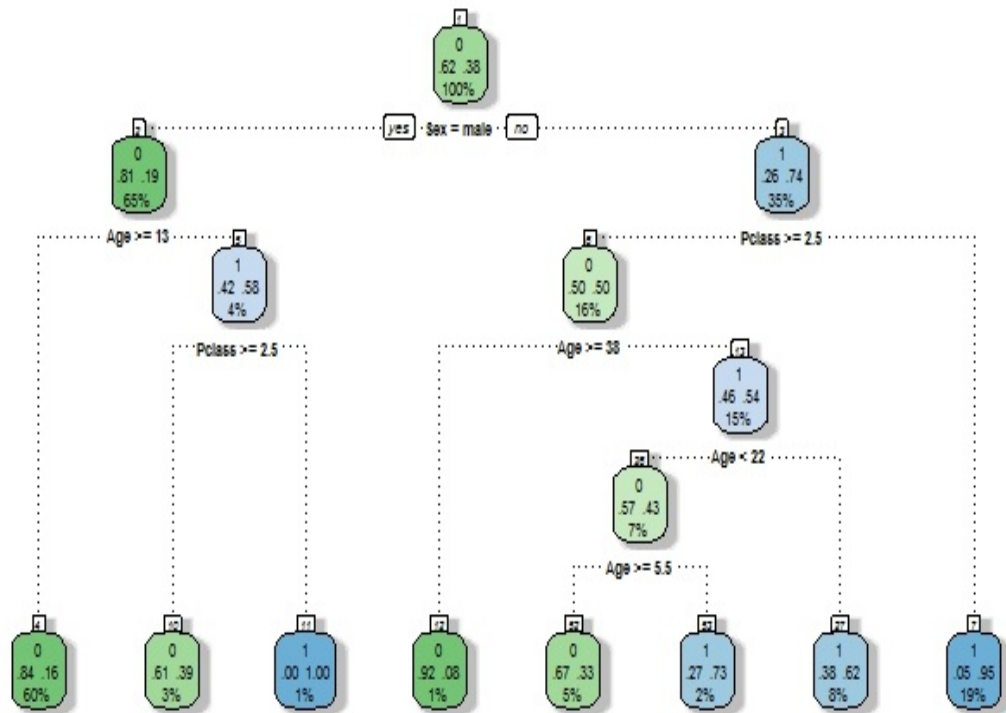


Figura 1 – Exemplo de estrutura de Árvore de Decisão

O critério utilizado para realizar as partições é o da utilidade do atributo para a classificação. Aplica-se, por este critério, um determinado ganho de informação a cada atributo. O atributo teste selecionado para o corrente nó é aquele que possui o maior ganho de informação, medido através de algum índice. A partir desta aplicação, inicia-se um novo processo de partição. Tradicionalmente o critério de partição é baseado no índice Gini, que mede o grau de heterogeneidade dos dados. Este índice num determinado nó é dado por:

$$\text{Índice Gini} = p(1 - p) \quad (2.4)$$

onde p é a frequência relativa da classe $k = 1$ em cada nó (FRIEDMAN; HASTIE; TIBSHIRANI, 2001, Cap. 9).

2.5 Bootstrap e Bagging

O método de *bootstrap* tem o objetivo de aproximar estimar a distribuição amostral de uma estatística de interesse utilizando as observações. O método *bootstrap* é utilizado para melhora na precisão de estimativas de amostra através de processos computacionais e é utilizado em uma gama de áreas e aplicações (EFRON; TIBSHIRANI, 1994)

Seguindo Friedman, Hastie e Tibshirani (2001, Cap. 9) definimos o conjunto de dados de treinamento como $\mathbf{Z} = (z_1, z_2, \dots, z_n)$ onde $z_i = (x_i, y_i)$. A ideia básica é tratar \mathbf{Z} como a população e utilizar reamostragem com reposição para construir sub-amostras \mathbf{Z}^b de tamanho n . Esta etapa é realizada um número grande de vezes, produzindo B amostras *bootstrap* $\mathbf{Z}^1, \dots, \mathbf{Z}^B$. Para cada amostra, o modelo é reestimado e, assim, obtemos uma amostra de estimadores, posteriormente utilizadas para construir sua distribuição.

O método de *Bagging* foi proposto por Breiman (1996) com o objetivo de melhorar o erro fora da amostra de árvores de decisão. O *bagging* utiliza-se de amostras *bootstrap* para construir uma coleção de árvores de decisão que são combinadas em um único classificador. Apesar de utilizarmos esta técnica com árvores de decisão, ela pode ser aplicada em diversos modelos base diferentes.

Dada uma amostra de treinamento, construímos B amostras *bootstrap* obtendo em cada amostra uma árvore de decisão $\hat{f}^b(x)$. A estimação por *bagging* é definida pela média das predições de todas as árvores ou a mais votada. No primeiro caso obtemos uma estimação da probabilidade enquanto no segundo a classe majoritária.

$$\hat{f}_{bag}(x) = \frac{1}{B} \sum_{b=1}^B \hat{f}^b(x). \quad (2.5)$$

A principal ideia da técnica *bagging* é utilizar a média de muitos modelos instáveis, porém aproximadamente não tendenciosos, para reduzir a variância do preditor final. Árvores de decisão são candidatas naturais para a aplicação de *bagging* pois essa técnica consegue capturar interações a partir de estruturas complexas nos dados, e caso cresça tão profundamente, possui erro relativamente pequeno (FRIEDMAN; HASTIE; TIBSHIRANI, 2001).

Para avaliar o erro fora da amostra quando aplicamos o *bagging*, ao invés de olhar para uma amostra de validação ou validação cruzada, recorreremos ao erro *out-of-bag* (FRIEDMAN; HASTIE; TIBSHIRANI, 2001, Cap. 15). O erro *out-of-bag* é um estimador não enviesado do erro fora da amostra e, para uma observação z_i , utiliza apenas as árvores estimadas usando amostras \mathbf{Z}^b que não continham esta observação. Mais precisamente,

$$\widehat{Err}_{oob}(f_{bag}) = \frac{1}{N} \sum_{i=1}^n \frac{1}{C_i} \sum_{b: z_i \notin \mathbf{Z}^b} I(y_i \neq \hat{f}^b(x_i)),$$

onde C_i é o número de amostras *bootstrap* que não contém a observação z_i .

2.6 Técnicas de balanceamento de amostras

Uma técnica muito comum quando se trata de classes com baixa frequência é a amostragem e os métodos que serão utilizados neste estudo são: *under-sampling* e *over-sampling*. O objetivo de ambos os métodos é tratar o desbalanceamento das classes, fazendo com que a classe menos frequente seja mais observada, tornando-a mais frequente. *Under-sampling* consiste em eliminar observações da classe mais frequente e *over-sampling* consiste em replicar a classe menos frequente (BUREZ; POEL, 2009).

Apesar de muito utilizadas, estas técnicas possuem alguns pontos negativos. Weiss (2004) destaca diversos itens que deverão ser observados ao utilizar estas técnicas de amostragem. *Under-sampling*, ao eliminar observações da classe mais frequente, pode perder efeitos de observações relevantes para a estimação do modelo. Além disso, *over-sampling*, ao replicar a classe menos frequente, pode causar *overfit* do modelo.

2.7 Random Forest

Random Forest foi introduzida por Breiman (2001) e representa uma modificação de *bagging*, na qual constrói-se uma grande coleção de árvores correlacionadas, no qual considera a média dos resultados. *Random forest* é bastante disseminada e possui implementações em diversos tipos de *softwares* e pacotes.

A ideia de *random forest*, demonstrada em seu algoritmo é de melhorar a redução da variância através do *bagging* reduzindo a correlação entre as árvores, sem aumentar muito o valor da variância. Isto é alcançado no processo de crescimento da árvore através da seleção aleatória das variáveis de entrada. Além disso, há uma facilidade de paralelização do processo dado a independência da estimação de cada árvore (FRIEDMAN; HASTIE; TIBSHIRANI, 2001).

1. Para $b = 1$ até B :

(a) Amostra *bootstrap* Z^* de tamanho N a partir da base de treinamento.

(b) Cresça a *random forest* T_b a partir da amostra *bootstrap*, repetindo recursivamente os seguintes passos para cada nó de cada árvore, até que o último nó seja alcançado.

i. Selecione m variáveis aleatoriamente a partir de p variáveis.

ii. Selecione a melhor variável de acordo com seu poder de discriminação da variável de interesse.

iii. Divida o nó a partir de dois nós filhos.

2. Reporte o *ensemble* de árvores $\{T_b\}_1^B$.

Para fazer a previsão em um novo ponto x :

Regressão: $\hat{f}_{rf}^B(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$.

Classificação: Seja $\hat{C}_b(x)$ a classe predita da b -ésima árvore de *random forest*.

Então $\hat{C}_b(x) =$ mais votado $\{\hat{C}_b(x)\}_1^B$.

Tabela 1 – Algoritmo *random forest* para regressão e classificação

Quando usada para a classificação, a *random forest* obtém um voto de classe de cada árvore e, em seguida, classifica-a usando o voto da maioria. Quando usadas para regressão, as previsões de cada árvore em um ponto alvo qualquer são simplesmente médias. Da mesma forma como apresentado por Burez e Poel (2009), facilmente um classificador probabilístico pode ser transformado em um classificador por voto determinando pontos de corte para definição de um classificador binário. Além disso, os autores fazem as seguintes recomendações:

- Para a classificação, o valor padrão para m é $\sqrt{\rho}$ e o número mínimo de nós é um.
- Para regressão, o valor padrão para m é $\rho/3$ e o número mínimo de nós é cinco.

sendo m o número de variáveis selecionadas em cada árvore. Na prática, o melhor valor para esses parâmetros dependerá do problema, e deverá ser tratado como parâmetro de ajuste.

Os gráficos da importância das variáveis podem ser construídos para *random forest* exatamente da mesma maneira que para modelos de *gradient-boosted*. Para toda divisão em cada árvore, a medida de importância da variável discriminante é determinada a partir de seu poder de divisão da classe de interesse e é acumulada em todas as árvores da floresta separadamente para cada variável. Enquanto *boosting* ignora algumas variáveis por completo, *random forest* as considera. Além disso, a metodologia *random forest* também utiliza as amostras *out-of-bag* para construir uma medida diferente da importância da variável, para medir a força de predição de cada variável (FRIEDMAN; HASTIE; TIBSHIRANI, 2001).

Quando a b -ésima árvore é gerada, as amostras *out-of-bag* são passadas para baixo da árvore, e a precisão de previsão é registrada. Em seguida, os valores para a j -ésima variável são aleatoriamente permutados nas amostras *out-of-bag* e a precisão é novamente calculada. A redução da precisão como resultado desta permuta é calculada em média em todas as árvores e é usada como uma medida da importância da variável j na *random forest* (FRIEDMAN; HASTIE; TIBSHIRANI, 2001).

A aleatorização efetivamente anula o efeito de uma variável, assim como estabelecer um coeficiente para zero em um modelo linear. Isto não mede o efeito sobre a predição se esta variável não estiver disponível, porque se o modelo fosse reescrito sem a variável, outras variáveis poderiam ser usadas como substitutos (FRIEDMAN; HASTIE; TIBSHIRANI, 2001).

2.8 *Balanced Random Forest*

Como proposto em Breiman (2001) o primeiro passo da *random forest* é selecionar uma amostra *bootstrap* a partir da amostra de treinamento. Quando *random forest* é aplicada em uma amostra extremamente desbalanceada, há uma probabilidade significativa de que sejam selecionados poucos elementos da classe menos frequente, resultando em uma árvore com baixa performance para predição da classe menos frequente. Uma boa maneira de resolver esse problema é utilizar um *bootstrap* estratificado, ou seja, selecionar amostras com reposição dentro de cada classe. Essa alternativa porém, não resolve o problema de classe desbalanceada por completo.

Como apresentado em Drummond, Holte et al. (2003), para árvores de classificação, tornando artificialmente as classes em tamanhos iguais através de *under-sampling* da classe majoritária ou *over-sampling* da classe minoritária é geralmente mais eficaz em relação a uma determinada medida de performance. No entanto, como apresentado no capítulo 2.6, pode-se ter alguns pontos negativos.

O algoritmo para *balanced random forest* é apresentado abaixo:

1. Para cada iteração em *random forest*, selecione uma amostra *bootstrap* da classe minoritária. Selecione aleatoriamente o mesmo número de casos, com substituição, da classe majoritária;
2. Induzir uma árvore de classificação dos dados para o tamanho máximo, sem poda. A árvore é induzida com o algoritmo CART (FRIEDMAN; HASTIE; TIBSHIRANI, 2001, Cap. 9), com a seguinte modificação: em cada nó, em vez de pesquisar através de todas as variáveis para a divisão ideal, basta pesquisar através de um conjunto de variáveis selecionadas aleatoriamente;
3. Repita os dois passos acima para o número de vezes desejado. Agregue as previsões do conjunto e faça a previsão final.

Para esse estudo, uma pequena modificação no item 1 foi realizada. Ao selecionar uma amostra *bootstrap* foram atribuídos pesos (η) para ambas as classes e balanceando a probabilidade de cada subconjunto da base de dados ser selecionada. Sendo assim, a cada *bootstrap*, a probabilidade de seleção da observação referente à classe minoritária será η vezes maior do que referente à classe majoritária. Portanto, o cálculo de η foi feito da seguinte forma:

$$\eta = \begin{cases} 1, & \text{para classe majoritária.} \\ \frac{N-\phi}{\phi}, & \text{para classe minoritária,} \end{cases} \quad (2.6)$$

onde N é o total de observações da amostra de treinamento e ϕ é o total de observações da classe minoritária. A quantidade de variáveis selecionadas aleatoriamente no item 2 foi de $\sqrt{N-1}$.

3 Fonte de Dados

A fonte de dados utilizada nesse estudo tem como origem os registros de clientes de uma grande operadora de planos de saúde brasileira. Foram utilizados apenas os clientes com as seguintes características: contratos firmados após a Lei 9.656/98, contratação pessoa física e clientes titulares. Após todo o tratamento, o conjunto de dados final utilizado no estudo possuía 75.386 clientes, sendo 4.508, ou seja 6%, clientes que realizaram o cancelamento do plano. Pode-se verificar que a baixa frequência de *churners* é parte integral da análise e apresenta um grande desafio na modelagem.

Os filtros aplicados na extração dos dados levam em consideração a padronização dos contratos dos clientes e definição de *churn* adotada. Os contratos firmados antes da Lei 9.656/98 possuem características particulares, modificadas após a implantação desta lei. Os contratos de pessoa jurídica não podem ser utilizados diretamente pois é apenas possível observar se o usuário foi excluído do contrato, mas não se escolheu sair do plano. Esta hipótese viola a definição de *churn* utilizada no trabalho, que requer o conhecimento se o usuário cancelou o serviço. Originalmente existem três formas de relação contratual: *titular*, que é o cliente contratante do plano, *dependente* ou *agregado*, que são os clientes inscritos pelo contratante. O usuário *titular* é o total responsável pelo contrato e seu cancelamento. Portanto, caso o *titular* efetue seu cancelamento, os dependentes e agregados também terão automaticamente seus contratos cancelados.

A base de dados foi construída a partir da extração dos clientes ativos na data-base do estudo (Dezembro de 2011), onde todas as informações cadastrais foram capturadas. A variável *churn* representa os clientes ativos na data-base e que cancelaram o plano de saúde em até um mês. Informações históricas foram construídas a partir da mesma data-base do estudo. As variáveis são apresentadas na tabela 2, divididas em três conjuntos: dados cadastrais, utilização do plano e interação financeira, representando um total de 54 variáveis; suas análises encontram-se no apêndice D. A base completa foi dividida em dois grupos, sendo 75% para treinamento dos modelos preditivos e 25% para teste, isto é *in-sample* e *out-of-sample*, respectivamente.

Sabe-se que variáveis macroeconômicas, tais como índice de desemprego e de inflação, podem influenciar o cliente à permanecer ou não com um contrato de plano de saúde ativo. Porém, devido natureza longitudinal dos dados extraídos, estes fatores não são contemplado no estudo. Para isso, regras de consolidações e extrações transacionais deverão ser realizadas e um novo conjunto de dados construído com base nessas novas informações. Isto está além do propósito deste estudo.

Categoria	Formato		
	Numérica	Catagórica	Dummy
Dados Cadastrais	Idade e Quantidade de anos de plano	Estado Civil, Gênero e Rede Assistencial	-
Utilização do Plano	Tempo desde a última consulta, exame, terapia e internação (30, 60, 90, 180 e 365 dias); Custo de utilização (3, 6, 9 e 12 meses); Custo e Quantidade de Exames e Terapias (Simples e Complexos); Custo e Quantidade de Consultas e Internações (12 meses)	-	Flag de internação (6 e 12 meses), Flag de utilização (3, 6, 9 e 12 meses)
Interação Financeira	Receita	-	Flag de atraso e de desconto

Tabela 2 – Lista de variáveis

As variáveis da categoria "dados cadastrais" foram capturadas a partir do registro do

cliente no sistema da empresa. A quantidade de anos de plano foi calculada pela diferença entre a data que o cliente foi incluído no sistema e a data-base do estudo. A variável estado civil e rede assistencial podem sofrer alterações durante o período de vigência do contrato, por exemplo o caso em que o cliente solteiro se casa e quanto o cliente solicita a mudança de plano. Para esses casos, também foram utilizados os últimos registros de acordo com a data-base do estudo.

As variáveis da categoria "utilização do plano" foram capturadas a partir do histórico referente à base de custo da empresa. O tempo desde o último procedimento (consulta, exame, terapia ou internação) foi calculado da seguinte forma: para cada cliente e para cada procedimento foi verificada a última data em que ele utilizou e calculada a diferença entre essa data e a data-base do estudo, determinando o tempo desde o último procedimento realizado. Após esse passo foi determinado um valor máximo de dias de apuração da utilização (30, 60, 90, 180 e 365 dias) e foi realizado um corte. Por exemplo, para a variável tempo desde a última consulta em 30 dias foi utilizado como corte 30 dias para os clientes que não utilizaram consulta nesse período e a diferença entre a data de utilização e a data-base do estudo para os que utilizaram. As demais variáveis foram computadas pela soma do custo ou da quantidade de procedimentos utilizados, assim como as variáveis *dummy* que representam uma indicação de utilização no período apurado. Além disso, a motivação em separar exames e terapias em simples e complexos é para tentar capturar a dependência de utilização do plano de saúde, devido ao fato dos exames complexos serem referentes ao diagnóstico de doenças mais graves.

As variáveis da categoria "interação financeira" foram capturadas a partir da base de pagamentos dos clientes para a empresa. A receita foi calculada a partir da soma dos valores pagos à operadora nos 12 meses anteriores a data-base do estudo. A variável *dummy* flag de atraso representa a apuração da diferença entre os dias de vencimento e pagamento das mensalidades nos 12 meses anteriores a data-base do estudo. Caso essa diferença seja zero ou positiva nenhum mês é computado como atraso, mas caso seja negativa o mês é computado como atraso. Ao final se há pelo menos uma diferença negativa o flag de atraso terá valor 1, e caso contrário o valor será igual a 0. Para cada pagamento da mensalidade há um registro do valor "cheio" da mensalidade e o valor com desconto. A variável *dummy* flag de desconto recebe o valor 1 caso haja valor com desconto na mensalidade nos 12 meses anteriores a data-base do estudo.

4 Resultados

O capítulo de resultados está dividido da seguinte forma: inicialmente são apresentadas as especificações dos modelos, detalhando a divisão dos grupos com características diferentes determinadas pela utilização ou não de técnicas de amostragem e as configurações utilizadas para estimação dos modelos; na segunda parte do capítulo apresentamos as métricas de avaliação da performance considerando cada configuração aplicada em cada grupo; e no final do capítulo serão apresentados as configurações com melhor performance encontradas para cada grupo de acordo com as métricas escolhidas e determinado o melhor método de acordo com os critérios estabelecidos.

4.1 Especificação dos modelos

As *random forests* utilizadas no estudo são divididas em quatro grupos:

1. Grupo 1: sub-amostras não balanceadas com o tamanho da amostra original.
2. Grupo 2: *over-sampling*. Sub-amostras balanceadas com o tamanho da amostra original.
3. Grupo 3: sub-amostras não balanceadas com o tamanho menor do que a amostra original.
4. Grupo 4: *under-sampling* ou *over-sampling*. Sub-amostras balanceadas com o tamanho menor do que a amostra original, dependendo do tamanho da sub-amostra.

Dentro de cada grupo utilizamos configurações diferentes, variando o número de árvores e tamanho das sub-amostras. São consideradas duas formas de atribuição da classe: maioria de votos e proporção. O primeiro resulta numa árvore de classificação onde a saída é a classe vencedora, enquanto no segundo a probabilidade de pertencer a cada classe. Para árvores de probabilidade são determinados pontos de corte para construção de uma classificação binária e de *matrizes de confusão*.

Neste trabalho são considerados seis valores para número de árvores no *ensemble* e seis valores para o tamanho da sub-amostra. Para o número de árvores utilizamos os valores 100, 200, 500, 1000, 2000, 5000. Para o tamanho da sub-amostra utilizamos os valores 200, 500, 1000, 5000, 10000, 15000. No total, considerando os quatro grupos especificados e as combinações entre número de árvores e tamanho de sub-amostras, foram construídas 336 formas de configurações.

Como em Nie et al. (2011) utilizamos a técnica *stepwise* para estimar o modelo com a combinação de variáveis para a regressão logística. Como modelo inicial, a covariável que possui maior correlação com a variável resposta foi inserida, e após as replicações de *stepwise*, o melhor modelo foi selecionado de acordo com a medida AIC.

4.2 Métricas de avaliação da performance

A escolha da métrica para mensuração da performance do modelo preditivo no caso de classes muito desbalanceadas é crítico para avaliação dos resultados. Um preditor que classifica todos os exemplos como a classe dominante ainda possuirá uma boa acurácia. Neste trabalho, utilizamos a matriz de confusão gerada pela classificação de *churners* para cálculo de estatísticas

que levam em consideração este fenômeno. Na situação onde o modelo retorna a probabilidade de pertencer a cada classe, usam-se diferentes valores de corte, escolhidos a priori em estudo reduzido, para construção da matriz de confusão. As métricas são calculadas *out-of-bag* para posteriores validações *out-of-sample*.

A partir da matriz de confusão apresentada na tabela 3 obtemos as métricas Verdadeiro Positivo (TP), Falso Positivo (FP), Verdadeiro Negativo (TN) e Falso Negativo (FN). Essas quantidades são utilizadas para calcular as métricas de performance usadas no trabalho. Utilizamos as métricas a seguir, também adotadas em Chen, Liaw e Breiman (2004), Weiss (2004), Burez e Poel (2009) e discutidas em mais detalhes em Bekkar, Djemaa e Alitouche (2013).

$$\begin{aligned} \text{True positive rate ou recall}(Acc^+) &= \frac{TP}{TP + FN}, \\ \text{True negative rate}(Acc^-) &= \frac{TN}{TN + FP}, \\ G\text{-mean}(G) &= (Acc^+ \times Acc^-)^{1/2}, \\ \text{Weighted accuracy}(W.Acc) &= \beta Acc^+ + (1 - \beta) Acc^-, \\ \text{Precision}(Prec) &= \frac{TP}{TP + FP}, \\ F\text{-measure}(F) &= \frac{2 \times Prec \times Acc^+}{Prec + Acc^+}. \end{aligned}$$

As métricas acima são frequente utilizadas para medir a performance em previsões de classificadores binários. As taxas Acc^+ e Acc^- representam respectivamente o número de casos positivos e negativos que são corretamente classificados. A métrica *g-mean* foi proposta por Kubat, Matwin et al. (1997) e representa o produto de ambas taxas (Acc^+ e Acc^-). Esta métrica indica o equilíbrio entre os desempenhos de classificação na classe majoritária e na minoritária. Caso haja uma baixa taxa de classificação dos casos positivos o valor de *g-mean* será baixo, mesmo com Acc^- alto. A métrica *precision* mede a proporção de falsos positivos classificados pela previsão, ou seja, quanto maior a *precision*, menor o número de clientes classificados incorretamente nessa classe. A *f-measure* também é muito utilizada neste tipo de estudo. Sua interpretação indica que o modelo performou melhor para a classe positiva quando possuir um valor alto.

Weighted accuracy é bastante utilizada quando se trabalha com classes desbalanceadas (CHEN; LIAW; BREIMAN, 2004). Para cada aplicação o valor de β pode ser ajustado para valores entre 0 e 1. Esta métrica é uma ponderação de Acc^+ e Acc^- , sendo importante para medir a acurácia do modelo e o equilíbrio entre as respectivas taxas. Neste estudo vimos na literatura que conquistar um novo cliente pode custar de 5 até 7 vezes mais do que mantê-lo, e por isso utilizamos o custo de até 5 vezes para construir a definição do valor de $\beta = 0.83$, sendo esse 5 vezes maior do que seu complementar.

Para qualquer classificador, há sempre um *trade off* entre aumentar o valor de Acc^+ ou de Acc^- , e conseqüentemente o mesmo se aplica para *recall* e *precisão*. No caso de construir análises em dados extremamente desequilibrados, muitas vezes a classe minoritária é de grande interesse. Em muitas aplicações, como em testes de remédios ou diagnósticos de doenças a escolha de um classificador de alta precisão de Acc^+ é primordial, mantendo a precisão razoável para Acc^- . Neste estudo o equilíbrio entre essas métricas é importante para que as empresas possam focar nos clientes que realmente necessitam de um tratamento diferencial para não cancelarem o produto, gerando com isso um aumento da lucratividade da empresa e uma redução da despesas em programas de retenção de cliente. Clientes *não churners* classificados

como *churners* poderiam representar um desperdício de tempo e de recursos e clientes *churners* classificados como *não churners* uma possível perda de receita futura.

	Previsão Positiva	Previsão Negativa
Real Positivo	TP (Positivo Verdadeiro)	FN (Falso Negativo)
Real Negativo	FP (Falso Positivo)	TN (Negativo Verdadeiro)

Tabela 3 – Matriz de Confusão

4.3 Comparação da performance

Inicialmente foram estimadas *random forest* para cada um dos grupos definidos na seção 4.1 com as configurações para os seis diferentes números de árvores e seis diferentes tamanhos de sub-amostras. Além disso, para cada configuração foram apuradas as probabilidades e classificações de ambas as classes *churners* e *não churners*. Para as *random forests* probabilísticas foram avaliados os resultados para diferentes pontos de corte. Ao final, foram construídas matrizes de confusão para todas estas configurações e apuradas as métricas definidas na seção 4.2. As tabelas completas com todas as 336 formas de configuração estão apresentadas no apêndice E. De posse dessas tabelas foram selecionadas as configurações vencedoras de cada grupo de acordo com as métricas *g-mean*, *weighted accuracy* e *f-measure* e apresentadas a seguir. A análise da regressão logística será feita posteriormente.

Um dos *insights* iniciais observados a partir dos resultados é de que não há diferenças significativas quando aumentado o número de árvores. Para árvores probabilísticas, os diferentes pontos de corte apresentaram resultados significativamente diferentes, inclusive quando comparado com árvores de classificação. Em relação ao tamanho de sub-amostras pode-se observar diferenças de performance. Além disso, de forma geral, *over-sampling* apresenta melhor performance quando comparado com *under-sampling*.

As configurações que obtiveram maiores valores em relação a métrica *g-mean* estão apresentados na tabela 4. Para essa métrica, o grupo que obteve melhor performance foi o grupo 3, com 2.000 árvores, com tamanho de sub-amostras de 15.000, considerando a árvore probabilística e 0,7 como ponto de corte. Ao compararmos *under-sampling* com *over-sampling*, verificamos melhor performance em todas as métricas com a utilização de *over-sampling*. Quando comparada suas configurações, *under-sampling* apresentou melhores resultados para poucas árvores e tamanho de sub-amostras menor. O grupo 1 apresentou resultados próximos ao grupo de melhor performance, demonstrando bons resultados para previsão de clientes *churners*.

Grupos	Número de árvores	Tamanho de sub-amostras	Ponto de corte	Acc^+	Acc^-	G	W.acc	Prec	F
1	2000	*	0,9	0,768979	0,918315	0,840336	0,793868	0,372889	0,502236
2	5000	*	0,7	0,625394	0,966679	0,777532	0,682275	0,439418	0,516165
3	500	1000	*	0,785353	0,856600	0,820203	0,797227	0,257015	0,387286
	2000	15000	0,7	0,762727	0,925968	0,840393	0,789934	0,394214	0,519781
4 under-sampling	200	1000	0,9	0,701399	0,885032	0,787884	0,732005	0,278158	0,398343
4 over-sampling	2000	15000	0,9	0,746948	0,918954	0,828500	0,775616	0,367943	0,493024

Tabela 4 – Resultados *random forest* considerando a métrica *g-mean*

As configurações que obtiveram maiores valores em relação a métrica *weighted accuracy* estão apresentados na tabela 5. Para essa métrica, o grupo que obteve melhor performance

foi o grupo 3, com 5.000 árvores, com tamanho de sub-amostras de 15.000, considerando a árvore probabilística e 0,9 como ponto de corte. Para esse grupo, observou-se um alto valor de Acc^+ , porém não demonstrou equilíbrio com o valor de Acc^- . Avaliando *under-sampling* e *over-sampling*, novamente a técnica de amostragem que obteve melhor performance em todas as métricas foi *over-sampling*.

Grupos	Número de árvores	Tamanho de sub-amostras	Ponto de corte	Acc^+	Acc^-	G	W.acc	Prec	F
1	2000	*	0,9	0,768979	0,918315	0,840336	0,793868	0,372889	0,502236
2	5000	*	0,7	0,625394	0,966679	0,777532	0,682275	0,439418	0,516165
3	2000	1000	0,7	0,922596	0,561922	0,720019	0,862484	0,117404	0,208301
	5000	15000	0,9	0,935695	0,549342	0,716949	0,871303	0,115939	0,206315
4 under-sampling	200	1000	0,9	0,701399	0,885032	0,787884	0,732005	0,278158	0,398343
4 over-sampling	100	15000	0,9	0,748735	0,910888	0,825841	0,775760	0,346705	0,473947

Tabela 5 – Resultados *random forest* considerando a métrica *weighted accuracy*

As configurações que obtiveram maiores valores em relação a métrica *f-measure* estão apresentados na tabela 6. Apesar dos maiores valores encontrados, os resultados não apresentaram altos valores para essa métrica. O vencedor foi grupo 3, com 5.000 árvores, com tamanho de sub-amostras de 15.000, considerando a árvore de classificação. Apesar disso, não foi a que obteve maior valor para Acc^+ . Para o mesmo grupo, porém considerando 500 árvores e tamanho de sub-amostra de 1.000, a métrica Acc^+ obteve maior valor obtendo o equilíbrio com a métrica Acc^- . Avaliando *under-sampling* e *over-sampling*, ambas não obtiveram boa performance apesar de altos valores de *f-measure*.

Grupos	Número de árvores	Tamanho de sub-amostras	Ponto de corte	Acc^+	Acc^-	G	W.acc	Prec	F
1	5000	*	0,7	0,561477	0,995036	0,747455	0,633736	0,877209	0,684698
2	5000	*	*	0,557675	0,994114	0,744575	0,630415	0,796885	0,656158
3	500	1000	*	0,785353	0,856600	0,820203	0,797227	0,257015	0,387286
	500	15000	*	0,610301	0,990090	0,777337	0,673599	0,795499	0,690701
4 under-sampling	2000	1000	0,7	0,399226	0,997838	0,631160	0,498995	0,921016	0,557009
4 over-sampling	100	15000	0,9	0,539744	0,995243	0,732923	0,615660	0,877541	0,668387

Tabela 6 – Resultados *random forest* considerando a métrica *f-measure*

Ao fim da análise dos resultados de *random forest*, podemos concluir que o ponto de corte foi importante para obtenção de boas métricas devido ao grande desbalanceamento dos dados. O tamanho de sub-amostras foi outra configuração importante para o estudo. Em geral, as métricas *precision* e *f-measure* obtiveram valores baixos, isto porque, para *precision*, o desbalanceamento das classes fez com que os valores de *false positive* aumentasse o denominador, e como *f-measure* utiliza a mesma métrica, seu valor também ficou baixo.

Considerando todas as métricas e análises feitas a partir dos resultados apresentados, foi selecionada a melhor configuração para posterior aplicação *out-of-sample*. O grupo selecionado foi 3 com tamanho de sub-amostra de 15.000. O número de árvores analisadas como vencedoras foram 500, 2.000 e 5.000 e para aplicação *out-of-sample* foi utilizada a configuração de 5.000 árvores levando em consideração um modelo mais robusto. Foi utilizado o modelo probabilístico e 0,9 como ponto de corte.

Aplicando a configuração escolhida para *random forest* podemos analisar o modelo como um todo. A avaliação da importância das variáveis nos ajuda a interpretar os resultados que obtivemos anteriormente. A figura 2 mostra as 20 principais variáveis do modelo e podemos verificar que a variável Receita possui um valor para o índice de gini bem maior que para as demais, demonstrando sua alta influência. Como já era esperado, a idade e o tempo de

plano foram variáveis importantes na estimação do modelo. As variáveis de utilização também se mostraram extremamente importantes para sua boa performance.

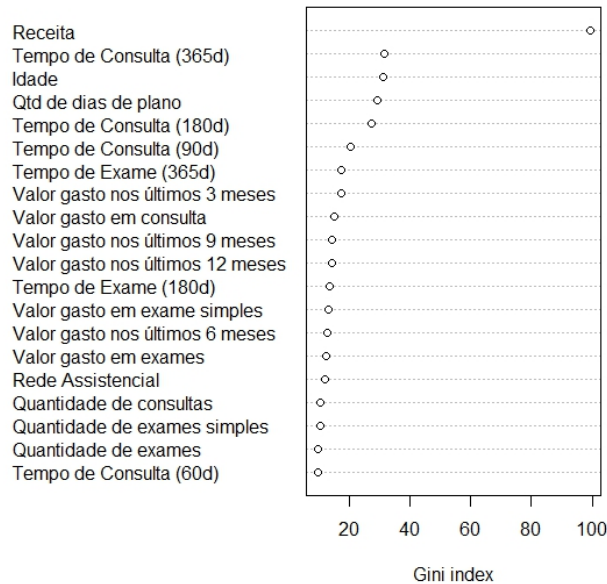


Figura 2 – Importância das variáveis

O modelo de regressão logística foi ajustado através da metodologia *stepwise* utilizando como critério de seleção a estatística AIC. Pode-se verificar através da tabela 8 que as variáveis importantes para a *random forest* foram selecionadas no modelo de regressão logística. Além disso, pelo p-valor, consideramos todas as variáveis, a um nível de significância $\alpha=5\%$, importantes para a previsão de *churn*. Pode-se comprovar também pela tabela 7 que a boa performance do modelo fez as métricas se tornarem equivalentes a *random forest*.

Modelo	Ponto de corte	Acc^+	Acc^-	G	W.acc	Prec	F
Regressão Logística	0,9	0,685323	0,9200451	0,794058	0,7244434	0,3512359	0,4644406

Tabela 7 – Métricas Modelo de Regressão Logística

A interpretação dos coeficientes da regressão logística demonstra que variáveis com coeficientes negativos indicam redução da probabilidade de *churn*, ao contrário dos coeficientes positivos. Como visto em *random forest* a variável receita também foi importante para regressão logística, indicando que clientes com maior receita tendem a cancelar menos o produto. Além disso, variáveis de utilização indicam uma tendência ao não cancelamento. Redes assistenciais mais caras indicam tendência ao cancelamento do produto, assim como se o cliente possui atraso da mensalidade.

A validação *out-of-sample* foi realizada a partir da base de teste dividida em etapas anteriores. A partir da configuração escolhida para *random forest* e das variáveis definidas pelo *stepwise* para regressão logística foram calculadas as métricas definidas para verificação da performance dos modelos, apresentados na tabela 9. Como era previsto a *random forest* captou melhor o comportamento dos dados e obteve melhor valor para todas as métricas que utilizamos para classificação dos modelos. Além disso, o equilíbrio entre as métricas Acc^+ e

Variável	Classes	Coefficiente	P-valor	Variável	Classes	Coefficiente	P-valor
Intercepto	*	-9.290e+00	<2e-16 ***	Valor gasto nos últimos 12 meses	*	-1.355e-04	5.70e-05 ***
Gênero	Masculino	-1.809e-01	1.89e-05 ***	Flag utilização (9 meses)	*	5.257e-05	1.68e-06 ***
Estado Civil	Casado	1.003e-01	0.093935 .	Valor gasto nos últimos 6 meses	*	1.958e-05	0.042314 *
Estado Civil	Viúvo	3.510e-01	0.038812 *	Valor gasto nos últimos 3 meses	*	-1.735e-05	0.054456 .
Estado Civil	Divorciado	4.415e-01	0.001524 **	Valor gasto em consultas	*	4.074e-04	3.67e-05 ***
Estado Civil	Separado	-7.253e-02	0.821537	Valor gasto em internação	*	1.013e-04	0.001127 **
Rede Assistencial	Rede 2	3.090e-01	8.28e-05 ***	Valor gasto em terapias complexas	*	9.995e-05	0.004098 **
Rede Assistencial	Rede 3	-2.418e-03	0.973439	Valor gasto em terapias simples	*	1.836e-04	0.092950 .
Rede Assistencial	Rede 4	8.524e-01	<2e-16 ***	Flag utilização (12 meses)	Utilizou	2.046e-01	0.062548 .
Rede Assistencial	Rede 5	1.731e+00	<2e-16 ***	Flag utilização (3 meses)	Utilizou	-3.575e-01	1.06e-07 ***
Rede Assistencial	Rede 6	2.124e+00	<2e-16 ***	Tempo de Consulta (90d)	*	2.723e-02	<2e-16 ***
Idade	*	4.529e-02	<2e-16 ***	Tempo de Consulta (180d)	*	1.143e-02	<2e-16 ***
Qtd de dias de plano	*	-1.104e-01	<2e-16 ***	Tempo de Consulta (365d)	*	-2.213e-03	3.59e-05 ***
Flag desconto na mensalidade	Com desconto	-1.777e-01	0.006774 **	Tempo de Exame (90d)	*	1.786e-02	1.16e-06 ***
Flag atraso na mensalidade	Com atraso	3.319e-01	6.39e-15 ***	Tempo de Exame (180d)	*	2.138e-03	0.019124 *
Receita	*	-1.677e-03	<2e-16 ***	Tempo de Internação (90d)	*	2.705e-02	0.004692 **
Quantidade de consultas	*	3.103e-02	1.63e-06 ***	Tempo de Internação (180d)	*	-5.734e-03	0.000485 ***
Quantidade de exames complexos	*	1.463e-01	0.010333 *	Tempo de Terapia (90d)	*	2.606e-02	0.002007 **
Quantidade de exames simples	*	6.434e-03	0.012894 *	Valor gasto em exames	*	-2.142e-04	0.072281 .
Quantidade de terapias simples	*	-9.425e-03	0.009198 **				

Tabela 8 – Resultado Modelo de Regressão Logística

Modelo	Acc^+	Acc^-	G	W.acc	Prec	F
<i>Random Forest</i>	0,7389034	0,9224206	0,8255784	0,7694896	0,3820882	0,5037081
Regressão Logística	0,6779809	0,9185784	0,7891632	0,7180804	0,3509009	0,4624518

Tabela 9 – Comparativo *out-of-sample random forest* e *logit*

Acc^- demonstra que as técnicas utilizadas nesse estudo foram eficazes para que a classe inferior obtivesse boa previsão de *churn* e comparadas com os valores apurados *in-sample* o resultado foi significativamente positivo.

5 Conclusão

CRM vem sendo tratado com maior importância pelas empresas e a retenção, como uma de suas etapas macro, contribui para a evolução de mecanismos para manutenção de clientes. A literatura é pouco desenvolvida na direção adotada nesse trabalho, quando observado o segmento de seguro saúde. Por isso, esse estudo é uma grande contribuição para a literatura devido a particularidades da base de dados e do segmento de atuação. Uma grande dificuldade encontrada a priori é a baixa frequência de clientes *churners*. Neste estudo, investigamos como melhorar a previsão ao tratarmos de classes extremamente desbalanceadas aplicando *balanced random forest* com diferentes configurações e regressão logística.

Para avaliar o impacto dessas técnicas, foram escolhidas algumas métricas e avaliados seus resultados para cada configuração. Se tratando de *random forests* foram selecionadas para cada grupo as configurações vencedoras para cada uma das métricas *g-mean*, *weighted accuracy* e *f-measure* e avaliados os resultados. A partir disso, não foram observadas diferenças significativas quando aumentado o número de árvores, apenas quando considerado tamanho de sub-amostras maiores.

Comparando *under-sampling* com *over-sampling*, observamos melhores resultados para *over-sampling*. Como o número de árvores não influenciou diretamente no resultado, a conclusão foi que ao utilizar um tamanho maior de sub-amostras para o treinamento da previsão, a *random forest* captou melhor as características dos clientes que determinam a discriminação entre *churners* e *não churners*.

Para trabalhos futuros sabe-se que informações macroeconômicas possuem forte relação com a perda de renda da população e conseqüentemente a escolha pela manutenção ou não do plano de saúde, dependendo de suas prioridades. Por isso, a inclusão de fatores externos pode contribuir para melhor captação do motivo e características que levam ao cancelamento. A análise exploratória nos mostrou diversas variáveis que aparentavam possuir diferentes distribuições. Uma outra abordagem seria a partir de mistura de especialistas, utilizando diferentes pesos a cada uma das distribuições selecionadas, inclusive com caudas pesadas, para tentar captar a distinção de comportamentos entre *churners* e *não churners*.

Referências

- ANS. *ANS Agencia Nacional de Saude Suplementar*. 2017. Disponível em: <<http://www.ans.gov.br/perfil-do-setor/dados-e-indicadores-do-setor>>.
- BEKKAR, M.; DJEMAA, H. K.; ALITOUICHE, T. A. Evaluation measures for models assessment over imbalanced data sets. *Journal Of Information Engineering and Applications*, v. 3, n. 10, 2013.
- BENOIT, D. F.; POEL, D. Van den. Improving customer retention in financial services using kinship network information. *Expert Systems with Applications*, Elsevier, v. 39, n. 13, p. 11435–11442, 2012.
- BOCK, K. W. D.; POEL, D. Van den. An empirical evaluation of rotation-based ensemble classifiers for customer churn prediction. *Expert Systems with Applications*, Elsevier, v. 38, n. 10, p. 12293–12301, 2011.
- BREIMAN, L. Bagging predictors. *Machine learning*, Springer, v. 24, n. 2, p. 123–140, 1996.
- BREIMAN, L. Random forests. *Machine learning*, Springer, v. 45, n. 1, p. 5–32, 2001.
- BUCKINX, W.; POEL, D. Van den. Customer base analysis: partial defection of behaviourally loyal clients in a non-contractual fmcg retail setting. *European Journal of Operational Research*, Elsevier, v. 164, n. 1, p. 252–268, 2005.
- BUREZ, J.; POEL, D. Van den. Crm at a pay-tv company: Using analytical models to reduce customer attrition by targeted marketing for subscription services. *Expert Systems with Applications*, Elsevier, v. 32, n. 2, p. 277–288, 2007.
- BUREZ, J.; POEL, D. Van den. Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, Elsevier, v. 36, n. 3, p. 4626–4636, 2009.
- CHEN, C.; LIAW, A.; BREIMAN, L. *Using random forest to learn imbalanced data*. [S.l.], july 2004. Disponível em: <xtf.lib.berkeley.edu/reports/SDTRWebData/accessPages/666.html>.
- COUSSEMENT, K.; POEL, D. Van den. Churn prediction in subscription services: An application of support vector machines while comparing two parameter-selection techniques. *Expert systems with applications*, Elsevier, v. 34, n. 1, p. 313–327, 2008.
- DRUMMOND, C.; HOLTE, R. C. et al. C4. 5, class imbalance, and cost sensitivity: why under-sampling beats over-sampling. In: CITESEER. *Workshop on learning from imbalanced datasets II*. [S.l.], 2003. v. 11.
- EFRON, B.; TIBSHIRANI, R. J. *An introduction to the bootstrap*. [S.l.]: CRC press, 1994.
- EIBEN, A.; KOUDIJS, A.; SLISSER, F. Genetic modelling of customer retention. In: SPRINGER. *European Conference on Genetic Programming*. [S.l.], 1998. p. 178–186.
- FARQUAD, M. A. H.; RAVI, V.; RAJU, S. B. Churn prediction using comprehensible support vector machine: An analytical crm application. *Applied Soft Computing*, Elsevier, v. 19, p. 31–40, 2014.
- FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. *The elements of statistical learning*. [S.l.]: Springer series in statistics Springer, Berlin, 2001.
- GUJARATI, D. N. *Basic econometrics*. [S.l.]: Tata McGraw-Hill Education, 2009.
- JR, D. W. H.; LEMESHOW, S.; STURDIVANT, R. X. *Applied logistic regression*. [S.l.]: John Wiley & Sons, 2013.
- KIM, K.; JUN, C.-H.; LEE, J. Improved churn prediction in telecommunication industry by analyzing a large network. *Expert Systems with Applications*, Elsevier, v. 41, n. 15, p. 6575–6584, 2014.

KOTLER, P. Administração de marketing: análise, planejamento, administração e controle. São Paulo: Ed. Atlas, 5ª edição, 1998.

KUBAT, M.; MATWIN, S. et al. Addressing the curse of imbalanced training sets: one-sided selection. In: NASHVILLE, USA. *ICML*. [S.l.], 1997. v. 97, p. 179–186.

LU, N. et al. A customer churn prediction model in telecom industry using boosting. *IEEE Transactions on Industrial Informatics*, IEEE, v. 10, n. 2, p. 1659–1665, 2014.

MORIK, K.; KÖPCKE, H. Analysing customer churn in insurance data – a case study. In: SPRINGER. *European Conference on Principles of Data Mining and Knowledge Discovery*. [S.l.], 2004. p. 325–336.

MOZER, M. C. et al. Predicting subscriber dissatisfaction and improving retention in the wireless telecommunications industry. *IEEE Transactions on neural networks*, IEEE, v. 11, n. 3, p. 690–696, 2000.

NIE, G. et al. Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, Elsevier, v. 38, n. 12, p. 15273–15285, 2011.

OWCZARCZUK, M. Churn models for prepaid customers in the cellular telecommunication industry using large data marts. *Expert Systems with Applications*, Elsevier, v. 37, n. 6, p. 4710–4712, 2010.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. [S.l.]: IME-USP São Paulo, 2004.

POEL, D. Van den; LARIVIERE, B. Customer attrition analysis for financial services using proportional hazard models. *European journal of operational research*, Elsevier, v. 157, n. 1, p. 196–217, 2004.

REICHHELD, F. F.; SASSER, W. E. Zero defections: Quality comes to services. *Harvard Business Review*, n. Sep–Oct, 1990.

ROCHA, A. da; CHRISTENSEN, C. *Marketing: Teoria E Prática No Brasil*. [S.l.]: Editora Atlas SA, 2000.

SU, J. et al. *Customer retention predictive modeling in HealthCare Insurance Industry*. Jacksonville, Florida, 2011.

WEI, C.-P.; CHIU, I.-T. Turning telecommunications call details to churn prediction: a data mining approach. *Expert systems with applications*, Elsevier, v. 23, n. 2, p. 103–112, 2002.

WEISS, G. M. Mining with rarity: a unifying framework. *ACM Sigkdd Explorations Newsletter*, ACM, v. 6, n. 1, p. 7–19, 2004.

XIE, Y. et al. Customer churn prediction using improved balanced random forests. *Expert Systems with Applications*, Elsevier, v. 36, n. 3, p. 5445–5449, 2009.

A Evolução do setor

Até o início da década de 60, a saúde do brasileiro era em grande parte financiada pelo sistema público, que possuía a maioria dos hospitais. E para lugares menos acessíveis, onde não existiam unidades próprias dos governos federal, estadual ou municipal, eram contratados serviços de hospitais privados para esses pacientes. Porém no início da década de 70, empresas privadas encontraram um mercado pouco assistido pelo sistema público e decidiram investir no setor. Sendo assim, desde esse período foi criado o conceito de Saúde Suplementar, pois a saúde passou a ser custeada, em grande parte, pela iniciativa privada (ROCHA; CHRISTENSEN, 2000).

Inicialmente as empresas não se preocupavam em fidelizar seus clientes. Além disso, as principais empresas ofereciam produtos com qualidade semelhantes, sem uma grande diferenciação uma das outras. Os clientes, como pessoas físicas, sempre possuíam pouco poder de negociação. Com isso, as empresas estabeleciam suas próprias regras em contrato e sem respeito aos direitos do consumidor. Desde então o cenário era controlado pela Superintendência de Seguros Privados (Susep), mas em 2000, para que o setor tivesse maior controle e regulação, foi criada a Agência Nacional de Saúde Suplementar (ANS) (ROCHA; CHRISTENSEN, 2000).

Em 1998 foi criada a Lei 9.656 que entrou em vigor com o objetivo de regular o setor de saúde suplementar no Brasil, estabelecendo regras importantes para a defesa dos direitos do consumidor, como teto para reajuste financeiro das mensalidades de planos individuais, relação entre faixas etárias de preço para proteção de clientes mais idosos, entre outras. Até os dias de hoje existem clientes com planos chamados não regulamentados, ou seja, contratados antes da Lei nº 9.656/98, os quais em Setembro de 2016 representavam cerca de 10% do total de clientes que possuem planos de saúde no Brasil (ANS, 2017).

Por conta de toda essa regulamentação, muitas operadoras deixaram de comercializar planos de saúde. A evolução do total de operadoras ativas no Brasil foi decrescendo linearmente, como podemos observar na Tabela 10, sendo que de Dez/06 à Set/06 houve uma redução de aproximadamente 32%. Todos esses fatores fazem com que a competição entre as empresas se torne cada vez mais acirrada e com isso os clientes passem a ter um tratamento diferenciado (ANS, 2017).

Ano	Total de operadoras com beneficiários
dez/06	1.610
dez/07	1.576
dez/08	1.521
dez/09	1.479
dez/10	1.411
dez/11	1.371
dez/12	1.320
dez/13	1.256
dez/14	1.217
dez/15	1.150
set/16	1.101

Tabela 10 – Evolução do total de operadoras ativas

B Estratégia para manutenção dos clientes

De acordo com Kotler (1998), o custo que as empresas possuem ao atrair novos consumidores é cinco vezes maior do que mantê-los satisfeitos. E por isso, atualmente grande parte das empresas investem no relacionamento com seus clientes de forma a torná-los mais próximos e aumentar a percepção positiva sob seus produtos. Segundo Kotler (1998), o processo que deverá ser utilizado pelas empresas com o objetivo de reduzir a perda de seus clientes é composto quatro etapas:

1. A empresa deve estabelecer e calcular uma medida de retenção de seus clientes, denominada taxa de retenção. Neste caso, considerando empresas de saúde, a taxa de renovação dos serviços pode ser considerada uma boa métrica para contratos Empresariais, onde a operadora oferece seus produtos para os funcionários das empresas contratantes, porém para contratos Individuais, onde não há renovação contratual, essa métrica deverá ser definida com base no planejamento estratégico das empresas;
2. A motivação do cancelamento do contrato por parte de seus clientes deverá ser sempre mapeada, para que a empresa consiga determinar se seus serviços estão adequados. Caso seja um cancelamento por uma qualidade ruim dos serviços, falta de atendimento em uma determinada ocasião ou até mesmo pelo cliente considerar os serviços caros demais, as empresas a partir desse mapeamento poderá decidir quais ações poderão ser feitas para mitigar os cancelamentos;
3. A partir das 1^o e 2^a etapas definidas, a empresa consegue calcular o lucro que seria obtido se o cliente não cancelasse o contrato. Nesta etapa a empresa deverá definir quais são os clientes lucrativos para seu negócio, pois se a relação entre a receita que o cliente paga e as despesas relacionadas à utilização dos serviços for negativa, a perda de determinados clientes pode se tornar uma vantagem;
4. Durante a última etapa deverá ser calculado o custo que seria empregado para realizar as ações de retenção de clientes e conseqüentemente o aumento da taxa de retenção. Esse custo deverá ser comparado com o lucro calculado na 3^a etapa, e caso seja inferior, o valor da diferença deverá ser designado às ações de mitigação de cancelamento de clientes.

De acordo com Reichheld e Sasser (1990), as empresas podem aumentar seus lucros entre 25 e 85% ao reduzir a taxa de cancelamento em 5%, e para isso as etapas definidas por Kotler (1998) são de extrema importância.

C Revisão da literatura para outros segmentos

Burez e Poel (2007) apresenta duas abordagens para gerenciamento do *churn* nas empresas: reativo e proativo. Numa abordagem reativa, a empresa geralmente aguarda uma solicitação de cancelamento por parte do cliente para que a partir desse momento ela possa agir oferecendo, caso seja de interesse, alguma vantagem para retenção dele. Já numa abordagem proativa, a empresa cria mecanismos para identificar os clientes mais propensos ao cancelamento antes mesmo dele tomar a atitude de fazê-la. Neste caso, modelos de predição são indicados e revelam características que identificam esses clientes mais propensos.

Wei e Chiu (2002) propuseram uma técnica de previsão de *churn*, para o mercado de Telecomunicações, que se baseava em informações contratuais do assinante e em alterações de padrões de suas chamadas, com o objetivo de identificar potenciais *churners* para um período de predição específico. A técnica proposta incorpora a abordagem de uma combinação de classes com vários classificadores pois há o desafio de trabalhar com distribuições estatísticas altamente distorcidas entre *churners* e não *churners*.

O trabalho de Poel e Lariviere (2004) trata do gerenciamento do relacionamento com o cliente (CRM) baseado no contexto de uma empresa europeia de serviços financeiros. A combinação de diversas variáveis preditoras, como idade, gênero e nível educacional, além de variáveis relacionadas ao comportamento real do cliente, torna o modelo em estudo mais abrangente. Utilizando um modelo de riscos proporcionais a partir de uma amostra aleatória, os autores sugerem como controle ao *churn* uma interação maior da empresa com seus clientes.

Farquard, Ravi e Raju (2014) utilizaram um conjunto de dados referente à clientes de cartão de crédito extremamente desbalanceado, onde apenas 6,76% eram *churners*. Eles propuseram uma abordagem híbrida, que possui três etapas de construção, para reparar uma grande desvantagem da técnica de SVM mais comumente utilizada, que é a dificuldade em interpretar o modelo a partir do conhecimento adquirida pela base de dados. A técnica de SVM-RFE (SVM - Recursive Feature Elimination) é utilizada na primeira etapa para redução do conjunto de recursos. Na segunda etapa, o conjunto determinado na etapa anterior é utilizado para obtenção do modelo SVM e seus vetores de suporte. A última etapa consistiu em utilizar a técnica de Naive Bayes Tree para obtenção das regras.

Coussement e Poel (2008) inicia o artigo dando uma grande importância aos projetos de CRM devido a concorrência intensiva entre empresas e mercados saturados, e por isso é crucial a construção de um modelo de previsão de *churn* para assegurar que seus melhores clientes não irão para a concorrência. A técnica de *machine learning* Support Vector Machine (SVM) é utilizada em um contexto de assinatura de um jornal. Além disso, o autor faz uma comparação com modelos de Regressão Logística e Random Florest.

Lu et al. (2014) determina a construção de dois *clusters* e utiliza no estudo o algoritmo de *boosting* para determinar pesos para cada um dos 2 *clusters* com o objetivo de aprimorar modelos de previsão de *churn*. O autor faz uma comparação entre um modelo de Regressão Logística e o modelo utilizando o algoritmo de *boosting* para cada *cluster* e sua conclusão é de que com o algoritmo de *boosting* a previsão ficou mais assertiva onde um dos *clusters* foi identificado com um alto risco de cancelamento.

Como proposta para modelagem de previsão de *churn*, foram apresentados dois classificadores baseados em rotação: *rotation forest* e *rotboost*. Bock e Poel (2011) também realizaram uma aplicação em quatro bases de dados e seus resultados mostraram que *rotation forest* teve maior desempenho quando comparadas à *rotboost*, considerando como medida de comparação a curva AUC (area under the curve), enquanto *rotboost* apresenta maior precisão. Além disso, foram feitas comparações com outras técnicas como: Análise de Componentes Principais (PCA), Análise de Componentes Independentes (ICA) e Projeções Aleatórias Esparsas (SRP).

D Análise Exploratória

Uma análise exploratória foi realizada para tentar entender o comportamento das variáveis em relação ao *churn*. Para manter a confidencialidade dos dados, neste capítulo foram apresentadas as análises de 3 das 54 variáveis utilizadas na modelagem: idade, receita e custo nos últimos 12 meses. Elas possuem, de acordo com a modelagem, grande importância para previsão dos clientes com alta chance de cancelamento do plano.

A figura 3 apresenta os gráficos de histograma e boxplot da variável idade para *churners* e *não churners*. Pode-se observar, a partir do gráfico de boxplot, que o grupo de *não churners* apresenta uma média e dispersão mais elevadas, e o grupo de *churners* apresenta uma distribuição dos dados negativamente assimétrica. O histograma nos mostra que existe uma mistura de diferentes distribuições, não apresentado apenas uma distribuição unimodal. Neste caso, modelos lineares podem ter dificuldades de captar esse tipo de comportamento.

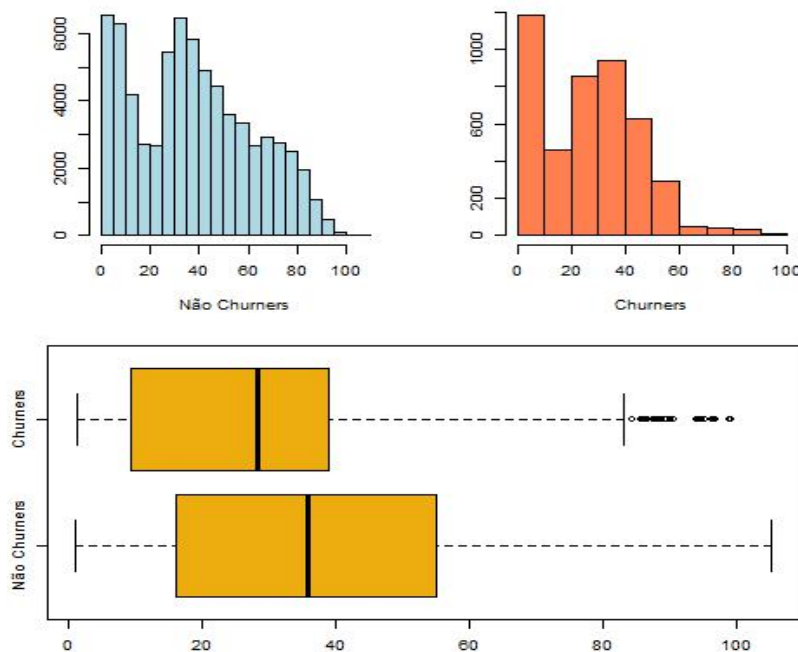


Figura 3 – Variável Exploratória: Idade do Cliente

É claro e evidente que quem paga mais pelo seguro saúde possui uma idade mais elevada ou busca um produto com mais qualidade. A figura 4 mostra que a variável apresenta distribuições bem diferentes para *churners* e *não churners* e em ambos os grupos uma mistura de distribuições unimodais pode ser visualizada. O boxplot apresenta uma cauda pesada para ambos os grupos e uma receita média maior para o grupo de *não churners*, além de possuir maior assimetria dos dados.

A variável custo de utilização apresenta grande concentração de valores baixos. Isso porque a consulta, que tem um custo inferior aos demais procedimentos, é a que possui maior frequência de utilização. Os valores extremos são referentes às poucas internações com extremo grau de complexidade que foram realizadas no período de análise.

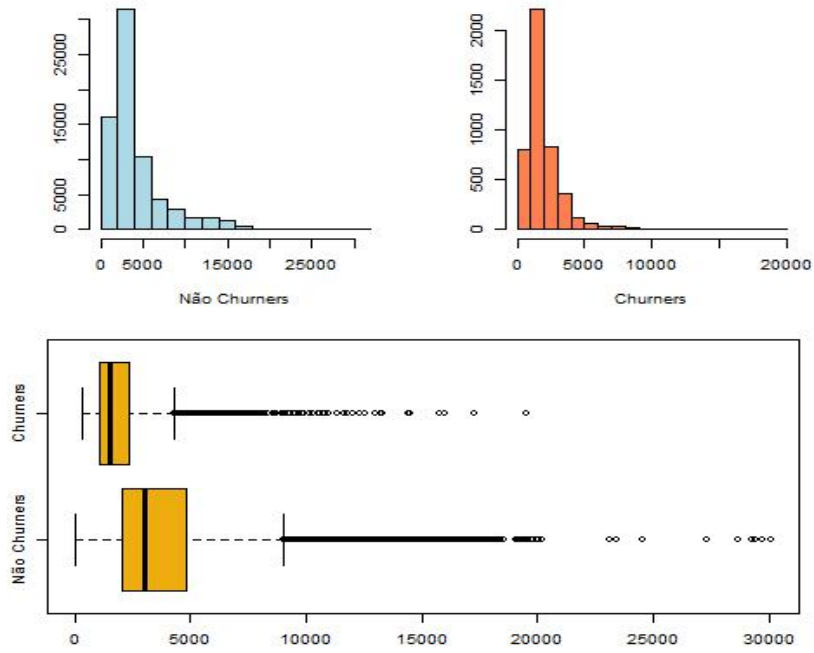


Figura 4 – Variável Exploratória: Receita paga pelo cliente

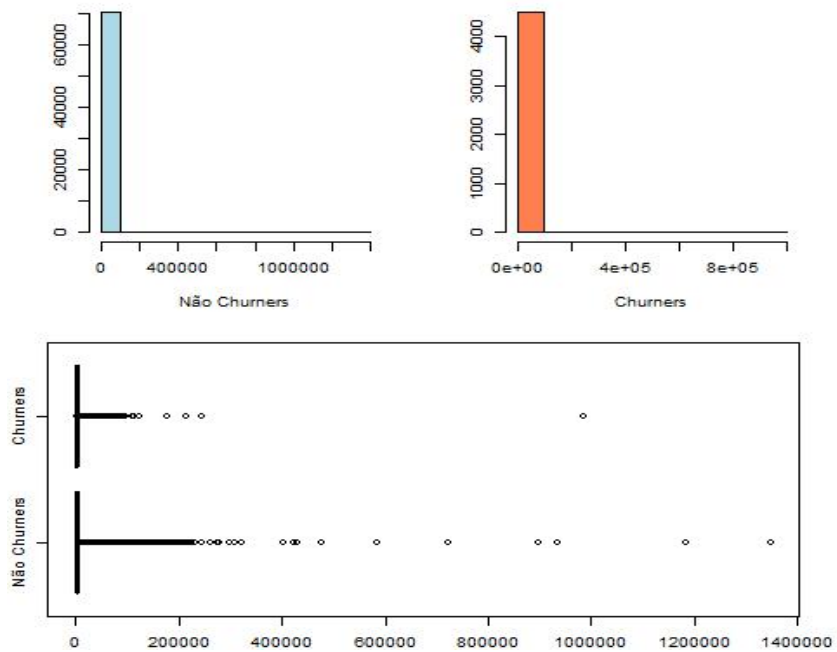


Figura 5 – Variável Exploratória: Custo de Utilização (12 meses)

E Tabelas completas *random forest*

Número de árvores	Ponto de corte	Acc^+	Acc^-	G	W.acc	Prec	F
100	*	0,4486	0,9981	0,6692	0,5402	0,9366	0,6067
100	0,5	0,4427	0,9983	0,6648	0,5353	0,9429	0,6025
100	0,7	0,5627	0,9935	0,7477	0,6345	0,8449	0,6755
100	0,9	0,7660	0,9035	0,8319	0,7889	0,3339	0,4651
200	*	0,4433	0,9983	0,6652	0,5358	0,9418	0,6028
200	0,5	0,4430	0,9983	0,6650	0,5355	0,9418	0,6026
200	0,7	0,5630	0,9946	0,7483	0,6349	0,8674	0,6828
200	0,9	0,7693	0,9114	0,8373	0,7930	0,3541	0,4849
500	*	0,4433	0,9984	0,6653	0,5358	0,9466	0,6038
500	0,5	0,4439	0,9983	0,6657	0,5363	0,9431	0,6036
500	0,7	0,5594	0,9947	0,7459	0,6319	0,8699	0,6809
500	0,9	0,7672	0,9162	0,8384	0,7920	0,3664	0,4960
1000	*	0,4466	0,9983	0,6677	0,5385	0,9446	0,6064
1000	0,5	0,4457	0,9984	0,6671	0,5378	0,9475	0,6062
1000	0,7	0,5588	0,9949	0,7456	0,6315	0,8734	0,6816
1000	0,9	0,7654	0,9185	0,8385	0,7909	0,3724	0,5010
2000	*	0,4463	0,9984	0,6675	0,5383	0,9469	0,6066
2000	0,5	0,4436	0,9984	0,6655	0,5361	0,9472	0,6042
2000	0,7	0,5615	0,9949	0,7474	0,6337	0,8736	0,6836
2000	0,9	0,7690	0,9183	0,8403	0,7939	0,3729	0,5022
5000	*	0,4466	0,9984	0,6677	0,5385	0,9470	0,6069
5000	0,5	0,4409	0,9984	0,6635	0,5338	0,9463	0,6015
5000	0,7	0,5615	0,9950	0,7475	0,6337	0,8772	0,6847
5000	0,9	0,7648	0,9197	0,8387	0,7906	0,3756	0,5038

Tabela 11 – Resultados Grupo 1

Número de árvores	Ponto de corte	Acc^+	Acc^-	G	W.acc	Prec	F
100	*	0,5857	0,9936	0,7629	0,6537	0,1076	0,1818
100	0,5	0,5729	0,9936	0,7545	0,6430	0,1392	0,2240
100	0,7	0,5729	0,9624	0,7425	0,6378	0,0268	0,0511
100	0,9	0,5833	0,7008	0,6394	0,6029	0,0035	0,0070
200	*	0,5347	0,9938	0,7290	0,6112	0,1883	0,2785
200	0,5	0,5497	0,9940	0,7392	0,6237	0,2065	0,3002
200	0,7	0,5762	0,9632	0,7450	0,6407	0,0426	0,0793
200	0,9	0,5762	0,6999	0,6350	0,5968	0,0054	0,0107
500	*	0,5803	0,9941	0,7595	0,6493	0,3962	0,4709
500	0,5	0,5483	0,9945	0,7384	0,6227	0,3963	0,4601
500	0,7	0,5767	0,9659	0,7464	0,6416	0,1007	0,1715
500	0,9	0,5881	0,7058	0,6442	0,6077	0,0131	0,0255
1000	*	0,5445	0,9940	0,7357	0,6194	0,5293	0,5368
1000	0,5	0,5448	0,9945	0,7360	0,6197	0,5506	0,5477
1000	0,7	0,6009	0,9662	0,7620	0,6618	0,1805	0,2776
1000	0,9	0,6100	0,7068	0,6566	0,6261	0,0251	0,0483
2000	*	0,5508	0,9942	0,7400	0,6247	0,6767	0,6073
2000	0,5	0,5512	0,9947	0,7404	0,6251	0,7046	0,6185
2000	0,7	0,6040	0,9666	0,7641	0,6644	0,2920	0,3937
2000	0,9	0,6139	0,7071	0,6588	0,6294	0,0456	0,0849
5000	*	0,5577	0,9941	0,7446	0,6304	0,7969	0,6562
5000	0,5	0,5191	0,9948	0,7186	0,5984	0,8052	0,6313
5000	0,7	0,6254	0,9667	0,7775	0,6823	0,4394	0,5162
5000	0,9	0,6439	0,7064	0,6744	0,6543	0,0839	0,1484

Tabela 12 – Resultados Grupo 2

Número de árvores	Tamanho de sub-amostras	Ponto de corte	Acc ⁺	Acc ⁻	G	W _{acc}	Prec	F
100	200	*	0,7839	0,7926	0,7882	0,7853	0,1927	0,3094
100	200	0,5	0,7877	0,7844	0,7861	0,7872	0,1875	0,3030
100	200	0,7	0,9291	0,4989	0,6808	0,8574	0,1048	0,1884
100	200	0,9	0,9961	0,1007	0,3167	0,8469	0,0654	0,1227
100	500	*	0,7907	0,8199	0,8052	0,7956	0,2171	0,3407
100	500	0,5	0,7839	0,8230	0,8032	0,7904	0,2185	0,3418
100	500	0,7	0,9232	0,5339	0,7021	0,8583	0,1112	0,1985
100	500	0,9	0,9926	0,1408	0,3738	0,8506	0,0680	0,1273
100	1000	*	0,7809	0,8496	0,8145	0,7923	0,2470	0,3753
100	1000	0,5	0,7809	0,8400	0,8099	0,7907	0,2357	0,3621
100	1000	0,7	0,9211	0,5537	0,7142	0,8599	0,1153	0,2050
100	1000	0,9	0,9899	0,1687	0,4087	0,8530	0,0700	0,1307
200	200	*	0,7821	0,7868	0,7844	0,7829	0,1881	0,3033
200	200	0,5	0,7687	0,8019	0,7851	0,7742	0,1968	0,3134
200	200	0,7	0,9330	0,4899	0,6761	0,8592	0,1036	0,1864
200	200	0,9	0,9976	0,0779	0,2788	0,8443	0,0640	0,1202
200	500	*	0,7767	0,8299	0,8029	0,7856	0,2239	0,3475
200	500	0,5	0,7797	0,8192	0,7992	0,7863	0,2141	0,3359
200	500	0,7	0,9259	0,5321	0,7019	0,8602	0,1111	0,1984
200	500	0,9	0,9932	0,1362	0,3677	0,8503	0,0677	0,1268
200	1000	*	0,7812	0,8518	0,8157	0,7930	0,2498	0,3785
200	1000	0,5	0,7770	0,8473	0,8114	0,7887	0,2433	0,3705
200	1000	0,7	0,9196	0,5608	0,7182	0,8598	0,1168	0,2073
200	1000	0,9	0,9908	0,1801	0,4225	0,8557	0,0709	0,1324
500	200	*	0,7800	0,7923	0,7861	0,7821	0,1918	0,3078
500	200	0,5	0,7803	0,7935	0,7869	0,7825	0,1927	0,3091
500	200	0,7	0,9372	0,4800	0,6707	0,8610	0,1022	0,1843
500	200	0,9	0,9964	0,0989	0,3139	0,8468	0,0653	0,1225
500	500	*	0,7865	0,8267	0,8064	0,7932	0,2228	0,3472
500	500	0,5	0,7806	0,8212	0,8007	0,7874	0,2162	0,3386
500	500	0,7	0,9271	0,5301	0,7011	0,8609	0,1108	0,1980
500	500	0,9	0,9940	0,1292	0,3584	0,8499	0,0673	0,1260
500	1000	*	0,7854	0,8566	0,8202	0,7972	0,2570	0,3873
500	1000	0,5	0,7842	0,8449	0,8140	0,7943	0,2421	0,3699
500	1000	0,7	0,9226	0,5595	0,7185	0,8621	0,1168	0,2074
500	1000	0,9	0,9923	0,1716	0,4127	0,8555	0,0703	0,1314
1000	200	*	0,7833	0,7940	0,7886	0,7851	0,1936	0,3105
1000	200	0,5	0,7830	0,7866	0,7848	0,7836	0,1882	0,3034
1000	200	0,7	0,9363	0,4891	0,6767	0,8618	0,1037	0,1868
1000	200	0,9	0,9967	0,0924	0,3034	0,8460	0,0649	0,1218
1000	500	*	0,7809	0,8237	0,8020	0,7880	0,2186	0,3416
1000	500	0,5	0,7812	0,8185	0,7996	0,7874	0,2138	0,3357
1000	500	0,7	0,9280	0,5302	0,7014	0,8617	0,1109	0,1982
1000	500	0,9	0,9923	0,1319	0,3618	0,8489	0,0673	0,1261
1000	1000	*	0,7854	0,8547	0,8193	0,7969	0,2545	0,3845
1000	1000	0,5	0,7851	0,8467	0,8153	0,7953	0,2443	0,3727
1000	1000	0,7	0,9226	0,5603	0,7190	0,8622	0,1170	0,2077
1000	1000	0,9	0,9917	0,1623	0,4012	0,8534	0,0696	0,1300
2000	200	*	0,7824	0,7951	0,7887	0,7845	0,1943	0,3113
2000	200	0,5	0,7836	0,7869	0,7852	0,7841	0,1884	0,3038
2000	200	0,7	0,9369	0,4843	0,6736	0,8614	0,1029	0,1855
2000	200	0,9	0,9964	0,0882	0,2965	0,8451	0,0646	0,1213
2000	500	*	0,7874	0,8236	0,8053	0,7935	0,2200	0,3439
2000	500	0,5	0,7848	0,8166	0,8005	0,7901	0,2128	0,3348
2000	500	0,7	0,9274	0,5322	0,7025	0,8615	0,1113	0,1987
2000	500	0,9	0,9949	0,1304	0,3602	0,8508	0,0674	0,1262
2000	1000	*	0,7854	0,8563	0,8201	0,7972	0,2566	0,3868
2000	1000	0,5	0,7812	0,8498	0,8148	0,7926	0,2472	0,3756
2000	1000	0,7	0,9226	0,5619	0,7200	0,8625	0,1174	0,2083
2000	1000	0,9	0,9920	0,1650	0,4046	0,8541	0,0698	0,1304
5000	200	*	0,7779	0,7995	0,7887	0,7815	0,1969	0,3142
5000	200	0,5	0,7806	0,7907	0,7856	0,7823	0,1906	0,3064
5000	200	0,7	0,9360	0,4874	0,6755	0,8612	0,1034	0,1863
5000	200	0,9	0,9970	0,0879	0,2960	0,8455	0,0646	0,1213
5000	500	*	0,7830	0,8248	0,8036	0,7899	0,2202	0,3437
5000	500	0,5	0,7845	0,8196	0,8018	0,7903	0,2154	0,3380
5000	500	0,7	0,9271	0,5312	0,7017	0,8611	0,1110	0,1983
5000	500	0,9	0,9932	0,1294	0,3585	0,8492	0,0672	0,1259
5000	1000	*	0,7857	0,8553	0,8197	0,7973	0,2553	0,3854
5000	1000	0,5	0,7839	0,8465	0,8146	0,7943	0,2439	0,3721
5000	1000	0,7	0,9226	0,5577	0,7173	0,8618	0,1164	0,2067
5000	1000	0,9	0,9914	0,1660	0,4057	0,8538	0,0698	0,1305

Tabela 13 – Resultados Grupo 3
Número de Sub-Amostras: 200, 500 e 1.000

Número de árvores	Tamanho de sub-amostras	Ponto de corte	Acc ⁺	Acc ⁻	G	W _{acc}	Prec	F
100	5000	*	0,7407	0,9334	0,8315	0,7728	0,4125	0,5299
100	5000	0,5	0,7324	0,9346	0,8273	0,7661	0,4142	0,5291
100	5000	0,7	0,8869	0,7128	0,7951	0,8579	0,1632	0,2757
100	5000	0,9	0,9750	0,2944	0,5358	0,8616	0,0803	0,1483
100	10000	*	0,6588	0,9729	0,8006	0,7112	0,6060	0,6313
100	10000	0,5	0,6538	0,9742	0,7981	0,7072	0,6155	0,6340
100	10000	0,7	0,8071	0,8418	0,8243	0,8129	0,2437	0,3744
100	10000	0,9	0,9461	0,4261	0,6349	0,8594	0,0943	0,1715
100	15000	*	0,5898	0,9889	0,7637	0,6563	0,7461	0,6588
100	15000	0,5	0,5777	0,9899	0,7562	0,6464	0,7570	0,6553
100	15000	0,7	0,7146	0,9186	0,8102	0,7486	0,3239	0,4457
100	15000	0,9	0,8183	0,5477	0,6695	0,7732	0,0898	0,1619
200	5000	*	0,7300	0,9377	0,8274	0,7646	0,4254	0,5375
200	5000	0,5	0,7291	0,9377	0,8268	0,7639	0,4251	0,5371
200	5000	0,7	0,8836	0,7112	0,7928	0,8549	0,1620	0,2738
200	5000	0,9	0,9809	0,2877	0,5313	0,8654	0,0800	0,1480
200	10000	*	0,6603	0,9743	0,8021	0,7126	0,6183	0,6386
200	10000	0,5	0,6505	0,9774	0,7974	0,7050	0,6449	0,6477
200	10000	0,7	0,8199	0,8493	0,8345	0,8248	0,2558	0,3899
200	10000	0,9	0,9497	0,4218	0,6329	0,8617	0,0940	0,1711
200	15000	*	0,5991	0,9892	0,7698	0,6641	0,7753	0,6759
200	15000	0,5	0,5872	0,9901	0,7625	0,6543	0,7852	0,6719
200	15000	0,7	0,7332	0,9253	0,8237	0,7652	0,3776	0,4984
200	15000	0,9	0,8707	0,5507	0,6925	0,8174	0,1070	0,1905
500	5000	*	0,7353	0,9407	0,8317	0,7696	0,4393	0,5501
500	5000	0,5	0,7318	0,9367	0,8279	0,7659	0,4220	0,5353
500	5000	0,7	0,8881	0,7078	0,7928	0,8580	0,1611	0,2727
500	5000	0,9	0,9821	0,2900	0,5337	0,8668	0,0804	0,1486
500	10000	*	0,6573	0,9766	0,8012	0,7105	0,6391	0,6481
500	10000	0,5	0,6570	0,9770	0,8012	0,7104	0,6438	0,6504
500	10000	0,7	0,8261	0,8472	0,8366	0,8296	0,2546	0,3892
500	10000	0,9	0,9586	0,4185	0,6334	0,8686	0,0943	0,1717
500	15000	*	0,6103	0,9901	0,7773	0,6736	0,7955	0,6907
500	15000	0,5	0,5990	0,9906	0,7703	0,6642	0,8003	0,6852
500	15000	0,7	0,7461	0,9243	0,8304	0,7758	0,3838	0,5068
500	15000	0,9	0,9155	0,5469	0,7075	0,8540	0,1132	0,2014
1000	5000	*	0,7365	0,9397	0,8319	0,7704	0,4356	0,5475
1000	5000	0,5	0,7291	0,9382	0,8271	0,7639	0,4270	0,5386
1000	5000	0,7	0,8848	0,7068	0,7908	0,8551	0,1601	0,2712
1000	5000	0,9	0,9818	0,2891	0,5328	0,8664	0,0802	0,1484
1000	10000	*	0,6582	0,9772	0,8020	0,7114	0,6455	0,6518
1000	10000	0,5	0,6579	0,9764	0,8015	0,7110	0,6378	0,6477
1000	10000	0,7	0,8249	0,8483	0,8366	0,8288	0,2557	0,3904
1000	10000	0,9	0,9595	0,4172	0,6327	0,8691	0,0942	0,1716
1000	15000	*	0,6052	0,9901	0,7741	0,6694	0,7948	0,6872
1000	15000	0,5	0,5969	0,9904	0,7689	0,6625	0,7972	0,6827
1000	15000	0,7	0,7600	0,9246	0,8383	0,7875	0,3889	0,5145
1000	15000	0,9	0,9265	0,5493	0,7134	0,8636	0,1149	0,2045
2000	5000	*	0,7324	0,9397	0,8296	0,7669	0,4342	0,5452
2000	5000	0,5	0,7333	0,9382	0,8294	0,7674	0,4282	0,5407
2000	5000	0,7	0,8842	0,7077	0,7911	0,8548	0,1604	0,2716
2000	5000	0,9	0,9809	0,2880	0,5315	0,8655	0,0801	0,1480
2000	10000	*	0,6633	0,9765	0,8048	0,7155	0,6406	0,6517
2000	10000	0,5	0,6588	0,9769	0,8023	0,7118	0,6435	0,6511
2000	10000	0,7	0,8297	0,8474	0,8385	0,8327	0,2557	0,3909
2000	10000	0,9	0,9577	0,4193	0,6337	0,8680	0,0943	0,1718
2000	15000	*	0,6020	0,9901	0,7720	0,6666	0,7926	0,6843
2000	15000	0,5	0,6014	0,9906	0,7718	0,6662	0,8022	0,6874
2000	15000	0,7	0,7627	0,9260	0,8404	0,7899	0,3942	0,5198
2000	15000	0,9	0,9333	0,5496	0,7162	0,8694	0,1157	0,2060
5000	5000	*	0,7330	0,9400	0,8301	0,7675	0,4357	0,5465
5000	5000	0,5	0,7282	0,9393	0,8270	0,7634	0,4310	0,5415
5000	5000	0,7	0,8854	0,7094	0,7925	0,8561	0,1614	0,2730
5000	5000	0,9	0,9815	0,2889	0,5325	0,8661	0,0802	0,1483
5000	10000	*	0,6579	0,9765	0,8015	0,7110	0,6384	0,6480
5000	10000	0,5	0,6576	0,9770	0,8016	0,7109	0,6436	0,6506
5000	10000	0,7	0,8339	0,8468	0,8403	0,8360	0,2558	0,3916
5000	10000	0,9	0,9601	0,4190	0,6342	0,8699	0,0945	0,1721
5000	15000	*	0,6040	0,9902	0,7734	0,6684	0,7960	0,6869
5000	15000	0,5	0,5984	0,9905	0,7699	0,6637	0,7989	0,6843
5000	15000	0,7	0,7589	0,9261	0,8383	0,7867	0,3934	0,5181
5000	15000	0,9	0,9357	0,5493	0,7169	0,8713	0,1159	0,2063

Tabela 14 – Resultados Grupo 3
Número de Sub-Amostras: 5.000, 10.000 e 15.000

Número de árvores	Tamanho de sub-amostras	Ponto de corte	Acc ⁺	Acc ⁻	G	W _{acc}	Prec	F
100	200	*	0,1426	0,9998	0,3776	0,2855	0,9756	0,2488
100	200	0,5	0,1069	0,9999	0,3269	0,2557	0,9836	0,1928
100	200	0,7	0,3054	0,9992	0,5524	0,4211	0,9589	0,4633
100	200	0,9	0,6207	0,8804	0,7392	0,6640	0,2469	0,3533
100	500	*	0,2712	0,9994	0,5206	0,3926	0,9650	0,4234
100	500	0,5	0,2441	0,9995	0,4940	0,3700	0,9704	0,3901
100	500	0,7	0,3626	0,9983	0,6016	0,4686	0,9298	0,5217
100	500	0,9	0,6645	0,8693	0,7600	0,6986	0,2430	0,3559
100	1000	*	0,3191	0,9990	0,5647	0,4325	0,9546	0,4784
100	1000	0,5	0,3063	0,9991	0,5532	0,4218	0,9572	0,4641
100	1000	0,7	0,3980	0,9970	0,6300	0,4979	0,8931	0,5507
100	1000	0,9	0,6999	0,8718	0,7811	0,7286	0,2564	0,3753
200	200	*	0,1736	0,9998	0,4166	0,3113	0,9782	0,2948
200	200	0,5	0,0634	0,9999	0,2518	0,2195	0,9816	0,1191
200	200	0,7	0,3087	0,9992	0,5554	0,4238	0,9584	0,4670
200	200	0,9	0,6392	0,8768	0,7486	0,6788	0,2468	0,3561
200	500	*	0,2787	0,9993	0,5277	0,3988	0,9600	0,4319
200	500	0,5	0,2852	0,9993	0,5339	0,4042	0,9628	0,4401
200	500	0,7	0,3790	0,9983	0,6151	0,4822	0,9333	0,5391
200	500	0,9	0,6743	0,8884	0,7740	0,7100	0,2762	0,3919
200	1000	*	0,3081	0,9991	0,5549	0,4233	0,9574	0,4662
200	1000	0,5	0,3180	0,9991	0,5636	0,4315	0,9553	0,4771
200	1000	0,7	0,3971	0,9976	0,6295	0,4972	0,9143	0,5538
200	1000	0,9	0,7014	0,8850	0,7879	0,7320	0,2782	0,3983
500	200	*	0,1137	0,9999	0,3372	0,2614	0,9845	0,2039
500	200	0,5	0,1072	0,9999	0,3274	0,2560	0,9809	0,1932
500	200	0,7	0,3269	0,9990	0,5714	0,4389	0,9531	0,4868
500	200	0,9	0,6392	0,8739	0,7474	0,6783	0,2425	0,3517
500	500	*	0,2682	0,9994	0,5178	0,3901	0,9667	0,4199
500	500	0,5	0,2474	0,9995	0,4973	0,3727	0,9674	0,3940
500	500	0,7	0,3665	0,9983	0,6049	0,4718	0,9312	0,5260
500	500	0,9	0,6728	0,8756	0,7675	0,7066	0,2546	0,3695
500	1000	*	0,3159	0,9991	0,5618	0,4297	0,9584	0,4751
500	1000	0,5	0,3177	0,9991	0,5634	0,4312	0,9570	0,4770
500	1000	0,7	0,3971	0,9980	0,6296	0,4973	0,9270	0,5561
500	1000	0,9	0,6898	0,8883	0,7828	0,7229	0,2807	0,3990
1000	200	*	0,1125	0,9999	0,3354	0,2604	0,9869	0,2020
1000	200	0,5	0,0962	0,9999	0,3101	0,2468	0,9818	0,1752
1000	200	0,7	0,3168	0,9990	0,5625	0,4305	0,9543	0,4756
1000	200	0,9	0,6314	0,8823	0,7464	0,6733	0,2532	0,3614
1000	500	*	0,2659	0,9994	0,5154	0,3881	0,9633	0,4167
1000	500	0,5	0,2706	0,9993	0,5200	0,3921	0,9629	0,4225
1000	500	0,7	0,3739	0,9983	0,6110	0,4780	0,9324	0,5338
1000	500	0,9	0,6719	0,8788	0,7684	0,7064	0,2593	0,3742
1000	1000	*	0,3188	0,9991	0,5644	0,4322	0,9571	0,4783
1000	1000	0,5	0,3138	0,9991	0,5599	0,4280	0,9573	0,4726
1000	1000	0,7	0,3986	0,9978	0,6307	0,4985	0,9209	0,5564
1000	1000	0,9	0,6966	0,8838	0,7847	0,7278	0,2746	0,3940
2000	200	*	0,1200	0,9999	0,3464	0,2666	0,9853	0,2139
2000	200	0,5	0,0866	0,9999	0,2943	0,2388	0,9831	0,1592
2000	200	0,7	0,3102	0,9991	0,5567	0,4250	0,9568	0,4685
2000	200	0,9	0,6436	0,8717	0,7490	0,6817	0,2406	0,3503
2000	500	*	0,2721	0,9994	0,5215	0,3933	0,9652	0,4245
2000	500	0,5	0,2653	0,9994	0,5149	0,3876	0,9632	0,4160
2000	500	0,7	0,3701	0,9983	0,6078	0,4748	0,9339	0,5301
2000	500	0,9	0,6740	0,8782	0,7694	0,7080	0,2590	0,3742
2000	1000	*	0,3194	0,9991	0,5649	0,4327	0,9563	0,4789
2000	1000	0,5	0,3162	0,9991	0,5620	0,4300	0,9576	0,4754
2000	1000	0,7	0,3992	0,9978	0,6312	0,4990	0,9210	0,5570
2000	1000	0,9	0,6943	0,8863	0,7844	0,7263	0,2783	0,3974
5000	200	*	0,1343	0,9998	0,3664	0,2785	0,9783	0,2361
5000	200	0,5	0,1012	0,9999	0,3181	0,2510	0,9827	0,1835
5000	200	0,7	0,3138	0,9991	0,5599	0,4280	0,9573	0,4726
5000	200	0,9	0,6401	0,8741	0,7480	0,6791	0,2430	0,3523
5000	500	*	0,2706	0,9994	0,5200	0,3921	0,9639	0,4226
5000	500	0,5	0,2667	0,9994	0,5163	0,3888	0,9634	0,4178
5000	500	0,7	0,3718	0,9983	0,6093	0,4763	0,9335	0,5318
5000	500	0,9	0,6788	0,8791	0,7724	0,7122	0,2617	0,3778
5000	1000	*	0,3180	0,9991	0,5636	0,4315	0,9570	0,4773
5000	1000	0,5	0,3162	0,9991	0,5620	0,4300	0,9576	0,4754
5000	1000	0,7	0,3992	0,9978	0,6312	0,4990	0,9210	0,5570
5000	1000	0,9	0,6954	0,8875	0,7856	0,7275	0,2808	0,4001

Tabela 15 – Resultados Grupo 4
Número de Sub-Amostras: 200, 500 e 1.000

Número de árvores	Tamanho de sub-amostras	Ponto de corte	Acc ⁺	Acc ⁻	G	W _{acc}	Prec	F
100	5000	*	0,3692	0,9986	0,6072	0,4741	0,9451	0,5309
100	5000	0,5	0,3772	0,9986	0,6137	0,4808	0,9434	0,5389
100	5000	0,7	0,4826	0,9958	0,6932	0,5681	0,8781	0,6229
100	5000	0,9	0,7279	0,9022	0,8104	0,7569	0,3197	0,4443
100	10000	*	0,3942	0,9985	0,6274	0,4949	0,9444	0,5562
100	10000	0,5	0,4061	0,9984	0,6367	0,5048	0,9407	0,5673
100	10000	0,7	0,5183	0,9952	0,7182	0,5978	0,8718	0,6501
100	10000	0,9	0,7434	0,9115	0,8232	0,7714	0,3467	0,4728
100	15000	0,5	0,4055	0,9985	0,6363	0,5043	0,9432	0,5671
100	15000	0,7	0,5275	0,9950	0,7245	0,6054	0,8686	0,6564
100	15000	0,9	0,7487	0,9109	0,8258	0,7758	0,3467	0,4739
100	15000	*	0,4165	0,9985	0,6449	0,5135	0,9453	0,5782
200	5000	*	0,3814	0,9986	0,6171	0,4842	0,9433	0,5431
200	5000	0,5	0,3698	0,9987	0,6077	0,4746	0,9459	0,5317
200	5000	0,7	0,4665	0,9965	0,6818	0,5548	0,8934	0,6129
200	5000	0,9	0,7285	0,9039	0,8114	0,7577	0,3237	0,4482
200	10000	*	0,4013	0,9985	0,6330	0,5008	0,9427	0,5630
200	10000	0,5	0,3965	0,9985	0,6293	0,4969	0,9447	0,5586
200	10000	0,7	0,5132	0,9954	0,7148	0,5936	0,8769	0,6475
200	10000	0,9	0,7443	0,9121	0,8239	0,7722	0,3485	0,4747
200	15000	0,5	0,4198	0,9985	0,6474	0,5162	0,9450	0,5813
200	15000	0,7	0,5397	0,9952	0,7329	0,6157	0,8775	0,6684
200	15000	0,9	0,7469	0,9180	0,8281	0,7755	0,3652	0,4906
200	15000	*	0,4147	0,9985	0,6435	0,5120	0,9470	0,5768
500	5000	*	0,3826	0,9986	0,6181	0,4852	0,9442	0,5445
500	5000	0,5	0,3718	0,9987	0,6094	0,4763	0,9462	0,5339
500	5000	0,7	0,4790	0,9963	0,6908	0,5652	0,8914	0,6232
500	5000	0,9	0,7321	0,9040	0,8135	0,7607	0,3252	0,4503
500	10000	*	0,4010	0,9985	0,6328	0,5006	0,9446	0,5630
500	10000	0,5	0,3954	0,9985	0,6283	0,4959	0,9445	0,5574
500	10000	0,7	0,5130	0,9957	0,7146	0,5934	0,8818	0,6486
500	10000	0,9	0,7419	0,9138	0,8234	0,7705	0,3521	0,4776
500	15000	0,5	0,4085	0,9985	0,6386	0,5068	0,9456	0,5705
500	15000	0,7	0,5272	0,9954	0,7245	0,6053	0,8798	0,6593
500	15000	0,9	0,7475	0,9159	0,8274	0,7756	0,3594	0,4855
500	15000	*	0,4153	0,9985	0,6439	0,5125	0,9451	0,5770
1000	5000	*	0,3781	0,9986	0,6145	0,4815	0,9463	0,5403
1000	5000	0,5	0,3730	0,9987	0,6104	0,4773	0,9478	0,5354
1000	5000	0,7	0,4754	0,9964	0,6883	0,5623	0,8927	0,6204
1000	5000	0,9	0,7315	0,9061	0,8141	0,7606	0,3298	0,4546
1000	10000	*	0,3998	0,9985	0,6318	0,4996	0,9444	0,5618
1000	10000	0,5	0,3951	0,9985	0,6281	0,4956	0,9438	0,5570
1000	10000	0,7	0,5088	0,9957	0,7118	0,5899	0,8827	0,6455
1000	10000	0,9	0,7386	0,9141	0,8217	0,7679	0,3520	0,4767
1000	15000	0,5	0,4135	0,9985	0,6426	0,5110	0,9462	0,5755
1000	15000	0,7	0,5266	0,9954	0,7240	0,6048	0,8792	0,6587
1000	15000	0,9	0,7461	0,9174	0,8273	0,7746	0,3633	0,4886
1000	15000	*	0,4162	0,9985	0,6447	0,5132	0,9465	0,5782
2000	5000	*	0,3778	0,9986	0,6142	0,4813	0,9435	0,5395
2000	5000	0,5	0,3748	0,9986	0,6118	0,4788	0,9445	0,5367
2000	5000	0,7	0,4781	0,9964	0,6902	0,5645	0,8932	0,6228
2000	5000	0,9	0,7306	0,9081	0,8145	0,7602	0,3342	0,4586
2000	10000	*	0,3983	0,9985	0,6307	0,4984	0,9442	0,5603
2000	10000	0,5	0,3948	0,9985	0,6278	0,4954	0,9438	0,5567
2000	10000	0,7	0,5127	0,9959	0,7145	0,5932	0,8872	0,6498
2000	10000	0,9	0,7386	0,9153	0,8222	0,7681	0,3552	0,4797
2000	15000	0,5	0,4117	0,9985	0,6412	0,5095	0,9453	0,5736
2000	15000	0,7	0,5290	0,9954	0,7257	0,6067	0,8780	0,6602
2000	15000	0,9	0,7469	0,9190	0,8285	0,7756	0,3679	0,4930
2000	15000	*	0,4192	0,9985	0,6470	0,5157	0,9469	0,5811
5000	5000	*	0,3775	0,9986	0,6140	0,4810	0,9449	0,5395
5000	5000	0,5	0,3757	0,9986	0,6125	0,4795	0,9446	0,5376
5000	5000	0,7	0,4793	0,9964	0,6911	0,5655	0,8944	0,6242
5000	5000	0,9	0,7306	0,9073	0,8141	0,7600	0,3323	0,4568
5000	10000	*	0,4004	0,9985	0,6323	0,5001	0,9439	0,5623
5000	10000	0,5	0,3948	0,9985	0,6278	0,4954	0,9431	0,5566
5000	10000	0,7	0,5109	0,9958	0,7132	0,5917	0,8841	0,6475
5000	10000	0,9	0,7401	0,9147	0,8228	0,7692	0,3540	0,4790
5000	15000	0,5	0,4093	0,9985	0,6393	0,5075	0,9450	0,5713
5000	15000	0,7	0,5278	0,9955	0,7249	0,6058	0,8821	0,6605
5000	15000	0,9	0,7437	0,9186	0,8265	0,7728	0,3659	0,4905
5000	15000	*	0,4186	0,9985	0,6465	0,5152	0,9462	0,5804

Tabela 16 – Resultados Grupo 4
Número de Sub-Amostras: 5.000, 10.000 e 15.000