Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Utilização de classificadores múltiplos com modelo de regressão em sobrevivência, na classificação de clientes.

Letícia Aparecida da Silva

Trabalho de Conclusão de Curso

Letícia Aparecida da Silva

Utilização de classificadores múltiplos com modelo de regressão em sobrevivência, na classificação de clientes.

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Letícia Aparecida da Silva e aprovado pela banca examinadora.

Aprovado em 30 de março de 2023.

Banca Examinadora:

- Prof. Dr. José Carlos Fogo (Orientador)
- Profª. Drª. Daiane Aparecida Zuanetti
- Profª. Drª. Tereza Cristina Martins Dias

Aos meus familiares e amigos por todo o apoio e incentivo.

Agradecimentos

Gostaria de agradecer primeiramente, aos meus pais Antônio e Edvana, que desde o ensino fundamental sempre me incentivaram e deram total apoio para que fosse atrás dos meus sonhos e realizações, e neste momento dedico toda minha trajetória futura a eles, sem isto não teria conseguido chegar até aqui.

Também agradeço ao meu irmão Vitor Hugo, que apesar de as vezes mais atrapalhar do que ajudar, sempre esteve do meu lado seja me fazendo passar raiva ou dar muitas risadas. Um agradecimento especial também a todos meus tios, avós, primos e familiares por todo o apoio prestado até o presente momento.

Agradeço também meus amigos André e Vitória, por desde os tempos do colégio me ajudarem a passar pelos momentos difíceis me alegrando sempre e sendo presentes em todas as mnhas conquistas, mesmo que de longe.

Um agradecimento especial ao meu namorado Victor, que há três anos vêm sendo o principal responsável por me incentivar e auxiliar tanto no ambiente acadêmico como também na vida, compartilhar meus momentos com você me foram responsáveis por me dar a oportunida de conseguir chegar até o final do ciclo.

Por fim, agradeço imensamente ao meu orientador, o professor Doutor José Carlos Fogo, que desde a orientação do meu projeto de iniciação científica me incentiva a procurar por coisas novas e a continuar trilhando um caminho dee aprendizado e esforço no dia a dia.

Agradeço a todos que compartilharam meu caminho acadêmico e que não foram citados aqui, porém cada pessoa teve seu papel imporante para as conquistas que vieram e virão.

Resumo

Atualmente, no relacionamento com clientes, as empresas e instituições apresentam interesse em identificar aqueles indivíduos propícios a abandonar um serviço, num processo identificado como *churn*, na tentativa de diminuir o prejuízo com possíveis rompimentos ou desligamentos destes clientes. Com isso, foram apresentados variados estudos com o intuito de prever este evento de interesse, através da utilização de classificadores múltiplos, aplicados através de aprendizado de máquina. Também é possível aplicar metodologias de análise de sobrevivência, a fim de estudar a taxa de falha nestes processos.

Neste trabalho o objetivo é avaliar métodos de classificação múltipla para classificação de indivíduos, diante de suas características pessoais, utilizando-se modelos de regressão em análise de sobrevivência. De maneira inicial, são aplicados os classificadores *bagging* e *boosting*. Sendo assim, são avaliados os resultados obtidos pelos modelos ajustados.

Palavras-chave: *análise de sobrevivência, churn, classificadores múltiplos, regressão.*

Abstract

Currently, in customer relationships, companies and institutions are interested in identifying those individuals prone to abandon a service, in a process identified as *churn*, in an attempt to reduce the damage with possible ruptures or disconnections of these customers. With this, several studies have been presented with the intention of predicting this event of interest, through the use of multiple classifiers, applied through machine learning. It is also possible to apply survival analysis methodologies, in order to study the failure rate in these processes.

This work aims to evaluate multiple classification methods for classifying individuals, given their personal characteristics, using regression models in survival analysis. Initially, the *bagging* and *boosting* classifiers will be applied. Then, the results obtained through the adjusted models will be evaluated.

Keywords: *multiple classifiers, survival analyses, churn, regression.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Na época atual, considerando o aumento da competitividade, muitas empresas colocaram como foco de mercado apenas a venda de produtos para novos clientes. Porém, diante desse cenário estas empresas atingiram um nível de saturação em que seria necessário encontrar novas estratégias de como se relacionar com o cliente. Sendo assim, a solução encontrada foi criar um segmento que ficou responsável em encontrar uma percepção melhorada das exigências e o que estes clientes esperam das empresas com os quais estão se relacionando. O intuito da estratégia adotada foi criar um vínculo duradoudo e com bom desempenho para satisfazer os dois lados (Assato, 2017). Esse segmento é chamado gestão de relacionamento (em inglês, *Costumer Relationship Management - CRM*).

O CRM é um processo de tratamento contínuo das informações do cliente que é de grande importância para a empresa, pois oferece benefícios e aprendizado contínuo para a mesma (Neslin *et al.*, 2006). Este processo utiliza de tecnologias com a intenção de automatizar, organizar e reunir as informações internas e externas do cliente em uma única base de dados, com o foco de entender melhor o comportamento destes clientes, sendo assim, capaz de intensificar os planos de negócios, como forma de evitar um processo de falha, que seria a quebra de contrato entre cliente e empresa, mantendo uma relação de fidelidade.

A priori, quando o CRM é utilizado importante ressaltar que se deve analisar todas as informações coletadas dos clientes de interesse. Por conta disso, com o objetivo de utilizar corretamente a gestão de relacionamento, a empresa que deseja tornar-se competitiva deve estruturar seus sistemas de bancos de dados e contratar profissionais da área. Em sequência, o processo natural performado é a organização dos dados, seleção das variáveis de interesse e desenvolvimento das análises estatísticas.

Sendo assim, outro recurso importante é o chamado *database marketing* (Nakano, 2017). Esta ferramenta utiliza processos estatísticos, criando modelos a partir de banco de dados, para caracterizar o comportamento dos clientes, com os mesmos intuitos já citados.

Uma forma de analisar informações dos clientes é através de modelos de regressão associados à análise de sobrevivência. Usualmente neste estudo é utilizado em análises médicas ou industriais, no qual se refere à técnicas estatísticas que são usadas para avaliação de tempos de falha de indivíduos ou objetos através de métodos estatísticos e computacionais. Neste trabalho são abordados diferentes modelos de sobrevivência aplicados à regressão.

Nesta análise consideram uma variável resposta dada pelo tempo de observação de um indivíduo e uma variável binária associada, assumindo o valor 1 quando esse tempo se refere à ocorrência de um evento de interesse, normalmente chamado de falha (e que nesse contexto será denominado *churn*) ou 0 quando a falha não ocorre, o que denotamos como censura.

Sendo assim, diversos métodos são propostos com o objetivo de aumentar a proporção de acerto dos modelos, tornando-os mais eficazes, como por exemplo, utilizando algoritmos de *machine learning* (Dias, 2012); (Santos, 2010).

Um procedimento que hoje está sendo muito utilizado a fim de se obter esta melhoria são os chamados de classificadores múltiplos. Segundo Santos (2010), estes classificadores múltiplos tem como objetivo um conceito intuitivo, no qual podemos pressupor que as opiniões dadas por um grupo de especialistas tendem a ser melhores se comparadas à opinião de apenas um isoladamente.

Para este trabalho, aplicamos os classificadores múltiplos *bagging* e *boosting*, que utilizam de recursos computacionais avançados, sendo largamente aplicados na atualidade.

Dito isso, os classificadores aplicados neste trabalho, são aqueles em que a probabilidade de falha, chamada de *churn*, que é a probabilidade de um cliente cancelar ou desistir de um serviço/produto, é ajustado através das metodologias obtidas pela análise de sobrevivência.

Portanto, este trabalho tem como objetivo principal elaborar algoritmos computacionais dos classificadores múltiplos em conjunto com modelos de sobrevivência e comparar os resultados obtidos contra as metodologias usuais como regressão logística, regressão por árvore e classificação por árvore.

Este trabalho está organizado da seguinte maneira. No próximo capítulo, apresenta-

mos definições e conceitos básicos de análise de sobrevivência assim como as principais metodologias que serão utilizadas nas análises deste estudo. No Capítulo 3 são apresentados os resultados que foram obtidos através das aplicações das metodologias presentes neste trabalho, assim como as interpretações dos mesmos, onde foram avaliados os diferentes métodos de classificação. No Capítulo 4, são apresentadas as conclusões obtidas após as diferentes análises aplicadas.

Capítulo 2

Metodologia

2.1 Análise de Sobrevivência

A metodologia de análise de sobrevivência se refere ao estudo de dados de tempos até a ocorrência de um evento relacionado a um índíviduo ou objeto, através de um conjunto de técnicas estatísticas. É mais amplamente utilizada na área médica, porém, é de grande utilidade na indústria, chamada de análise de confiabilidade, no qual é feito o estudo da duração de componentes até o seu tempo de falha, quando os mesmos deixam de operar corretamente, por exemplo. De maneira geral o tempo de vida de um indivíduo ou componente é contabilizado até um evento de interesse (Fogo, 2007).

Sendo assim, o tempo de origem equivale ao momento em que o indivíduo entra no estudo, podendo ser o início de um tratamento, ocorrência de uma adversidade, início de um mal funcionamento etc. Também temos a medida de tempo final, que pode ser o instante da cura de uma doença, ou até mesmo o momento da morte de um paciente, a quebra do sistema, rompimento de um contrato etc. (Collett, 2015).

Neste contexto definimos tempo de vida, denotado como t, é contabilizado desde a entrada do indivíduo no estudo até a ocorrência do evento em questão. No caso deste estudo, este tempo será contabilizado até a ocorrência do evento de interesse que será o rompimento do vínculo entre o cliente e a empresa, denominado *churn*, ou até o encerramento do estudo.

Podemos ter o caso em que um indivíduo que está no estudo não tenha solicitado o rompimento do contrato, nesta situação, chamamos este evento de censura, e será melhor definido nas próximas seções.

2.1.1 Censura

Uma característica importante de dados de tempos de vida é a presença de observações incompletas, também chamadas de censuras. Definimos uma variável indicadora $c_i = 1$, representando uma falha, ou seja, a ocorrência do evento de interesse e $c_i = 0$, $i = 1, 2, \ldots, n$, representando uma censura.

Um evento é classificado como censura quando, por algum motivo, o tempo até o evento de interesse não é observado para alguns indivíduos presentes no estudo (Lawless, 2011). Esta ocorre quando não se consegue observar os tempos de sobrevivência de todos os indivíduos. As causas desta perda de informação podem ser:

- Ocorrência de eventos por alguma causa não relacionada com o estudo;
- Término do estudo antes que o evento de interesse aconteça;
- Abandono do estudo; etc

Apesar da censura gerar problemas na análise, já que não temos disponíveis os tempos de vida para alguns dos indivíduos, estes tempos devem ser incluídos nas análises, pois carregam a informação de que até aquele determinado instante de tempo este indivíduo ainda estava no estudo e não havia ocorrido a falha (Fogo, 2007).

Esquemas de Censura

- **Censura de tipo I**: aqui, o estudo ou experimento é feito até um tempo limite pré-fixado L. Desta forma, os indivíduos que tiveram falhas observadas dentro deste tempo serão registrados, e aqueles que ainda continuarem vivos ou funcionando terão seus tempos de vida censurados. Uma característica deste tipo de censura é que o número de falhas é aleatório, ou seja, podem ocorrer qualquer quantidade de falhas dentro do tempo limite. A Figura 2.1 ilustra o acontecimento de censuras do tipo I.

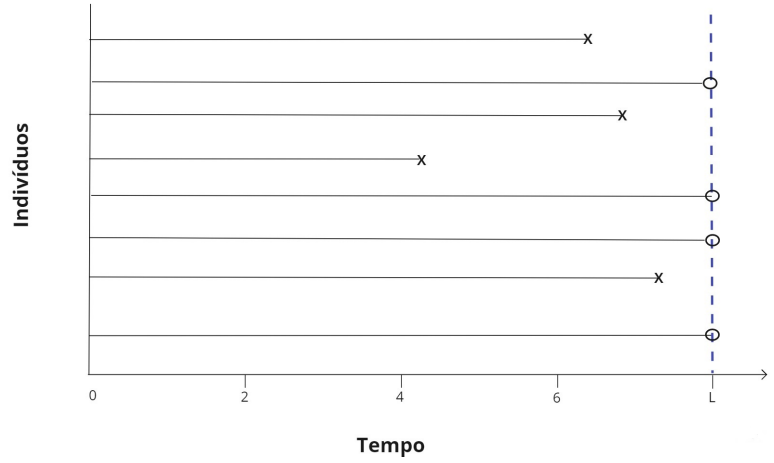

Figura 2.1: Exemplo de caso onde ocorre a censura do tipo I.

- **Censura do tipo II**: para este tipo de censura, são utilizados n unidades em teste, em que o experimento será conduzido até que se atinja um número r fixado inicialmente, de falhas que podem ocorrer. Sendo assim, as $n - r$ unidades que não apresentaram o evento de interesse após a ocorrência das r falhas serão censuradas como tendo seus tempos de vida t_r, onde este é o tempo de falha do r-ésimo indivíduo (Fogo, 2007). A Figura 2.2 ilustra a ocorrência de censuras tipo II.

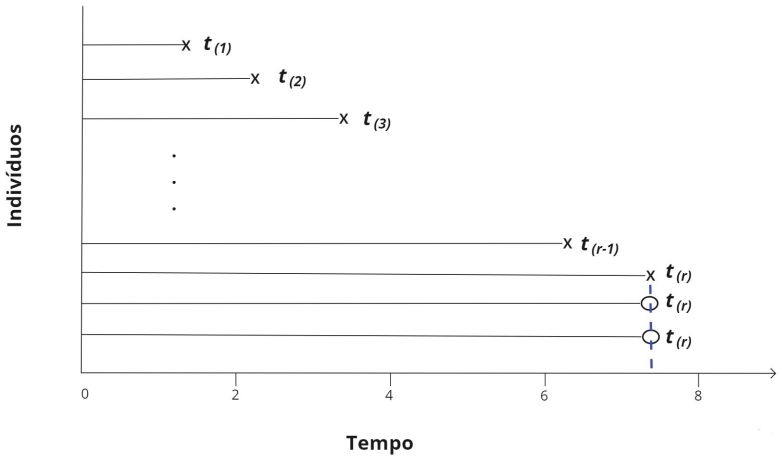

Figura 2.2: Exemplo de caso onde ocorre a censura do tipo II.

- **Censura Aleatória**: esta censura ocorre quando, em alguns tipos de experimento, não se tem o número adequado de indivíduos para o estudo logo no início, sendo necessário a entrada posterior de novos indivíduos em tempos aleatórios, após o início do estudo. Outra caracterização de censura aleatória se dá quando um indivíduo deixa o estudo sem que tenha apresentado o evento de interesse: por abandono, ocorrência da falha por uma outra causa ou qualquer outro motivo que o impeça

de continuar. Esta censura se assemelha com a Censura tipo I, pois seu número de falhas também será aleatório, porém não se têm o mesmo tempo de início em todos os indivíduos (Fogo, 2007). A ocorrência de censura aleatória está ilustrada na Figura 2.3.

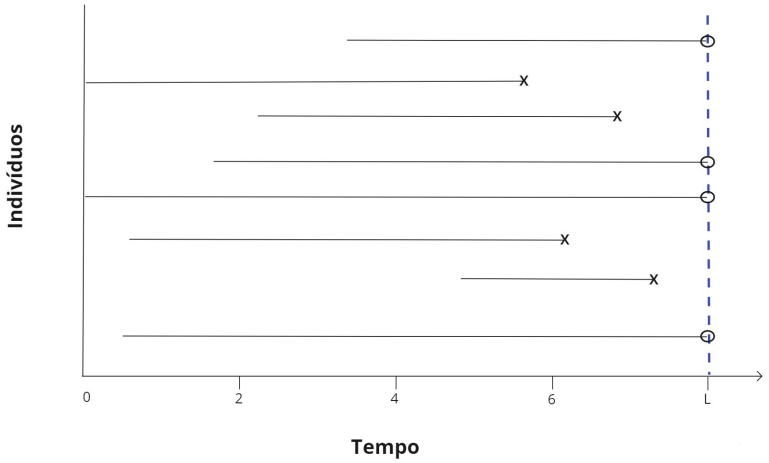

Figura 2.3: Exemplo de caso onde ocorre a censura aleatória.

2.1.2 Função de Sobrevivência

Podemos considerar o tempo de vida t, como uma observação de uma variável aleatória contínua e não negativa definida como T. Esta possui início determinado pelo instante em que o indivíduo começou a fazer parte do estudo até a ocorrência do evento de interesse. (Cintra, 2021). A função de distribuição de T é definida como (Casella e Berger, 2002):

$$F(t) = P(T \leq t). \tag{2.1}$$

E com função densidade de probabilidade dada por

$$f(t) = \frac{\partial F(t)}{\partial t}. \tag{2.2}$$

A função de sobrevivência define a probabilidade de um indivíduo não sofrer a falha até um determinado tempo t (Cintra, 2021). Em termos probabilísticos (Colosimo e Giolo, 2006),

$$S(t) = P(T > t). \tag{2.3}$$

2.1.3 Função de Risco

A função de risco representa a probabilidade de ocorrência do evento de interesse, condicionada à sobrevivência no início de um pequeno intervalo de tempo, $[t, t + dt)$. A definição da função de risco é

$$h(t) = \lim_{dt \to 0} \frac{P(t < T < t + dt | T > t)}{dt}. \tag{2.4}$$

Um ponto particularmente importante é que existe uma relação entre as funções de sobrevivência, de risco e a densidade, no qual estas podem ser escritas como uma sendo função das outras, como definido abaixo (Colosimo e Giolo, 2006):

$$f(t) = S(t)h(t). \tag{2.5}$$

Temos também uma expressão que chamaremos de Função de risco acumulada, que será escrita como (Colosimo e Giolo, 2006),

$$H(t) = \int_0^t h(u)d(u). \tag{2.6}$$

Utilizando a relação (2.6) podemos reescrever a função de sobrevivência de forma que esta seja resultante da função de risco acumulado,

$$S(t) = \exp\{-H(t)\}. \tag{2.7}$$

Desta forma, podemos reproduzir a relação (2.5) utilizando o resultado acima de forma que a função densidade de probabilidade da variável aleatória T seja apenas descrita através da função de risco.

$$f(t) = h(t) \exp\{-H(t)\}. \tag{2.8}$$

Dependendo de como se comporta o modelo utilizado, a função de risco pode ser crescente ou decrescente, na prática utilizadas pelo modelo de *Weibull*, dada por:

$$h(t) = \lambda \delta t^{\delta - 1}. \tag{2.9}$$

Estes modelos serão melhor definidos nas seções seguintes. Outras possibilidades de função de risco também têm sido exploradas, sendo elas:

- Função de risco com ponto de máximo ($h(t)$ unimodal);
- Função de risco em formato de banheira.

Além disso, a função de risco pode ser considerada em intervalos distintos de tempo, como por exemplo o formato de banheira, em que têm-se três fases bem definidas: decrescente, constante (praticamente) e crescente.

2.1.4 Função de Verossimilhança

O ajuste do modelo se faz pela função de verossimilhança, que é definida por:

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} [f(t_i)]^{c_i} [S(t_i)]^{1-c_i}, \tag{2.10}$$

em que $\boldsymbol{\theta}$ é o vetor de parâmetros do modelo e c_i é o indicador de falha.

Utilizando a relação (2.5), podemos escrever a função de verossimilhança como (Fogo, 2007):

$$L(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} [(h(t)]^{c_i} [S(t_i)]^{c_i} [S(t_i)]^{1-c_i}$$

$$= \prod_{i=1}^{n} [(h(t)]^{c_i} S(t_i). \tag{2.11}$$

Feito isso, considerando a expressão (2.8), podemos ainda escrever

$$L(\boldsymbol{\theta}|D) = \left[\prod_{i=1}^{n} [h(t_i)^{c_i}]\right] exp\left\{-\sum_{i=1}^{n} \int_0^{t_i} h(u)d(u)\right\}. \tag{2.12}$$

2.1.5 Modelos de Regressão

Na metodologia estatística, são utilizados modelos de regressão com o intuito de explicar a relação entre duas ou mais variáveis. Estes modelos servem, primeiramente, para explicar a relação de um conjunto de covariáveis com a variável resposta Y (ou vetor de

respostas), assim como também para realizar a predição de novas observações através de características já conhecidas presentes em um vetor $\mathbf{Z}$ de covariáveis, chamadas de variáveis preditoras.

Sendo assim, o interesse deste trabalho será estudar, através de modelos de regressão, a influência que as características dos clientes presentes no banco de dados exercem sobre os tempos de vida dos indivíduos.

Existem diversas abordagens para os modelos de regressão, como por exemplo modelos semi-paramétricos, utilizando o modelo de riscos proporcionais de Cox Collett (2015) e Carvalho *et al.* (2011), também temos modelos paramétricos encontrados em Lawless (2011), inclusive estes modelos podem ser expostos através de processos de contagem apresentado em Gill (1984).

A influência das covariáveis para predizer a probabilidade na ocorrência do *churn*, num espaço de tempo determinado, é representada por meio da modelagem da função de risco $h(t)$.

Considere um conjunto de dados de tempos de vida com n indivíduos e p covariáveis. Seja T uma variável aleatória representando o tempo de se ocorrer um determinado evento, o modelo de regressão estabelece que a função de risco de um indivíduo, no tempo t, dado o vetor de covariáveis $\mathbf{z} = (z_1, z_2, \ldots, z_p)^t$ é

$$h(t_i|\mathbf{z}_i) = h_0(t_i)g(\mathbf{z}_i|\boldsymbol{\beta}), i = 1, 2, \ldots, n, \tag{2.13}$$

sendo $\boldsymbol{\beta} = (\beta_1,\ \beta_2,\ \ldots,\ \beta_p)^t$ o vetor de parâmetros a serem estimados, $\mathbf{z}_i$ o vetor de covariáveis do indivíduo i, $g(.)$ uma função positiva de covariáveis e $h_0(t)$ conhecida como função de risco de base, sendo a função de risco de um indivíduo para o qual $g(\mathbf{z}) = 1$ (Collett, 2015).

Uma propriedade bastante utilizada do modelo (2.13) se dá pela proporcionalidade da razão entre os riscos de dois indivíduos ou componentes, isto é, eles não são dependentes do tempo. Por exemplo, a razão de riscos com variáveis $\mathbf{z}_i$ e $\mathbf{z}_j$ é dada por (Fogo, 2007),

$$\frac{h(t_i|\mathbf{z}_i)}{h(t_j|\mathbf{z}_j)} = \frac{g(\mathbf{z}_i|\boldsymbol{\beta})}{g(\mathbf{z}_j|\boldsymbol{\beta})}. \tag{2.14}$$

Por conta da propriedade (2.14) assume-se que $g(\mathbf{z}_i|\boldsymbol{\beta})$ deve ser não negativa, uma escolha apropriada seria fazer $g(\mathbf{z}_i) = \exp\left\{\boldsymbol{\beta}^t\mathbf{z}_i\right\}$. Sendo assim, a função de risco para o

indivíduo pode ser reescrita por (Wienke (2010))

$$h(t_i|\mathbf{z}_i) = h_0(t_i)\exp\left\{\boldsymbol{\beta}^t\mathbf{z}_i\right\} \tag{2.15}$$

$$= h_0(t_i)\exp\left\{\beta_1 z_{i1} + \beta_2 z_{i2} + \cdots + \beta_p z_{ip}\right\}, \tag{2.16}$$

em que β_0 é incorporado na função de risco de base.

Utilizando as funções (2.7) e (2.13), têm-se que a função de sobrevivência pode ser escrita como

$$S(t|\mathbf{z}) = \exp\left\{-\int_0^t h_0(u)g(\mathbf{z}|\boldsymbol{\beta})du\right\}, \tag{2.17}$$

na qual temos que a função de sobrevivência de T, dada as covariáveis de $\mathbf{z}$ é

$$S(t|\mathbf{z}) = [S_0(t)]^{g(\mathbf{z}|\boldsymbol{\beta})}, \tag{2.18}$$

em que, $S_0(t)$ é a sobrevivência no instante t, obtida a partir da função de risco de base.

Existem alguns modelos de regressão em sobrevivência como os modelos paramétricos, em especial os modelos exponencial e *Weibull*, onde assumimos que os tempos de sobrevivência possuem uma distribuição de probabilidade, indexada por um ou mais parâmetros, assim como também temos os modelos de regressão semi-paramétricos, como por exemplo o modelo de Cox. Nessas situações têm-se o interesse em estimar os efeitos das covariáveis sobre o tempo de vida nos indivíduos.

Dito isso, a seguir serão definidos os modelos paramétricos exponencial e *Weibull*, como também o modelo semi-paramétrico de Cox.

Modelo Exponencial

Apesar de simples, o modelo de regressão exponencial é historicamente o mais utilizado na análise de sobrevivência, principalmente na área industrial. Este deverá ser utilizado quando assume-se que o risco é constante ao longo do tempo.

Neste modelo têm-se que a função densidade de probabilidade de T é igual á

$$f(t|\lambda) = \lambda e^{-\lambda t},\ t \geq 0 \text{ e } \lambda > 0.$$

Uma forma de expressar o modelo de regressão exponencial é dado pela relação $\lambda(\mathbf{z}) = g(\mathbf{z}|\boldsymbol{\beta})$ e, dado que o risco do modelo é constante e igual a $\lambda(\mathbf{z})$, têm-se que $h(t|\mathbf{z}) = g(\mathbf{z}|\boldsymbol{\beta})$

(Fogo, 2007). Considerando a relação (2.16), o modelo de regressão exponencial é definido por

$$h(t_i|\mathbf{z}_i) = h_0(t_i) \exp\{\boldsymbol{\beta}'\mathbf{z}_i\}, \tag{2.19}$$

em que $h_0(t_i)$ é a função de risco de base do i-ésimo indivíduo (Fogo, 2007).

Modelo *Weibull*

Este modelo utiliza a distribuição de *Weibull* da forma $f(t) = \lambda\delta t^{\delta-1}\exp\{-\lambda t^{\delta}\}$, com $t \geq 0$ e $\lambda, \delta > 0$, tendo um parâmetro de escala λ e um de forma δ. Sendo assim, a função de risco para o modelo de *Weibull* é dada por

$$h(t|\mathbf{z}) = \lambda\delta t^{\delta-1} = \delta t^{\delta-1} g(\mathbf{z}|\boldsymbol{\beta}). \tag{2.20}$$

A função de risco pode ser reescrita como

$$h(t|\mathbf{z}) = \delta t^{\delta-1} \exp\{\boldsymbol{\beta}'\mathbf{z}\}, \tag{2.21}$$

considerando que $g(\mathbf{z}|\boldsymbol{\beta}) = \exp\{\boldsymbol{\beta}'\mathbf{z}\}$.

Uma outra abordagem possível para modelos de regressão, considerando o princípio da proporcionalidade, é estimar os efeitos que as covariáveis $\mathbf{z}$ exercem sobre os tempos de sobrevivência sem a necessidade de se fazer alguma suposição sobre os tempos de vida, como acontece nos modelos paramétricos vistos anteriormente. Esta metodologia ficou conhecida como modelo de regressão de Cox (Cox, 1972).

Este modelo não assume nenhuma distribuição sobre $h_0(t)$ e também é muito utilizado em dados com censura, além de que incorpora muito bem covariáveis com dependência no tempo.

2.2 Classificadores Múltiplos

Uma metodologia muito utilizada hoje em dia seria a classificação de indivíduos ou componentes, na qualidade de bom ou ruim, com o intuito de por exemplo, classificar o cliente como bom ou mal pagador, no ambiente de crédito, classificar um produto como tendo bom funcionamento, na indústria ou até mesmo em classificar um cliente a fim de

fazer com que ele permaneça na carteira de assinantes. Existem algumas metodologias estatísticas que realizam este procedimento como por exemplo a classificação por árvore e regressão logística.

Após alguns estudos, porém, percebeu-se que a real função que representa o problema em análise pode não estar presente no espaço de modelos possíveis em um problema de classificação de maneira individual. Desta forma, foi iniciada a utilização dos *classificadores múltiplos* com o intuito de evitar esta situação, já que os mesmos são realizados através de uma combinação da decisão tomada a partir dos individuais (Dias, 2012).

Segundo Santos (2010), o conceito de classificadores múltiplos têm fundamento em um conceito intuitivo, podemos imaginar que opiniões de um grupo de especialistas para resolver um problema, tende a ser de qualidade superior à opinião de apenas um especialista. A ideia geral é que os classificadores múltiplos são constituídos de um conjunto de classificadores individuais. Teoricamente, o conjunto de modelos possíveis será aumentado, assim pode-se alcançar um resultado melhor (Dias, 2012).

Neste estudo, os classificadores múltiplos utilizados serão os métodos *bagging* e *boosting*, adaptados à função de regressão em sobrevivência, a fim de comparar quais deles apresentam melhores resultados, assim como comparar estes resultados com os já obtidos em estudos anteriores.

Normalmente, na aplicação dessas técnicas, o conjunto de dados é divido em duas partes denominadas *treinamento* e *validação*. Esta divisão é realizada de maneira aleatória em porcentagens que podem variar de 75% e 25%, 70% e 30%, 60% e 40% e até mesmo em duas partes de 50%, sendo as duas primeiras as mais comuns. O conjunto de treinamento é utilizado para desenvolver e aprimorar o classificador e o de validação que, como o próprio nome diz, é utilizado para avaliar, ou validar o classificador desenvolvido com o conjunto de treino.

2.2.1 *Bagging*

Também conhecido como *bootstrap aggregating*, o método *bagging* tem como objetivo diminuir a variabilidade dos dados com o propósito de melhorar a classificação de um problema, já que cria classificadores gerando B amostras de tamanho n com reposição, geradas a partir de um conjunto de treinamento. Com esta divisão do conjunto de dados em treinamento e validação evita-se a ocorrência do *overffiting*, que é melhor definido na subseção 2.2.3.

Esta metodologia utiliza a técnica de *bootstrap*, que consiste em uma reamostragem utilizada quando precisa-se quantificar incertezas associadas a estimadores (James *et al.*, 2013). Desta maneira, cada classificador será executado por uma amostra aleatória diferente das demais, atavés de réplicas do conjunto de treinamento, obtidas com reposição (Opitz e Maclin, 1999), (James *et al.*, 2013).

O processo de reamostragem implica em amostras com elementos repetidos ou faltantes, que podem ocasionar classificadores diferentes e com erro possivelmente maior do que em um classificador que utilizou todos os elementos. Contudo, ao realizar a combinação de diversos classificadores individuais, obtém-se a vantagem de que cada classificador apresenta um bom desenvolvimento com uma particularidade do conjunto de dados, complementando um ao outro. Ao combinar diferentes classificadores, eles podem apresentar um erro menor do que o classificador único (Silva, 2021).

O método *bagging* é iniciado a partir da repartição do banco de dados, tamanho N, em conjunto de treinamento e conjunto de validação. O conjunto treinamento é onde será aplicado o método e o conjunto de validação será utilizado para avaliação das predições obtidas pelas metodologias.

Considere B amostras de *bootstrap* que foram obtidas através do conjunto de treinamento de tamanho n, em cada uma destas amostras será aplicada a metodologia de classificação de interesse (regressão logística, árvore de decisão, regressão em sobrevivência) que irá gerar um classificador $g_B(\mathbf{x})$. Feito isso, ao final de cada procedimento, pegam-se todos os classificadores gerados a partir de todas as amostras e realiza-se a combinação dos mesmos através da média ou moda. Após a sua obtenção, o classificador final $g(\mathbf{x})$ é aplicado no conjunto de validação para que seja checada a sua eficácia.

Para melhor entendimento do processo, segue a descrição do algoritmo de maneira simplificada:

- **Algoritmo *bagging* de aprendizagem.**

  Legenda:

  T - Conjunto de treinamento de tamanho n;

  B - número total de reamostras.

  1. De $j = 1$ até B faça:

     $T_{(j)}$ como sendo a j-ésima amostra com reposição retirada do conjunto **T**;

  2. O classificador individual $g^n(x)$ é treinado em cada amostra $T_{(j)}$ gerada;

3. Obtém-se o classificador múltiplo para regressão ou classificação, respectivamente através de:

$$g_{final}(\mathbf{x}) = \frac{1}{B}\sum_{j=1}^{B} g_j^n(\mathbf{x})$$

ou

$$g_{final}(\mathbf{x}) = moda\left[g_1^n(\mathbf{x}), g_2^n(\mathbf{x}), \ldots, g_B^n(\mathbf{x})\right].$$

A Figura 2.4 ilustra o procedimento do *bagging*.

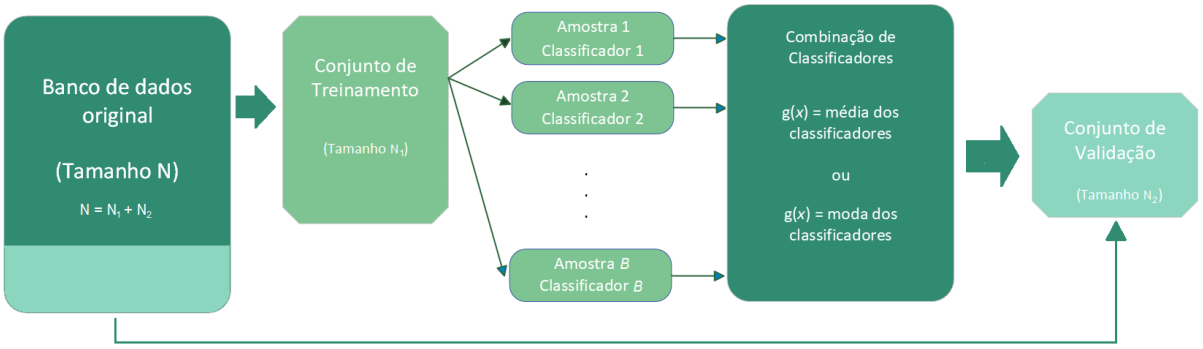

Figura 2.4: Procedimento iterativo da metodologia *bagging*.

2.2.2 *Boosting*

Proposto por Freund e Schapire (1997), e assim explicado por eles, o *boosting* é um classificador que contribui para melhorar a performance de qualquer algoritmo de *machine learning*. Apresentando diversas variações, a que será utilizada neste estudo é a *Adaboost*, que pode ser aplicado em classificações binárias, pois o objetivo do estudo é realizar a classificação de *churn*.

O *boosting* gera vários classificadores de maneira sequencial, onde, em toda iteração em função dos classificadores anteriores, altera-se a distribuição do conjunto treinamento, segundo Dias (2012). Conforme Liska *et al.* (2012) e Friedman *et al.* (2000), este método, utilizando a variação *Adaboost* opera como um algoritmo interferindo na amostra de treinamento, e gerando em cada interação uma distribuição sobre as observações da amostra, atribuindo pesos maiores às observações que foram classificadas erroneamente no passo anterior. Este processo se repeta para uma sequência de amostras, sendo que o classificador final é uma combinação linear dos classificadores individuais.

Freund e Schapire (1997) apresentaram que o erro diminui de maneira exponencial de acordo com o número de iterações. De maneira resumida, primeiramente são atribuídos para todos os indivíduos do conjunto treinamento um peso igual a n^{-1}. Em seguida, é feito um treino no classificador seguindo a distribuição da m-ésima iteração, calcula-se o erro fazendo uma nova distribuição, que diminuirá o peso das classificações corretas de forma a reprimir as incorretas. Este processo será repetido M vezes, realizando uma atualização nos valores dos erros e pesos, de maneira que, os novoc classificadores fiquem mais eficazes cada vez que forem ajustados (Nakano, 2017).

Para melhor entendimento do processo, segue descrição do algoritmo:

- **Algoritmo** *Adaboost* **de aprendizagem.** (Liska *et al.*, 2012); (Alfaro *et al.*, 2013), (Nakano, 2017).

  Seja o conjunto de treinamento $T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$ e o conjunto de classes[1] $C = \{-1, 1\}$, o classificador do tipo *boosting* é dado por:

$$\text{sinal}\,[F(\mathbf{x})] = \text{sinal}\left[\sum_{m=1}^{M} c_m f_m(\mathbf{x})\right], \tag{2.22}$$

  em que:
  - M é o número de iterações num processo iterativo de refinamento do classificador;
  - $f_m(\mathbf{x}) \in C$, é um classificador base (árvore, regressão logística, etc...);
  - c_m, $m = 1, 2, \ldots, M$, são constantes obtidas em cada etapa do procedimento.

  A partir do conjunto de treinamento T, o procedimento *AdaBoost* ajusta o classificador $f_m(\mathbf{x})$ iterativamente, dando pesos menores aos itens classificados corretamente e maiores àqueles classificados errados. Os pesos são ajustados em cada etapa de forma que, ao final do processo, obtêm-se o classificador combinado, dado em (2.22) (Liska et al.,2012).

  Para melhor entendimento do processo, segue descrição do algoritmo:

  1. Seja o conjunto de treinamento $T = \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)\}$, em que $\mathbf{x}_i \in \mathbf{R}^p$ e $y_i \in C$. Considere, inicialmente, pesos iguais para as observações $\mathbf{x}_i$:

$$w_i^m = \frac{1}{n}, \quad i = 1, 2, \ldots, n.$$

[1] O conjunto das classes pode, ainda, ser definido como $C = \{0, 1\}$

2. Fixe o número de iterações M e, para $m = 1, 2, \ldots, M$, faça:

   a) Ajuste $f_m(\mathbf{x})$ considerando os pesos w_i^m para cada observação do conjunto de treinamento;

   b) Calcule os erros ε_m e as constantes c_m, fazendo:

$$\varepsilon_m = \frac{\sum_{i=1}^{n} w_i^m I\left[y_i \neq f_m(\mathbf{x}_i)\right]}{\sum_{i=1}^{n} w_i^m},$$

$$c_m = \frac{1}{2} \ln\left(\frac{1-\varepsilon_m}{\varepsilon_m}\right),$$

   O peso ε_m representa a média ponderada dos erros, com pesos $\mathbf{w}^t = (w_1, \ldots, w_n)$.

   c) Atualize os pesos w_i fazendo:

$$w_i^{m+1} = \begin{cases} \dfrac{w_i^m \mathrm{e}^{-c_m}}{z^m}, & \text{se} \quad y_i = f_m(\mathbf{x}_i) \quad \text{(classificação correta)}, \\ \\ \dfrac{w_i^m \mathrm{e}^{c_m}}{z^m}, & \text{se} \quad y_i \neq f_m(\mathbf{x}_i) \quad \text{(classificação errada)}, \end{cases}$$

   em que $z^m = \sum_{i=1}^{n} w_i^m \exp\left\{-c_m\, y_i\, f_m(\mathbf{x_i})\right\}$, é a constante de normalização.

   Em cada iteração, os pesos $w_i, i = 1, 2, \ldots, n$, são ajustados de forma que as observações classificadas erradas tenham uma ponderação maior.

3. Obter o classificador combinado

$$F(\mathbf{x}) = \sum_{m=1}^{M} c_m f_m(\mathbf{x}).$$

   A classificação final, portanto, é dada por.

$$\hat{y}_i = \text{sinal}\left[F(\mathbf{x}_i)\right], \quad i = 1, 2, \ldots, n.$$

De maneira simplificada, a Figura 2.5 ilustra o procedimento da metodologia.

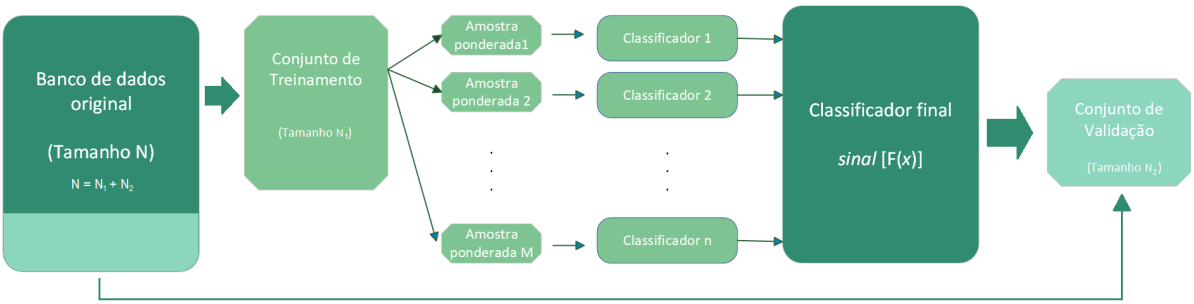

Figura 2.5: Procedimento iterativo da metodologia *AdaBoost*.

2.2.3 *Overffiting*

O interesse de uma predição é a busca de resultados que apresentam maior número de acertos possíveis. Contudo, quando o número de variáveis for relativamente grande, pode ocorrer um problema com o classificador que tende a se adaptar de maneira específica aos dados, o que ocasiona um desempenho inconsistente no modelo, tendendo a reduzir a taxa de acerto quando futuramente aplicado em dados diferentes. Este acontecimento é chamado de *overfitting* (James *et al.* (2013)).

O *overfitting* seria um processo de superajuste, ou seja, o método de classificação se ajustou demais ao conjunto de treinamento, sem prever o que aconteceria em termos preditivos, podendo ocorrer incompatibilidades nas classificações.

Este problema pode ser evitado através da verificação de que o modelo não necessita apresentar todas as variáveis presentes no banco de dados para a modelagem. Um método utilizado para evitar a ocorrência deste problema é a chamada *data splitting*, no qual é realizada a separação do conjunto de dados original em duas partes chamadas de treinamento, conjunto este que é utilizado para o ajuste dos parâmetros, e validação, ou teste, onde será realizada a avaliação do modelo. Esta divisão casualmente é realizada de forma em que 60% dos dados estejam no treinamento e os 40% restantes na validação.

Capítulo 3

Resultados

Neste capítulo iremos apresentar a aplicação das técnicas abordadas no capítulo anterior. O conjunto de dados escolhido para isso foi o presente em Assato (2017), originalmente retirado do software *IBM SPSS* (2010). Os dados se referem às informações do cadastro de clientes em uma empresa de telecomunicações.

Todas as técnicas foram aprimoradas manualmente dentro do *software* R Core Team (2020) e os códigos se encontram no apêndice deste trabalho. O processo foi elaborado da seguinte forma: primeiro uma análise descritiva que percorreu todas as variáveis com o intuito de verificar o comportamento dos dados e analisar possíveis perfis presente nos clientes. Posteriormente os dados foram preparados para o ajuste de modelos de regressão em sobrevivência e após este, também ajustados para o classificador múltiplo *bagging*.

Por fim foi feito uma comparação entre os melhores modelos ajustados, para validarmos os resultados obtidos no que foi considerado mais adequado.

A seguir, é apresentado uma descrição resumida do conjunto de dados.

3.1 Descrição do Conjunto de dados

O banco de dados chamado de Telecom contém 1000 observações no total, tendo como variável resposta $Y = churn$, no qual representa o cancelamento do contrato de clientes da empresa de telecomunicações. A seguir, na Tabela 3.1, uma breve descrição das variáveis presentes no banco de dados.

O conjunto de dados é composto por 14 variáveis, sendo sete contínuas e sete categóricas. No contexto da análise de sobrevivência, a variável *churn* é tomada como o indicador de censura e o tempo de contrato (Cliente duração), como a variável tempo. As demais

serão consideradas todas covariáveis na modelagem da regressão.

Neste sentido, o *churn* = 1, ou seja, a quebra do contrato pelo cliente, é considerado como "falha" e, portanto, *churn* = 0 é uma "censura". Nesse trabalho iremos avaliar a utilização da função de sobrevivência estimada como um possível classificador para o *churn*.

A partir disto, é realizada uma análise descritiva das variáveis, a fim de conhecer melhor as características do banco de dados.

Tabela 3.1: Descrição das variáveis.

Variável	**Categoria**	**Descrição**
Cliente duração		Tempo do contrato como cliente (mês)
Região	Região 1 Região 2 Região 3	Região de residência do cliente
Idade		Idade do cliente (anos)
Estado Civil	Casado Não-Casado	Estado civil do cliente
Residência Duração		Tempo na residência (anos)
Renda		Renda anual da família (milhar)
Educação	Médio Incompleto Médio Completo Superior Incompleto Superior Completo Pós-Graduação Incompleta	Nível educacional do cliente
Emprego Duração		Tempo no último emprego (anos)
Aposentado	Sim Não	Aposentado ou não
Sexo	Masculino Feminino	Sexo do indivíduo
Residentes		Número de pessoas que moram na residência
Receitas		Receita no último mês geradas pelo cliente
Cliente Categoria	Serviço Básico Serviço Eletrônico Serviço Plus Serviço Total	Categoria do serviço do cliente
Churn	Não ou 0 Sim ou 1	Se o cliente encerrou o contrato no último mês

3.2 Análise Descritiva

Primeiramente, foi construída uma tabela contendo as informações sumarizadas referentes às variáveis contínuas do banco.

Tabela 3.2: Descritiva das variáveis contínuas.

	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Cliente Duração	1,00	17,00	34,00	35,53	54,00	72,00
Idade	18,00	32,00	40,00	41,00	51,00	77,00
Residência Duração	0,00	3,00	9,00	11,55	18,00	55,00
Renda	9,00	29,00	47,00	77,53	83,00	1668,00
Emprego Duração	0,00	3,00	8,00	10,99	17,00	47,00
Residentes	1,00	1,00	2,00	2,33	3,00	8,00
Total Receitas	0,00	17,00	38,50	52,86	78,59	269,45

A partir da Tabela 3.2 tem-se que a variáel Cliente Duração tem valor médio de $35,53$ anos, isto é, em média o tempo de contrato do cliente na empresa telefônica é em torno de 35 meses. Assim, também é possível observar que os valores mínimos e máximos são bem distantes, sendo iguais a 1 e 72 meses respectivamente, o que dá indícios que esta variável têm grande variabilidade.

Para a idade do indivíduo, pode-se notar que o mínimo é 18 anos, o que faz sentido dado que a idade mínima para que uma pessoa possa legalmente assinar um contrato é esta. A média e mediana estão bem próximas entre si, sendo a primeira no valor de $41,68$, o que pode-se dizer os contratantes desta empresa costumam ter uma idade em torno de 42 anos.

Também percebe-se que o tempo médio que os indivíduos no estudo ficam em uma residência é de aproximadamente 11 anos, dados os tempos de vida de uma pessoa, pode-se dizer que em média, os indivíduos mudam de residência depois de 11 anos morando nelas. Tendo também um valor máximo alto de 55 anos.

A variável renda do indivíduo aresenta valores de média e mediana distantes, o que indica que apesar do valor médio ser em torno de $77,53$ unidades de milhar, a maior parte dos indivíduos apresentam renda próximos de 47 unidades em milhar. Também apresentam valores mínimos e máximos bem distantes, indicando uma grande variabilidade dos dados, assim como o terceiro quartil e o valor máximo também têm uma diferença significativa.

Para avaliar o tempo em anos, da duração do último emprego do cliente, pode-se notar que existe uma grande variabilidade nos dados, dado que há uma diferença entre seu valor mínimo e máximo. Também percebe-se que os valores de média e mediana estão próximos, sendo assim, em média o cliente fica em torno de 11 anos no mesmo emprego, aproximadamente.

O número de residentes na residência do cliente, fica em média, em torno de 2 a

3 pessoas, sendo esse último o valor do terceiro quartil. Este resultado é possível ser observado na mediana também, onde nota-se que a maioria das residências têm apenas 2 residentes. O valor máximo desta variável é igual a 8, um valor relativamente alto, dado as características da variável.

Por fim, para analisar as receitas geradas no último ano pelo cliente, é notável que esta apresenta grande variabilidade, dados seus valores mínimos e máximos, assim como seus valores de média e mediana também não estão tão próximos.

Estes resultados comentados serão verificados posteriormente, na forma de análise de *boxplots*.

No próximo passo, serão apresentados tabelas de frequência para as variáveis categóricas.

Tabela 3.3: Frequência das variáveis.

Variável	**Categoria**	**Quantidade**	**Frequência**
Região	Região 1	322	32,2%
	Região 2	334	33,4%
	Região 3	344	34,4%
Estado Civil	Casado	495	49,5%
	Não-Casado	505	50,5%
Educação	Médio Incompleto	204	20,4%
	Médio Completo	287	28,7%
	Superior Incompleto	209	20,9%
	Superior Completo	234	23,4%
	Pós-Graduação Incompleta	66	6,0%
Aposentado	Sim	47	4,7%
	Não	953	95,3%
Sexo	Masculino	483	48,3%
	Feminino	517	51,7%
Cliente Categoria	Serviço Básico	266	26,6%
	Serviço Eletrônico	217	21,7%
	Serviço Plus	281	28,1%
	Serviço Total	236	23,6%

Através da Tabela 3.3 é possível notar que para a variável região, a distribuição dos clientes nas três regiões está bem semelhante, tendo a Região 3 com um pouco mais de clientes.

Também nota-se que para as características sexo e estado civil do cliente não há uma distinção da frequência de cada classe, sendo elas bem próximas, portato tem-se um público bem dividido.

Para as frequências dos níveis educacionais, é possível observar que existem mais clien-

tes com ensino médio completo, seguidos por superior completo. Para o nível educacional pós-graduação incompleta têm-se o menor valor de frequência, apenas 6% dos clientes neste nível educacional.

Com relação à aposentadoria, pode-se obeservar que a maioria dos clientes não são aposentados, num percentual de mais de 95%.

Por fim, a distribuição dos serviços oferecidos pela empresa ao cliente, também têm uma distribuição de frequência bem semelhante, sendo o serviço Plus o que apresenta relativamente maior aderência dos clientes.

Com o intuito de confirmar as suposições observadas pelas tabelas descritivas, serão apresentados gráficos no estilo Boxplot, para as variáveis contínuas, e gráficos de barras de frequência para as categóricas, através da Figura 3.1,

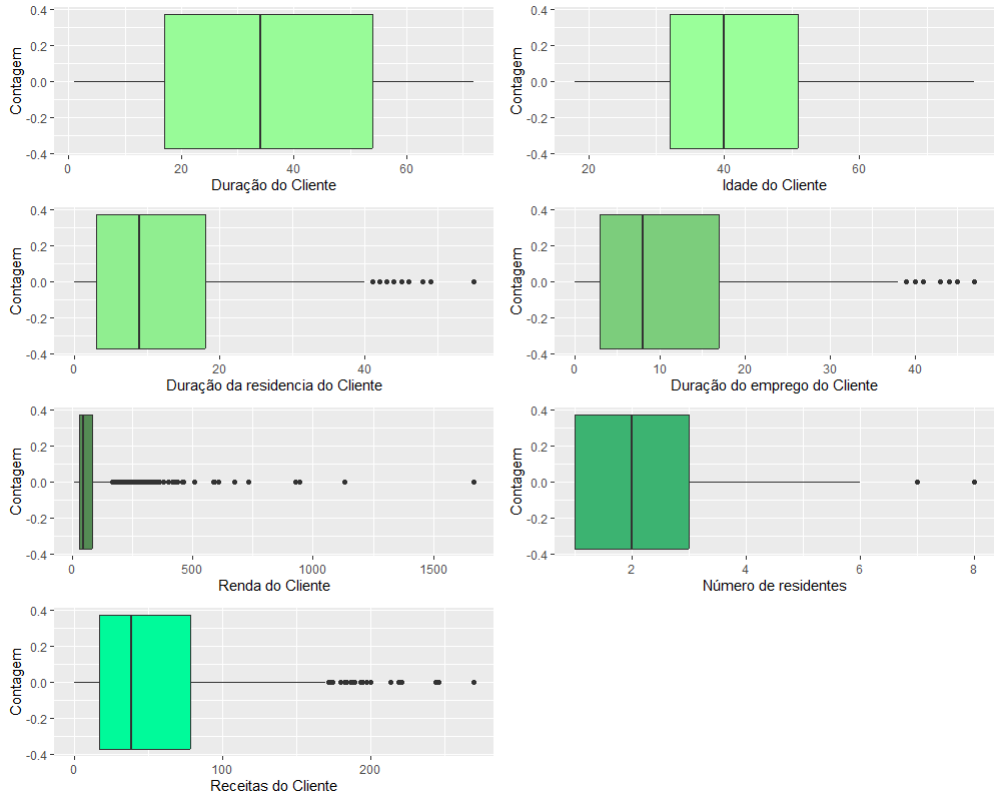

Figura 3.1: Boxplots referentes às variáveis contínuas.

Através dos *boxplots* acima, é possível confirmar algumas visões adquiridas na análise descritiva inicial, onde, o tempo de duração do contrato do cliente apresenta uma grande variabilidade, porém quase que simétrica entre o primeiro e terceiro quartil.

O tempo de duração da residência e a duração do emprego do cliente apresentam alguns valores discrepantes, o que também foi possível observar na tabela descritiva, pelos valores altos de máximo distantes do terceiro quartil.

A variável renda do cliente é a que apresenta maiores quantidades de valores outliers, o que pode ser visto no terceiro gráfico à esquerda. Isto se deve ao fato já observado de que esta têm grande variabilidade e também apresentara valores bem grandes, se comparados à mediana.

Para o número de residentes, nota-se que a caixa do Boxplot é aparentemente simétrica, mostrando que o primeiro quartil, mediana e terceiro quartil são próximos, apesar de apresentarem dois valores discrepantes.

Por fim, serão apresentados os gráficos de barras referentes às frequências das variáveis qualitativas.

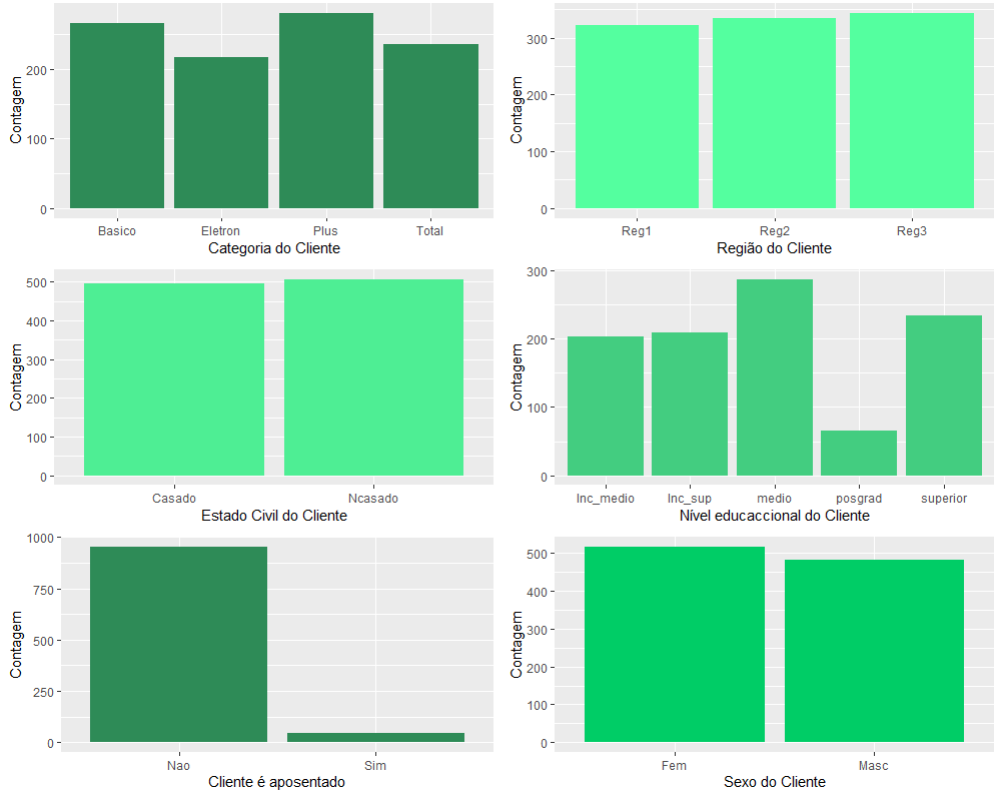

Figura 3.2: Graficos de barras referentes às variáveis categóricas.

Como já dito anteriormente, através da Figura 3.2, as variáveis que apresentam uma diferença significativa em sua distribuição são as referentes à aposentadoria e ao nível escolar do cliente, esta última tendo suas maiores diferenças nas classes de ensino médio completo e pós graduação incompleta.

Para as outras variáveis, as frequências aparentam ser bem semelhantes, sendo bem homogêneas.

Será realizada uma avaliação da variável resposta, através do cálculo da real proporção de cada classificação de *churn*, na Tabela 3.4.

Tabela 3.4: Porcentagem de *churn* no banco de dados.

Churn	%
Não	72, 6
Sim	27, 4

Dada a distribuição da variável, nota-se um desbalanceamento na resposta, onde a maior proporção são de clientes que não cancelaram o contrato, $72,6\%$.

Desta forma, podemos caracterizar os clientes como tendo tempo médio de contrato de 36 meses, na maioria assinantes dos planos Plus e Básico, com pelo menos o ensino médio completo e não aposentados. Com idade média de 42 anos, em sua maioria morando a no máximo 18 meses em residências com famílias de até três pessoas; metade destes têm até 10 anos de emprego, com uma renda mediana anual de $38,5$, concentrada em até 83 mil; não estão concentrados em nenhuma região em particular e, além disso, há um equilíbrio entre os assinantes relação aos sexos masculino e feminino e, também, entre os casados e não casados. Aproximadamente 73% destes ainda mantinham o contrato à época do levantamento.

Sendo assim, teminada a análise descritiva, como sequência será iniciado o ajuste dos modelos e metodologias. Porém, antes de dar início aos ajustes de fato, como têm-se presente nos dados variáveis categóricas é necessário transformá-las em variáveis *dummies*, isto é, mudar as categorias para que elas virem variáveis binárias, rerferentes à cada uma.

3.3 Variáveis *Dummies*

Estas são variáveis binárias que assumem valores 0 ou 1, criadas com o intuito de representar uma variável qualitativa que apresenta duas ou mais categorias. Em um caso de variável com $d = 3$ ou mais classes, é necessário criar $d - 1$ dummies, deixando uma delas como referência.

No processo de modelagem, a utilização da variável *dummy* permite que se encontre a diferença do valor esperado entre as diferentes categorias, ou seja, o coeficiente do modelo será o valor médio que a categoria representa em relação à referência não representada em nenhuma variável *dummy*.

Dado isto, com as seis variáveis qualitativas presentes no estudo, foi realizada esta transformação, da maneira como descrito abaixo:

- **Categoria do Cliente**: foi escolhido como classe de referência o Serviço Básico. Sendo assim, criou-se três variáveis da seguinte forma:

$$ClienCategEletronico = \begin{cases} 1; \text{se a categoria do serviço for Eletrônico,} \\ 0; \text{caso contrário;} \end{cases}$$

$$ClienCategPlus = \begin{cases} 1; \text{se a categoria do serviço for Plus,} \\ 0; \text{caso contrário;} \end{cases}$$

$$ClienCategTotal = \begin{cases} 1; \text{se a categoria do serviço for Total,} \\ 0; \text{caso contrário;} \end{cases}$$

  Quando todas as três variáveis forem iguais à 0, o indivíduo em questão é do Serviço Básico.

- **Região**: foi escolhido como classes de referência a Região 1. As *dummies* são:

$$Regiao2 = \begin{cases} 1; \text{se o região do cliente for a 2,} \\ 0; \text{caso contrário;} \end{cases}$$

$$Regiao3 = \begin{cases} 1; \text{se o região do cliente for a 3,} \\ 0; \text{caso contrário;} \end{cases}$$

  Quando as duas variáveis forem iguais à 0, a região do cliente é a 1.

- **Estado Civil**:

$$EstCivilNCasado = \begin{cases} 1; \text{se o estado civil do cliente for não casado,} \\ 0; \text{caso contrário,} \end{cases}$$

  ou seja, neste caso, se a variável for igual a zero, o estado civil do cliente é casado.

- **Nível Educacional**: para esta, utilizou-se como categoria referência o nível médio incompleto.

$$EducacaoMedio = \begin{cases} 1; \text{se o nível educacional for médio completo,} \\ 0; \text{caso contrário;} \end{cases}$$

$$EducacaoSupIncompleto = \begin{cases} 1; \text{se o nível educacional for superior incompleto,} \\ 0; \text{caso contrário;} \end{cases}$$

$$EducacaoSuperior = \begin{cases} 1; \text{se o nível educacional for superior completo,} \\ 0; \text{caso contrário;} \end{cases}$$

$$EducacaoPosGrad = \begin{cases} 1; \text{se o nível educacional for Pós-graduação completo,} \\ 0; \text{caso contrário;} \end{cases}$$

aqui, quando todas essas variáveis forem iguais a zero, o cliente tem nível médio completo.

- **Aposentado**:

$$AposentadoSim = \begin{cases} 1; \text{se o cliente é aposentado,} \\ 0; \text{caso contrário;} \end{cases}$$

Para esta variável, se o seu valor for igual a zero, quer dizer que o cliente não é aposentado.

- **Sexo**:

$$SexoMasc = \begin{cases} 1; \text{se o sexo do cliente é masculino,} \\ 0; \text{caso contrário.} \end{cases}$$

Por fim, se a variável assumir valor zero, significa que o cliente é do sexo feminino.

Finalizado o processo de criação das variáveis tipo *dummy*, estas foram adicionadas ao banco de dados, em que serão utilizadas no lugar de suas respectivas variáveis categóricas

originais.

Após isto, será iniciado o processo de modelagem utilizando regressão em sobrevivência com a função "Survreg" do *software* R Core Team (2020), que faz o ajuste do modelo. Os resultados para este procedimento são apresentados na seção seguinte.

3.4 Ajustes dos Modelos

Para dar início ao proxímo passo, que é o ajuste do modelo de regressão em sobrevivência, foram testados diversos modelos diferentes, alterando as variáveis presentes no mesmo.

Para ajuste, é necessário que sejam definidas as variáveis que irão indicar o tempo de sobrevivência e a censura. Estas foram definidas como a variável cliente duração e *churn*, respectivamente.

Como serão utilizadas metodologias de regressão e *machine learning*, para obtermos melhores resultados também foi realizada uma divisão no banco de dados em treinamento e validação. Esta divisão foi realizada de forma aleatória, em que a proporção de dados em cada uma foi de 75% e 25%, respectivamente.

Inicialmente, foi utilizado o conjunto de treinamento e realizado um processo de seleção, como primeiro passo ajustando um modelo com todas as variáveis presentes no banco de dados e retirando aquelas que se apresentavam menos significativas, de acordo com o p-valor apresentado e também foram testados modelos retirando aleatoriamente algumas variáveis. Assim, foi-se analisando o valor do erro que os modelos ajustados apresentavam nas predições e a proporção de predição de *churn* respectiva, independente se esta foi um erro ou um acerto, a fim de se obter valores próximos da proporção real apresentada no banco de dados.

Após toda esta análise, foram separados seis modelos que melhor se ajustaram, sendo eles descritos a seguir:

- **Modelo 1**: contém todas as variáveis, sendo elas, Cliente Categoria Eletronico, Cliente Categoria Plus, Cliente Categoria Total, Regiao2, Regiao3, Idade, Estado Civil Não Casado, Residência duração, Renda, Educacao Superior Incompleto, Educação Médio Completo, Educação Pós Graduação Incompleto, Educação Superior Completo, Emprego Duração, Receita, Aposentado Sim, Sexo Masculino.

- **Modelo 2**: Cliente Categoria Eletronico, Cliente Categoria Plus, Cliente Categoria Total, Residência duração, Renda, Educacao Superior Incompleto, Educação Médio Completo, Educação Pós Graduação Incompleto, Educação Superior Completo, Emprego Duração, Aposentado Sim, Sexo Masculino.

- **Modelo 3**: Cliente Categoria Eletronico, Cliente Categoria Plus, Cliente Categoria Total, Estado Civil Não Casado, Residência duração, Renda, Educacao Superior Incompleto, Educação Médio Completo, Educação Pós Graduação Incompleto, Educação Superior Completo, Emprego Duração, Receita.

- **Modelo 4**: Regiao2, Regiao3, Idade, Estado Civil Não Casado, Residência duração, Renda, Educacao Superior Incompleto, Educação Médio Completo, Educação Pós Graduação Incompleto, Educação Superior Completo.

- **Modelo 5**: Cliente Categoria Eletronico, Cliente Categoria Plus, Cliente Categoria Total, Regiao2, Regiao3, Idade, Renda, Emprego Duração.

- **Modelo 6**: Cliente Categoria Eletronico, Cliente Categoria Plus, Cliente Categoria Total, Estado Civil Não Casado, Residência duração, Educacao Superior Incompleto, Educação Médio Completo, Educação Pós Graduação Incompleto, Educação Superior Completo, Emprego Duração.

Um dos processos de ajuste do modelo é obter os valores preditos e após isto comparar com os reais para o cálculo do erro. Como dito anteriormente, o processo de obter as predições de um ajuste de regressão em sobrevivência de acordo com o objetivo deste trabalho não está pré calculado no pacote utilizado.

Desta maneira, foi necessário implementar manualmente todo o processo de predição, através das fórmulas e metodologias apresentadas na Seção 2 deste estudo. A metodologia de predição consiste, resumidamente, em estimar as probabilidades de sobrevivência através do modelo ajustado. Após isto deve-se escolher um ponto de corte, em que, se o valor da estimação for maior que este ponto de corte o indivíduo será classificado como 0 (não *churn*), e 1 caso contrário.

Isto é, quando maior o valor da probabilidade de sobrevivência, quer dizer que o indivíudo tem maiores chances de não cancelar o contrato. No entando, ao decorrer das análises foram encontrados algumas dúvidas sobre qual valor de corte escolher para o

ajuste. Dado isto, foi decidido por ajustar os seis diferentes modelos escolhidos utilizando cinco pontos de corte diferentes, nos valores de $0,5, 0,67, 0,7, 0,75$ *e* $0,8$.

Serão avaliados os diferentes valores de corte na estimação tanto para a sobrevivência quanto para os resultados do classificador múltiplo *bagging*, utilizando os diferentes modelos. As métricas usadas para avaliação da predição foram: o erro total de predição, que é dado pelo total de erros da predição dividido pelo total de observações e a porcentagem de predição de *churn* que é o total de itens preditos como *churn* dividido pelo total de observações do modelo ajustado.

Sendo assim, após o ajuste dos seis modelos descritos, foi decidido por utilizar como ponto de corte o valor $0,7$, pois este apresenta valores razoáveis do erro global de predição e também é o que apresenta proporções de predição de *churn* mais próximas do real.

Dado isto, temos abaixo uma comparação dos modelos propostos, aplicados ao ponto de corte $0,7$ utilizando a regressão em sobrevivência e o classificador múltiplo *bagging*. Este último foi implementado com as mesmas variáveis presentes em cada modelo descrito anteriormente, utilizando análise de sobrevivência como classificador. Foram utilizadas 200 reamostras e o classificador final foi obtido através da média dos classificadores individuais.

- **Sobrevivência**

Tabela 3.5: Comparação dos modelos utilizando a regressão em sobrevivência

Modelo	Erro Total	Predição de *churn*
Modelo 1	36,9%	22,3%
Modelo 2	36,5%	23,2%
Modelo 3	37,6%	23,4%
Modelo 4	37,8%	24,0%
Modelo 5	39,9%	22,8%
Modelo 6	38,0%	23,6%

- *Bagging*

Tabela 3.6: Comparação dos modelos utilizando o *bagging*.

Modelo	Erro Total	Predição de *churn*
Modelo 1	48,9%	32,4%
Modelo 2	48,6%	29,7%
Modelo 3	48,8%	32,8%
Modelo 4	48,9%	29,7%
Modelo 5	49,8%	30,6%
Modelo 6	49,0%	30,1%

Dado os resultados das Tabelas 3.5 e 3.6, o modelo que apresentou menor valor de erro global nas duas metodologias foi o modelo 2. Ele também apresentou a proporção de predição de *churn* bem próximo do real.

Dado estes dois resultados, se dará sequência na análise com o modelo 2. Nas seções a seguir, serão apresentados os resultados para a regressão em sobrevivência e esta aplicada ao classificadores múltiplo *bagging*.

3.4.1 Modelo de Sobrevivência

Foi realizado o ajuste das variáveis presentes no modelo escolhido, utilizando a função "Survreg", inicialmente no conjunto de treinamento. Os resultados estão na Tabela 3.7.

Tabela 3.7: Modelo de sobrevivência final ajustado.

	Valores	Erro Padrão	p-valor
Intercepto	3,485	0,224	$2,00\,\mathrm{e}^{-16}$
Cliente Categoria (Serviço Eletrônico)	0,882	0,182	$1,30\,\mathrm{e}^{-6}$
Cliente Categoria (Serviço Plus)	0,823	0,194	$2,20\,\mathrm{e}^{-5}$
Cliente Categoria (Serviço Total)	0,447	0,165	0,006
Residência Duração	0,051	0,009	$2,70\,\mathrm{e}^{-8}$
Renda	-0,001	0,001	0,132
Educação (Superior Incompleto)	-0,229	0,233	0,326
Educação (Médio)	-0,272	0,223	0,222
Educação (Pós Graduação Completa)	-0,320	0,292	0,274
Educação (Superior)	-0,541	0,231	0,019
Emprego Duração	0,058	0,011	$1,70\,\mathrm{e}^{-7}$
Aposentado Sim	0,253	0,626	0,686
Sexo Masculino	0,022	0,123	0,858

Pela Tabela 3.7 é possível notar que o modelo apresenta, ainda, variáveis não significativas, se considerarmos um nível de significância de $0,05$, isto é, elas têm valores p's maiores do que o nível de significância, porém quando retiradas estas variáveis do modelo, este apresentava valores de erros maiores, portanto foi decidido por mantê-las no ajuste. É importante ressaltar, que para as variáveis categóricas que foram transformadas em *dummies*, se pelo menos uma das categorias for significativa, deve-se manter todas as demais no modelo ajustado.

Feito isto, a seguir será apresentada a tabela de confusão para a predição de *churn* no conjunto treinamento.

Observando a tabela de confusão (Tabela 3.8), percebe-se que o modelo classifica bem as observações que são censuras, porém há um acerto bem pequeno na classificação do

churn (classe 1).

Tabela 3.8: Tabela de confusão do modelo de regressão em sobrevivência (modelo 2) no conjunto treinamento.

		Classe Real	
		0	1
Classe	0	428	148
Predita	1	126	48

Apesar de já apresentados anteriormente, como a amostra que foi selecionada foi a mesma, os valores de erro global e proporção de predição estão na Tabela 3.9.

Tabela 3.9: Erros na predição do do modelo de regressão em sobrevivência (modelo 2), no conjunto de treinamento

Erro global de predição	Proporção de predição de *churn*
$36,5\%$	$23,2\%$

No conjunto treinamento, obteve-se um valor de erro razoável, em torno de 36%, assim como a proporção de predição de *churn* relativamente próxima da real.

3.5 *Bagging*

Após ajustado o modelo de sobrevivência, será realizada a implementação da metodologia de classificação *bagging*. Primeiro, será definido um total de 200 reamostragens, sendo que, para o conjunto de treinamento teremos 750 observações e, para a validação, 250. É importante ressaltar, que as bases treinamento e validação são escolhidas aleatoriamente, permanecendo fixas desde o início do ajuste.

Finalizado o ajuste do modelo, tem-se os valores ajustados dos parâmetros, que são calculados pela média dos ajustes das 200 reamostras do *bootstrap*, para cada variável preditora, expostas na tabela abaixo.

Estes valores são obtidos através das estimativas médias do *bagging*, foram apresentados apenas para realização de uma comparação com os valores encontrados utilizando o modelo de regressão em sobrevivência.

A partir desta tabela, o modelo final ajustado, contém as covariáveis: Cliente Categoria Eletronico, Cliente Categoria Plus, Cliente Categoria Total, Residência duração,

Tabela 3.10: Parâmetros ajustados utilizando o classificador múltiplo *bagging*.

	Valores
Intercepto	5,014
Cliente Categoria (Serviço Eletrônico)	-0,002
Cliente Categoria (Serviço Plus)	0,037
Cliente Categoria (Serviço Total)	0,000
Residência Duração	0,013
Renda	0,001
Educação (Superior Incompleto)	0,252
Educação (Médio)	-0,023
Educação (Pós Graduação Completa)	-0,251
Educação (Superior)	-0,145
Emprego Duração	-0,011
Aposentado Sim	0,0001
Sexo Masculino	-0,135

Renda, Educacao Superior Incompleto, Educação Médio Completo, Educação Pós Graduação Incompleto, Educação Superior Completo, Emprego Duração, Aposentado Sim, Sexo Masculino. Assim como também, sendo o valor do intercepto referente ao parâmetro λ do modelo.

Sendo assim, após o ajuste do modelo será apresentada a tabela de confusão, comparando os valores preditos do *churn* com os reais da variável resposta.

Tabela 3.11: Tabela de confusão pelo método de classificador múltiplo *bagging* no conjunto treinamento.

		Classe Real	
		0	1
Classe	0	357	169
Predita	1	197	27

A Tabela 3.11 mostra que, assim como no modelo anterior, este também apresenta valores altos de acertos na classe 0, porém uma baixa proporção de acertos na categoria 1.

É importante ressaltar que, a metodologia *bagging* realiza reamostras dentro do conjunto a ser utilizado, como não é possível controlar estas amostras, o resultado do erro global e da proporção de predição de *churn* foram diferentes do encontrado na comparação dos modelos (ver Tabela 3.6), sendo eles expostos abaixo.

É possível observar que, comparado com o erro global obtido utilizando apenas a regressão em sobrevivência, o valor para o *bagging* é maior. Esta diferença do erro, será

Tabela 3.12: Erros na predição do método de *bagging*, no conjunto de treinameno

Erro global de predição	Proporção de predição de *churn*
$48,8\%$	$29,8\%$

melhor explorada após o ajuste no conjunto de validação.

Na Figura 3.3, será avaliado a evolução do erro ao decorrer do classificador múltiplo, a fim de verificar possíveis comportamentos atípicos durante a execução do classificador.

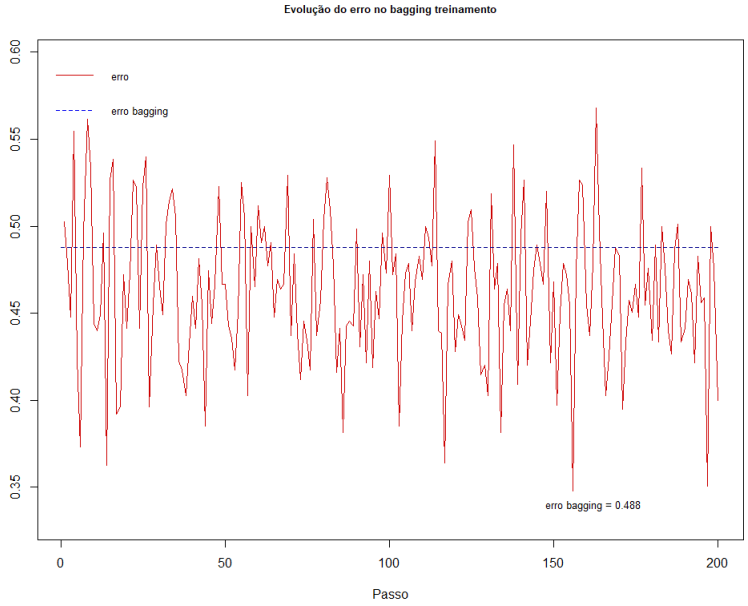

Figura 3.3: Evolução do erro no conjunto treinamento

Nota-se que existem alguns picos no qual o erro do classificador no decorrer das 200 iterações do *bagging* foi pequeno se comparado aos outros, abaixo de 40%. Porém, não se observa nenhum comportamento que seja considerado atípico, já que existe um padrão aleatório na evolução do erro.

Agora, o modelo ajustado será aplicado utilizando o conjunto de validação, para realizar uma melhor avaliação dos resultados.

Após o ajuste, foram obtidos a tabela de confusão e os valores dos erros globais e proporção de predição de *churn*.

Pela Tabela 3.13, assim como no conjunto treinamento, houve um valor bem baixo de acertos, apenas três, em relação aos clientes que cancelaram o contrato. Por sua vez, assim como no treinamento, o modelo realiza uma boa predição em relação aos clientes que não são *churn*.

Tabela 3.13: Tabela de confusão pelo método de classificador múltiplo *bagging* no conjunto validação.

		Classe Real	
		0	1
Classe	0	118	75
Predita	1	54	3

Tabela 3.14: Erros na predição do método de *bagging*, no conjunto de validação

Erro global de predição	Proporção de predição de *churn*
$51,6\%$	$22,8\%$

Pela Tabela 3.14, nota-se que o modelo erra em torno de $51,6\%$, um valor relativamente alto, dado que neste modelo classificou erroneamente pouco mais da metade das observações.

Irá se avaliar também, a evolução do erro no conjunto de validação.

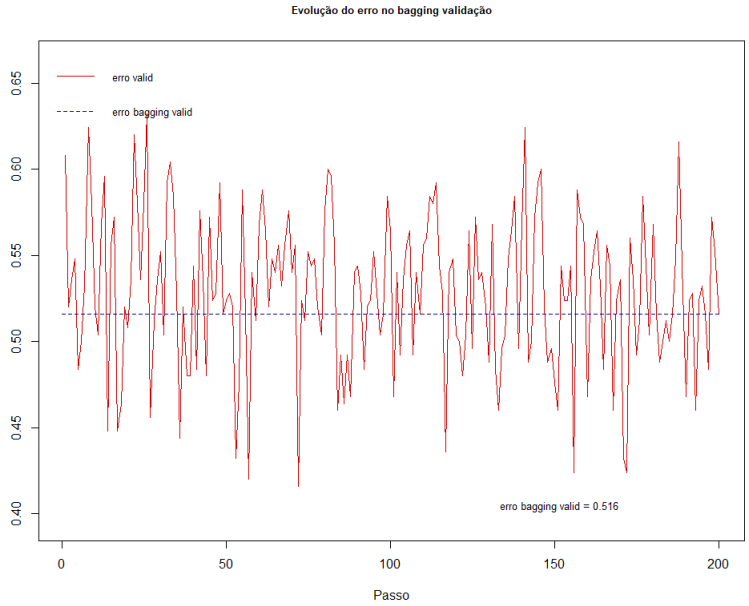

Figura 3.4: Evolução do erro no conjunto validação.

Pode-se notar que apesar dos erros apresentarem alguns picos mais baixos, assim como no conjunto de treinamento, não foi encontrado nenhum comportamento atípico na evolução do erro no conjunto validação e observa-se um padrão aleatório na evolução do erro.

Dado os resultados obtidos pelos ajustes dos modelos, também será realizado teste de Kolmogorv-Smirnov, com o intuito de verificar se estes são capazes de diferenciar bem quais clientes serão *churn* ou não.

3.5.1 Teste *KS*

Para a realização deste teste foi utilizada a função "KStest" do R. Serão testados o conjunto treinamento e validação.

Treinamento

Tabela 3.15: Resultado do teste *KS* para o conjunto de treinamento.

D	P-valor
$0,28$	$5,56\,\mathrm{e}^{-10}$

Através da Tabela 3.15, considerando um nível de significância $\alpha = 0,05$ (5%), vemos que o teste apresenta um p-valor menor do que α, sendo assim tem-se indícios de que o ajuste no treinamento é capaz de diferenciar bem cada classe da variável resposta.

Para melhor vizualizar este resultado, será apresentado a curva da distribuição empírica, através da Figura 3.5.

Validação

Tabela 3.16: Resultado do teste *KS* para o conjunto de validação.

D	P-valor
$0,44$	$1,32\,\mathrm{e}^{-9}$

Através da Tabela 3.16, considerando também um nível de significância de $\alpha = 0,05$, nota-se que o p-valor menor do que α, sendo assim tem-se indícios de que o ajuste no validação também é capaz de diferenciar cada classe da variável resposta.

Para melhor vizualizar este resultado, será apresentado a curva da distribuição empírica, na Figura 3.6.

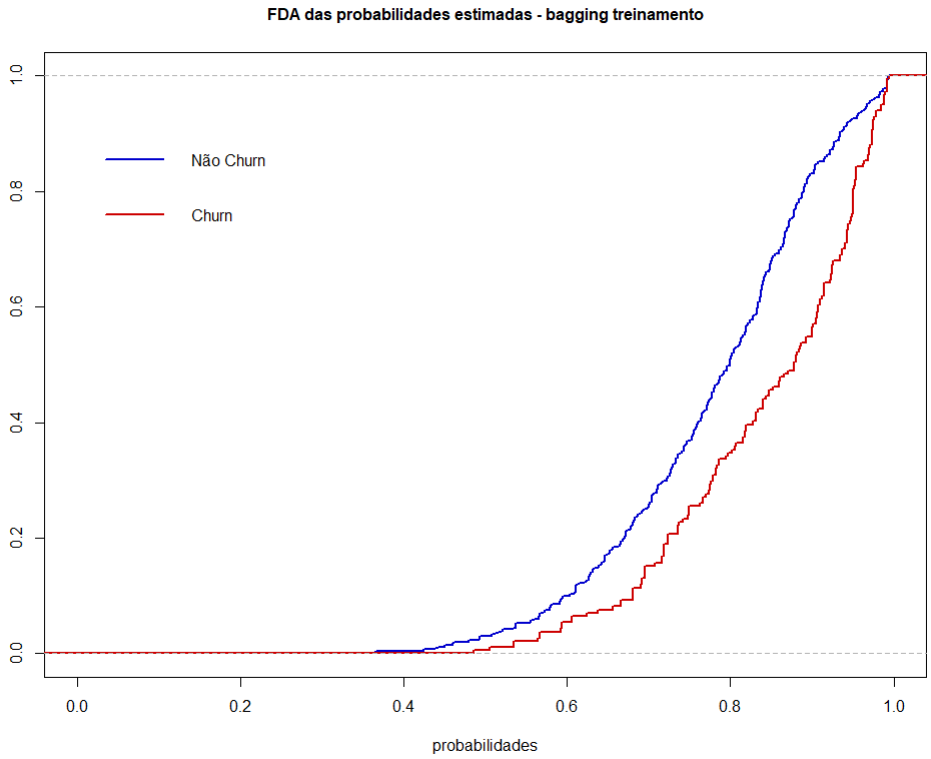

Figura 3.5: Distribuição empírica do conjunto treinamento.

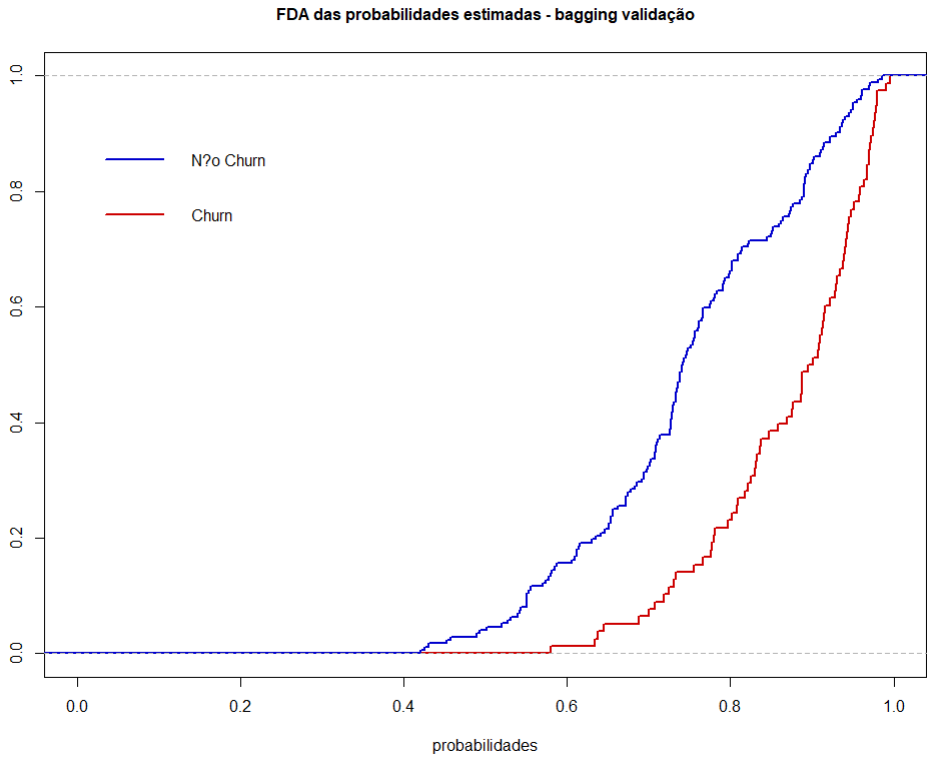

Figura 3.6: Distribuição empírica do conjunto validação.

3.6 *Boosting*

Como dito no início deste capítulo, foi necessário realizar a implementação manual das metodologias de classificação para que fosse possível realizar as predições do modelo de regressão em sobrevivência, pois as mesmas não estão pré implementadas no *software* R Core Team (2020).

Para o classificador múltiplo, o *bagging*, foi possível realizar esta implementação, porém para a aplicação do *boosting*, após estudos sobre o mesmo, foi notado uma certa complexidade em seu desenvolvimento. Dado isto, não foi possível a utilização do procedimento *boosting* com a técnica de sobrevivência.

A ideia inicial seria a realização da classificação assumindo um ponto de corte para a função de sobrevivência, porém devido à presença de censuras, o modelo é ajustado utilizando o método de máxima verossimilhança, ao invés de mínimos quadrados. Dado este fato, a ponderação pelos pesos gerados pelo *boosting* torna o ajuste demasiado complexo, não sendo possível a sua realização.

3.7 *Score* de *Brier*

O *score* de *Brier* (Brier *et al.*, 1950) tem como intuito realizar ponderações nos modelos ajustados em cada reamostra, usando como preditor, no *bagging*, uma média ponderada ao invés da média simples. O *score* de *Brier* é calculado pela seguinte expressão:

$$s^{(k)} = \sum_{i=1}^{n_t} (\hat{S}^{(k)}(t_i) - c_i)^2, \quad k = 1, 2, \ldots, n_B, \tag{3.1}$$

em que $\hat{S}^{(k)}(t_i)$ é o valor predito da sobrevivência do indivíduo i, no instante t_i, c_i é a sua real classificação, sendo 0 = censura, 1 = falha ou *churn*, n_t é o total de indivíduos na reamostra, oriundos do conjunto de treinamento e n_B é o número de reamostras. Realiza-se esta expressão para calcular a diferença das estimativas das sobrevivências com a real classificação de cada indívíduo. Isto significa se este valor for muito alto, quer dizer que o erro de predição também é elevado, caso contrário, o modelo está realizando predições próximas à real.

Dado isto, os valores das ponderações, que serão aplicadas ao *bagging*, são dadas por:

$$w^{(k)} = \frac{\frac{1}{s^{(k)}}}{\sum_{j=1}^{n_B} \frac{1}{s^{(j)}}}, \quad k = 1, 2, \ldots, n_B. \tag{3.2}$$

Nessa situação, os pesos são obtidos a partir dos recíprocos do *score* de *Brier* para que as classificações corretas tenham maior peso nas ponderações. A fórmula de classificação final é exposta a seguir.

$$g_{Brier}(\boldsymbol{X}) = \sum_{k=1}^{n_B} w_k g_k^n(x). \tag{3.3}$$

Sendo assim, esta metodologia foi aplicada no modelo final utilizado, considerando o mesmo ponto de corte para que seja possível realizar as comparações. Foram obtidos os seguintes resultados.

Tabela 3.17: Tabela de confusão pelo método de *scores* de *Brier* no conjunto treinamento.

		Classe Real	
		0	1
Classe	0	356	167
Predita	1	198	29

Através da Tabela 3.17, é possível notar que esta ficou bem semelhante à encontrada utilizando apenas o classificador *bagging*. Mostrando que o modelo comete muitos acertos ao realizar a predição de clientes não *churn*. Um ponto também, é o baixo volume de observações classificadas corretamente de clientes que cancelaram o contrato.

Observando agora os valores de erros de predição e proporção de predição de *churn*.

Tabela 3.18: Erros na predição no conjunto de treinameno, utilizando os *scores* de *Brier*.

Erro global de predição	Proporção de predição de *churn*
$48,6\%$	$30,2\%$

Nota-se que, novamente, os valores presentes na Tabela 3.18 estão semelhantes aos da Tabela 3.12. Mesmo assim, ainda se considera um valor de erro alto para um modelo de predição.

Como próximo passo, será ajustado o modelo utilizando conjunto de validação. Eis os resultados obtidos:

Tabela 3.19: Tabela de confusão pelo método de classificador múltiplo *bagging* no conjunto validação.

		Classe Real	
		0	1
Classe	0	117	75
Predita	1	55	3

Novamente, se observa uma certa semelhança entre os modelos que utilizaram o *score* de *Brier* e apenas o *bagging*, olhando para as Tabelas 3.19 e 3.13.

Tabela 3.20: Erros na predição do método de *score* de *Brier*, no conjunto de validação.

Erro global de predição	Proporção de predição de *churn*
$52,0\%$	$23,2\%$

Pela Tabela 3.20, nota-se que o modelo erra em torno de $52,0\%$, um valor considerado alto também, para uma classificação.

Um ponto que foi percebido é que a aplicação dos scores de Brier no modelo não realizou nenhuma mudança significativa, apresentando resultados muito parecidos com os obtidos pelo *bagging*, além de estarem apresentando valores de erros multo altos.

Uma possível causa para o problema dos valores de erros, pode ser que seja a alta proporção de observações censuradas no banco de dados, aproximadamente 73%. Para verificar se realmente a presença de muitas censuras estaria resultando em erros altos, foi pensado em utilizar o mesmo banco de dados, porém diminuindo a proporção de observações censuradas presentes no mesmo.

Sendo assim, através de uma escolha aleatória foram selecionados apenas 25% do total de censuras presentes na base total e adicionadas às não censuradas, criando uma nova base de dados com total de 367 observações. Também foram testados outros valores de proporção de censura, utilizando valores maiores ou menosres do que o apresentado, porém os resultados obtidos não tiveram uma diferença significativa entre eles.

3.8 Proporção de Censura

Dado a alteração na proporção de censura total da base, as novas proporções de *churn* no banco de dados são:

Tabela 3.21: Porcentagem de *churn* no novo banco de dados.

Churn	%
Não	24, 8
Sim	75, 2

Foi realizado também um outro ajuste nas observações censuradas, para que o conjunto fosse melhor caracterizado pela censura do tipo I. Neste tipo de censura, existe um tempo limite observado, em que após este tempo as observações que não apresentaram o evento de interesse são classificadas como censura.

Com isto, avaliamos o tempo máximo encontrado de permanência de um cliente com contrato na empresa dentro do banco de dados e o valor encontrado foi de 72 meses, realizou-se uma transformação nos tempos das obsevações que não apresentaram *churn*, para que tivessem valor igual a 72. Dado isto, substituímos os tempos das observações que apresentavam censura para este valor máximo, a fim de verificar se realizando essa mudança, apresentariam resultados diferentes nas taxas de erro de predições.

Aqui também foram considerados outros valores para o tempo limite, como 36 e 48 meses, porém não foram encontradas diferenças significativas entre os resultados. Foram aplicados as metodologias *bagging* e *score* de *Brier* utilizando apenas os dados com alteração na proporção de censura e apenas com a mudança do tempo limite e também um os dados contendo as duas alterações simultâneamente.

Os resultados a seguir referem-se aos ajustes dos modelos utilizando o banco de dados contendo as duas alterações conjuntamente. Foi realizado o ajuste dos modelos para verificarmos se realmente o alto valor de observações censuradas e a característica da censura interferem no resultado final.

3.8.1 *Bagging*

Iniciando pelo classificador múltiplo bagging, a tabela de confusão obtida no conjunto treinamento foi a seguir.

Uma situação a ser observada na Tabela 3.22 é que agora o modelo não está realizando

Tabela 3.22: Tabela de confusão pelo *bagging* no conjunto treinamento, com alteração no banco de dados.

		Classe Real	
		0	1
Classe	0	0	103
Predita	1	63	109

predições corretas das observações que não são *churn*. Em contrapartida, aumentou-se as observações preditas corretamente quando há a desistência do contrato.

Estes valores serão melhor vizualizados pela tabela abaixo.

Tabela 3.23: Erros na predição no conjunto de treinameno, utilizando *bagging*, com alteração no banco de dados.

Erro global de predição	Proporção de predição de *churn*
$60,4\%$	$62,5\%$

Dado a tabela do erro, vemos que este apresentou um valor alto, de $60,4\%$ e também houve um aumento na proporção de predição de *churn* no modelo, o que já era de se esperar dado o que foi visto pela tabela de confusão.

Agora, serão apresentados os resultados do ajuste utilizando o conjunto de validação.

Tabela 3.24: Tabela de confusão pelo método de classificador múltiplo *bagging* no conjunto validação, com alteração no banco de dados.

		Classe Real	
		0	1
Classe	0	9	23
Predita	1	19	39

A tabela de confusão (3.24), nos mostra que apesar de serem poucas observações, diferente do ajuste no treinamento, a validação acerta algumas predições de clientes que não são *churn*. Também é notado que a maioria das predições estão presentes na categoria de *churn*.

Observando os erros de predição, é possível observar que o valor do erro cai bastante, se comparado ao conjunto treinamento, apesar de continuar sendo um valor relativamente alto. A proporção de predição de *churn* também fica próxima da real.

Comparado ao ajuste sem realizar as alterações, utilizando apenas o *bagging*, é observado que no conjunto validação, que é onde deve-se comparar os modelos, o banco de

Tabela 3.25: Erros na predição do método *bagging*, no conjunto de validação, com alteração no banco de dados.

Erro global de predição	Proporção de predição de *churn*
$46,7\%$	$64,4\%$

dados com ajuste de censura e tempo apresenta resultados relativamente melhores, sendo eles com erros mais baixos.

Agora, irá se avaliar os resultados calculando as predições pelo *score* de Brier.

3.8.2 *Score* de *Brier*

A tabela de confusão do modelo no conjunto treinamento é apresentada a seguir.

Tabela 3.26: Tabela de confusão pelo *score* de *Brier* no conjunto treinamento, com alteração no banco de dados.

		Classe Real	
		0	1
Classe	0	0	104
Predita	1	63	108

Assim como ocorreu no ajuste inicial, sem a modificação do banco de dados, os resultados obtidos pela Tabela 3.26 são semelhantes aos obtidos utilizando apenas o classificador *bagging*. É importante ressaltar que também não ocorreu predição de não *churn* de maneira correta.

A seguir, se avaliará os valores dos erros e proporção de predição de *churn* no conjunto treinamento.

Tabela 3.27: Erros na predição no conjunto de treinamento, utilizando *score* de *Brier*, com alteração no banco de dados.

Erro global de predição	Proporção de predição de *churn*
$60,7\%$	$62,2\%$

Como já observado pela tabela de confusão, os valores obtidos foram bem próximos dos ajuste pelo *bagging*, apresentando também um erro bem elevado e proporção de predição de *churn* relativamente próxima da real.

Após isto, será ajustado o modelo no conjunto validação, sendo apresentados a tabela de confusão e os valores de proporção de predição de *churn* e do erro total do modelo.

Tabela 3.28: Tabela de confusão pelo *score* de *Brier* no conjunto validação, com alteração no banco de dados.

		Classe Real	
		0	1
Classe	0	9	23
Predita	1	19	39

Tabela 3.29: Erros na predição do *score* de *Brier*, no conjunto de validação, com alteração no banco de dados.

Erro global de predição	Proporção de predição de *churn*
$46,7\%$	$64,4\%$

Pelas Tabelas 3.28 e 3.29, têm-se que os valores obtidos foram idênticos aos do modelo utilizando o *bagging*, sendo assim, as conclusões obtidas através dele serão as mesmas da metodologia citada, no conjunto de validação.

A partir dos resultados obtidos após a alteração no banco de dados, percebe-se que na validação realmente ocorreu um decaimento do erro nas predições, apesar dos mesmos ainda serem altos. Iso pode ser um indício de que a alta proporção de censura nos dados pode estar interferindo na predição do modelo, porém seria necessário realizar mais testes para verificar esta suposição.

Capítulo 4

Conclusão

Neste estudo foi realizada a aplicação de classificadores múltiplos para predição de *churn*, utilizando modelos de regressão em análise de sobrevivência. Inicialmente seriam aplicados os classificadores *bagging* e *boosting*, porém como não existiam pacotes no *software* R Core Team (2020) utilizado que realizavam estas predições através da metodologia de interesse, foi necessária a aplicação manual destas técnicas. Além disso, dado a complexidade da aplicação do *boosting* não foi possível elaborar sua aplicação neste estudo.

Entretanto, foi pesquisado uma metodologia que, assim como o *boosting*, realiza ponderações na classificação, o chamado *score* de Brier. Outro ponto importante encontrado no decorrer das análises, foi a presença de alta proporção de variáveis censuradas no banco de dados, por conta disso, surgiram algumas dúvidas, em que talvez esse alto valor estaria impactando nos resultados finais, por isto, foi realizado uma modificação no banco de dados a fim de criar uma amostra em que houvesse menos variáveis censuradas.

Dado isto, foram analisados diversos modelos também utilizando diferentes pontos de corte para a classificação das observações e foi decidido pelo que apresentava os melhores resultados. Esta avaliação se deu através dos valores do erro global de predição, que é a taxa de quanto o modelo errou na classificação das observações. Outro ponto avaliado foi a proporção de predição de *churn* que o modelo apresentou no total, a fim de verificar se esta está próxima do valor real desta proporção.

Foram apresentados os resultados obtidos utilizando apenas o modelo de regressão em sobrevivência, como também pelo *bagging* e *score* de Brier, primeiramente sem a alteração no banco de dados. Foi possível observar que os modelos apresentavam taxas de erros de predições muito grandes, se comparados aos obtidos utilizando outras metodologias com o mesmo banco de dados, como regressão logística Nakano (2017) e classificação por árvore

(Silva, 2021).

Esta conclusão se mantém mesmo após a modificação realizada no banco de dados, apesar de ter-se observado que a taxa de erro apresenta valores melhores, se comparados com os resultados utilizando os dados originais. Porém, mesmo assim continuam sendo taxas altas.

Para trabalhos futuros, entendo que seria interessante levar em consideração a incidência de censuras nos dados. Uma forma de corrigir este problema seria com a aplicação de técnicas de ponderações, por exemplo. Assim como,a implementação da metodologia *boosting*, que apresenta um desenvolvimento mais complexo, seria um passo interessante para o estudo. Em particular, acreditamos que ambos os estudos são relevantes para os avanços da pesquisa nessa área.

Referências Bibliográficas

Alfaro, E.; Gamez, M.; Garcia, N. adabag: An R package for classification with boosting and bagging. *Journal of Statistical Software*, v. 54, p. 1–35, 2013.

Assato, M. Modelos de sobrevivência em marketing: o valor do tempo de vida de clientes., 2017.

Brier, G. W. et al. Verification of forecasts expressed in terms of probability. *Monthly weather review*, v. 78(1), p. 1–3, 1950.

Carvalho, M. S.; Andreozzi, V. L.; Codeço, C. T.; Campos, D. P.; Barbosa, M. T. S.; Shimakura, S. E. *Análise de sobrevivência: teoria e aplicações em saúde*. SciELO-Editora FIOCRUZ, 2011.

Casella, G.; Berger, R. L. *Statistical inference*. Duxbur y advanced series, 2nd edition ed., 2002.

Collett, D. *Modelling survival data in medical research*. CRC press, 2015.

Colosimo, E. A.; Giolo, S. R. *Análise de sobrevivência aplicada*. Editora Blucher, 2006.

Cox, D. R. Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, v. 34(2), p. 187–202, 1972.

Dias, A. A. D. *Previsão do incumprimento no crédito a empresas com classificadores múltiplos*. Tese (Doutorado), Universidade Tecnica de Lisboa (Portugal), 2012.

Fogo, J. C. *Modelo de regressao para um processo de renovaçao Weibull com termo de fragilidade*. Tese (Doutorado), Universidade de São Paulo, 2007.

Freund, Y.; Schapire, R. E. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences*, v. 55(1), p. 119–139, 1997.

Friedman, J.; Hastie, T.; Tibshirani, R. Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors). *The annals of statistics*, v. 28(2), p. 337–407, 2000.

Gill, R. D. Understanding cox's regression model: a martingale approach. *Journal of the American Statistical Association*, v. 79(386), p. 441–447, 1984.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An introduction to statistical learning*, volume 112. Springer, 2013.

Lawless, J. F. *Statistical models and methods for lifetime data*. John Wiley & Sons, 2011.

Liska, G. R.; de Menezes, F. S.; Cirillo, M. Â.; Vivanco, M. J. F. Classificação de dados em modelos com resposta binária via algoritmo boosting e regressão logística. *Revista Matemática e Estatística em Foco-ISSN*, v. 2318, p. 0552, 2012.

Nakano, C. Database marketing: predição logística de churn com classificadores múltiplos., 2017.

Neslin, S. A.; Gupta, S.; Kamakura, W.; Lu, J.; Mason, C. H. Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of marketing research*, v. 43(2), p. 204–211, 2006.

Opitz, D.; Maclin, R. Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, v. 11, p. 169–198, 1999.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2020. URL `https://www.R-project.org/`.

Santos, R. *Um método para segmentação de preditores*. Tese (Doutorado), Doctorate Thesis, Centro de Informática, Universidade Federal de Pernambuco ..., 2010.

Silva, L. A. Aplicação de árvores de decisões na detecção de churn, associadas à técnicas de machine learning com classificadores múltiplos., 2021.

Wienke, A. *Frailty models in survival analysis*. Chapman and Hall/CRC, 2010.

Apêndice

A - Códigos

```
library(dplyr)
library(readxl)
library(survival)
library(fastDummies)
library(ggplot2)
library(xtable)# para tabela sair em latex
library(gridExtra)

#########################################################
## Preditor linear do survreg
#########################################################

um = rep(1,750)
xa=cbind(um,train$ClienCateg_Eletronico, train$ClienCateg_Plus,
         train$ClienCateg_Total, train$residduracao,train$renda,
         train$Educacao_Sup_Incompleto, train$Educacao_Medio,
         train$Educacao_PosGradua, train$Educacao_Superior,
         train$emprduracao, train$Aposentado_Sim, train$Sexo_Masc)

lambh = exp(xa%*%modelo_weibull_10$coefficients)
deltah = 1/modelo_weibull_10$scale

S.ajus = exp(-(tempo/lambh)^deltah)
S.ajus

pred.churn = um
pred.churn[which(S.ajus>0.7)] <- 0
```

```
tabela_sobrev=table(pred.churn,cens)

erro_sobrev <- (tabela_sobrev[1,2]+tabela_sobrev[2,1])/750
churn.pred_sobrev <- (tabela_sobrev[2,1]+tabela_sobrev[2,2])/750

tabela_sobrev
erro_sobrev
churn.pred_sobrev

#########################################################
## Bagging modelo de sobrevivência
#########################################################
S.corte <- 0.7

nt <- 750 #n do conjunto
nb <- 200 #n do bagging
um = rep(1,nt)
taxa <- numeric(nb)
taxa.acum <- numeric(nb)
taxa.pred <- numeric(nb)
erro.total <- 0
npar <- 13
k <- 1
param.ajus <- matrix(numeric(nb*npar),nb)
escala.ajus <- numeric(nb)
for(i in 1:nb){
  amostra <- sample(1:nt,nt,replace=T)
  tempo.b <- tempo[amostra]
  cens.b <- cens[amostra]
  var1 <- ClienCateg_Eletronico[amostra]
  var2 <- ClienCateg_Plus[amostra]
  var3 <- ClienCateg_Total[amostra]
```

```
var4 <- residduracao[amostra]
var5 <- renda[amostra]
var6 <- Educacao_Sup_Incompleto[amostra]
var7 <- Educacao_Medio[amostra]
var8 <- Educacao_PosGradua[amostra]
var9 <- Educacao_Superior[amostra]
var10 <- emprduracao[amostra]
var11 <- Aposentado_Sim[amostra]
var12 <- Sexo_Masc[amostra]
modelo_weibull_b <- survreg(Surv(tempo.b, cens.b) ~ var1+var2+var3+
                            var4+var5+var6+var7+var8+var9+var10+
                            var11+var12, dist='weibull')
param.ajus[k,] <- modelo_weibull_b$coefficients
escala.ajus[k] <- modelo_weibull_b$scale
xa=cbind(um,var1,var2,var3,var4,var5,var6,var7,var8,var9,var10,
         var11,var12)
lambh = exp(xa%*%modelo_weibull_b$coefficients)
deltah = 1/modelo_weibull_b$scale
S.ajus = exp(-(tempo.b/lambh)^deltah)
pred.churn = um
pred.churn[which(S.ajus>S.corte)] <- 0
tabela <- table(pred.churn, cens.b)
if(dim(tabela)[1]!=2){tabela <- rbind(tabela,c(0,0))}
erro <- tabela[1,2]+tabela[2,1]
churn.pred <- tabela[2,1]+tabela[2,2]
erro.total <- erro.total+erro
tx.erro <- erro/nt
tx.pred <- churn.pred/nt
taxa[k] <- tx.erro
taxa.pred[k] <- tx.pred
n.acum <- k*nt
taxa.acum[k] <- erro.total/n.acum
k <- k+1 }
```

```
param.b <- apply(param.ajus,2,mean) # média das colunas dos
                                      parametros ajustados
round(param.b,5)
escala.b <- mean(escala.ajus)
escala.b
xa=cbind(um,train$ClienCateg_Eletronico,
         train$ClienCateg_Plus, train$ClienCateg_Total,
         train$residduracao,train$renda, train$Educacao_Sup_Incompleto,
         train$Educacao_Medio, train$Educacao_PosGradua,
         train$Educacao_Superior, train$emprduracao,
         train$Aposentado_Sim, train$Sexo_Masc)

erro.b <- 0
churn.pred.b <- 0
tx.erro.b <- 0
tx.pred.b <- 0
final.b <- 0
lamb.b = exp(xa%*%param.b)
delta.b = 1/escala.b
S.b = exp(-(tempo/lamb.b)^delta.b)
pred.churn.b = um
pred.churn.b[which(S.b>S.corte)] <- 0
tabela <- table(pred.churn.b, cens)
tabela

erro.b <- tabela[1,2]+tabela[2,1]
churn.pred.b <- tabela[2,1]+tabela[2,2]
tx.erro.b <- erro.b/nt
tx.pred.b <- churn.pred.b/nt
final.b <- c(tx.erro.b, tx.pred.b)
names(final.b) <- c("Erro bagging", "Predição churn bagging")
final.b
```

```
mn <- 0.95*min(taxa, tx.erro.b)
mx <- 1.05*max(taxa, tx.erro.b)

plot(c(1,nb), c(mn,mx), type = "n", main = "Evolução do
     erro no bagging treinamento", xlab = "Passo",ylab = " ",
     cex.main=0.8)
lines(taxa, col="red3")
legend("bottomright", legend=paste("erro bagging =",
      round(tx.erro.b,4)), bty="n", cex=0.8)
lines(c(0,nb), c(tx.erro.b,tx.erro.b), lty=2, col="blue4")
legend("topleft", c("erro", "erro bagging"), lty=c(1,2),
       col=c("red3","blue2"), bty="n",cex=0.8)

#####################################################
## predicao conjunto de validacao
#####################################################
tempo.valid = valid$clienduracao
cens.valid <- valid$churn01
cens.valid
nv <- 250
vum <- rep(1,nv)

v.xa=cbind(vum,valid$ClienCateg_Eletronico, valid$ClienCateg_Plus,
           valid$ClienCateg_Total, valid$residduracao,valid$renda,
           valid$Educacao_Sup_Incompleto, valid$Educacao_Medio,
           valid$Educacao_PosGradua, valid$Educacao_Superior,
           valid$emprduracao, xalid$Aposentado_Sim, valid$Sexo_Masc)

lamb.valid = exp(v.xa%*%param.b)
delta.valid = 1/escala.b
S.valid = exp(-(tempo.valid/lamb.valid)^delta.valid)
pred.churn.valid = vum
```

```
pred.churn.valid[which(S.valid>S.corte)] <- 0
tabela.valid <- table(pred.churn.valid, cens.valid)
tabela.valid

taxa.v <- numeric(nb)
taxa.pred.v <- numeric(nb)
war1 <- valid$ClienCateg_Eletronico
war2 <- valid$ClienCateg_Plus
war3 <- valid$ClienCateg_Total
war4 <- valid$residduracao
war5 <- valid$renda
war6 <- valid$Educacao_Sup_Incompleto
war7 <- valid$Educacao_Medio
war8 <- valid$Educacao_PosGradua
war9 <- valid$Educacao_Superior
war10 <- valid$emprduracao
war11 <- valid$Aposentado_Sim
war12 <- valid$Sexo_Masc

k <- 1
vxa=cbind(vum,war1,war2,war3,war4,war5,war6,war7,war8,war9,
          war10,war11,war12)
for(i in 1:nb){
  lambh.v = exp(vxa%*%param.ajus[i,])
  deltah.v = 1/escala.ajus[i]
  S.ajus.v = exp(-(tempo.valid/lambh.v)^deltah.v)
  pred.churn.v = vum
  pred.churn.v[which(S.ajus.v>S.corte)] <- 0
  tabela.v <- table(pred.churn.v, cens.valid)
  if(dim(tabela.v)[1]!=2){tabela.v <- rbind(tabela.v,c(0,0))}
  erro.v <- tabela.v[1,2]+tabela.v[2,1]
  churn.pred.v <- tabela.v[2,1]+tabela.v[2,2]
  # erro.total.v <- erro.total.v+erro.v
```

```
  tx.erro.v <- erro.v/nv
  tx.pred.v <- churn.pred.v/nv
  taxa.v[k] <- tx.erro.v
  taxa.pred.v[k] <- tx.pred.v
  #  n.acum <- k*nt
  #  taxa.acum[k] <- erro.total/n.acum
  k <- k+1
}

erro.valid <- tabela.valid[1,2]+tabela.valid[2,1]
churn.pred.valid <- tabela.valid[2,1]+tabela.valid[2,2]
tx.erro.valid <- erro.valid/nv
tx.pred.valid <- churn.pred.valid/nv
final.valid <- c(tx.erro.valid, tx.pred.valid)
names(final.valid) <- c("Erro bagging.validaçao", "Predição
                        churn bagging.validação")
final.valid

mn <- 0.95*min(taxa.v)
mx <- 1.05*max(taxa.v)

plot(c(1,nb), c(mn,mx), type = "n", main = "Evolução do erro
     no bagging validação", xlab = "Passo",ylab = " ", cex.main=0.8)
lines(taxa.v, col="red3")
legend("bottomright", legend=paste("erro bagging valid =",
       round(tx.erro.valid,4)), bty="n", cex=0.8)
lines(c(0,nb),c(tx.erro.valid,tx.erro.valid), lty=2, col="blue4")
legend("topleft", c("erro valid", "erro bagging valid"),
       lty=c(1,2), col=c("red3","blue2"), bty="n",cex=0.8)
```

```
################################
## Treinamento
################################
score.bag_t0 <- S.ajus[cens.b==0]
score.bag_t1 <- S.ajus[cens.b==1]

ks.test(score.bag_t0, score.bag_t1)

################################
## Gráfico f.d.a. empíricas
################################
Fy.bag.t0 <- ecdf(score.bag_t0)
Fy.bag.t1 <- ecdf(score.bag_t1)
plot(Fy.bag.t0, xlim=c(0,1), ylim=c(0,1), main = "FDA das probabilidades
     estimadas - bagging treinamento", do.points=F, col="blue3",
     xlab="probabilidades", ylab="", cex.main=1, lwd=2, verticals=T)
par(new=T)
plot(Fy.bag.t1, xlim=c(0,1), ylim=c(0,1), main=" ", do.points=F, axes=F,
     col="red3", xlab=" ", ylab=" ", cex.main=1, lwd=2,
     verticals=T)
legend(0, 0.95, legend=c("Não Churn", "Churn"), col=c("blue3","red3"),
       lwd=2, bty="n")



################################
## Validação
################################
score.bag_v0 <- S.ajus.v[cens.valid==0]
score.bag_v1 <- S.ajus.v[cens.valid==1]

ks.test(score.bag_v0, score.bag_v1)
```

```
################################
## Grafico f.d.a. empíricas
################################
Fy.bag.v0 <- ecdf(score.bag_v0)
Fy.bag.v1 <- ecdf(score.bag_v1)
plot(Fy.bag.v1, xlim=c(0,1), ylim=c(0,1), main = "FDA das probabilidades
     estimadas - bagging validação", do.points=F, col="red3",
     xlab="probabilidades", ylab="", cex.main=1, lwd=2, verticals=T)
par(new=T)
plot(Fy.bag.v0,xlim=c(0,1), ylim=c(0,1), main=" ", do.points=F, axes=F,
     col="blue3", xlab=" ", ylab="", cex.main=1, lwd=2, verticals=T)
legend(0, 0.95, legend=c("Não Churn", "Churn"), col=c("blue3","red3"),
      lwd=2, bty="n")
```