Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de Sobrevivência e Previsão de *Churn* de Clientes de Seguros de Vida do Banco do Brasil

Jose Maria Amorim Araújo

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Orientador
Prof. Dr. Gladston Luiz da Silva

Brasília
2022

Ficha catalográfica elaborada automaticamente,
com os dados fornecidos pelo(a) autor(a)

AA663a

Araújo, Jose Maria Amorim
Análise de Sobrevivência e Previsão de Churn de Clientes de Seguros de Vida de Uma Instituição Financeira Brasileira / Jose Maria Amorim Araújo; orientador Gladston Luiz Silva. -- Brasília, 2022.
57 p.

Dissertação (Mestrado - Mestrado Profissional em Computação Aplicada) -- Universidade de Brasília, 2022.

1. Análise de Sobrevivência. 2. Seguro de Vida. 3. Churn. 4. Regressão de Cox. 5. Kaplan-Meier. I. Silva, Gladston Luiz, orient. II. Título.

Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Ciência da Computação

Análise de Sobrevivência e Previsão de *Churn* de Clientes de Seguros de Vida do Banco do Brasil

Jose Maria Amorim Araújo

Dissertação apresentada como requisito parcial para conclusão do
Mestrado Profissional em Computação Aplicada

Prof. Dr. Gladston Luiz da Silva (Orientador)
PPCA/UnB

Prof.a Dr.a Maristela T. Holanda
PPCA/UnB

Prof.a Dr.a Ana Paula Bernardi da Silva
UCB

Prof. Dr. Marcelo Ladeira
Coordenador do Programa de Pós-graduação em Computação Aplicada

Brasília, 26 de janeiro de 2022

Dedicatória

Dedico este trabalho aos meus amados pais, José Pereira Araújo e Maria das Graças Amorim Araújo, por serem minha inspiração cotidiana, minha base forte. A minha linda família, minha esposa Natália e minha filha Agnes, que ao longo deste período de mestrado estiveram ao meu lado, apoiando e sendo meu porto seguro.

Agradecimentos

Em primeiro lugar agradeço ao Grande Deus, pelo dom da Vida, sem Ele não conseguiria este título. Não foi fácil essa jornada, porém, muito emocionante, cheias de conhecimentos, amizades, e com grandes desafios.

A resiliência é a palavra que define esta trajetória de estudo, não esperávamos uma pandemia na metade do mestrado, alguns colegas desistiram, outros trancaram suas matrículas, tempos difíceis. Entretanto, ao final, a palavra é gratidão a Deus por ter conseguido.

Sou grato a meus amados pais e irmãos, que um dia sonharam com esta conquista, e me deram todo o apoio necessário para que eu conseguisse realizar este belo sonho.

Agradeço muito a minha esposa Natália, que esteve ao meu lado nestes dias árduos, suportando horas de estudos sem que eu esteja ao seu lado, contudo, me deu forças e apoio incondicional. A minha amada filha Agnes que nasceu durante este mestrado, foi e está sendo uma das maiores alegrias da minha vida, ter este ser lindo ao meu lado.

Ao meu orientador prof. Dr. Gladston, meus sinceros agradecimentos, pela sua orientação, dedicação e seus conselhos, certamente levarei seus conselhos para a vida. Agradeço à renomada UnB e aos queridos professores deste brilhante mestrado (Dr. Ladeira, Dra. Maristela, enfim todos), pela dedicação e pelos conhecimentos a nós passados.

Agradeço ao Sr. Alberto Marques e Alessandra, amigos de longos anos que me orientaram e incentivaram a concorrer a este mestrado, lembro das tardes em sua residência, conversando sobre como seria importante conseguir um título de mestre pela UnB.

Agradeço à minha instituição onde trabalho, pelo apoio, em especial ao meu gerente Paulo André (PA).

E por fim e não menos importante, agradeço a querida amiga Carine, colega de trabalho e estudo, que dedicou um pouco de seu tempo a me orientar os caminhos para chegar ao mestrado na UnB, acredite, você fez a diferença. Aos colegas de trabalhos e estudo, Roberto Mourão e Sueli que também fizeram parte desta conquista.

Resumo

A modernização e a digitalização de produtos e serviços no setor financeiro avançam de forma rápida e eficiente. A cada dia, novas empresas digitais como *fintechs* trazem produtos inovadores e diferenciados ao mercado com foco centrado na experiência do cliente, fazendo com que grandes atores desse setor, tais como instituições financeiras e seguradoras sintam-se ameaçados em um cenário cada vez mais competitivo. Nesse contexto, a ocorrência de rotatividade de clientes, mais conhecido como *churn*, é cada vez maior, tornando-se um problema para essas empresas. A empresa Alfa, configura-se como um destes atores com uma considerável fatia do mercado de seguros no Brasil, cuja rotatividade de clientes mostra-se como um grande desafio a ser superado. O presente estudo tem por objetivo implementar um modelo capaz de fazer uma análise de sobrevivência e predição de rotatividade de clientes contratantes de seguros de vida de forma assertiva, a fim de possibilitar a tomada de decisão e a elaboração de campanhas que contribuam para a diminuição do *churn* e, consequentemente, para uma maior fidelização dos clientes ao produto seguro de vida. Para isso, foram coletados dados de clientes no período de cinco anos, que corresponde ao período de vigência dos contratos de seguros de vida. Utilizou-se o estimador de Kaplan-Meier para traçar a curva de sobrevivência, e a regressão de Cox para efetuar a predição de rotatividade dos clientes contratantes de seguros de vida, além de possibilitar traçar o perfil dos clientes que têm maiores probabilidades de cancelarem seus contratos de seguros. Os resultados mostraram que o estimador de Kaplan-Meier obteve bons resultados traçando a curva de sobrevivência e possibilitando a geração da tabela de vida para os clientes, na predição de *churn*, a regressão de Cox convergiu de forma muito boa aos dados, chegando ao índice de concordância de 0.61, além de traçar curvas preditivas por clientes. Assim, a combinação do estimador de Kaplan-Meier e a regressão de Cox, mostrou-se uma união poderosa na criação de análise de sobrevivência e predição de *churn* com dados de clientes de seguros de vida.

Palavras-chave: análise de sobrevivência, seguro de vida, *churn*, regressão de cox, Kaplan-Meier

Abstract

The modernization and digitalization of products and services in the financial sector is advancing quickly and efficiently. Every day, new digital companies such as fintechs bring innovative and differentiated products to the market with a focus on customer experience, making major players in this sector, such as financial institutions and insurance companies, feel threatened in an increasingly competitive scenario. In this context, the occurrence of customer turnover, better known as churn, is increasing, becoming a problem for these companies. The company Alfa is one of these players with a considerable share of the insurance market in Brazil, whose customer turnover is a major challenge to be overcome. The present study aims to implement a model capable of performing an analysis of survival and prediction of turnover of customers contracting life insurance in an assertive way, in order to enable decision-making and the elaboration of campaigns that contribute to the reduction of the churn and, consequently, for greater customer loyalty to the life insurance product. For this, data were collected from customers over a period of five years, which corresponds to the period of validity of life insurance contracts. The Kaplan-Meier estimator was used to trace the survival curve, and the Cox regression to predict the turnover of life insurance customers, in addition to making it possible to trace the profile of customers who are more likely to cancel their insurance contracts. The results showed that the Kaplan-Meier estimator obtained good results by tracing the survival curve and allowing the generation of the life table for the customers, in the churn prediction, the Cox regression converged very well to the data, reaching the index of agreement of 0.61, in addition to drawing predictive curves by customers. Thus, the combination of the Kaplan-Meier estimator and the Cox regression proved to be a powerful union in creating survival analysis and churn prediction with life insurance customer data.

Keywords: Survival analysis, life insurance, churn, cox regression, Kaplan-Meier

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas e Siglas

AUC Area Under the Curve.

CRISP-DM CRoss Industry Standard Process for Data Mining.

IBM International Business Machines.

IC Index Concordance.

KDD Knowledge Discovery in Databases.

KM Kaplan Meier.

KNN K-Nearest Neighbors.

ROC Receiver Operating Characteristic.

SVM Support Vector Machine.

TP True Positive.

Capítulo 1

Introdução

Atualmente, o mercado de seguros de vida no Brasil está cada vez mais rentável, no qual verifica-se a consolidação de grandes instituições financeiras, tais como bancos e seguradoras, assim como o surgimento de novas corretoras de seguros. Entretanto, o crescimento desse mercado está ocorrendo num período de crise econômica brasileira, onde percebe-se o envelhecimento da população, e cujo sistema previdenciário nacional caminha para a falência.

1.1 Contextualização

Devido ao surgimento de pequenas empresas, as *fintechs*, e ao nível de exigência pelos clientes, o mercado de seguros está se tornando cada vez mais concorrido. Esse fenômeno é importante para as empresas, assim como para a comunidade como um todo, na medida em que gera qualidade para os produtos, e diminuição nos preços desses seguros. Porém, isso causa preocupação para grandes empresas de seguros, ao ponto que acontece uma intensa jornada na busca e na fidelização dos clientes contratantes de seguros, ocasionando maiores custos e, consequentemente, menores lucros.

Esta instituição financeira figura entre os gigantes de vendas de seguros no Brasil, com uma considerável fatia do mercado, chegando ao quarto lugar no *ranking* geral das melhores seguradoras do Brasil [5], e em terceiro lugar quanto a seguros de pessoas ou seguro de vida. Esse setor é considerado muito importante e estratégico; prova disso, é que o banco criou uma subsidiária chamada empresa Alfa – braço da instituição para a gestão de seguros, previdência e serviços. Nesse processo de segregação dos setores, e como estratégia de mercado, foi criado um grupo segurador chamado empresa Beta – uma fusão entre a instituição financeira e a empresa Beta de Seguros.

A empresa Alfa possui um processo de gestão dos clientes que contratam seus seguros de vida. Entretanto, muitos contratos são cancelados diariamente, tendo havido uma

redução de 400 milhões de reais, na comparação do 4º trimestre de 2020 com 1º trimestre de 2021, conforme mostra Figura 1.2. O fato de o cliente cancelar seu contrato, configura-se como *churn* – métrica que indica quantos clientes deixaram determinado produto. Portanto, é essencial o gerenciamento eficaz do *churn*, que é o processo sistemático de tentar reter ativamente os consumidores, antes de eles deixarem a empresa [6]. Contudo, reter e fidelizar esses clientes torna-se vital e configura um enorme desafio.

A retenção de clientes é uma das questões mais relevantes para as empresas [7]. Dada a complexidade do tema, torna-se importante o estudo especializado do *churn* nas empresas, considerando que é mais barato e menos desgastante manter os clientes fidelizados na base, do que conquistar novos clientes [8]. Conforme afirma F. Reichheld, o aumento de 5% na retenção de clientes produz um incremento de 25% nos lucros da empresa [9]. Para o referido estudo, foram utilizadas as bases de dados de uma instituição financeira brasileira, onde obtêm-se bases consistentes e consolidadas, com informações relevantes em cerca de milhões de registros gerados diariamente.

1.2 Justificativa

Um recente editorial publicado pelo portal da CNseg, mostra um panorama do mercado segurador no Brasil, bem como suas perspectivas e o crescimento esperado para esse setor no ano de 2021. Para o presidente da CNseg, Márcio Coriolano, "(...) passado o período de turbulência e incertezas políticas, e diante da perspectiva de aprovação de reformas estruturais e microeconômicas, o cenário é otimista" [10]. Esse otimismo se reflete no pujante crescimento obtido pelo setor nos primeiros quatro meses de 2021, chegando a patamares de 15,5% em relação ao ano de 2020, com destaque para o segmento de cobertura de pessoas, ou seguros de vida, com crescimento de 18,5%, com faturamento de R$ 54,5 bilhões no ano de 2021, denotando um mercado em franca expansão, extremamente lucrativo e concorrido [11].

Entretanto, conforme mostrado na Figura 1.1, no ano de 2020 houveram quedas nas vendas e nos lucros desse setor; isso é dado pelo efeito causado pela pandemia do COVID-19, iniciada em meados de março de 2020, que promoveu um elevado número de desemprego, fechamento de comércios e paralisação de atividades essenciais para o setor, deixando assim, as pessoas sem perspectivas de adquirirem ou manterem seus seguros. Contudo, é visível neste mesmo gráfico, a retomada deste setor em meados do mês de agosto, chegando ao seu pico de alta de outubro para novembro de 2020 [11].

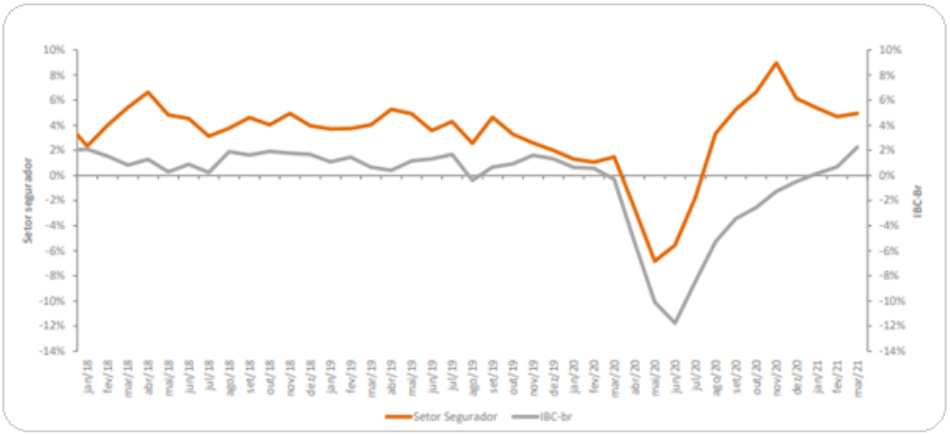

Figura 1.1: Gráfico do crescimento do setor de seguros nos últimos 39 meses [9].

O forte crescimento do mercado de seguros no Brasil e o acirramento da concorrência na busca e retenção de clientes, explica a entrada de empresas inovadoras, como as *startups* chamadas *insurtechs*, que são plataformas *online* focadas no mercado de seguro, e estas estão mais sensíveis aos anseios da sociedade [10]. Contudo, um estudo mostra que os novos clientes de uma empresa são os clientes perdidos de outras [12]; fato evidenciado a partir do momento em que muitas empresas crescem mais que as outras, vendendo os mesmos produtos, apesar de não ser uma tarefa tão simples.

Entretanto, o mercado de seguros de vida é um dos que mais crescem no Brasil [11], segundo a recente divulgação de resultados da empresa Alfa referente ao 1º trimestre de 2021, obtendo crescimento de 10,7% de lucro em relação ao mesmo período de 2020, um total de 977 milhões de reais, considerando-se um resultado excelente devido às condições adversas, como a pandemia em que a sociedade se encontra. Contudo, os resultados referentes aos prêmios emitidos de seguros de vida no referido trimestre, tiveram um modesto aumento de 7,7% e uma redução de 15,7% em relação ao 4° trimestre de 2020. Comparando os resultados do 1º trimestre de 2021 em relação ao 4º trimestre de 2020, apresenta-se uma queda grande nos números de prêmios[1] emitidos e, em consequência, a redução dos lucros recebidos, conforme mostrado na Figura 1.2. Esse resultado indica a enorme quantidade de contratos cancelados, ou seja, cada parcela não emitida configura um cliente que sai da base da empresa, aumentando o problema de *churn* de clientes para o produto, mesmo com o ritmo elevado de vendas [2].

[1]Prêmio refere-se ao valor pago pelo cliente ao contratar o produto seguro de vida.

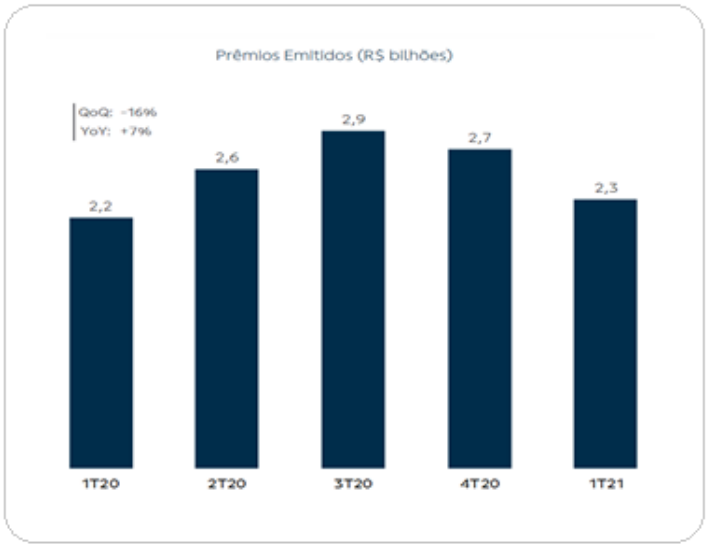

Figura 1.2: Série de prêmios emitidos durante o ano de 2020 e o 1º trimestre de 2021 [2].

Conforme a Figura 1.2, há um ritmo acelerado nas vendas; porém, não adianta envidar forças para vendas se não houver um gerenciamento eficiente da rotatividade de clientes. Para o presente estudo, buscou-se implementar um modelo de análise de sobrevivência e previsão de *churn* de clientes contratantes de seguros de vida, que auxilie os gestores na busca de melhores estratégias, campanhas de marketing e tomada de decisão, visando a fidelização dos clientes à empresa, e consequentemente, a diminuição dos cancelamentos dos contratos de seguros.

A empresa Alfa dispõe de poucos estudos e mecanismos eficientes no gerenciamento do *churning* que sejam específicos para cada produto. A forma como busca reter seus clientes ainda é empírica, sem uma gestão diferenciada por produtos e serviços, não considerando as especificidades de cada cliente. É evidente que estudos, modelos analíticos e inteligência artificial, ajudam nas tomadas de decisões da gestão da retenção e fidelização de clientes [13]. Portanto, esse estudo tem como objetivo propor um modelo capaz de gerar inteligência baseado em dados, para uso dos gestores na condução eficiente do *churning* e na melhoria dos resultados, possibilitando aumentar a retenção dos clientes e maior fidelidade ao produto.

A complexidade e a relevância do estudo do tema *churn* de clientes, dá-se em três aspectos: primeiro, quanto à proeminência para as empresas neste mercado de elevada concorrência, como estratégia de sobrevivência no mercado [10]; segundo, quanto à queda dos lucros ao passar dos meses na empresa Alfa, conforme Figura 1.2; terceiro, quanto às publicações com contribuições científicas, que apesar dos estudos sobre o tema, não exploram a questão da rotatividade de indivíduos em relação a contratos de seguros de

vida, conforme levantamento efetuado nas bases de dados *Web Of Science*, Scorpus e IEEE *xplorer*.

1.3 Objetivos

Esta seção contempla os objetivos geral e específicos do presente estudo, e apresenta os resultados a serem atingidos, limitando o escopo do projeto e descrevendo as etapas para a implementação da pesquisa.

1.3.1 Objetivo Geral

O presente trabalho tem como objetivo propor um modelo de análise de sobrevivência e previsão de *churn* de clientes contratantes de seguros de vida na empresa Alfa.

A principal contribuição da solução proposta neste estudo, será a criação de um modelo capaz de traçar a curva de vida dos clientes, e prever, de maneira proativa, os clientes que pretendem cancelar seus seguros de vida. Para isso, fez-se uma combinação dos métodos do estimador de Kaplan-Meier e a regressão de Cox, buscando identificar *insights* de negócios e aumentar a rentabilidade desse produto junto a instituição financeira brasileira.

1.3.2 Objetivos Específicos

Para atingir o objetivo geral, foram definidos os seguintes objetivos específicos:

1. Selecionar as variáveis mais importantes, de acordo com o negócio;
2. Obter *insights* na análise das distribuições das variáveis;
3. Traçar a curva de sobrevivência e a tabela de vida, utilizando o estimador de Kaplan-Meier;
4. Traçar os riscos proporcionais de *churn* por cliente;
5. Prever o tempo de cancelamento de contrato de seguros pelo cliente.

1.4 Modelo de Referência CRISP-DM

Para a implementação do referido estudo, foi utilizado como referência o modelo CRISP-DM, passando por todas as fases, sendo: compreensão do negócio, compreensão dos dados, preparação dos dados, modelagem, avaliação e implantação [1]. A Figura 1.3 apresenta as etapas e subetapas do processo que compõe esse modelo. É fundamental que, em estudo

de caso em que se utiliza a mineração de dados, como foi desenvolvido neste trabalho, tenha-se como norte um modelo amplamente utilizado e consistente.

Para embasamento da implementação do modelo, foram realizadas pesquisas bibliográficas nas mais referenciadas plataformas de periódicos, conforme especificado no item 2.4 deste estudo. Entretanto, em uma revisão do estado da arte no tema previsão de *churn* de clientes, foram encontrados poucos estudos que abordassem a predição de *churn* de seguros de vida, que é o assunto deste artigo.

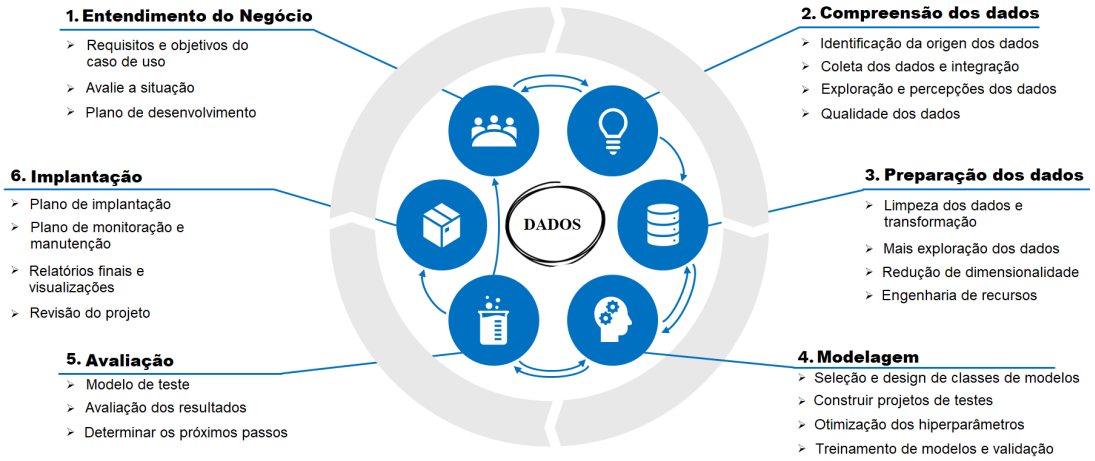

Figura 1.3: Fases e processos do modelo de referência CRISP-DM [1].

A Figura 1.3 mostra a estrutura das fases do modelo CRISP-DM. Consta a seguir o detalhamento dessa estrutura [1]:

1. A primeira fase de entendimento dos dados é primordial para o sucesso do estudo. Nesta etapa buscam-se os detentores do conhecimento referente ao negócio, para extrair os requisitos e objetivos do projeto, avaliar a situação, e efetuar o planejamento da implementação do estudo.

2. Quanto à segunda fase, consiste em conhecer as bases de dados, analisar de onde serão extraídos os dados, e contempla toda a parte de coleta e exploração dos dados obtidos, além do melhor entendimento e retirada das primeiras percepções dos dados. Esta fase também abrange a verificação da qualidade e da consistência desses dados para o estudo a ser desenvolvido.

3. A terceira fase, refere-se à parte de preparar os dados para entrada na fase da modelagem. Nesta etapa, é feita a limpeza dos dados, ou seja, é uma fase de exploração e transformação desses dados, como: retirada de dados com valores ausentes, verifi-

cação de *outliers*, transformação de variáveis *target*, categorização, além de efetuar a escolha das variáveis entrantes no estudo e redução da dimensionalidade.

4. A fase da modelagem é onde se constrói o modelo planejado nas fases anteriores. Neste momento, procede-se à seleção das ferramentas e linguagens a serem utilizadas; é feita a parte de otimização de parâmetros dos métodos, treinamentos de modelos e, ao final, a validação do modelo desenvolvido.

5. Na quinta fase, após a implementação e validação do modelo, é feita a avaliação dos resultados obtidos a partir dos dados; nesta etapa é testado o modelo para que sejam analisados os resultados e avaliados quanto aos objetivos do estudo.

6. Ao final das cinco etapas, chega-se à fase de colocar o modelo em produção; nesse ponto, são efetuados os planos de implantação, monitoramento e manutenção do modelo, além dos relatórios finais e geração de gráficos para visualizações, e, por fim ,revisar todo o projeto desenvolvido.

1.5 Estrutura do Trabalho

O presente estudo está organizado em cinco capítulos, iniciando com a Introdução, que corresponde ao Capítulo 1, objetivando fazer uma contextualização do tema rotatividade de clientes contratantes de seguros de vida, mostrando as motivações, o problema a ser resolvido, bem como os objetivos do estudo realizado.

Seguindo, no Capítulo 2, apresenta-se a Contextualização Negocial, contendo os conceitos do contexto negocial e análise de sobrevivência, técnicas e modelos de aprendizagem de máquinas, referentes ao tema de *churn* de clientes, além de apresentar o referencial teórico, especificando os trabalhos que foram fundamentais para a elaboração deste estudo.

O Capítulo 3 refere-se à Revisão da Literatura, apresentando os principais autores e estudos que foram utilizados como bases para esta pesquisa, focando nos trabalhos que usaram o estimador de Kaplan-Meier e Regressão de Cox, que foram as técnicas utilizadas para o desenvolvimento do modelo proposto neste trabalho.

No Capítulo 4, será apresentado o Estudo de Caso, bem como a metodologia de pesquisa utilizada para o desenvolvimento do trabalho, além do detalhamento das fases de um projeto de mineração de dados, considerando que foi adotado o modelo bastante usado pelo mercado que é o CRISP-DM, que contempla as fases: Entendimento do Negócio, Compreensão dos Dados, Preparação dos Dados, Modelagem, Avaliação e Implantação. Ao seguir as fases, o presente modelo mostrou-se eficaz em todo o processo de desenvol-

vimento da pesquisa, tanto na estruturação, quanto no entendimento dos resultados do estudo.

Para finalizar, no Capítulo 5 serão apresentadas as conclusões e considerações sobre os resultados obtidos no estudo, e as possibilidades de evolução do modelo para trabalhos futuros.

Capítulo 2

Contextualização Negocial

Neste capítulo são apresentados o contexto negocial, onde acontece o fenômeno da rotatividade de clientes, bem como os conceitos e características do produto, as técnicas utilizadas, além das pesquisas que nortearam a implementação do estudo, efetuadas nas mais renomadas bases de dados referentes ao termo *churn* de clientes. Os referencias teóricos sobre o tema desta pesquisa foram elencados de sites, plataformas web, livros e estudos acadêmicos, relacionados ao contexto geral sobre *churn* de clientes.

2.1 O Negócio

O mercado de seguros continua aquecido com entradas de novas empresas e maiores faturamentos na venda de seguros. Diariamente, centenas de apólices de seguros são emitidas por diversas seguradoras existentes no Brasil, dentre as quais, a empresa Alfa – braço do ramo de seguros da instituição financeira brasileira. Porém, diante das expressivas vendas de seguros, há também um número bastante expressivo e crescente de cancelamento dos contratos de seguros, fenômeno conhecido como *churn* de clientes, principalmente quando se refere a seguros de vida [2]. Os clientes podem cancelar seus contratos por diversos motivos; assim sendo, grandes empresas seguradoras, atualmente vêm sofrendo com essa rotatividade e gastam cada vez mais recursos financeiros no processo de retenção e aquisição de novos clientes [13].

2.2 A Empresa Alfa

A empresa Alfa, é uma empresa do conglomerado de uma instituição financeira brasileira, que detém cerca de 66% das suas ações, e configura-se, portanto, como acionista majoritário. Compõem o grupo de empresas participadas [3]:

- **Empresa A**, atua nos ramos de seguros de vida, residencial, habitacional e de veículo;
- **Empresa B**, atua no ramo de previdência privada;
- **Empresa C**, atua no ramo de títulos de capitalização;
- **Empresa D**, atua no ramo de planos odontológicos;
- **Empresa E**, atua nos ramos de seguros viagem, seguro celular e saúde.

Dessa forma, a empresa Alfa apresenta-se como um conglomerado de empresas que atua em diversos ramos do mercado financeiro, conforme sua composição, mostrada na Figura 2.1 [2].

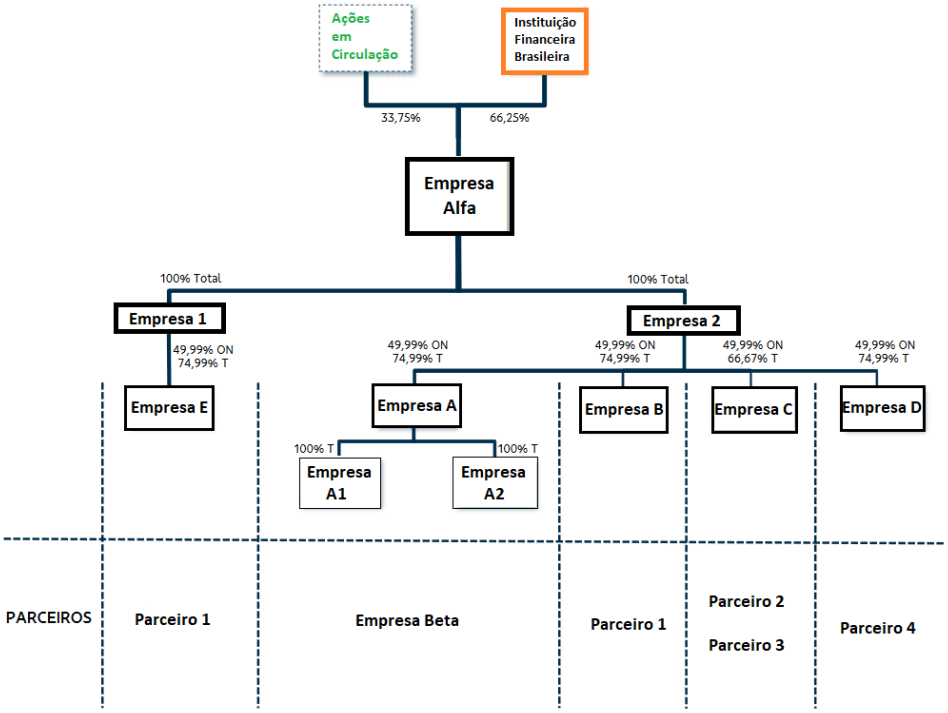

Figura 2.1: Estrutura da organização da empresa Alfa [2]

2.3 Seguros de Vida

Seguro de Vida é um seguro do ramo de Seguros de Pessoas, que consiste em um produto desenvolvido e comercializado por empresas corretoras de seguros, com o objetivo de prover garantias mediante o pagamento de um valor – conhecido como prêmio – a uma empresa, para recebimento de uma indenização ao segurado ou aos seus beneficiários, de

acordo com as condições negociadas em contratos e pactuados entre uma empresa corretora e uma pessoa, assegurado as coberturas e limites dos capitais segurados, considerando que este seguro pode ser contratado de forma individual ou coletivo [14].

Os contratos de seguros de vida, no âmbito da corretora empresa Alfa, têm duração de 5 anos, podendo ser renovados de acordo com o desejo do cliente, que pode escolher entre as modalidades de seguros de vida A, B e o C, dependendo da necessidade, dos objetivos e benefícios escolhidos pelo segurado. Como descrito acima, a empresa Alfa efetuou uma reformulação em seu portfólio de seguros de vida, onde de um total de mais de 6 tipos de seguros se resumiu a apenas 3 [3], conforme Figura 2.2.

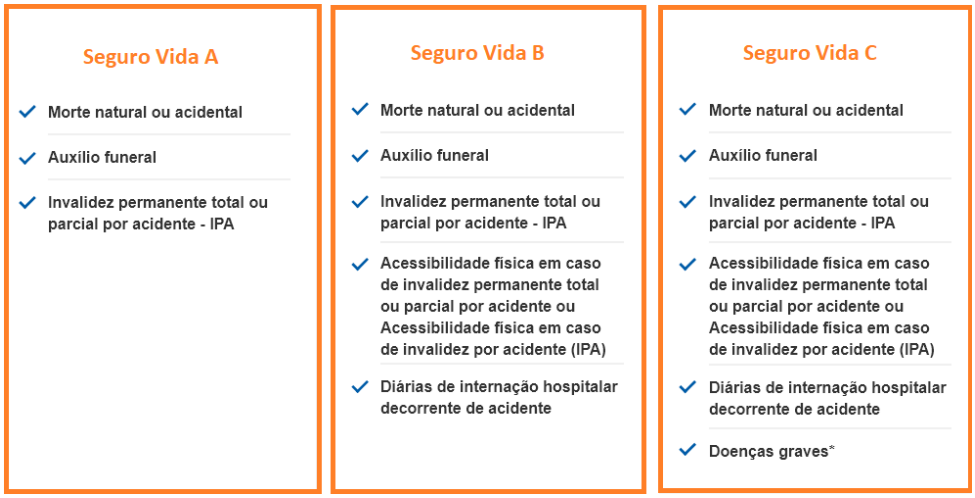

Figura 2.2: Modalidades de seguros de vida ofertados pela empresa Alfa [3].

2.4 Rotatividade de Clientes

A rotatividade de clientes ou *churn* de clientes, é um fenômeno que ocorre quando uma determinada pessoa deixa de fazer negócios com uma empresa ou serviço que detinha contrato. É também conhecido como atrito entre clientes, movimento em que um indivíduo concretiza o evento de saída ou cancelamento de contratos pactuados, para iniciar um novo pacto em outra empresa, ou simplesmente abandonar a empresa com a qual tinha relacionamento [15].

O *churn* de clientes, é considerada como uma métrica bastante utilizada nas mais variadas empresas, tanto de pequeno, quanto de médio e grande porte. A retenção de clientes torna-se, a cada dia, um item de extrema importância para a sobrevivência dos negócios de muitas empresas; dessa forma, essas empresas não medem esforços para manter seus clientes fidelizados e satisfeitos com seus produtos, considerando que reter um cliente torna-se menos oneroso do que ganhar um novo cliente [7].

Portanto, os cancelamentos de contratos de serviços ou cancelamentos de contratos de produtos, configuram-se como grandes preocupações de muitas empresas, criando verdadeiros desafios para manter seus clientes em suas bases, consumindo seus produtos e gerando rentabilidade. Porém, quando ocorre o *churn*, esse processo gera saídas de clientes e, como consequência, a diminuição da rentabilidade, e a redução da fatia de mercado que uma empresa domina.

Para obter a métrica da taxa de *churn* de uma determinada empresa, é necessário efetuar o seguinte cálculo de divisão, como descrito na Figura 2.3:

$$\text{Taxa de } \textit{Churn} = \frac{\text{Número de cliente que cancelaram seus contratos no mês}}{\text{Número de clientes no início do mês}}$$

Figura 2.3: Representação da fórmula da taxa de rotatividade.

Uma das ações mais importantes dentro de uma organização, é tentar antever quantos de seus clientes cancelarão seus contratos ou deixarão de se relacionar com ela. Por essa razão, é primordial que ocorra uma gestão eficiente e preditiva da rotatividade de clientes, considerando a sobrevivência dos negócios da instituição ao longo dos anos. Portanto, medir a taxa de *churn* e gerir essa rotatividade com uso de modelos analíticos eficientes, é crucial para a sobrevivência de muitos negócios [15].

2.5 Referencial Teórico

Nesta sub-seção apresenta-se o referencial teórico, onde consta os principais trabalhos e autores que embasaram o desenvolvimento do presente estudo.

2.5.1 Mineração de Dados

A mineração de dados constitui uma etapa do processo de descoberta do conhecimento, conhecido como KDD, refere-se a um campo amplo onde, em níveis abstratos, esse processo se ocupa no desenvolvimento de técnicas e métodos para dar sentido aos dados. Constitui uma estrutura composta por várias etapas, que busca encontrar informações úteis e padrões ocultos nos dados. Nesse sentido, a descoberta de conhecimento se torna crucial devido à quantidade enorme e crescente de dados [16].

O KDD se divide em 5 etapas bem definidas e complementares; assim, uma etapa gera insumo para outra em um processo contínuo de busca de conhecimento [16], conforme descrito na Figura 2.4:

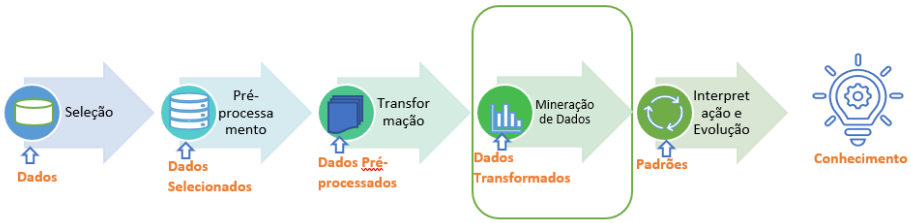

Figura 2.4: Etapas dos processos do KDD.

Apresentada na Figura 2.4, a mineração de dados é uma etapa muito importante do processo de descoberta do conhecimento; porém, não é um processo recente – surgiu no final dos anos 1980, deu grandes passos durante os anos 1990 e continua seu crescimento no novo milênio [16]. Consiste no processo de coleta, limpeza, processamento, análise e obtenção de percepções úteis sobre dados; dentro dessas etapas existem as subetapas muito importantes que permeiam todo o fluxo de mineração de dados [17].

Em tempos modernos, todos os sistemas automatizados, de certa forma geram milhões de dados diariamente, podendo esses dados serem utilizados em estudos e análises, a fim de obter informações e *insights* importantes. Assim sendo, minerar dados pode ser compreendido como o resultado do processo de evolução natural da tecnologia da informação [17].

2.5.2 Análise de Sobrevivência

Análise de sobrevivência ou análise de sobrevida, constitui um ramo da estatística que visa estimar o tempo de duração esperada até a ocorrência de um determinado evento [18]. Tais eventos podem ser denominados ocasião da falha, quando se refere a pessoas ou organismos, falhas de sistemas automatizados, cancelamento de contratos de prestação de serviços, dentre outros tipos de eventos. Esta modalidade de análise está sendo usada para diversos tipos de estimativas de sobrevivência, porém, as mais comuns são referentes a estudos e pesquisas médicas para estimar o tempo de vida de determinados pacientes que são submetidos a diferentes tratamentos ou uso de medicamentos, conforme estudos efetuados por diversos autores [19][20][21].

Para um estudo robusto de análise de sobrevivência, um dos pontos mais importantes é definir o que configura o evento de interesse [18]; portanto, para o presente estudo foi

definido como evento de interesse, o momento em que um determinado cliente realiza o ato de cancelar seu contrato de seguro de vida, deixando de fazer parte da base de clientes da empresa. Dessa forma, a análise de sobrevida visa estimar o tempo de sobrevivência e o ritmo que um determinado cliente contratante de seguro de vida leva até o momento em que este cancela seu seguro, configurando como a finalização do contrato pelo cliente.

2.5.2.1 Censura

Um fator de extrema importância na análise de sobrevivência é a ocorrência de censura dos dados estudados, essa ocorrência possibilita uma maior robustez à pesquisa, considerando que retira do estudo dados de tempos incompletos. Como o estudo de sobrevida compreende um espaço de tempo, há ocorrências que já existiam antes do início do estudo, considerados como dados censurados à esquerda, e existirão também dados que não conhecemos o desfecho, considerando que continua ativo no final do estudo, sendo elencados como censura à direita. O dado sem início e/ou fim conhecido/s, será censurado do estudo.

Por definição, censura é o evento que ocorre quando temos informações sobre o tempo de sobrevivência individual, mas não sabemos exatamente o tempo total de sobrevivência deste indivíduo, ou seja, não temos o evento de morte ou falha. Assim, a censura ocorre de 3 maneiras [18], conforme segue:

1. Quando o indivíduo ou objeto não experimenta o evento de falha ou morte antes do fim do estudo, o tempo de sobrevivência é maior ou igual ao tempo de sobrevivência observado, configurando como censura à direita.

2. Quando o indivíduo ou objeto perde o acompanhamento durante o período de estudo, o tempo de sobrevivência é menor ou igual ao tempo de sobrevivência observado, configurando como censura à esquerda.

3. Quando o indivíduo ou objeto é retirado do estudo por eventos adversos ao evento de interesse, que é a morte ou falha, configura-se como uma censura aleatória.

Portanto, a notação para a censura dá-se como apresentado abaixo, onde os tempos censurados são descritos com 0, e os não censurados, denotados por 1, o $T =$ ao tempo de sobrevivência e $C =$ ao evento da censura, onde T tem que ser maior que C.

$$\delta = \begin{cases} 0, & \text{se o tempo de sobrevivência é censurado,} \\ 1, & \text{para tempo de sobrevivência não censurado, ou seja, } T > C. \end{cases}$$

Para o presente estudo, será utilizada apenas a censura à direita, considerando a natureza e as características da pesquisa.

2.5.3 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier (KM), foi desenvolvido pelos pesquisadores Edward Kaplan e Paul Meier em 1958 [22]. É uma técnica muito utilizada pela comunidade acadêmica para traçar a curva de sobrevivência de indivíduos ao longo do tempo. Este estimador visa efetuar uma análise de sobrevivência onde os intervalos de tempo são alterados de acordo com o acontecimento do evento de interesse, que pode ser a falha de uma peça mecanizada ou até a morte de uma pessoa. Assim, o tempo até a ocorrência de um evento de interesse, pode ser definido como uma variável de duração no curso do estudo para cada sujeito, tendo um início e um fim acontecendo em qualquer lugar da linha do tempo na duração do estudo [4], como demonstrado na Figura 2.5:

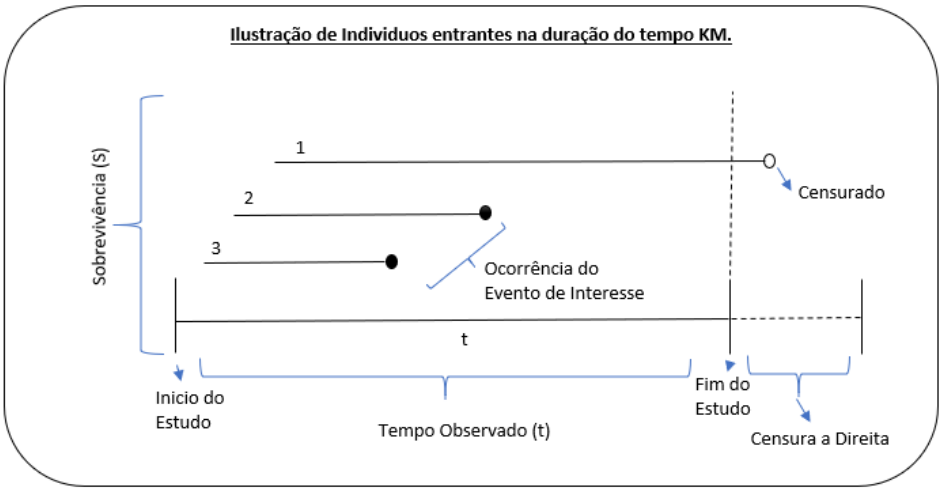

Figura 2.5: Ilustração dos indivíduos no estudo de sobrevida de Kaplan-Meier.

A Figura 2.5, apresenta o processo de indivíduos entrantes no estudo de sobrevida em momentos distintos, e sua finalização também em momentos distintos, considerando que a finalização se dá pelo acontecimento do evento de interesse, como é o caso de falha ou morte do indivíduo, ou pela sua censura do estudo. Dessa forma, nessa ilustração, os indivíduos 2 e 3 falharam, representados por círculos cheios, e no indivíduo 1 não foi observado o evento de interesse até o final do estudo. Assim, esse indivíduo não contou para o cálculo da geração da curva de sobrevivência até o final do estudo, e por fim foi censurado [18], representado pelo círculo vazio. Portanto, no processo da análise de sobrevida de KM, cada indivíduo pode ser caracterizado por três variáveis distintas [4]:

1. Seu tempo de duração;

2. Seu estado final no fim do estudo (falho ou censurado);

3. Em qual grupo de estudo esse indivíduo está inserido;

Para este artigo utilizaram-se amostras extraídas de forma aleatória de tamanhos N. Dessa forma, a estimativa do limite do produto (PL) pode ser calculada efetuando a organização dos tempos de sobrevida em ordem crescente, conforme ($0 < t1 < t2 < t3.... < tn$), independente do momento em que cada indivíduo entrou no estudo [22]. Assim, os membros que sobreviveram ou não falharam no tempo (t), configuram-se como censura, onde esses elementos são considerados nos cálculos da função da probabilidade de sobrevivência até o momento em que ocorrer a falha; portanto, a censura acontece quando temos informações referentes ao tempo de sobrevivência individual, porém, não sabemos exatamente quanto tempo de sobrevida este indivíduo possui [18].

Considerando o universo de problemas de estimativas, é notadamente inviável ou impossível efetuar medições completas de todos os itens de uma amostra aleatória [22]. Portanto, baseado nessa proposição, este estudo visa determinar a distribuição dos tempos de sobrevida até que ocorra o evento de interesse – neste caso, o cancelamento do contrato de seguros de vida. Dessa forma, muitos dos clientes contidos na amostra não chegarão ao evento esperado, ocorrendo a perda de contato antes da falha; por isso, essas observações são consideradas incompletas, sendo necessário estipular um prazo adequado para gerar os resultados da análise de sobrevida. Neste estudo, a amostra extraída pode ser considerada incompleta; assim, a estimativa correspondente é dada como uma função escalonada com descontinuidade do tempo dos eventos observados [22].

2.5.3.1 Função de Sobrevivência

O estimador de Kaplan-Meier é definido como um método não paramétrico, onde sua função de sobrevivência não depende da sua distribuição de probabilidades, isso é caracterizado quando a classe de distribuições admissíveis que melhor se ajustar aos dados do estudo, será a classe de todas as distribuições [22]. Portanto, na criação de uma análise de sobrevida, utilizando como estimador o KM, os possíveis resultados só podem acontecer com duas situações [4]:

- O - Ocorrer a censura;
- 1 - Ocorrer a falha ou morte;

Conforme apresentado na Figura 2.6.

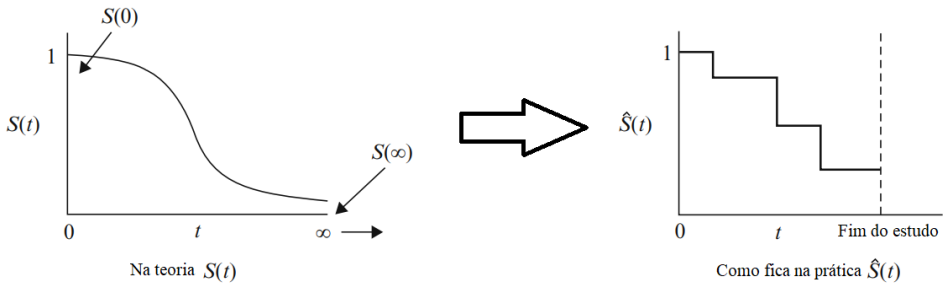

Figura 2.6: Estimador da curva de sobrevivência [4].

A Figura 2.6 ilustra a teoria e a prática de como fica representada a curva de sobrevida de KM, após aplicação da fórmula nos dados da amostra [18], onde:

$S(t)$ – Refere-se à função de sobrevivência no tempo t.

t – Tempo de duração dos indivíduos durante o estudo,

1 – Ocorrência do evento de interesse (falha)

0 – A censura (quando não foi possível observar o evento de interesse)

Para uma análise mais robusta, é primordial o estudo e o entendimento da formulação por trás da representação gráfica que o estimador KM traça para os dados em estudo [18]. Duas importantes funções são apresentadas:

1. $S(t)$ **Função de Sobrevivência**, objetiva dar a probabilidade em que a variável aleatória T ultrapasse o tempo estimado em t, ocupando-se no evento de não falhar. Assim, essa função pode ser descrita como $P(t) = Pr(T > t)$ [18], onde:

   Pr = Probabilidade relativa

   T = Tempo de sobrevivência (somatório dos tempos individuais)

   t = Valor específico de T, tempo de cada indivíduo.

2. $h(t)$ **Função de Risco**, objetiva dar o potencial instantâneo por unidade de tempo para que o evento aconteça, considerando que o indivíduo não falhou até o tempo t, ocupando-se no evento de falha. Dessa forma, quanto maior o risco médio, pior o impacto sobre a sobrevida [18].

   h – Risco de sobrevivência

   (t) – Tempo de sobrevivência do indivíduo.

Para calcular os estimadores de KM, usa-se a seguinte fórmula 2.1:

$$S(t) = \prod_{t=0}^{j} \frac{l_j - i}{l_j} \tag{2.1}$$

Onde:

$S(t)$ Sobrevivência até o tempo t

$\prod_{t=0}^{j}$ Somatório da divisão de $\frac{l_j - i}{l_j}$

i – Pode ser, 1 se ocorrer o evento de falha ou 0 se for censurado.

l_j – Número de indivíduos exposto no início do estudo.

2.5.4 Outros Estimadores

Além do estimador de Kaplan-Meier, existem outros estimadores que são utilizados como alternativas, que é o caso do estimador de Nelson-Aalen e o de Breslow. O estimador de Nelson-Aalen foi desenvolvido por Nelson em 1972 e Aalen em 1978 [23]; assim como o KM, este é também um estimador não paramétrico, onde sua função utiliza a taxa de risco cumulativa, não requerendo variáveis explicativas, ou seja, não mede os efeitos de outras variáveis; somente da variável de tempo ou de censura de forma ordenada [24]. Um outro importante estimador é o de Breslow, de 1980, que propôs a junção do estimador de Neslon-Aalen com sua função acumulativa, e a função de relação de casos contínuos. Dada essa junção, o estimador é conhecido como Nelson-Aalen-Breslow [25]. Entretanto, para o presente estudo, optou-se pelo estimador de Kaplan-Meier, por se adequar melhor a grandes conjuntos de dados – neste caso, mais de 917 mil registros – enquanto o estimador de Nelson-Aalen é mais utilizado para pequisas com conjunto de dados menores [18].

2.5.5 Modelo de Riscos Proporcionais de Cox

A Regressão de Cox ou Modelo de Risco Proporcional de Cox, desenvolvida pelo estatístico britânico Dr. David R. COX em 1972 [26], tornou-se um dos mais populares e estudados modelos estatísticos, que tem por finalidade a modelagem de dados de tempos de vida censurados com covariáveis. Um dos fatores que tornou esse modelo tão importante é o fato de ele ser semi-paramétrico, possibilitando maior flexibilidade no desenvolvimento da análise de sobrevivência. Um outro fator que potencializa a popularidade do modelo, é o de ser considerado robusto por apresentar resultados muito parecidos com modelos paramétricos corretos, levando em consideração o fato de que não é fácil ter essa certeza referente à escolha do modelo paramétrico correto [18].

Por ser um modelo semi-paramétrico, ele se baseia em uma estrutura de riscos proporcionais, ao ponto que fica desconhecida sua linha de base [27]. Assim, o modelo assume que as taxas de falhas são proporcionais, considerando que o risco de falha das variáveis são constantes ao longo do tempo. Exemplificando essa suposição, podemos dizer que o risco do cancelamento de um contrato por um cliente com estado civil casado, em relação a outro cliente com estado civil solteiro, seja constante ao longo do tempo do estudo [26].

Normalmente os modelos de sobrevivências de riscos proporcionais requerem duas etapas [26], a primeira refere-se à construção da função de risco da linha de base, onde compreende a mudança do risco do evento por tempo (medido em unidades) ao longo do período nos níveis da linha de base, em relação às covariáveis. E a segunda, refente ao parâmetros de efeitos, mostrando como o risco varia em relação a covariáveis explicativas, exemplo dessas variáveis são idade do indivíduo, gênero, estado civil, dentre outras.

Cox, verificou que, se os riscos proporcionais forem válidos ou considerado válidos, será possível estimar os parâmetros de efeitos, desconsiderando a função de risco. Portanto, o modelo proporcional de Cox, parte dos seguintes pressupostos [26][27]:

- As covariáveis agem multiplicando os efeitos nos riscos, configurando a parte paramétrica do modelo;
- A razão dos riscos é constante ao longo do tempo do estudo, gerando os riscos proporcionais;
- Os tempos de ocorrência do evento são independentes.

Partindo dos pressupostos descritos acima, e por ser semi-paramétrico, o modelo de Cox é considerado como um método bastante flexível e versátil na modelagem de dados de sobrevivência [27]. O método já modela diretamente a função de risco; em contrapartida os modelos paramétricos são mais rígidos, exigindo que já exista suposição de distribuição para os tempos de sobrevivência [27]. Para o presente estudo, será utilizado o modelo semi-paramétrico de Cox para aproveitar o máximo dessa flexibilidade e versatilidade que o modelo oferece na análise de riscos proporcionais de sobrevida. Assim, busca-se a estimação dos efeitos das covariáveis, sem efetuar nenhuma suposição quanto à sua distribuição do tempo de vida [26].

Atualmente existem diversos estudos na área médica que se utilizam do modelo de regressão de Cox para predizer o evento de interesse buscado pelo pesquisador.

2.5.5.1 Função de Cox

O modelo desenvolvido por Cox em 1972, possui a propriedade que diferentes indivíduos têm funções de risco que são proporcionais [28]. Dessa maneira, esse modelo fornece uma fórmula para o risco no tempo t para certo indivíduo, que seja não negativa e com uma determinada especificação de um conjunto de covariáveis explicativas, representadas por X. Com essas, gera-se um vetor de variáveis preditoras para modelar a previsão do risco de um indivíduo. Assim, sua fórmula pode ser descrita da seguinte forma $h(t/X1, X2, ..., Xk) = h_0(t).exp(\beta 1X1 + \beta 2X2 + ...+)$ e pode ser dividida em duas partes [18]:

1. Refere-se à função de risco de linha de base, representada pela fórmula 2.2:

$$h_0(t) \tag{2.2}$$

   Interessante notar que a função 2.2 não envolve as variáveis explicativas X, tornando-se uma função independente em relação às variáveis preditoras. Essa primeira fórmula, configura-se como parte paramétrica do modelo; por isso, o modelo é descrito como semi-paramétrico, considerando ser uma função não especificada [18].

2. Referente à expressão exponencial do somatório linear, onde a soma é sobre as p variáveis explicativas X, conforme abaixo:

$$e^{\sum_{i=0}^{p} \beta_i X_i} \tag{2.3}$$

   Diferente da função de risco de base 2.2, essa expressão 2.3, já envolve as variáveis preditoras X, porém fica, independente do tempo, representada por t. Portanto o X, representa as variáveis preditoras independentes do tempo, visto que essa característica que configura a suposição de riscos proporcionais de Cox [18].

Portanto, a representação formal da fórmula do modelo proporcional de Cox, fica como segue:

$$h(t, X) = h_0(t) \cdot e^{\sum_{i=0}^{p} \beta_i X_i} \tag{2.4}$$

Onde essa equação 2.4, representa a junção das duas expressões 2.2 e 2.3, dado que, a função de risco de linha de base vai traçar o risco dos indivíduos em função do tempo, e a função exponencial, que é o somatório envolvendo a função de máxima verossimilhança parcial β_i, com as variáveis preditoras X_i. Uma propriedade importante da fórmula do modelo de COX, é a suposição de que se todo X for igual a zero, o modelo se reduz à função de linha base 2.2, assim a parte exponencial se torna elevada a 0, que é igual a 1 [26][27].

2.5.5.2 Função de Máxima Verossimilhança Parcial

A função de máxima verossimilhança parcial é dada que, uma amostra com n indivíduos, com número de falhas ou mortes distintas representadas por k, onde k seja maior ou igual a n, considerando os tempos $t1 < t2 < t3...tk$ [26]. Esta função parte do pressuposto que os tempos de vida são contínuos, considerando que não exista tempos empatados nas

observações. Desta forma, para grandes conjuntos de dados com um número considerável de tempos empatados, o mais recomendado é que seja utilizado o modelo de Cox [29][30].

Portanto, a verossimilhança parcial parte do argumento de que a probabilidade condicional da i-ésima observação vir a falhar no tempo t, conhecendo as observações que estarão sob o risco em t_i [31], a sua notação fica conforme descrita 2.5:

$$\frac{\lambda_i(t)}{\sum_{j \in R(t_i)} \lambda_j(t)} = \frac{\lambda_0(t) \cdot exp\{x_i'\beta\}}{\sum_{j \in R(t_i)} \lambda_0(t) \cdot exp\{x_i'\beta\}} = \frac{exp\{x_i'\beta\}}{\sum_{j \in R(t_i)} exp\{x_i'\beta\}} \tag{2.5}$$

Conforme notação 2.5, condicionando as falhas e censuras no tempo t_i, desaparece o componente não-paramétrico $\lambda_0(t)$ e o $R(t_i)$ corresponde ao conjunto de observações sob risco em t_i.

Assim, a função de verossimilhança parcial não prever a possibilidade de observações empatadas; ela considera que os valores dos tempos de sobrevivência são contínuos [26][31].

Considerando isso, ela é composta pelo produto dos termos descritos na notação 2.5, com o indicador de falha ou morte denotado por $\lambda_0(t)$, seguindo da fórmula 2.6:

$$L(\beta) = \prod_{i=1}^{k} \frac{exp\{x_i'\beta\}}{\sum_{j \in R(t_i)} exp\{x_i'\beta\}} = \prod_{i=1}^{n} \left(\frac{exp\{x_i'\beta\}}{\sum_{j \in R(t_i)} exp\{x_i'\beta\}} \right)^{\delta_i} \tag{2.6}$$

Para se obter uma função que leve em consideração os tempos empatados, é necessário ajustes na fórmula descrita acima 2.6. Essa alteração para que possa considerar observações empatadas, foi proposta por Breslow em 1972 [32]. Porém, para o presente estudo será utilizada a função sem a presunção de tempos empatados.

Capítulo 3

Revisão da Literatura

O presente Capítulo é destinado a apresentar o estado da arte, conceitos e técnicas relacionadas à análise de sobrevivência e predição de falhas ao longo do tempo.

Esta revisão da literatura foi feita através de buscas de trabalhos nas bases de dados *Web of Science*, Scopus, e *Google Scholar*, foram selecionados os estudos dentre os 10 mais bem ranqueados pelo nível de relevância de cada base. Dessa forma, a revisão foi estruturada em três seções, sendo a primeira referente aos estudos com técnicas variadas utilizadas para análise de sobrevivência e predição de falhas, a segunda apresenta os trabalhos que utilizaram as técnicas de Kaplan-Meier e Regressão de Cox, e a terceira traz as considerações dos trabalhos estudados relativos ao tema deste estudo.

3.1 Análise de sobrevivência Técnicas Diversas

O estudo da análise de sobrevivência é um tema importante, considerando a quantidade de trabalhos buscados nas 4 bases de dados, que passam de três mil publicações, pesquisado com o título *"survival analysis"*, sem restrição de tempo. Nos artigos estudados, foram encontrados diversos temas e objetivos diferentes para propor soluções para uma gama de problemas, conforme os parágrafos a seguir.

Os autores Azeem, Usman e Fong [33], em seu estudo sobre previsão de *churn* de clientes de empresas de telecomunicações, propuseram desenvolver um modelo de previsão de *churn* mais adequado, capaz de ajudar a identificar os clientes com maior probabilidade de abandono, ou seja, tentar prever a rotatividade do cliente. Para isso, os autores aplicaram alguns métodos de aprendizado de máquina, como os classificadores: Rede Neural, Regressão Linear, C4.5, SVM, AdaBoost, *Gradient Boosting* e *Random Forest*, nos dados reais desta empresa, e fizeram uma comparação dos resultados para verificar a acurácia de cada método nos dados, e assim chegou-se à conclusão de que os classificadores *fuzzy* (FuzzyNN, VQNN, OWANN e FuzzyRoughNN) se mostraram mais adequados, com

maior poder preditivo, e com maior precisão para identificar os clientes que tendem a *churners*, com uma taxa positiva verdadeira (TP) de 98% e AUC 0,68, enquanto os demais modelos ficaram abaixo de 69% e 0,52, respectivamente.

Gaur e Dubey [34], focaram os estudos na comparação de diferentes técnicas de mineração de dados para entender qual dessas técnicas tem maior poder e eficiência na previsão do *churn* de clientes. Basearam-se em dados reais de uma empresa de telecomunicações, e aplicaram essas técnicas para construir modelos de classificação com Regressão Logística, SVM, Floresta Aleatória e Árvore Aumentada de Gradiente. Após a aplicação dos modelos, os resultados foram comparados e concluíram que a árvore *Gradient boosted* (com AUC 84,57%) teve maior poder preditivo e eficiência entre os quatro modelos (com AUC abaixo de 82,86%).

Já os autores Zhang, Li, Mo e Tan [35], propuseram uma combinação de modelos superficiais, como a regressão logística, com sofisticada engenharia de recursos e aprendizado de máquina profundo. Para os autores do estudo, o objetivo deste novo modelo, denominado DSM (*Deep Shallow Model*) para prever a rotatividade para o setor de seguros, é buscar aproveitar e combinar os pontos fortes de cada uma das duas metodologias para chegar a uma solução ideal em eficiência e alto poder preditivo de rotatividade de clientes. De acordo com os autores, o DSM foi superior a todos os outros modelos comparados, sendo CNN, LSTM, *Stochastic Gradient Descent, Linear Discriminant Analysis, Quadratic Discriminant Analysis, Gaussian Naive Bayes, AdaBoost, Random Forest* e *Gradient Tree Boosting*.

Em outra abordagem referente ao *churn* no campo de recursos humanos, Yigit e Shourabizadeh [36], aplicam técnicas de mineração de dados de classificação bem conhecidas, que são Árvore de Decisão, Regressão Logística, SVM, KNN, Random *Random Forest* e *Naive Bayes*, em dados de funcionários para prever Rotatividade de funcionários da IBM. Este estudo mostra que as pesquisas no campo de rotatividade de indivíduos não se resume a clientes de empresas, mais também, em outras áreas de negócios, como relações humanas.

Dulhare e Ghori [37], em seu estudo, mostram uma combinação de técnicas para prever a rotatividade de clientes; os autores criaram um modelo híbrido, juntando os métodos *Aximatic Fuzzy Set* AFS com agrupamento DBSCAN paralelo e, com isso, aplicaram aos dados para efetuar predição de *churn* de clientes e concluíram que esse modelo proposto é eficiente tanto em tempo de processamento, quanto em manipulação de volume crescente de dados.

Em sua pesquisa, os autores Perianez et al [38] propuseram a criação de modelos de previsão para antecipar quando um determinado jogador de jogos sociais de celulares iria deixar de jogar. Para tanto, aplicaram as técnicas de Kaplan-Meier para traçar a curva

de sobrevivência dos jogadores de jogos sociais, e também utilizaram a regressão de Cox para fazer previsão da rotatividade desses jogadores ao longo do tempo em que estavam jogando. Assim, como resultado do estudo possibilitou que os autores criassem perfis de jogadores que teriam maiores probabilidades de deixar o jogo.

3.2 Análise de sobrevivência com Klapan-Meier e Regressão de Cox

As buscas de artigos relacionados ao tema do presente estudo, foram restringindo a pesquisa para o tema *"survival analysis kaplan-meier and cox regression"*, sem restrição de tempo. Chegou-se, então, aos artigos nos parágrafos abaixo.

Os autores Abdelsalam, Elbashir e SaadEldeen [39], buscaram o desenvolvimento de um modelo de sobrevivência utilizando as técnicas de Kaplan-Meier e regressão de Cox para traçar a curva de sobrevivência em pacientes com câncer. No estudo, eles utilizaram o estimador de Kaplan-Meier para estimar a função de sobrevivência e o tempo médio dos pacientes com câncer de mama; assim, os pacientes foram agrupados pelo fator do estágio da doença e utilizaram o Teste Log-Rank para comparar os resultados das curvas de sobrevida por grupos de pacientes. A regressão de Cox foi usada para determinar quais fatores estariam afetando o tempo de vida dos pacientes, e assim poder avaliar os riscos relativos. Como resultado, os autores concluíram que não há muita diferença nas curvas de sobrevivência dos pacientes do sexo masculino para o feminino; os fatores prognósticos que influenciam são a idade, o estágio da doença, metástase à distância, parede torácica e hormônio. Os pacientes possuem tempo de vida global de 4.429 dias e média de sobrevivência de 2.085 dias, além disso, os pacientes do grupo no estágio IV da doença tem 4,53 vezes mais risco de morte do que os pacientes que estão no estágio I, com índice de confiança de 95%.

Fidelizar clientes é um dos desafios de muitas empresas. O fenômeno da rotatividade apresenta-se com maior evidência nas indústrias de telecomunicações, como mostra o autor Tristan [40], em cujo estudo foi desenvolvida uma análise de sobrevivência e predição de *churn*, de clientes de uma empresa de telecomunicação americana. Para isso, foram utilizadas as técnicas do estimador de Kaplan-Meier para traçar as probabilidades de sobrevida e a regressão de Cox para identificar os riscos proporcionais de *churn* dos clientes, objetivando identificar com antecedência os potenciais clientes que a abandonariam a empresa, possibilitando que os gestores efetuem ações para a retenção desses clientes. Um outro ponto importante estudado, foi a busca da identificação do efeito das covariáveis sobre a sobrevivência, e o risco da rotatividade entre as subcovariáveis. Com essas infor-

mações, podem-se classificar os clientes que possuem maiores riscos de abandonarem a empresa.

Em seu estudo, os autores Tsai et al [41] desenvolveram uma estrutura sistemática para previsão de *churn*, além de fornecer a compreensão e as respostas do fenômeno conhecido como rotatividade de clientes. O estudo traz conceitos importantes sobre *churn* de clientes, classificação e as principais abordagens para detecção da rotatividade de clientes nas empresas. Os pesquisadores deste artigo foram bem-sucedidos em suas propostas, considerando a ajuda para o desenvolvimento da presente dissertação, visto que, trouxe pontos de vantagens e benefícios para a implementação de estrutura para o estudo da rotatividade de clientes, conforme segue [41]:

- Fornece uma maneira mais abrangente de previsão de rotatividade de clientes, não apenas para prever a rotatividade individual, mas também a rotatividade de clientes em grupo, para prever a rotatividade em três estágios, o que pode descobrir mais padrões ocultos de rotatividade do cliente por meio das informações do rastreamento, análise e medição do comportamento do cliente residencial, habitacional e de veículo;

- Fornece um mecanismo para a empresa entender melhor seus clientes e olhar para frente em sua rotatividade, incluindo por que, como e onde está a rotatividade e, portanto, poder melhorar o conhecimento do cliente e aprimorar produtos e serviços personalizados de acordo com os gostos e necessidades dos clientes.

- Fornece suporte para a decisão responsiva da empresa às situações de possível rotatividade do cliente e sua intenção pode ser entendida pelas empresas com foco nas necessidades e requisitos do cliente. Portanto, a empresa pode concentrar ou utilizar melhor seus recursos na segmentação de clientes que teriam alto impacto na receita se fossem cancelados; capitalização.

3.3 Considerações

Os artigos [33] [34] [35] [36] [37] e [38], foram importantes para este trabalho, no sentido de entender as variadas técnicas que a comunidade acadêmica utiliza para propor modelos na busca de melhores resultados quanto à análise de sobrevivência e predição de *churn*; porém, esses modelos descritos nos estudos, mostram como resultado apenas se um determinado indivíduo possui ou não probabilidade de sobreviver a determinado evento em suas análises de sobrevivência.

Um dos principais objetivos propostos neste estudo foi o de traçar curvas de sobrevivência e efetuar predição de *churn* para gerar os riscos relativos e possibilitar dizer de

forma mais precisa o tempo em que um determinado indivíduo chegará ao evento de interesse, que, para este trabalho, é o cancelamento do seu contrato de seguro. Nesse sentido, os estudos [39] [40] e [41], foram utilizados como base para a implementação do modelo, e ajudaram no entendimento das técnicas e formas de desenvolvimento.

A inovação trazida neste estudo, é a possibilidade de prever, de forma mais precisa, o tempo de falha de um determinado indivíduo, ou seja, a possibilidade de conhecer os riscos relativos e predizer em meses o tempo em que o cliente irá cancelar seus contratos de seguros. Esse ponto não foi encontrado nos estudos das pesquisas constantes nas seções 3.1 e 3.2; todavia, o presente trabalho traz, como ganho para a comunidade acadêmica e empresarial, esse aspecto importante da análise preditiva de sobrevida dos indivíduos.

Capítulo 4

Estudo de Caso

Este capítulo apresenta o estudo de caso realizado nesta pesquisa, na qual foram seguidas as etapas do modelo CRISP-DM, a saber: Entendimento do negócio, Compreensão dos dados, Preparação dos dados, Modelagem, Avaliação e Implantação do modelo proposto. Os resultados parciais alcançados são apresentados a seguir.

4.1 Entendimento do Negócio

A empresa Alfa atualmente comercializa diversos produtos e serviços em seu portfólio, desde seguros a planos odontológicos. Em seu leque de seguros, dispõe de diversas modalidades [42], conforme apresentado na Figura 4.1.

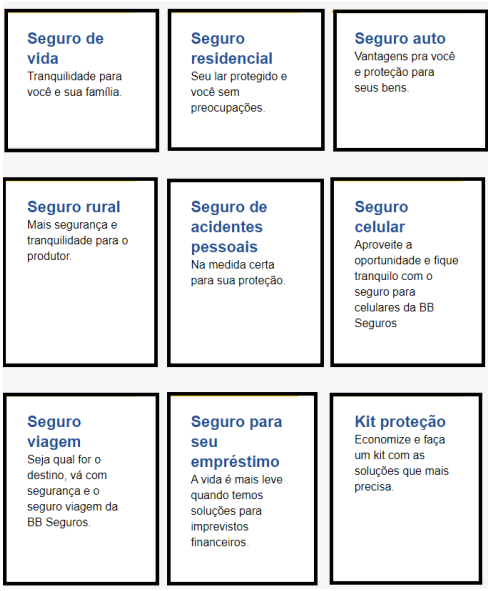

Figura 4.1: Tipos de seguros comercializado pela empresa Alfa.

A empresa Alfa funciona como uma corretora e gestora do produto seguro de vida, e tem como parceira nesse segmento a empresa Beta, conforme apresentado no organograma da empresa, na Figura 2.1. Este produto é comercializado principalmente nas redes de agências da instituição financeira brasileira, que é detentor da maioria das ações da empresa. Além das agências, esses seguros também são vendidos por corretores autônomos e empresas especializadas autorizadas para comercializar o produto.

O presente estudo focou em Seguros de vida, onde antes existiam mais de oito submodalidades de seguros. A reestruturação do portfólio de Seguros de Vida, resultou em apenas três tipos de seguros, conforme apresentado na Figura 4.2. Este produto tem vigência de cinco anos, podendo ser renovado automaticamente ao final de cada vigência, e podendo ser cancelada a qualquer momento a pedido do cliente.

A Figura 4.2, apresenta os preços que atualmente são praticados no mercado, com preço médio de R$ 20,06. O preço deste seguro é composto pela quantidade de coberturas existente em cada plano e, mediante aumento da quantidade, o preço aumenta. Além disso, ao final de cada ano de contrato, há um reajuste no preço das coberturas, conforme avança a idade do segurado. Para a contratação do seguro, é necessário que o segurado esteja em perfeitas condições de saúde e tenha idade entre 18 e 70 anos de idade [43].

Figura 4.2: Planos e preços do seguro de vida.

4.2 Compreensão dos Dados

Os dados utilizados para a realização deste estudo foram extraídos de três bases de dados estruturadas, distintas, segundo o modelo relacional, por meio da ferramenta de administração de dados *Dbeaver*.

A Tabela 4.1, descreve as bases de dados e quantidades de colunas existentes em cada uma.

Tabela 4.1: Descrição das bases de dados.

Bases	Descrição das Tabelas	Quant. de Colunas
1	Contratos de Seguros	96
2	Cadastros de Clientes	36
3	Dimensão Municípios	15

A base 1 contém mais de 10 milhões de registros e 96 colunas, com diversos formatos e tamanhos, contendo informações do cliente e do produto contratado por ele. Para facilitar a consulta dos dados desta tabela, foram utilizadas as chaves primárias, a saber: código do produto, modalidade do produto, estado da proposta, e período compreendido de 5 anos até a data de 31/12/2019.

A base 2 contém dados cadastrais de mais de 20 milhões de clientes da instituição financeira brasileira, com grande quantidade de registros, porém com menor número de colunas que a base 1 – apenas 36. Abrange, além das informações relativas ao cadastro do cliente, dados de relacionamento com o banco, tais como: quantidade de produtos contratados, tempo de banco e rendimento do cliente.

Já a base 3, representa a dimensão município, e contém dados de todas as cidades do Brasil. É uma tabela relativamente pequena, porém muito importante para a complementação dos dados utilizados para esse estudo. O interessante nessa base de dados é que, além das informações como nome, localização e região, contém ainda a população e a geolocalização dos municípios.

Para fazer a junção das três bases de dados, utilizou-se as seguintes combinações: código do cliente e código do município, conforme apresentado na Tabela 4.2.

Tabela 4.2: Forma de junção das bases de dados.

Atributo	Base	Operador	Atributo	Base
Código do Cliente	base 1	Igual	Código do Cliente	base 2
Código do Município	base 2	Igual	Código do Município	base 3

Para a extração dos dados foi utilizada a linguagem *Structured Query Language* (SQL), cujos dados foram armazenados em arquivos Excel, no formato *Comma-Separated Values* (CSV), visando facilitar o tratamento dos dados e desenvolvimento do modelo.

Com a extração dos dados, foi possível obter algumas informações, conforme a Figura 4.3, que apresenta a porcentagem de clientes por sexo. Pelo gráfico, verifica-se que não há diferença significativa em relação ao gênero dos clientes que contratam seguros de vida.

Figura 4.3: Proporção entre clientes do sexo feminino e masculino contratantes de seguros de vida.

Outra informação interessante pode ser observada na Figura 4.4, que apresenta os percentuais de clientes por região. A maioria (41%) dos clientes que contratam seguros de vida são da Região 4 (Sudeste), seguidos pela Região 2 (Nordeste) com 22%, Região 5 (Sul) com 19%, Região 3 (Centro-Oeste) com 12% e, por último, a Região Norte com 7%.

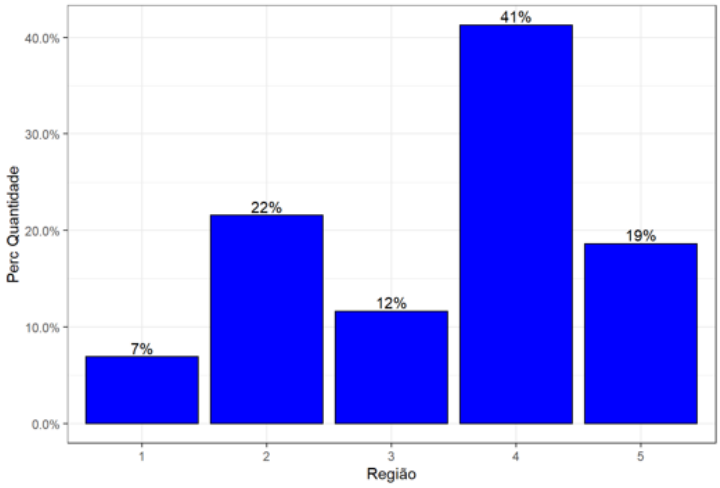

Figura 4.4: Proporção de clientes por região.

Neste estudo, as informações foram extraídas levando-se em consideração o estado da proposta de seguro, isto é, clientes que ainda constam como ativos no sistema, com estado 1, e os que cancelaram seus contratos e seguros, estado 3, o estado 2 refere-se a outro estado do contrato que não foi estudado neste trabalho. Quanto ao estado 3, esse cancelamento é configurado como *churn* de clientes. A Figura 4.5, apresenta esses dois grupos, tendo a maioria dos contratos como ativos na base de dados.

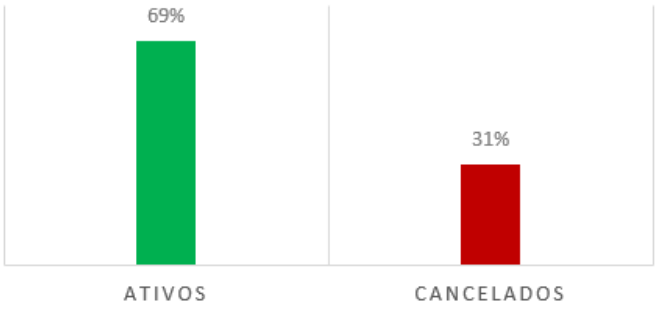

Figura 4.5: Proporção de contratos ativos e cancelados.

Foram excluídos os registros cujas variáveis Sexo, Região, e Renda do cliente, apresentaram valores ausentes.

4.3 Preparação dos Dados

Preparar os dados é uma das tarefas mais importantes no processo da mineração de dados, sendo trabalhosa e crucial para o sucesso do modelo [1].

Como exposto na Seção 3.2, os dados foram extraídos de três tabelas distintas de um *data warehouse* da instituição financeira brasileira, utilizando a linguagem de manipulação de dados SQL. Arquivos CSV foram utilizados para os processos de compreensão e preparação dos dados, visando à implementação do modelo.

Assim, para a realização desta etapa foram realizados os seguintes passos:

1. Limpeza e transformação dos dados.
2. Exploração e análise descritiva dos dados.
3. Escolha das variáveis e redução da dimensionalidade.

4.3.1 Limpeza e Transformação dos Dados

Como as bases estão estruturadas com poucos valores faltantes ou poucos dados discrepantes, classificados como *outliers*, esta etapa não foi trabalhosa. No entanto, teve-se cuidado no processo de limpeza e transformação dos dados, considerando o risco do não entendimento do negócio, ou distorção da representação das variáveis transformadas no contexto negocial.

No processo de limpeza foram identificados valores ausentes, valores negativos e discrepantes da média do atributo, nas seguintes variáveis: Sexo, Escolaridade, Região, Renda

Mensal, Estado civil e Natureza da ocupação. Por serem poucos registros, cerca de 120, optou-se por retirá-los da amostra utilizada para o estudo. Quanto aos valores negativos constantes na amostra, como a natureza do negócio não permite valores negativos para nenhum atributo, tais registros também foram retirados do estudo.

O somatório das colunas de todas as tabelas chega a 147 e o número de registros após o processo de limpeza totaliza 917.700. Desse total, 637.727 são relativos a contratos de seguros ativos e 279.973 estão cancelados, conforme percentuais mostrados na Figura 4.5. Cada registro na tabela configura um contrato de seguro de vida, ou seja, cada contrato refere-se a um cliente.

Após o processo de limpeza dos dados, foram feitas transformações de todos os atributos, conforme indicado nas Tabelas 4.3 e 4.4:

Tabela 4.3: Antes

Descrição das Variáveis	**Tipo**	**Domínio**
Estado civil	Inteira	1 a 12
Sexo	Inteira	1 e 2
Natureza da ocupação	Inteira	1 a 999
Escolaridade	Inteira	1 a 9
Quantidade de anos de relacionamento	Inteira	1 a 59
Quantidade de filhos	Inteira	1 a 25
Renda mensal	Decimal	1 a $+\infty$
Idade	Inteira	18 a 77
Valor de risco do contrato	Decimal	1 a $+\infty$
Valor de prêmio do contrato	Decimal	1 a $+\infty$
Canal de venda	Inteira	1 a 5

Tabela 4.4: Depois

Descrição das Variáveis	**Tipo**	**Domínio**
Estado civil	Categórica	1 a 4
Sexo	Categórica	1 e 2
Natureza da ocupação	Categórica	0 a 6
Escolaridade	Categórica	1 a 5
Faixa de anos de relacionamento	Categórica	1 a 8
Quantidade de filhos	Categórica	0 a 1
Faixa salarial	Categórica	1 a 5
Faixa etária	Categórica	1 a 9
Faixa de risco do contrato	Categórica	1 a 7
Faixa de prêmio do contrato	Categórica	1 a 8
Forma de venda	Categórica	1 a 3

As Tabelas 4.3 e 4.4, apresentam o antes e o depois das transformações das variáveis. Foram alterados tanto a descrição, quanto o tipo e o domínio. As variáveis foram categorizadas para melhor adequação aos algoritmos utilizados no estudo. Para efetuar a transformação, foram utilizadas a ferramenta Excel e R *Studio*com linguagem R.

Todos os atributos foram categorizados de acordo com a estrutura de negócio que melhor se adequou aos modelos utilizados neste estudo. A Tabela 4.5, apresenta todos os domínios das variáveis do conjunto de dados, com no mínimo duas e no máximo doze categorias.

Tabela 4.5: Domínios dos Atributos

Descrição das Variáveis	Quantidade	Descrição dos Domínios
Estado civil	4	1 - Solteiro(a) 2 - Casado(a) 3 - Divorciados/Separados 4 - Viúvo(a)
Sexo	2	1 - Masculino 2 - Feminino
Natureza da ocupação	7	0 - Outros 1 - Servidor Público Concursado 2 - Servidor Público Não Concursado 3 - Empregado Setor Privado 4 - Profissional Liberal 5 - Empresário 6 - Aposentado ou Pensionista
Escolaridade	5	1 - Ensino Fundamental 2 - Ensino Médio 3 - Superior Completo 4 - Pós-graduação/Mestre/Doutor 5 - Superior em andamento
Faixa de anos de relacionamento (anos)	8	1 - Até 1 2 - 2 a 5 3 - 6 a 7 4 - 8 a 10 5 - 11 a 15 6 - 16 a 20 7 - 21 a 25 8 - Acima de 25
Quantidade de filhos	2	0 - Sem Filhos 1 - Com Filhos (1 ou mais)
Faixa salarial (R$)	5	1 - ATÉ 0,5 SALÁRIO MINIMO (<= 520) 2 - 0,5 A 1 SALÁRIO MINIMO (521 - 1039) 3 - 1 A 1,5 SALARIO MINIMO (1040 - 1.559) 4 - 1,5 A 2 SALÁRIOS MINIMO (1.560 - 2.078) 5 - 2 A 3 SALÁRIOS MINIMO (2.079 - 3.117) 6 - 3 A 5 SALÁRIOS MINIMO (3.118 - 5.195) 7 - 5 A 7 SALÁRIOS MINIMO (5.196 - 7.273) 8 - 7 A 10 SALÁRIOS MINIMO (7.274 - 10.390) 9 - 10 A 13 SALÁRIOS MINIMO (10.391 - 13.507) 10 - 13 A 18 SALÁRIOS MINIMO (13.508 - 18.702) 11 - 18 A 25 SALÁRIOS MINIMO (18.703 - 25.975) 12 - ACIMA 23 SALÁRIOS MINIMO (>= 25.976)
Faixa etária (anos)	9	1 - Até 25 2 - 26 a 30 3 - 31 a 35 4 - 36 a 40 5 - 41 a 45 6 - 46 a 50 7 - 51 a 55 8 - 56 a 60 9 - Acima 60
Faixa de risco do contrato (R$)	7	1 - Até 35.000 2 - 35.001 a 40.000 3 - 40.001 a 60.000 4 - 60.001 a 100.000 5 - 100.001 a 150.000 6 - 150.001 a 250.000 7 - Acima de 250.000
Faixa de prêmio do contrato (R$)	8	1 - Até 200,00 2 - 200,01 a 400,00 3 - 400,01 a 600,00 4 - 600,01 a 800,00 5 - 800,01 a 1000,00 6 - 1000,01 a 2000,00 7 - 2000,01 a 4000,00 8 - Acima de 4000,00
Forma de venda	3	1 - Ofertado Presencial 2- Ofertado Virtual 3 - Espontâneo

Para o atributo *Forma de Pagamento*, também foi necessário efetuar transformação devido à diversidade de canais de vendas onde o banco comercializa seus produtos. Dessa forma, cinco canais de venda foram organizados para três tipos de comercialização ou forma de venda, conforme descrito na Tabela 4.5.

4.3.2 Exploração e Análise Descritiva dos Dados

Passada a fase de limpeza, o processo de exploração e análise descritiva dos dados configurou-se como uma sub-etapa do processo de preparação dos dados, sendo efetuado pelo uso do R *Studio* com a linguagem R e do editor *Jupyter Notebook* com a linguagem *Python*.

As Figuras 4.6 a 4.11 apresentam as análises descritivas das variáveis preditoras que foram utilizadas no estudo. Como descrito acima, esta análise foi desenvolvida com a linguagem R, utilizando as bibliotecas ggplot2, Hmisc, dplyr, ggalt e *scales*. Os gráficos estão descritos em percentuais das categorias que estão apresentadas na Tabela 4.5, de domínios das variáveis.

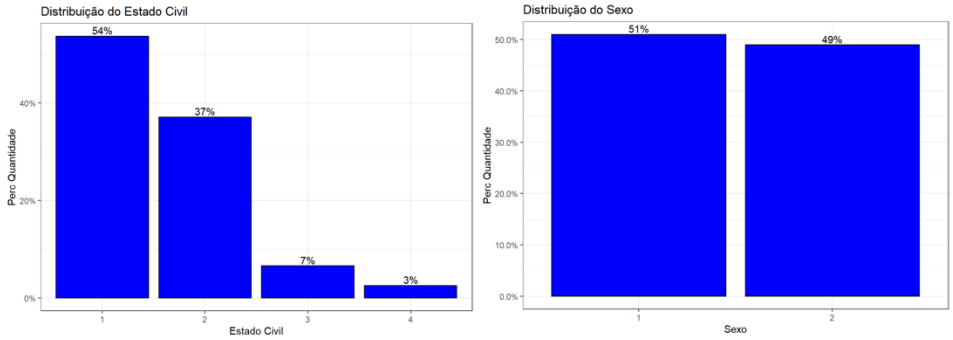

Figura 4.6: Distribuição das variáveis Estado Civil e Sexo.

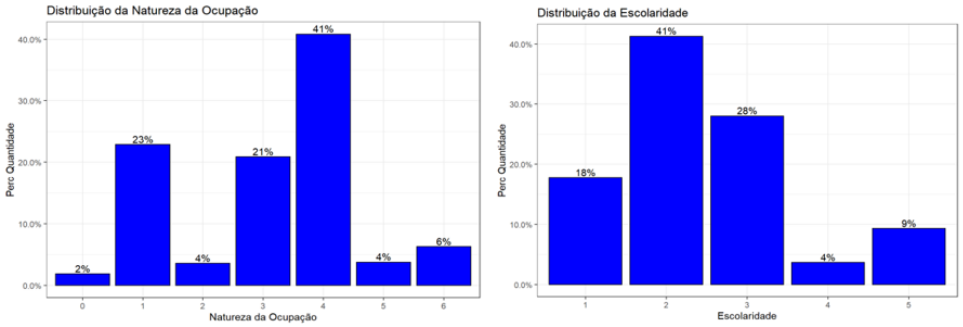

Figura 4.7: Distribuição das variáveis Natureza da Ocupação e Escolaridade.

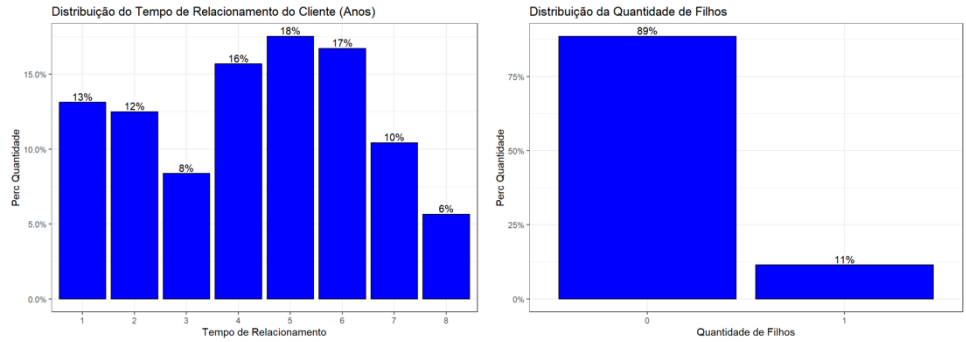

Figura 4.8: Distribuição das variáveis Tempo de Relacionamento e Quantidade de filhos.

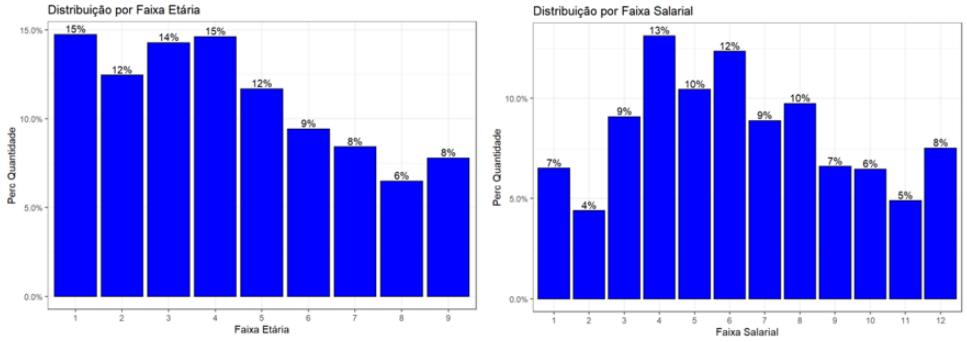

Figura 4.9: Distribuição das variáveis Faixa Etária e Faixa Salarial.

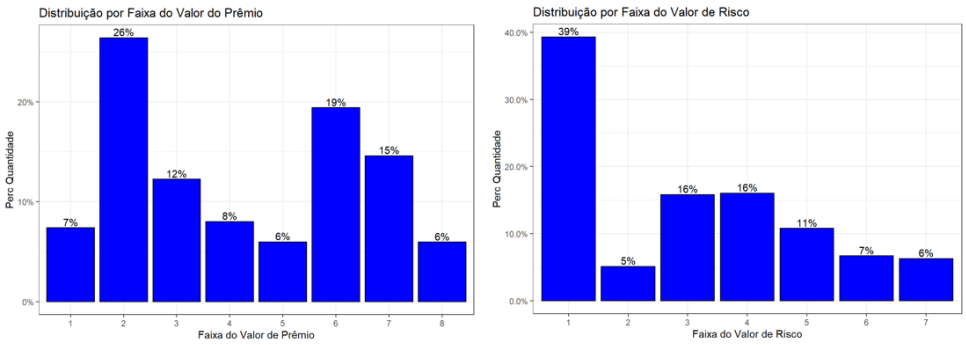

Figura 4.10: Distribuição das variáveis Faixa Valor de Prêmio e Faixa de Risco.

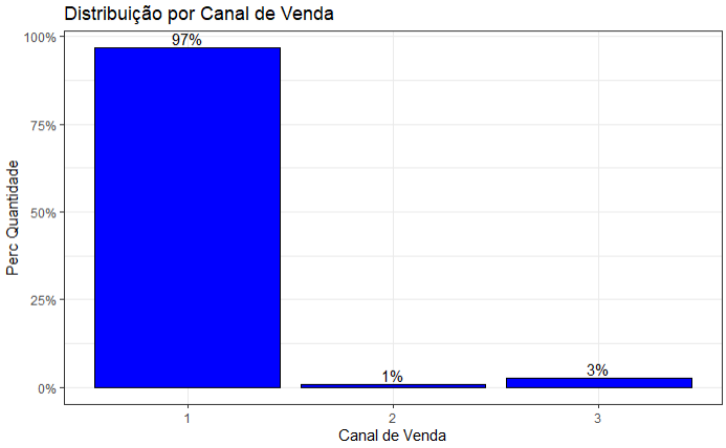

Figura 4.11: Distribuição das variáveis Canal de Venda.

4.3.3 Escolha das Variáveis e Redução da Dimensionalidade

A seleção das variáveis foi feita de acordo com a necessidade do estudo, colocando como cerne as orientações dos gestores do negócio, isto é, sua relevância foi baseada no contexto negocial e na importância de cada atributo em estudo. Portanto, seguiram-se os seguintes passos para a escolha das variáveis:

1. Reuniu-se com os gestores do produto, para entender a necessidade e as variáveis de interesses do negócio; nesse ponto foram elencadas as variáveis de maior interesse para a criação do estudo.

2. Após a escolha das variáveis, foi realizada reunião com os analistas técnicos, para definir as bases de dados para extração dos dados; nesse passo foram definidas as tabelas onde se encontram as variáveis, e os relacionamentos entre as tabelas para extração das variáveis.

3. Com os dados extraídos, num total de 147 variáveis, foi feito um processo de limpeza e extração apenas das variáveis escolhidas para entrada no estudo, e assim, gerado o *dataset* com as 11 variáveis estabelecidas pelos gestores do negócio, conforme apresentado na Figura 4.12.

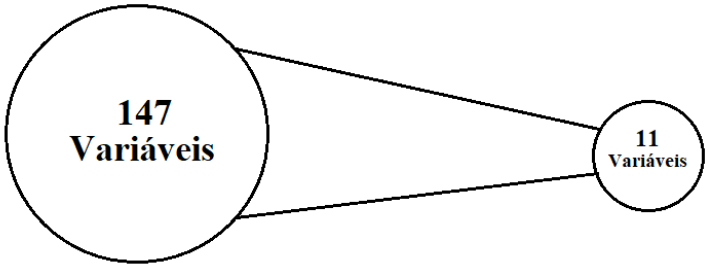

Figura 4.12: Redução da dimensionalidade.

Foram extraídas 147 variáveis das bases e reduzidas para 11 variáveis, a serem utilizadas no estudo. Com as variáveis preditoras selecionadas, para a implementação dos algoritmos de análise de sobrevivências e predição de *churn*, foi necessária a criação de dois campos a partir dos dados, que são a quantidade de meses de contrato e variável de interesse. A Figura 4.13, mostra como ficou a distribuição do tempo de contrato de seguros em faixas de meses.

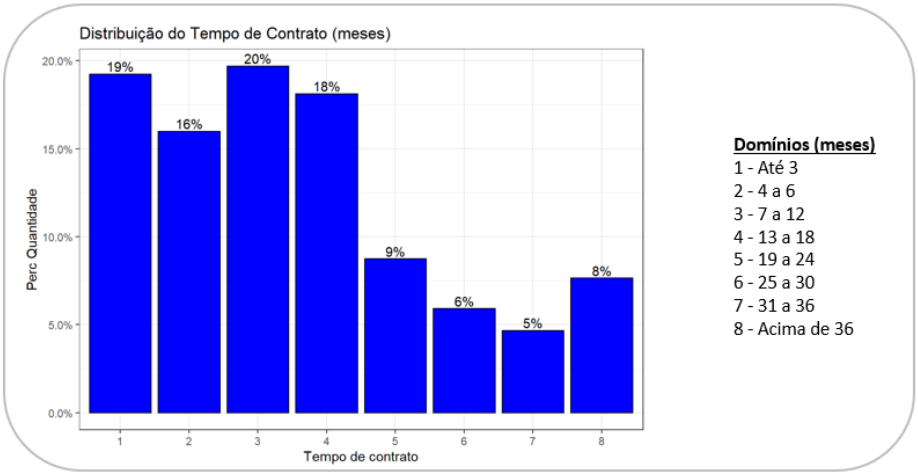

Figura 4.13: Distribuição e domínios da variável tempo de contrato de seguros.

A variável de interesse, ou *target*, foi rotulada como indicador de *churn* e transformada em variável binária, de acordo com estado da proposta de seguros. Para o estado de contrato ativo, a instituição financeira utiliza como código (01) e para contratos cancelados pelo cliente utiliza o (03). Após a transformação, esses novos valores ficaram como ativos (0) e cancelados (1), ajuste necessário para a implementação das técnicas utilizadas neste estudo.

Para treinamento do modelo, foi utilizada a Regressão de Cox nos dados categorizados, o que implicou na transformação de todas as variáveis preditoras em variáveis binárias, ou *dummies*.

4.4 Modelagem

O modelo proposto neste estudo é complementar em seus resultados e foi divido em duas partes:

1. **Análise de Sobrevivência**. Recorrendo ao estimador de Kaplan-Meier, nesta primeira parte utilizou-se apenas duas variáveis no algoritmo do estimador para gerar os resultados, que foram os atributos de tempo de contrato do cliente e indicador de *churn*.

2. **Previsão de *Churn***. Nesta parte, foi treinado o modelo com a Regressão de Cox, para gerar riscos proporcionais para cada categoria das variáveis, possibilitando efetuar previsão da rotatividade individual dos clientes.

4.4.1 Análise de Sobrevivência

Para o processo de Análise de Sobrevivência, foi utilizado o estimador de Kaplan-Meier, implementado na ferramenta *Jupyter Notebook*, com a linguagem *Python*, usando a biblioteca *Lifelines*, e importando o estimador a partir do método *KaplanMeierFitter*. Além da biblioteca específica do método, foram utilizadas as bibliotecas Pandas, para importar e manipular os dados, Numpy, para algumas transformações de atributos, e Matplotlib, para geração dos gráficos dos resultados. A aplicação desse método possibilitou traçar a curva de sobrevivência e a tabela de vida.

Para o processo de Análise de Sobrevivência foi utilizado o estimador de Kaplan-Meier, implementado na ferramenta *Jupyter Notebook*, com a linguagem *Python* usando a biblioteca *Lifelines* e importando o estimador a partir da *KaplanMeierFitter*. Além da biblioteca específica do método, foram utilizadas as bibliotecas Pandas, para importar e manipular os dados, Numpy para algumas transformações de atributos e Matplotlib para geração dos gráficos dos resultados. A aplicação desse método possibilitou traçar a curva de sobrevivência e a tabela de vida.

A Figura 4.14, apresenta o resultado da aplicação do estimador de Kaplan-Meier:

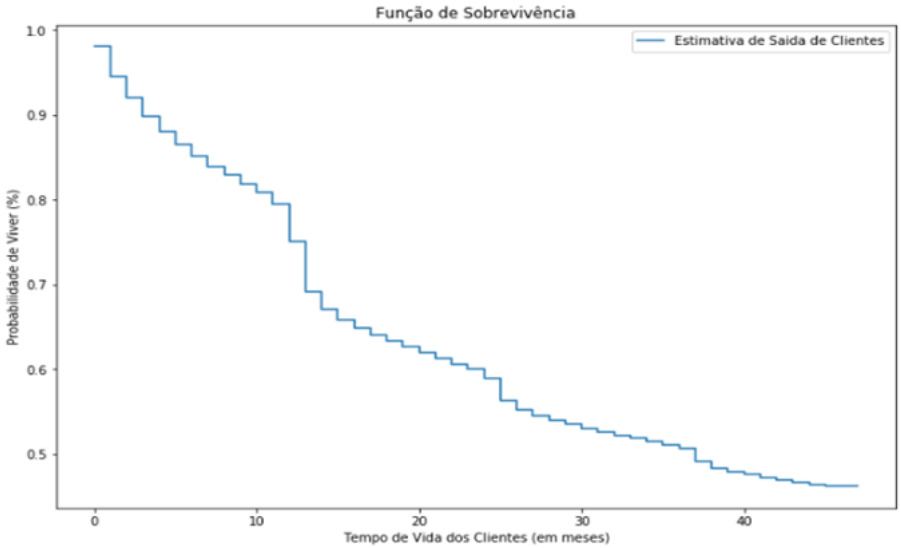

Figura 4.14: Curva de sobrevivência de Kaplan-Meier.

A Figura 4.14, apresenta o resultado da aplicação do estimador de Kaplan-Meier aos dados, como pode ser visto, o resultado mostra a função de sobrevivência de Kaplan-Meier em um gráfico conhecido como "Gráfico de Escada", onde cada degrau revela o número de ocorrência, ou seja, o número de falhas, enquanto o tamanho do degrau representa a proporcionalidade do número de ocorrências, de forma que, quanto maior o degrau, maior o número de falha no tempo (t) definido em meses. A escala horizontal, eixo X, representa o número de meses constantes no estudo, e o eixo Y representa a probabilidade de sobrevivência do indivíduo em percentuais.

Conforme o gráfico contido na Figura 4.14, a curva de vida de todos os clientes contratantes de seguros de vida obtidos, ao passar dos meses, a probabilidade de sobrevivência vai decaindo. Para se chegar no resultado apresentado no gráfico, o cálculo se baseia na quantidade de contratos que foram cancelados, relativamente aos que estão ativos na base, conforme explicado na seção 2.6.

Portanto, percebe-se que há uma queda acentuada no degrau nos meses 12 e 13, queda que se mantém constante até os meses 24 e 25, mais precisamente, e vai decaindo até ocorrer as quedas mais acentuadas nos meses 36 e 37. Com isso, pode-se concluir que há grandes cancelamentos de contratos quando ocorrem os aniversários dos contratos, ou seja, ao completar um ano de contrato, quando acontece a renovação, os clientes têm maior probabilidade de cancelar seus contratos de seguros de vida.

Com este resultado, podemos estimar as probabilidades de *churn* de clientes e fornecer *insights* para os gestores do produto, visando a melhor estratégia de retenção desses clientes que têm maiores probabilidades de cancelarem seus seguros.

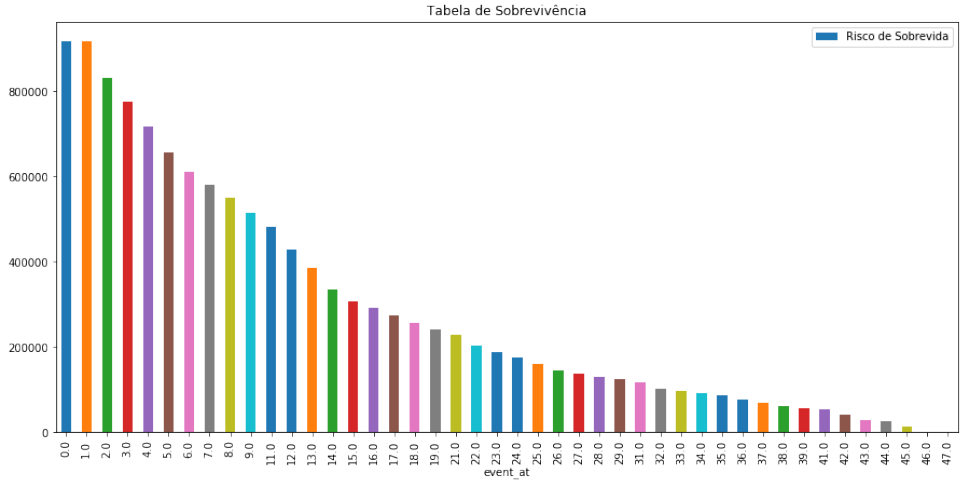

Figura 4.15: Gráfico da curva de sobrevivência.

A Figura 4.15, apresenta o gráfico da curva de sobrevivência de Kaplan-Meier, refere a uma outra visão da curva de vida apresentada na Figura 4.14, porém, esse gráfico permite entender melhor as probabilidades de cancelamentos de seguros, bem como os meses em que ocorrem maiores possibilidades da rotatividade.

event_at	removed	observed	censored	entrance	at_risk	probability
0.0	0	0	0	917700	917700	1.000000
1.0	85770	50377	35393	0	917700	1.000000
2.0	56095	21395	34700	0	831930	0.906538
3.0	58827	18735	40092	0	775835	0.845412
4.0	61971	14208	47763	0	717008	0.781310
5.0	44669	11368	33301	0	655037	0.713781
6.0	31300	9663	21637	0	610368	0.665106
7.0	29171	8226	20945	0	579068	0.630999
8.0	36535	6926	29609	0	549897	0.599212
9.0	30804	6188	24616	0	513362	0.559401
11.0	54237	13365	40872	0	482558	0.525834
12.0	43682	24344	19338	0	428321	0.466733
13.0	49913	30232	19681	0	384639	0.419134
14.0	27927	9920	18007	0	334726	0.364744
15.0	15795	5674	10121	0	306799	0.334313
16.0	18254	4114	14140	0	291004	0.317101
17.0	15853	3527	12326	0	272750	0.297210
18.0	15984	3042	12942	0	256897	0.279936
19.0	12056	2595	9461	0	240913	0.262518
21.0	26855	4602	22253	0	228857	0.249381
22.0	13792	2169	11623	0	202002	0.220118
23.0	12606	2094	10512	0	188210	0.205089
24.0	14800	3032	11768	0	175604	0.191352
25.0	16408	7172	9236	0	160804	0.175225
26.0	7394	2655	4739	0	144396	0.157346
27.0	6834	1698	5136	0	137002	0.149288
28.0	6293	1430	4863	0	130168	0.141842
29.0	6634	1124	5510	0	123875	0.134984
31.0	15319	1896	13423	0	117241	0.127755
32.0	6099	748	5351	0	101922	0.111062
33.0	3793	634	3159	0	95823	0.104416
34.0	4986	671	4315	0	92030	0.100283
35.0	11049	644	10405	0	87044	0.094850
36.0	7989	759	7230	0	75995	0.082810
37.0	6442	1992	4450	0	68006	0.074105
38.0	4638	930	3708	0	61564	0.067085
39.0	4293	516	3777	0	56926	0.062031
41.0	12259	712	11547	0	52633	0.057353
42.0	12500	229	12271	0	40374	0.043995
43.0	2990	195	2795	0	27874	0.030374
44.0	11986	125	11861	0	24884	0.027116
45.0	12744	47	12697	0	12898	0.014055
46.0	149	0	149	0	154	0.000168
47.0	5	0	5	0	5	0.000005

Figura 4.16: Tabela de Vida de KM.

A Figura 4.16 apresenta uma tabela com os resultado o estimador de KM, demonstrando os detalhes do resultado obtido aplicando a função de KM aos dados. A primeira coluna refere-se aos meses dos contratos, a segunda os indivíduos removidos do estudo, a terceira as ocorrências de falhas observadas, a quarta os contratos censurados, a quinta a quantidade de dados de entrada, a sexta os indivíduos com risco de falha, isto é, os clientes que estão exposto ao risco de cancelar seu contrato de seguro, e por último a coluna de probabilidade.

Percebe-se que a probabilidade de falhas aumenta com o passar do tempo, com a

probabilidade de sobrevivência diminuindo, de tal forma que os indivíduos que chegarem aos seis meses de contrato têm 66% de chance de continuar com seus contratos ativos e 34% de chance de cancelar seus seguros. Da mesma forma, com um ano de contrato há 46% de chance continuar com contratos ativos e, ao final de três anos, essa chance chega apenas a 0,08%. Dessa forma, ao final dos 47 meses do estudo, o cliente tem uma probabilidade irrisória de continuar com seu contrato de seguro ativo, levando ao cancelamento deste contrato.

O detalhamento dos resultados torna-se extremamente importante para entender a dinâmica dos eventos observados e da probabilidade de sobrevivência ao longo dos meses, fornecendo uma visão ampliada das ocorrências e possibilitando um maior entendimento dos resultados do estudo.

Uma outra forma de usar o estimador de Kaplan-Meier é a possibilidade de fazer comparações das curvas de sobrevivências das variáveis qualitativas, assim como comparar as categorias que os atributos possuem.

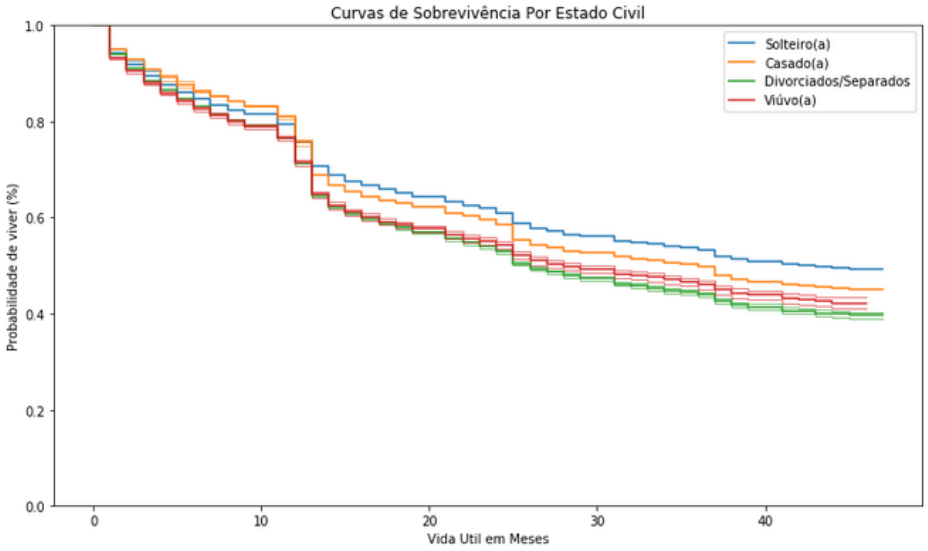

Figura 4.17: Curva de sobrevivência por Estado Civil

A Figura 4.17 mostra a curva de sobrevivência comparando as categorias do atributo Estado Civil, cujos clientes declarados como solteiros possuem maior probabilidade de continuar com seus contratos ativos, enquanto os divorciados e separados têm maiores probabilidade de cancelarem seus seguros de vida.

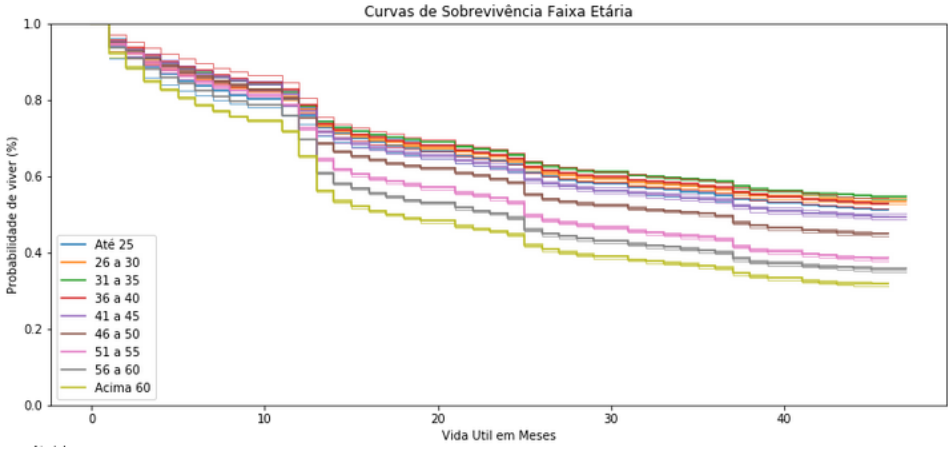

Figura 4.18: Curva de sobrevivência por Faixa Etária.

Comparando-se as curvas de vida por faixa etária, conforme Figura 4.18 percebe-se que os clientes na faixa de 31 a 35 anos têm maior probabilidade de manterem seus contratos ativos, enquanto os clientes nas faixas etárias acima de 60 anos têm maior probabilidade de cancelarem seus contratos. Isso mostra um contrassenso, pois os clientes dessa faixa etária seriam os que mais precisariam do seguro. Porém, como os valores dos seguros ficam mais caros à medida que o cliente envelhece, a tendência é de que os mesmos não mantenham seus seguros ativos.

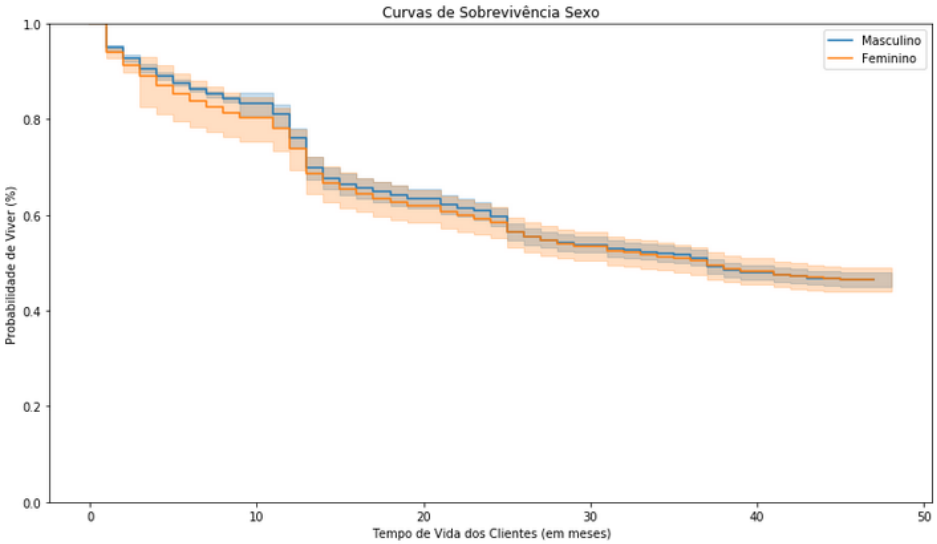

Figura 4.19: Curva de sobrevivência por Sexo.

Na comparação por sexo, Figura 4.19, não houve muita discrepância em relação ao gênero de cada cliente, visto que a curva se manteve praticamente igual para ambos os

sexos ao longo dos meses de estudo.

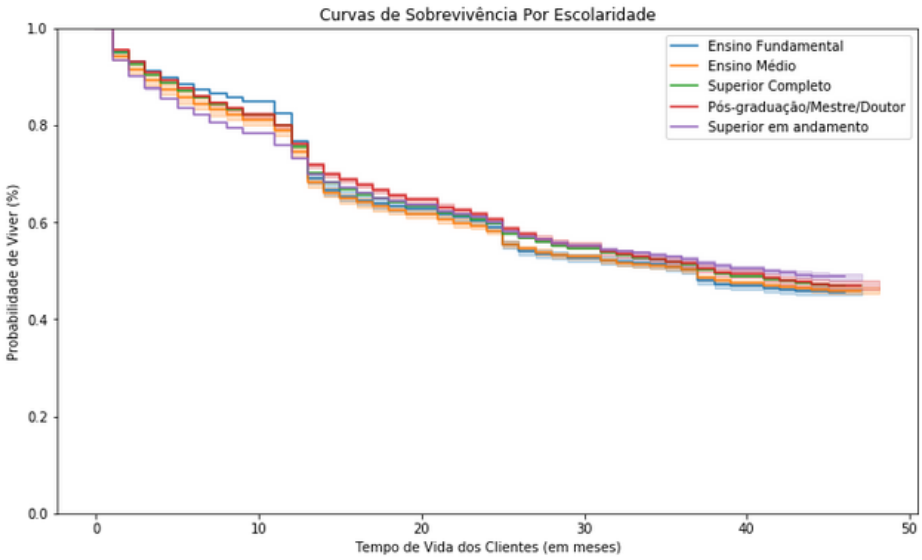

Figura 4.20: Curva de sobrevivência por Escolaridade.

Uma outra comparação importante foi relativa ao nível de Escolaridade, Figura 4.20, que apresentou pequenas divergências entre as curvas de vida das categorias. Foi observado uma maior probabilidade de clientes com nível de escolaridade superior manterem seus contratos de seguro de vida ativos, em relação aos clientes com apenas ensino fundamental.

4.4.2 Predição de Churn de Clientes

Para fazer a predição de *churn* dos clientes, foi utilizada a regressão múltipla de Cox, conforme descrito na Subseção 2.9. Refere-se a uma técnica flexível no processo de modelar a análise de sobrevivência e predição de falhas ao longo do tempo; consiste em predizer as taxas de falhas entre as covariáveis gerando os coeficientes: a razão de taxa de falhas ou risco relativo [27].

No processo de implementação da técnica foi necessária a preparação dos dados para possibilitar uma melhor convergência no treinamento do modelo, para isso seguiram-se os passos descritos na Subseção 3.3. Com isso, nossa base de dados do modelo utilizou 917.700 registros com 11 variáveis categóricas.

Antes da aplicação do modelo, procedeu-se ao tratamento dos dados para transformar as categorias em variáveis *dummy*, utilizando o método *get_dummies* da biblioteca Pandas do *python*. O ajuste deu-se para que a regressão se adequasse e houvesse convergência na descida do gradiente.

Para a implementação do modelo foi utilizado o editor *Jupyter Notebook* com a linguagem *python* versão 3.7 e biblioteca *lifelines* estendendo o método *CoxPHFitter*. O treinamento levou 5 iterações para conversão bem-sucedida, onde foi utilizada a variável MESES_CONTRATO, contendo a duração dos contratos em meses e a variável IND_SAIDA, com as indicações dos eventos de cancelamento ou não dos contratos de seguros.

Os resultados da aplicação da regressão de Cox estão apresentados na Figura 4.21, abaixo, que contém os coeficientes resultantes para cada característica do cliente, seguidos do exponencial do coeficiente, que corresponde ao risco proporcional de sobrevivência gerado; após, vem o erro padrão encima dos coeficientes, o valor z, o p *value*, e os intervalos de confiança inferior e superior para os coeficientes de risco. O exponencial do coeficiente mostrado na Figura 4.21, refere-se ao risco relativo, ou seja, as chances que um cliente tem de sobrevivência caso ele tenha determinada característica, como os clientes que estão na faixa etária 9, que possuem até 2,44 vezes mais chances de cancelar seu contrato de seguro quando comparado aos de outras faixas.

	coef	exp(coef)	se(coef)	z	p	log(p)	lower 0.95	upper 0.95	
CPF	-0.00	1.00	0.00	-3.26	<0.005	-6.80	-0.00	-0.00	*
ESTADO_CIVIL_2	-0.03	0.97	0.00	-6.89	<0.005	-25.88	-0.04	-0.02	***
ESTADO_CIVIL_3	0.03	1.03	0.01	3.91	<0.005	-9.31	0.02	0.05	***
ESTADO_CIVIL_4	-0.19	0.83	0.01	-15.34	<0.005	-120.68	-0.21	-0.16	***
SEXO_2	-0.05	0.95	0.00	-11.79	<0.005	-72.16	-0.06	-0.04	***
NATUREZA_OCUPACAO_2	-0.10	0.90	0.01	-8.30	<0.005	-36.77	-0.13	-0.08	***
NATUREZA_OCUPACAO_3	-0.06	0.94	0.01	-9.12	<0.005	-44.01	-0.08	-0.05	***
NATUREZA_OCUPACAO_4	-0.18	0.83	0.01	-30.19	<0.005	-459.25	-0.20	-0.17	***
NATUREZA_OCUPACAO_5	-0.29	0.75	0.01	-27.50	<0.005	-381.62	-0.32	-0.27	***
NATUREZA_OCUPACAO_6	-0.07	0.94	0.01	-7.82	<0.005	-32.90	-0.08	-0.05	***
NATUREZA_OCUPACAO_7	-0.05	0.96	0.01	-3.07	<0.005	-6.15	-0.07	-0.02	*
ESCOLARIDADE_2	0.05	1.05	0.01	8.59	<0.005	-39.30	0.04	0.06	***
ESCOLARIDADE_3	-0.06	0.94	0.01	-9.57	<0.005	-48.33	-0.08	-0.05	***
ESCOLARIDADE_4	-0.06	0.94	0.01	-5.26	<0.005	-15.73	-0.08	-0.04	***
ESCOLARIDADE_5	0.14	1.15	0.01	15.07	<0.005	-116.48	0.12	0.15	***
FX_ANOS_REL_2	0.26	1.30	0.01	27.20	<0.005	-373.58	0.24	0.28	***
FX_ANOS_REL_3	0.20	1.22	0.01	18.66	<0.005	-177.24	0.18	0.22	***
FX_ANOS_REL_4	0.09	1.10	0.01	9.08	<0.005	-43.65	0.07	0.11	***
FX_ANOS_REL_5	0.05	1.05	0.01	4.48	<0.005	-11.81	0.03	0.07	***
FX_ANOS_REL_6	-0.01	0.99	0.01	-1.26	0.21	-1.57	-0.03	0.01	
FX_ANOS_REL_7	-0.07	0.94	0.01	-5.83	<0.005	-19.01	-0.09	-0.04	***
FX_ANOS_REL_8	-0.12	0.89	0.01	-9.69	<0.005	-49.46	-0.15	-0.10	***
FX_QUANT_FILHOS_2	-0.02	0.98	0.01	-3.06	<0.005	-6.11	-0.03	-0.01	*
FX_SALARIAL_2	0.01	1.01	0.01	0.78	0.43	-0.84	-0.02	0.04	
FX_SALARIAL_3	-0.02	0.98	0.01	-1.51	0.13	-2.03	-0.04	0.01	
FX_SALARIAL_4	-0.12	0.89	0.01	-11.87	<0.005	-73.19	-0.14	-0.10	***
FX_SALARIAL_5	-0.22	0.80	0.01	-20.45	<0.005	-212.30	-0.25	-0.20	***
FX_SALARIAL_6	-0.33	0.72	0.01	-30.99	<0.005	-483.91	-0.35	-0.31	***
FX_SALARIAL_7	-0.40	0.67	0.01	-36.32	<0.005	-663.46	-0.43	-0.38	***
FX_SALARIAL_8	-0.46	0.63	0.01	-42.26	<0.005	-inf	-0.48	-0.44	***
FX_SALARIAL_9	-0.56	0.57	0.01	-47.73	<0.005	-inf	-0.58	-0.54	***
FX_SALARIAL_10	-0.64	0.53	0.01	-54.03	<0.005	-inf	-0.66	-0.61	***
FX_SALARIAL_11	-0.68	0.51	0.01	-54.61	<0.005	-inf	-0.70	-0.65	***
FX_SALARIAL_12	-0.84	0.43	0.01	-72.09	<0.005	-inf	-0.86	-0.81	***
FX_ETARIA_2	-0.05	0.96	0.01	-4.79	<0.005	-13.32	-0.06	-0.03	***
FX_ETARIA_3	-0.04	0.96	0.01	-4.36	<0.005	-11.25	-0.06	-0.02	***
FX_ETARIA_4	0.03	1.03	0.01	2.78	0.01	-5.21	0.01	0.05	*
FX_ETARIA_5	0.18	1.19	0.01	14.48	<0.005	-107.70	0.15	0.20	***
FX_ETARIA_6	0.36	1.43	0.01	26.27	<0.005	-348.51	0.33	0.38	***
FX_ETARIA_7	0.56	1.76	0.02	36.61	<0.005	-673.86	0.53	0.60	***
FX_ETARIA_8	0.72	2.05	0.02	41.85	<0.005	-inf	0.68	0.75	***
FX_ETARIA_9	0.89	2.44	0.02	46.06	<0.005	-inf	0.85	0.93	***
FX_VALOR_RISCO_2	-0.05	0.95	0.01	-5.98	<0.005	-19.93	-0.07	-0.04	***
FX_VALOR_RISCO_3	-0.02	0.98	0.01	-3.74	<0.005	-8.61	-0.04	-0.01	**
FX_VALOR_RISCO_4	0.12	1.13	0.01	16.76	<0.005	-143.50	0.11	0.14	***
FX_VALOR_RISCO_5	0.28	1.33	0.01	31.01	<0.005	-484.42	0.26	0.30	***
FX_VALOR_RISCO_6	0.43	1.54	0.01	36.89	<0.005	-684.45	0.41	0.45	***
FX_VALOR_RISCO_7	0.68	1.97	0.01	47.01	<0.005	-inf	0.65	0.71	***
FX_VALOR_PREMIO_2	-0.05	0.95	0.01	-4.22	<0.005	-10.64	-0.07	-0.03	***
FX_VALOR_PREMIO_3	0.02	1.02	0.01	1.19	0.23	-1.45	-0.01	0.04	
FX_VALOR_PREMIO_4	0.00	1.00	0.02	0.05	0.96	-0.04	-0.03	0.03	
FX_VALOR_PREMIO_5	0.09	1.09	0.02	5.01	<0.005	-14.41	0.05	0.12	***
FX_VALOR_PREMIO_6	0.14	1.15	0.02	8.00	<0.005	-34.33	0.11	0.18	***
FX_VALOR_PREMIO_7	0.24	1.27	0.02	11.74	<0.005	-71.58	0.20	0.28	***
FX_VALOR_PREMIO_8	0.27	1.31	0.02	11.23	<0.005	-65.71	0.22	0.32	***
CANAL_VENDA_AGRUP_2	-0.31	0.73	0.03	-9.87	<0.005	-51.27	-0.38	-0.25	***
CANAL_VENDA_AGRUP_3	0.53	1.70	0.01	48.86	<0.005	-inf	0.51	0.55	***

```
---
Signif. codes: 0 '***' 0.0001 '**' 0.001 '*' 0.01 '.' 0.05 ' ' 1

Concordance = 0.61
Likelihood ratio test = 39951.10 on 57 df, log(p)=-inf
```

Figura 4.21: Resultado da aplicação da Regressão de Cox aos dados.

Uma medida importante do modelo, bastante sensível à censura, é o índice de concordância. Esta medida avalia um ponto chave em uma análise de sobrevivência, que é a precisão da classificação do tempo previsto, avaliando as classificações relativas dos tempos de eventos do sujeito no estudo. Ela é considerada uma generalização de AUC, outra função de perda comum [44].

Os resultados podem ser interpretados da seguinte forma [44]:

- Índice de concordância de 0.5, é o resultado esperado para previsões aleatórias;
- Com resultado de 1.0 é a concordância perfeita;
- Resultado 0.0, é considerada a anti-concordância perfeita; nestes casos é necessário multiplicar as previsões por -1 para obter 1.0.

Portanto, um modelo de sobrevivência normalmente obtém índice de concordância entre 0.55 e 0.75. Os resultados que este intervalo compreende são considerados muito bons para os padrões de modelos de previsões em análises de sobrevivência utilizando a regressão de Cox ajustada [44].

O presente estudo mostrou-se bastante eficiente nas previsões, chegando ao resultado de índice de concordância de 0.61, conforme Figura 4.21. Isso mostra o poder de precisão com que o modelo se ajusta aos dados, para fornecer os coeficientes dos riscos relativos para cada característica dos indivíduos entrantes no estudo.

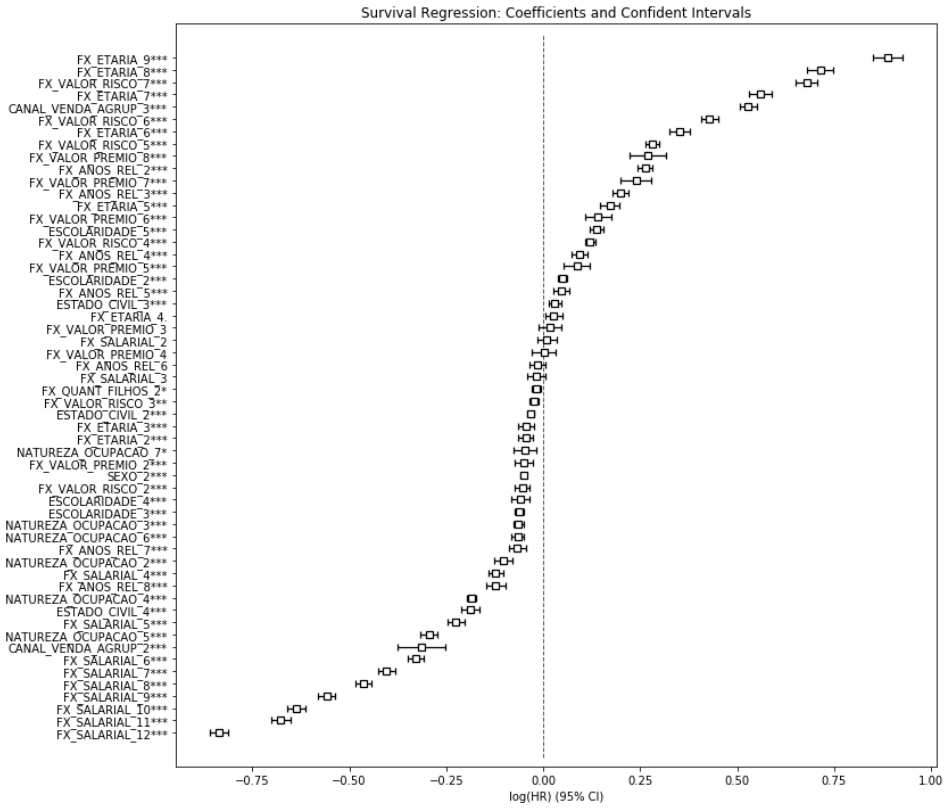

Figura 4.22: Intervalos de confiança dos coeficientes de riscos proporcionais de COX.

A Figura 4.22 apresenta o intervalo de confiança para os coeficientes relativos, mostrando a classificação desses coeficientes entre positivos e negativos na escala de log(HR), com precisão de 95% no intervalo de confiança.

Tabela 4.6: Estimativa dos Coeficientes de Regressão de Cox

Variáveis	**Categorias**	**Coeficiente(EP)**	**HR**	**IC (95%)**
Faixa Etária	Acima dos 60 Anos	0,89	2,44	(0,85 0,92)
Estado Civil	Divorcial/Separado	0,03	1,03	(0,02 0,05)
Sexo	Feminino	-0,05	0,95	-(0,06 0,04)
Natureza da Ocupação	Aposentado ou Pensionista	-0,05	0,95	-(0,07 0,02)
Escolaridade	Superior em Andamento	0,14	1,15	(0,12 0,15)
Faixa de Anos de Relacionamento	2 a 5 Anos	0,26	1,30	(0,24 0,28)
Faixa da Quant. de Filhos	Com Filhos	-0,02	0,98	-(0,03 0,01)
Faixa Salarial	0,5 a 1 Salário Mínimo	0,01	1,01	-(0,02 0,04)
Faixa Valor de Risco	Acima de 250 mil reais	0,68	1,97	(0,65 0,71)
Faixa Valor de Prêmio	Acima de 4 mil reais	0,27	1,31	(0,22 0,32)
Canal de venda	Canal Virutal	0,53	1,70	(0,51 0,55)

Os riscos proporcionais traçados e apresentados através dos coeficientes da regressão de Cox, nos permitem criar um perfil do cliente, reunindo as características que obtiveram as maiores probabilidades de cancelar seu contrato de seguro de vida. A Tabela 4.6 mostra as características de um cliente com maiores riscos proporcionais de acontecer o *churn*. Dessa forma, um cliente com idade acima dos 60 anos, divorciado ou separado, do sexo feminino, que seja aposentado, com nível de escolaridade superior em andamento, com pouco tempo de relacionamento – entre 2 a 5 anos, com filhos, com valor salarial menor que 1 salário mínimo, e que geralmente contrata seguros de vida com valores de prêmio acima de 4 mil reais e com valor de risco de vida de 250 mil reais nos canais virtuais, representa risco elevado de cancelar seus produtos.

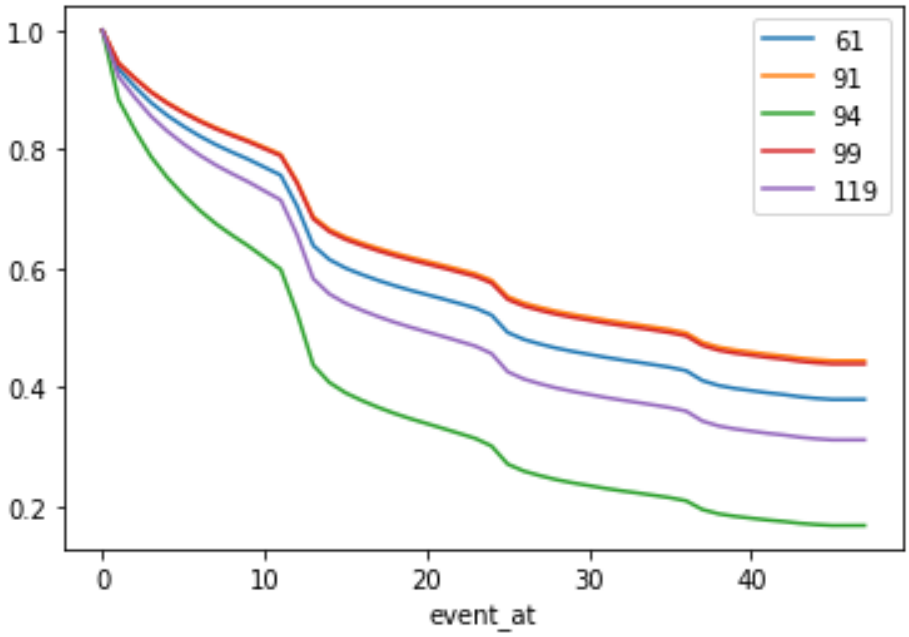

Figura 4.23: Análise individual de clientes contratantes de seguros de vida.

A regressão de Cox possibilita traçar curvas de sobrevivência de forma estratificada. Neste estudo foram selecionados, de forma aleatória, 5 clientes dentre os que tinham os maiores valores de contratos de seguros, para que seja traçada a curva de sobrevivência. A Figura 4.23 mostra uma análise de sobrevivência desses clientes, podemos perceber que o cliente 91 e o 99 possuem as menores probabilidades de cancelarem seus seguros de vida, enquanto com o cliente 94 as chances de ocorrer o evento de *churn* são maiores.

Um ponto importante nessa análise é a possibilidade de estratificação a nível de cliente, podendo gerar análises personalizadas por clientes, e assim podendo sugerir possíveis ações e campanhas para a retenção desses contratantes de seguros por parte dos gestores, representando importante insumo para o gerenciamento do *churn* dentro da instituição, gerando uma gama de opções de uso dessa informação no processo de retenção e fidelização desses clientes.

4.5 Avaliação

Esta seção é destinada ao processo de avaliação do modelo implementado no estudo. Dessa forma, esse passo é importante e essencial para garantir robustez e qualidade ao modelo, através da validação dos resultados.

4.5.1 Validação dos Resultados

Ao final do treinamento do modelo, a biblioteca *python* fornece uma medida muito importante: o IC, que se refere à pontuação de concordância, também conhecida como índice c do ajuste. O índice C é uma generalização do ROC AUC para dados de sobrevivência, incluindo censura. O IC é uma medida que mostra a precisão preditiva do modelo ajustado aos dados de treinamentos [44].

A Figura 4.21 mostra os resultados do treinamento do modelo com a regressão de Cox ajustado aos dados, e contém a medida IC que foi dada ao nosso modelo, que ficou com 0.61, representando alta precisão do modelo, conforme a documentação da biblioteca [44]. Assim, podemos dizer que nosso modelo teve uma excelente medida de precisão no processo de convergência dos dados.

Há outras formas de validar o desempenho e a precisão do modelo, como por exemplo, a técnica de validação cruzada (*cross-validation*), fornecida pela própria biblioteca *lifelines*. Contudo, tal técnica é mais utilizada para comparações de modelos e para estimativas modeladas em grupos de indivíduos, que não são de interesse deste estudo [44].

4.6 Implantação

O modelo desenvolvido neste estudo será disponibilizado para a divisão de retenção de clientes da empresa, e sua implantação seguirá as seguintes fases:

1. Implantação em ambiente de homologação, cujos resultados serão aprovados pelos gestores.

2. Implantação em ambiente de produção, com dados extraídos diariamente.

3. Validação das informações na produção pelos gestores do produto.

Os resultados serão apresentados em formato de gráficos e tabelas de forma sumarizada, traçando as curvas de sobrevivência geral e individual, no portal da empresa Alfa e na plataforma, visando fornecer informações a todos os funcionários.

Capítulo 5

Conclusões e Trabalhos Futuros

Este capítulo destina-se a apresentar as conclusões obtidas através dos resultados do modelo aplicado aos dados, bem como os próximos passos para o aprimoramento e aperfeiçoamento do modelo em trabalhos futuros.

5.1 Conclusões

A retenção e fidelização dos clientes são vitais para uma empresa. O estudo da rotatividade de clientes atualmente se torna estratégico e fundamental para a sobrevivência das instituições frente a enorme concorrência mundial.

O *churn* já é realidade de muitas empresas, principalmente instituições financeiras, que travam verdadeiras guerras nas mídias sociais em busca de reter ou conquistar novos clientes, para se manterem competitivas e como estratégia de sobrevivência, considerando que ao consumir seus produtos e serviços, o cliente se configura como o cerne da empresa.

O presente trabalho focou no estudo da rotatividade de clientes contratantes de seguros de vida de uma instituição financeira brasileira. Como resultado, foi proposto um modelo de análise capaz de gerar a curva de sobrevivência dos clientes, além de previsões das chances de cancelamento de seguros de vida, que leva à saída do cliente da base de dados da empresa.

Foram utilizados dados de clientes contratantes de seguros de vida em um período de 5 anos, tempo de vigência de um contrato de seguro. O modelo foi desenvolvido combinando os resultados do estimador de Kaplan-Meier com a Regressão múltipla de Cox, chegando a respostas bastante interessantes no processo de análise de sobrevivência e de predição de saída de clientes.

Para tanto, foi traçada a curva de sobrevivência de todos os clientes com o estimador de Kaplan- Meier, que fornece uma gama de informações aos gestores, possibilitando-lhes visualizar pontos de maior índice de cancelamentos de seguros e datas possíveis para

intervenção junto ao cliente na busca de retê-lo. Em seguida, foram traçadas curvas por características dos clientes, além da tabela de vida, consolidando os *insights* obtidos a partir da curva de sobrevida geral.

Para efetuar a previsão de *churn* foi utilizado a Regressão de Cox, onde treinou-se o modelo nos dados e obteve-se resultado satisfatório, chegando ao índice de concordância de 0.61. Esse resultado apresenta boa convergência do modelo aos dados, mostrando o poder de precisão da classificação dos tempos de sobrevivência, que o modelo obteve se ajustando aos dados do treinamento. Além disso, foram traçados os coeficientes, possibilitando a criação do perfil dos clientes com maior risco de cancelamento dos seus contratos de seguros. Por fim, foi obtida a curva de sobrevida por cliente, possibilitando um estudo estratificado e exclusivo para cada um.

A combinação dessas duas técnicas revelou-se interessante e relevante para o estudo da análise de sobrevivência e predição de *churn* para clientes contratantes de seguros de vida. Um dos principais pontos apresentado por esse modelo é a possibilidade de prever o tempo em que determinado cliente poderá cancelar seu contrato de seguro.

5.2 Trabalhos Futuros

Para trabalhos futuros vislumbra-se a possibilidade de buscar dados de pesquisas junto ao cliente e dados macroeconômicos para juntar aos dados constantes nesse estudo, podendo estratificar esses clientes por região, e assim ter uma visão mais clara quanto ao fenômeno de rotatividade no país.

Outro ponto a ser realizado é a implementação deste modelo antes e depois de uma campanha de *marketing* e retenção de clientes, para verificar sua eficácia e eficiência.

Referências

[1] Chapman, Pete, Randy Kerber, Tom Khabaza, Thomas Reinartz, Colin Shearer e Rüdiger Wirth: *CRISP-DM 1.0 step-by-step data mining guide*. janeiro 1999. x, 5, 6, 31

[2] Seguridade, Portal BB: *Apresentação de resultados 1t21*, 2021. `http://www.bbseguridaderi.com.br/`, [Online; accessed 02-maio-2021]. x, 3, 9, 10

[3] Seguros, Portal BB: *Quem somos*, 2021. `https://www.bbseguros.com.br/seguradora/seguros/`, [Online; accessed 21-maio-2021]. x, 9, 11

[4] Rich, Jason, John Neely, Randal Paniello, Courtney Voelker, Brian Nussenbaum e Eric Wang: *A practical guide to understanding kaplan-meir curves*. Otolaryngology–head and neck surgery : official journal of American Academy of Otolaryngology-Head and Neck Surgery, 143:331–6, setembro 2010. x, 15, 16, 17

[5] Sincor, Portal: *Ranking das seguradoras*, 2020. `https://www.sincor.org.br/wp-content/uploads/2021/05/ranking-das-seguradoras-2020.pdf`, [Online; accessed 02-julho-2021]. 1

[6] Hoffman, K. D., Bateson J. E. G: *Essentials of Services Marketing*. 1997. 2

[7] Burez, Jonathan e Dirk Van den Poel: *Handling class imbalance in customer churn prediction*. Expert Systems with Applications, 36:4626–4636, junho 2008. 2, 11

[8] Murphy, Emmett C., Murphy Mark A: *Leading on the Edge of Chaos: the 10 Critical Elements for Success in Volatile Times.* 2002. 2

[9] Reichheld, F.: *Prescription for cutting costs*, setembro 2001. `https://www.bain.com/contentassets/2598a2341fed40eba41954ee442ead22/bb_prescription_cutting_costs.pdf`, [Online; accessed 25-fevereiro-2021]. 2

[10] CNSeg, Portal: *Artigo crescimento seguro*, 2021. `http://midias.cnseg.org.br/data/files/DC/76/58/C4/DF89A61069CEB5A63A8AA8A8/_BR_%20FORBES%20BRASIL%20MAIO%202019%20-%20p64%20a%20p82.pdf`, [Online; accessed 02-julho-2021]. 2, 3, 4

[11] CNSeg, Portal: *Conjuntura cnseg nº 46*, 2021. `https://cnseg.org.br/data/files/AA/34/32/78/7F91A710F7E56E973A8AA8A8/Conjuntura%20046_edtorial_v4.pdf`, [Online; accessed 02-junho-2021]. 2, 3

[12] Wenjie, Bi, Meili Cai, Mengqi Liu e Guo Li: *A big data clustering algorithm for mitigating the risk of customer churn.* IEEE Transactions on Industrial Informatics, 12:1–1, junho 2016. 3

[13] Burez, Jonathan e Dirk Van den Poel: *Crm at a pay-tv company: Using analytical models to reduce customer attrition by targeted marketing for subscription services.* Expert Syst. Appl., 32:277–288, fevereiro 2007. 4, 9

[14] Susep, Portal: *Seguros de pessoas*, 2021. `http://www.susep.gov.br/menu/informacoes-ao-publico/planos-e-produtos/seguros/seguro-de-pessoas`, [Online; accessed 12-maio-2021]. 11

[15] Galetto, Molly: *What is customer churn?* 2016. 11, 12

[16] Fayyad, Usama, Gregory Piatetsky-Shapiro e Padhraic Smyth: *From data mining to knowledge discovery in databases.* AI Magazine, 17:37–54, março 1996. 12, 13

[17] Aggarwal, Charu C.: *Data Mining.* 2015, ISBN 978-3-319-14142-8. 13

[18] Kleibaum, David G., Klein Mitchel: *Survival Analisys: A SelfLearning Text.* Third edição, 2010, ISBN 978-3-319-14142-8. 13, 14, 15, 16, 17, 18, 19, 20

[19] Liu, Pei, Bo Fu, Simon Yang, Ling Deng, Xiaorong Zhong e Hong Zheng: *Optimizing survival analysis of xgboost for ties to predict disease progression of breast cancer.* IEEE Transactions on Biomedical Engineering, PP:1–1, maio 2020. 13

[20] Wijetilake, Navodini, Dulani Meedeniya, Charith Chitraranjan, Indika Perera e Mobarakol Islam: *Glioma survival analysis empowered with data engineering—a survey.* IEEE Access, 9:43168–43191, março 2021. 13

[21] Cawley, Gavin, Nicola Talbot, gj Janacek e Michael Peck: *Sparse bayesian kernel survival analysis for modeling the growth domain of microbial pathogens.* IEEE transactions on neural networks / a publication of the IEEE Neural Networks Council, 17:471–81, abril 2006. 13

[22] Kaplan, EL e P Meier: *Nonparametrics estimates for incomplete observations.* Journal of the American Statistical Association, 53:457–480, janeiro 1958. 15, 16

[23] Aalen, Odd: *Nonparametric inference for a family of counting processes.* Ann. Statist., 6, julho 1978. 18

[24] Nelson, Wayne: *Theory and applications of hazard plotting for censored failure data.* Technometrics, 42:12–25, fevereiro 2000. 18

[25] Oakes, David, N. Breslow e N. Day: *Statistical methods in cancer research, vol 1: The analysis of case- control studies.* Journal of the Royal Statistical Society. Series A (General), 145:264, janeiro 1982. 18

[26] Cox, David: *Regression models and life table.* Journal of the Royal Statistical Society. Series B, 34, janeiro 1972. 18, 19, 20, 21

[27] Jullum, Martin e Nils Hjort: *What price semiparametric cox regression?* Lifetime Data Analysis, 25, julho 2019. 18, 19, 20, 45

[28] Lee, Elisa T., Wang John Wenyu: *Statistical Methods for Survival Data Analysis.* Third edição, 2003, ISBN 978-0471458548. 19

[29] Prentice, R. L., Gloeckler L. A.: *Regression analysis of grouped survival data with application to breast cancer data.* Biometrics, páginas 57–67, março 1978. 21

[30] Lawless, Jerald: *Statistical Models and Methods for Lifetime Data: Lawless/Statistical.* novembro 2002, ISBN 9780471372158. 21

[31] [30] Colosimo, Enrico Antonio. Giolo, Suely Ruiz.: *Análise de Sobrevivência Aplicada.* First edição, 2006, ISBN 978-8521203841. 21

[32] [31] Breslow, N.E.: *Contribution to discussion of paper by d. r. cox.* Journal of the Royal Statistical Society, páginas 216–217, 1972. 21

[33] Azeem, Muhammad, Muhammad Usman e A. Fong: *A churn prediction model for prepaid customers in telecom using fuzzy classifiers.* Telecommunication Systems, 66, dezembro 2017. 22, 25

[34] Gaur, Abhishek e Ratnesh Dubey: *Predicting customer churn prediction in telecom sector using various machine learning techniques.* páginas 1–5, dezembro 2018. 23, 25

[35] Zhang, Rong, Weiping Li, Wei Tan e Tong Mo: *Deep and shallow model for insurance churn prediction service.* páginas 346–353, junho 2017. 23, 25

[36] Yiğit, İbrahim e Hamed Shourabizadeh: *An approach for predicting employee churn by using data mining.* setembro 2017. 23, 25

[37] Dulhare, Uma e Ifrah Ghori: *An efficient hybrid clustering to predict the risk of customer churn.* páginas 673–677, janeiro 2018. 23, 25

[38] Perianez, Africa, Alain Saas, Anna Guitart e Colin Magne: *Churn prediction in mobile social games: Towards a complete assessment using survival ensembles.* páginas 564–573, outubro 2016. 23, 25

[39] Abdelsalam, Nisreen, Murtada Elbashir e SaadEldeen SaadEldeen: *Applying cox regression in time to event data.* páginas 1–5, agosto 2018. 24, 26

[40] Lim, Tristan: *Applying survival analysis for customer retention: A u.s. regional mobile service operator.* páginas 338–342, maio 2020. 24, 26

[41] Tsai, Tien Yu, Chin Teng Lin e Mukesh Prasad: *An intelligent customer churn prediction and response framework.* páginas 928–935, novembro 2019. 25, 26

[42] Seguros, Portal BB: *Produtos e serviços.*, 2021. `"https://www.bb.com.br/pbb/pagina-inicial/voce/produtos-e-servicos/seguros#/"`, [Online; accessed 21-maio-2021]. 27

[43] Seguros, Portal BB: *Seguro de vida/condições.*, 2021. "https://www.bbseguros.com.br/seguradora/seguros/para-voce/seguro-vida/seguro-vida/condicoes.jsp", [Online; accessed 21-julho-2021]. 28

[44] Lifelines, Docs: *Lifelines*, 2021. "https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html", [Online; accessed 21-Outubro-2021]. 47, 48, 51