UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

ISABELE ALVES PEREIRA

APLICAÇÃO DAS TÉCNICAS DE ANÁLISE DE SOBREVIVÊNCIA NA ESTIMAÇÃO DO TEMPO DE *CHURN* EM *E-COMMERCES* BRASILEIROS

PRESIDENTE PRUDENTE

2023

ISABELE ALVES PEREIRA

APLICAÇÃO DAS TÉCNICAS DE ANÁLISE DE SOBREVIVÊNCIA NA ESTIMAÇÃO DO TEMPO DE *CHURN* EM *E-COMMERCES* BRASILEIROS

Relatório Final para Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina Trabalho de Conclusão de Curso.

Orientador: Prof. Dr. Mário Hissamitsu Tarumoto.

PRESIDENTE PRUDENTE

2023

P436a

Pereira, Isabele Alves
Aplicação das técnicas de análise de sobrevivência na estimação do tempo de churn em e-commerce brasileiros / Isabele Alves Pereira. -- Presidente Prudente, 2023
77 p. : il., tabs., mapas

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Mário Hissamitsu Tarumoto

1. Análise de sobrevivência. 2. Censura. 3. Churn. 4. E-commerce. 5. Modelo de regressão. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

ISABELE ALVES PEREIRA

APLICAÇÃO DAS TÉCNICAS DE ANÁLISE DE SOBREVIVÊNCIA NA ESTIMAÇÃO DO TEMPO DE *CHURN* EM *E-COMMERCES* BRASILEIROS

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____
Prof. Dr. Mário Hissamitsu Tarumoto.

_____
Prof. Dr. Fernando Antonio Moala.

_____
Prof. Dr. Manoel Ivanildo Silvestre Bezerra

Presidente Prudente, 5 de fevereiro de 2023.

AGRADECIMENTOS

Mais um ciclo se encerra na minha vida e outro que começa. Novos objetivos, novas metas e novos desafios vem pela frente. Em todos os caminhos que passei algumas pessoas fizeram de cada passo meu.

Agradeço a todos os meus professores que transmitiu parte de seus conhecimentos. Foram a partir dos seus recursos e ferramentas que consegui evoluir a cada dia.

Agradeço também ao professor Drº Mário Hissamitsu Tarumoto que com paciência me orientou e ensinou.

Em especial ao meu amigo João Pedro Martins que tive o prazer de conhecer na graduação e de longe é a pessoa mais inteligente e preciosa que tive na minha vida.

Aos meus pais que me deram força em toda a graduação, me dando força para não desistir, que ofereceram educação da melhor qualidade, conselhos que levarei para a vida toda. Os grandes amigos que fiz na graduação, Lais Gimenes, Beatriz Lourenção, Juliana Ricci, Amanda Vasconcelos, Estéfhanie Talacimon, Juan Ruiz, Victor Mendes, Matheus Chiarion, Victor Kassab e Henrique Marchetti que me ouviram reclamar.

A todos que de alguma forma me ajudaram e acreditaram em mim, agradeço imensamente.

RESUMO

Este estudo refere-se a análise preditiva de *churn* em *e-commerces* brasileiros sob a ótica da análise de sobrevivência e teve como objetivo principal, entender quando os clientes estarão com uma maior propensão de *churn* em *e-commerces* brasileiros usando análise de sobrevivência. Os dados utilizados neste estudo foram os dados da Olist, loja de departamentos do mercado brasileiro que conecta pequenos canais de vendas. O conjunto de dados são de comercio eletrônico brasileiro de pedidos feitos entre 2016 a 2018. Por meio das extrações e manipulações das informações foi possivel construir um modelo de analise de sobrevivência baseado no ticket médio do consumidor, preço médio do frete, se a compra é parcelada, tipo de pagamento, sua localização e a categoria do produto comprado, para prever quando os clientes estarão em risco de *churn*.

Palavras-chave: Análise de Sobrevivência, Censura, *Churn*, *E-commerce*. Modelos de Regressão.

Abstract

This study refers to the predictive analysis of churn in Brazilian e-commerces from the perspective of survival analysis and had as its main objective to understand when customers will be most likely to churn in Brazilian e-commerces using survival analysis. The data used in this study was data from Olist, a department store in the Brazilian market that connects small sales channels. The dataset is Brazilian e-commerce from orders placed between 2016 and 2018. Through information extraction and manipulation, we built a survival analysis model built based on the average consumer's ticket, average shipping price, whether the purchase is in installments, payment type, their location, and category of product purchased, to predict when customers will be at risk of churn.

Keywords: Survival Analysis, Censoring, Churn, E-commerce. Regression Models.

Sumário

1 Introdução

Observa-se que no cenário atual, que a sociedade está rodeada por empresas que competem pela oferta de serviços e produtos, em que na sua maioria os produtos são semelhantes ou iguais. Isso se dá pela abundância e facilidade ao acesso de informações de tais ofertas proporcionadas pelo ambiente digital. Dessa maneira, em 1990, nos Estados Unidos surgiu o *e-commerce,* em que dá abreviação do inglês, significa comércio virtual, ou seja, compras e vendas são feitas através da *internet*, com o surgimento de grandes companhias, como por exemplo a *amazon*, *alibaba* e *ebay*. No Brasil o *e-commerce* teve aparição em 1995 com a *Booknet*, loja de livros que foi comprada pela Submarino em 1999*.* Desse modo, o comércio eletrônico brasileiro tem pouco mais de duas décadas de idade. Desde então, o varejo digital brasileiro seguiu em constante crescimento. Assim, pode-se enfatizar que a esfera *on-line* de comércio é um segmento ágil e rápido, considerando que os consumidores têm o poder de escolha do melhor produto de acordo com os seus critérios, como o preço, qualidade, marca ou outros motivos. Consequentemente tais descrições podem fazê-lo buscar por outras companhias, aumentando bastante a competitividade entre elas.

Todavia, as companhias desempenham um esforço grande na criação de estratégias de *marketing,* tal como envios de *e-mails, sms (*serviços de mensagens curtas*)* e *pushs* (empurrar) com descontos e/ou produtos de afinidade para o consumidor. O intuito destas estratégias é manter um bom relacionamento, evitando assim um fenômeno chamado de *churn,* que é a desistência voluntária ou involuntária do consumidor da utilização de seus serviços ou compra de produtos. Por conseguinte, é necessária uma nova aquisição de usuários, uma vez que, existe a perda deles. Em geral, a aquisição de novos consumidores é muito mais cara do que a manutenção dos mesmos, diminuindo a receita da empresa.

Portanto, é imprescindível uma empresa ter uma boa interação com seus clientes. Além de otimizar a rentabilidade e aumento de vendas, o seu maior foco deve estar direcionado ao time de *customer success* (sucesso do cliente), no qual o seu principal objetivo é a gestão do sucesso do cliente, captação de novos

consumidores, retenção dos mesmos e até a identificação do risco de *churn*. Entretanto, torna-se cada vez mais caro para as companhias, o investimento na angariação de novos compradores, na qual o sucesso não é garantido, portanto, manter o cliente é essencial, evitando-se o *churn.*

A partir desta problemática, empresas tomam conscientização do *churning*, posto isto, as companhias podem usufruir deste fenômeno para retirar *insights e* assim reduzir a taxa de cancelamento. Deste modo, esta pesquisa teve o intuito de fazer um estudo aprofundado sobre este problema e desenvolver planos de ação junto com a área de *marketing* para evitar a rotatividade de consumidores e reter cada vez mais clientes.

Para a realização desta pesquisa, a base de dados que foi utilizada é de um sistema chamado Olist, que é uma loja de departamentos dentro de *marketplaces*, onde há diferentes lojas, como por exemplo, o mercado livre, americanas e entre outros, em que lojistas podem se inscrever, cadastrar os seus produtos e vendê-los nos grandes *e-commerces* brasileiros.

Nesta base de dados, o objetivo é tentar identificar o tempo entre a a última compra de cada cliente e a data de aquisição da base de dados, que foi o dia 24/07/2020, caracterizando uma base de dados identificando o tempo entre as compras.

Dessa forma, essa pesquisa terá como base, a utilização das técnicas de análise de sobrevivência para identificar quando os usuários estarão com uma maior propensão à *churn,* como um modo de predição de quando poderá deixar de ser cliente.

Na sequência, será apresentada algumas definições dos termos utilizados neste relatório.

1.1 E-commerce

O *e-commerce* significa lojas virtuais ou comercio eletrônico, isto é, toda compra e venda, são feitas por meio da *internet.* O advento do comercio virtual iniciou após o surgimento de grandes companhias, no qual algumas fazem

sucesso até o momento atual, como, *Amazon, Alibaba, eBay* entre outros. Essas empresas revolucionaram o conceito de compra em todo mundo e modificaram o comportamento do consumidor.

Novos modelos de negócios surgem: supermercados potencializam suas operações online e aplicativos de conveniência ganham espaço. Grandes marcas iniciam operações próprias e o setor ganha maior diversidade (Ebit|Nilsen, 2019, ed. 39).

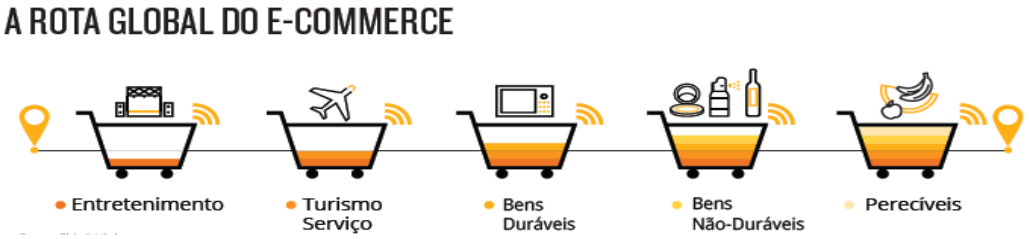
Destarte, com a expansão da virtualização dos meios de compra há uma corrida acirrada para que os estabelecimentos físicos se encaixem neste espaço.

1.1.1 *E-commerce* no Brasil

Sobredito, na década de 90, enquanto as lojas virtuais de outros países já estavam em constante crescimento, o Brasil ainda estava um pouco tímido de entrar neste meio eletrônico, sendo que uma das causas disso, pode ser o pouco desenvolvimento da *internet* naquela época. Todavia, após o aparecimento da *booknet,* loja online de livros, todo este cenário se configurou completamente. Em 1999 a Submarino adquiriu a *booknet,* devido a sua notoriedade, e o objetivo dos empresários era de expandir para outros países da América Latina, que em 2002 foi um grande exemplo para outras empresas, pois obteve equilíbrio entre receitas e despesas. Após isso, outros estabelecimentos foram adentrando ao comercio virtual, como Ponto Frio, Americanas, Casas Bahia e principalmente o Mercado Livre, tal que é um dos maiores *e-commerces* da América Latina.

A partir deste momento, o *e-commerce* brasileiro se consolidou neste cenário e tem se expandido a cada ano. Segundo o relatório da Nielsen (Ebit|Nilsen, 2019, ed. 39) o Brasil está em uma posição representativa na rota global do *e-commerce,* representado na figura 1*,* por meio dos bens não-duráveis, no entanto, tem grande espaço em outras categorias, como entretenimento, turismo e serviços duráveis.

Figura 1: A Rota Global do *e-commerce.*



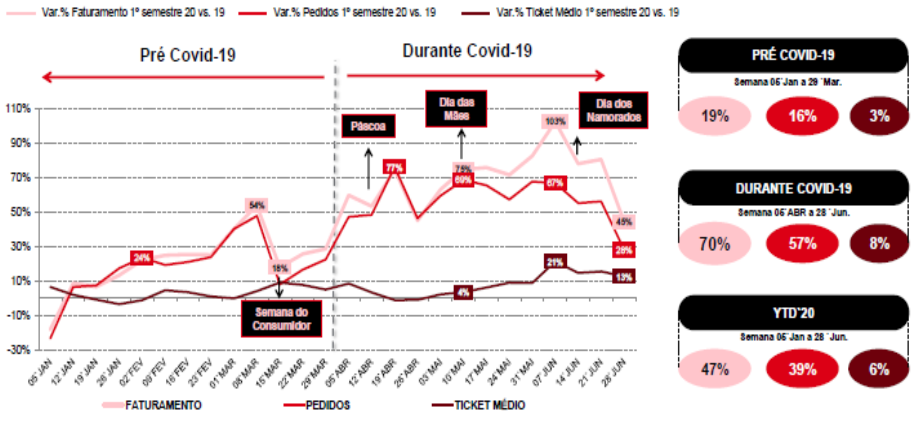
Fonte: Ebit|Nilsen Nilsen Webshoppers 39, 2019.

1.1.2 Impacto do Covid-19 no *e-commerce* brasileiro

Com o advento da pandemia do covid-19 e o aumento dos casos e óbitos foi necessário o fechamento de 75,2 mil lojas físicas em 2020. Com o avanço da crise sanitária do coronavírus impedindo a produtividade do comercio físico local, o comercio *on-line* ganha bastante espaço e adquire um crescimento impulsionado pela alta de pedidos.

Uma confirmação para essa expansão seria o crescimento de 47% no faturamento só no primeiro semestre do ano de 2020, com 40% a mais de consumidores em relação ao primeiro semestre do ano de 2019, isto é, 41 milhões de novos clientes e aproximadamente 90,8 milhões de pedidos

Figura 2: Crescimento a partir de abril durante a pandemia.

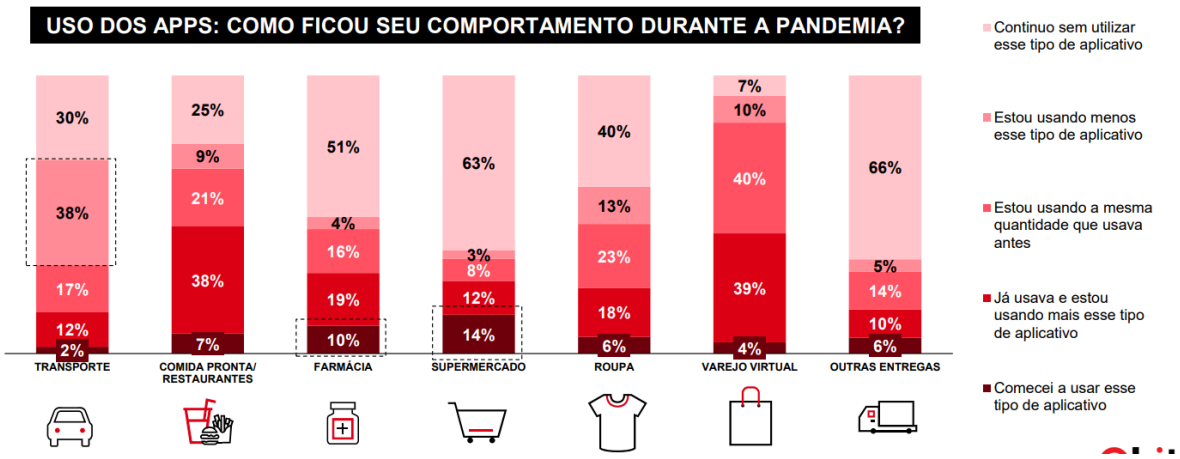


Fonte: Ebit | Nilsen Webshoppers 42, 2020.

Além disso, o comportamento do consumidor se modificou bastante ao buscar outro meio de compras devido as condições no momento de pandemia. Esse novo método de compra seria por aplicativos mobile, ou seja, pelo celular. De acordo com o relatório da Ebit | Nielsen 2020 (Ebit | Nielsen, 2020, ed 42), 72% dos consumidores brasileiros começaram a usar, ou estão usando mais, aplicativos de delivery durante a pandemia.

Como é demonstrado na figura 3, alguns estabelecimentos tradicionais, como farmácias e supermercados precisaram adaptar o seu comércio à tecnologia para atender as necessidades dos consumidores já que eles estavam limitados de sair de suas casas, pois houve um aumento de 10% de consumidores que passaram a usar aplicativos de farmácias e um crescimento de 14% dos consumidores que começaram a usar a tecnologia para fazer compras em supermercados.

Figura 3: Seguimentos que tiveram um aumento de novos compradores na pandemia.

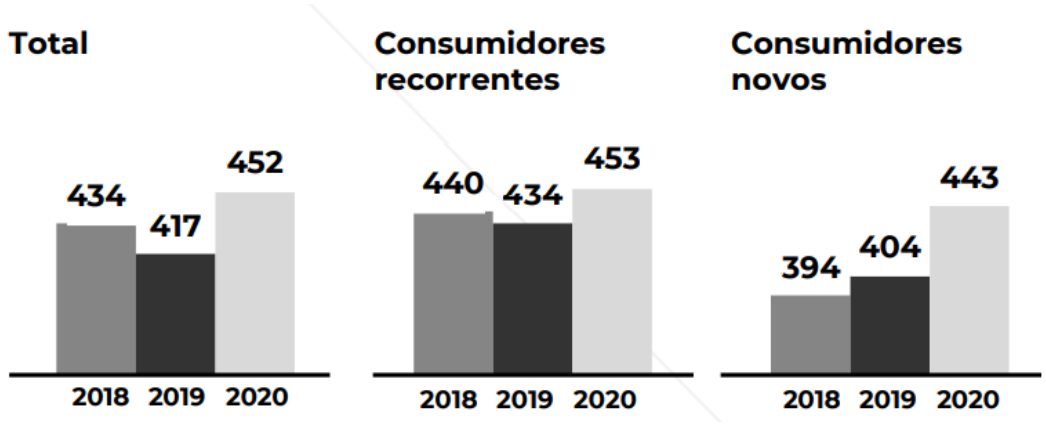


Fonte: Ebit | Nilsen Webshoppers 42, 2020.

Com isso, os comerciantes buscavam aprimorar os seus meios de vendas oferecendo para os consumidos diferentes benefícios diferentes meios de pagamentos, cupons de descontos, entregas dentro do prazo, variedades de produtos e entre outros.

O poder de compra dos consumidores novos e recorrentes também aumentaram.

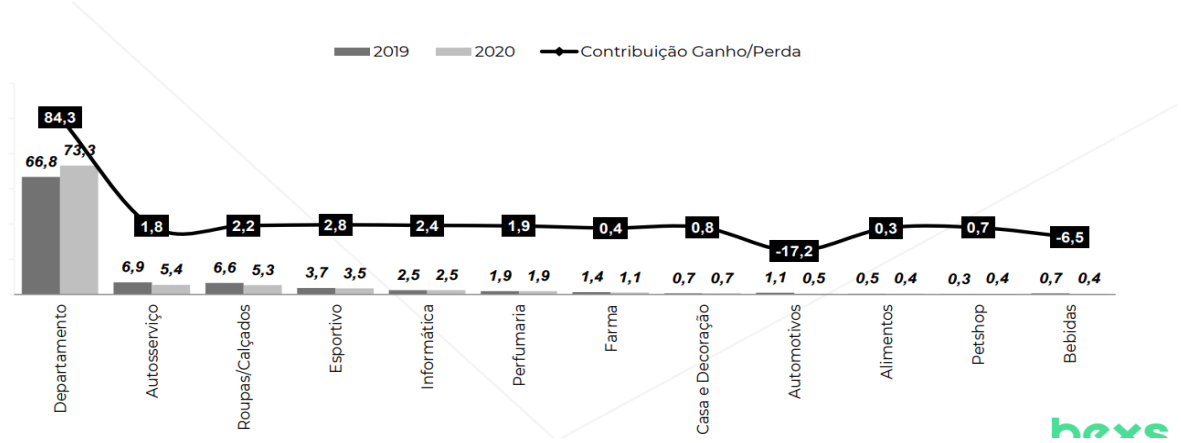
Figura 4: Evolução do ticket médio[1] R$ por pedido em 2020.



Fonte: Ebit | Nilsen Webshoppers 43, 2021.

Apesar de existirem vários segmentos no comércio virtual o que mais se destacou no cenário pandêmico de 2020 foram as lojas de departamento, como é mostrado na figura 5, que mostra a relevância do aumento do faturamento nas lojas de departamento e entres outros segmentos. Assim, o segmento das lojas de departamento obteve uma contribuição de 84,5% e uma importância de 73,3% no faturamento no ano de 2020.

Figura 5: Faturamento do *e-commerce* nos segmentos.



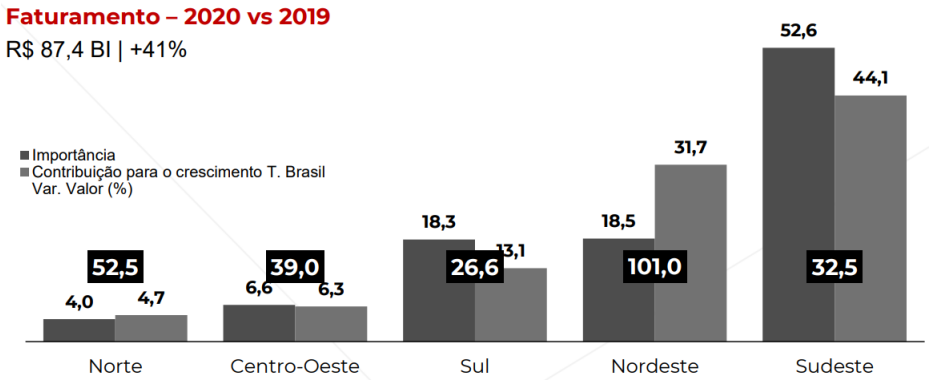
Fonte: Ebit | Nilsen Webshoppers 43, 2021.

Algo que é importante mencionar é que o consumidor leva em consideração que sua taxa de entrega seja com um preço baixo, dentro do prazo e sem problemas com extravio ou avaria.

[1] Ticket médio representa o valor médio de vendas por cliente.

Em relação ao faturamento regional, houve alta de 41%, R$ 87 bilhões, no ano de 2020 em relação ao ano de 2019 no faturamento em todas as regiões do Brasil, como pode ser verificado na Figura 6.

Figura 6: Crescimento do *e-commerce* nas regiões.



Fonte: Ebit | Nilsen Webshoppers 43, 2021.

A região Sudeste é a região que mais contribuiu para este crescimento, com 44,1% de contribuição para crescimento do faturamento de 2020 versus 2019, com 52,6% de importância nos números totais do País e com 32,5% de variação no faturamento de 2020 versus 2019.

Com a pandemia, 2020 fechou com 29% a mais de consumidores, sendo que 13 milhões são novos. Sendo assim, 83% desses novos consumidores mencionaram que voltariam a comprar.

Os consumidores não estão limitados apenas ao seu computador para fazer suas compras de forma virtual. Os mesmos podem utilizar o meio *mobile*, o seu aparelho celular, isso se denomina, *m - commerce.* O meio *mobile* oferece uma forma mais ágil para o consumidor e leva para o dono do *e-commerce* a missão de prestar facilidade para o comprador no momento da compra do produto.

Por fim, em comparação ao ano de 2020, que obteve um faturamento de 143,6 bilhões de reais (Ebit | Nielsen, 2021, ed 43), o *e-commerce* brasileiro no ano de 2021 cresceu 27%, tendo um fechamento de 182,7 bilhões de reais em vendas (Ebit | Nielsen, 2022, ed 45)

1.1.3 O que é um MarketPlace?

Marketplace são diversas lojas online em que é permitido a venda de diferentes produtos, semelhante a um shopping center, porém de modo virtual. Este tipo de plataforma possibilita a venda de produtos de forma econômica, rápida e fácil, tudo de forma digital.

Em uma pesquisa sobre *marketplaces* realizada pela Nielsen em setembro de 2019 mostrou que 32% dos consumidores declararam não saber o que é um *marketplace* (Ebit|Nilsen, 2020, ed. 42). Isso dificulta a entrada de novos varejos nesse meio digital.

Para aqueles varejistas que obtinham *marketplaces* no primeiro semestre de 2020 tiveram participação de 78% do faturamento do *e-commerce*. A quantidade de pedidos no primeiro semestre de 2020 foram 64 milhões, enquanto no primeiro semestre de 2019 foram 42 milhões de pedidos, uma variação de 52%. Além disso, o ticket médio do primeiro semestre de 2020 foi de R$466,00, sendo que no primeiro semestre de 2019 havia sido de R$455,00. A partir dessas informações, observa-se que o *e-commerce* com *marketplace* teve um faturamento de 30 bilhões de reais, um aumento de 56% em relação ao primeiro semestre de 2019.

1.1.4 *Customer-Churn*

Para que uma empresa cresça e seja referência no mercado, é necessário um modelo de negócio, planos traçados, diferencial dentre outras companhias e o mais importante, conhecer o cliente. Além de saber quem é o público-alvo, é preciso mantê-lo sempre consumindo o produto no qual é levado ao mercado.

Assim, o *customer churn* ou simplesmente *churning* é uma taxa em que indica o cancelamento de clientes em um certo período, isto é, a taxa denota quantos clientes cancelaram um tal serviço em um mês, bimestre, trimestre etc. Essa métrica tem uma grande importância, pois se um cliente cancela, o faturamento da empresa é afetado. Há diversos motivos do porquê o consumidor abandone uma empresa como, o cliente pode falir, o produto perdeu a qualidade,

o produto não tem mais valor para o cliente e tantas outras causas que precisam ser estudadas.

Salienta – se que, o principal objetivo deste estudo é identificar quando os usuários estarão com uma maior propensão à *churning,* utilizando as técnicas de análise de sobrevivência. Como mencionado anteriormente, a base de dados que será utilizada será de um sistema chamado Olist, onde há pedidos de 2016 a 2018, em que diz respeito à uma loja de departamentos dentro de *marketplaces*, onde há diferentes lojas como mercado livre, americanas e entre outros, em que lojistas podem se inscrever, cadastrar os seus produtos e vende-los nos grandes *e-commerces* brasileiros.

2 Fundamentação Teórica

2.1 Análise de Sobrevivência

A análise de sobrevivência é uma das áreas da estatística que mais cresceu nas últimas duas décadas do século passado. A razão desse crescimento foi o desenvolvimento e aprimoramento de técnicas estatística combinando com computadores cada vez mais velozes. (COLOSIMO; GIOLO, 2006, p. 1).

A área de análise de sobrevivência e confiabilidade tem uma grande importância tanto no meio clínico quanto no financeiro. O seu principal objetivo é avaliar a variável de interesse que no qual se caracteriza pelo tempo até a ocorrência de um evento, ou seja, o tempo de falha. Assim, para o estudo em foco, o propósito é estudar o tempo até o *churn* de clientes que compram em *e-commerces* brasileiros e identificar clientes que terão uma probabilidade maior de continuar comprando e aqueles que terão uma maior propensão ao *churn*. Todo o embasamento teórico foi extraído do livro de análise de sobrevivência aplicada, dos autores Enrico Antônio Colosimo e Suely Ruiz Giolo, 1ª edição, ano de 2006.

2.1.1 Tempo de Falha

O tempo de falha é caracterizado pelo acontecimento do evento de interesse, em que muitas das vezes são indesejáveis, como a morte de um paciente. O tempo de falha do estudo em questão é o tempo até o *churn* do consumidor. Os dados a serem utilizados, é o tempo da última compra e a data de aquisição da base de dados, portanto, o início do estudo é a data da última compra e o término para cada cliente, é a data de aquisição da base. Para este estudo, o início é a última compra do cliente e a falha é a data de obtenção dos dados. Desta forma, neste estudo, estamos estimando o tempo de não *churn*. Ademais, existem diferente escalas de medidas, para cada situação, como na engenharia, que é definido por ciclos, quilometragem ou outro tipo de circunstâncias. Neste caso, é o tempo em dias.

2.1.2 Censuras

Os dados de sobrevivência censurados consistem em observações incompletas e ou parciais. Os motivos de existir censuras em estudos de sobrevivência seria pelo não acontecimento do evento de interesse ou por outras razões. No entanto, mesmo que os dados sejam censurados eles devem ser utilizados na pesquisa estatística, pois com eles podem fornecer informações importantes e com a falta delas os resultados podem ser viciados. Diante disso, na pesquisa, as observações que serão censuradas dependerão do comportamento do tempo de permanência de cada cliente. Assim, foi definido como censura se o consumidor fez apenas uma compra em todo o período em que ele faz parte da base de dados.

Existem diferentes tipos de censuras em estudos clínicos. A censura tipo I é definido pelo encerramento do estudo após um período estabelecido, sendo que todos os indivíduos que não falharam, são considerados como censura. Por outro lado, a censura tipo II é definida após uma quantidade de falhas pré-definida ter acontecido. Há também a censura aleatória, que é descrita quando um indivíduo sai do estudo ou vai ao óbito por uma razão diferente da estudada. O dado de censura aleatória é representado por duas variáveis aleatórias, T, que

determina o tempo de falha e C, uma variável aleatória representando a censura, no qual, é independente de T. Assim, eles são observados como

$$t = \min(T, C) \tag{1}$$

e

$$\delta = \begin{cases} 1, T \leq C \\ 0, T > C \end{cases}. \tag{2}$$

Além disto, a censura à direita se denota quando o tempo do evento de interesse está à direita do tempo registrada, já a censura à esquerda, significa que o tempo registrado é maior que o tempo de falha, portanto, a falha já aconteceu. E por fim, censura intervalar, do tipo generalizado, o evento de interesse ocorre dentro de um intervalo de tempo. Como o tempo de falha T_i não é conhecido, logo $T_i \in (L_i, U_i]$, caso o evento não ocorra, é denominada de censura intervalar.

2.1.3 Representação dos Dados

Os dados de sobrevivência podem ser representados por (t_i, δ_i), no qual t_i é o tempo de falha ou de censura e δ_i indica falha ou censura, assim,

$$\delta_i = \begin{cases} 1 \; se \; t_i \text{ é } um \; tempo \; de \; falha \\ 0 \; se \; t_i \text{é } um \; tempo \; censurado \end{cases}, \; i = 1, \ldots, n \tag{3}$$

Se existirem covariáveis, como $\boldsymbol{X}_i = (sexo, idade, tratamento)$ no $i - ésimo$ indivíduo, os dados são denotados por $(t_i, \delta_i, \boldsymbol{X}_i)$, se os dados são do tipo intervalar, $(l_i, u_i, t_i, \delta_i, \boldsymbol{X}_i)$, tal que l_i é o limite inferior e u_i, limite superior.

2.1.4 Função de Sobrevivência

A função de sobrevivência é caracterizada pela probabilidade de um indivíduo sobreviver até um certo tempo t, ou seja, a probabilidade de um consumidor permanecer comprando até um tempo t. É descrito por,

$$S(t) = P(T \geq t). \quad (4)$$

Logo, a probabilidade de uma observação não sobreviver até o tempo t será descrita como

$$F(t) = 1 - S(t). \quad (5)$$

2.1.5 Função de Taxa de Falha ou de Risco

A taxa de falha é definida pela probabilidade de uma falha, ou *churn*, ocorrer em um certo intervalo de tempo $[t_1, t_2)$. Sua função de sobrevivência é dada como

$$S(t_1) - S(t_2). \quad (6)$$

Portanto, a taxa de falha em um intervalo $[t_1, t_2)$ também pode ser expressa através da probabilidade de que falha ocorra neste intervalo, dado que não ocorreu antes t_1, dividido pelo comprimento do intervalo. Assim sendo,

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}. \quad (7)$$

Generalizando a expressão (7), o intervalo é redefinido com $[t, t + \Delta t)$ tem uma nova forma

$$\lambda(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}. \quad (8)$$

Devemos assumir que Δt deve ser bem pequeno, e $\lambda(t)$ se caracteriza pela taxa de falha no tempo t condicional à sobrevivência até o tempo t. A função de taxa de falha $\lambda(t)$ pode ser definida de outra forma

$$\lambda(t) = lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \quad (9)$$

2.1.6 Função de Taxa de Falha Acumulada

A função de risco acumulada nos dá a taxa de falha acumulada de um indivíduo. A mesma, é definida como

$$\Lambda(t) = \int_0^t \lambda(u)\mathrm{du}. \tag{10}$$

Esta função não tem um modo de interpretação, porém é utilizada para avaliação da função da taxa de falha $\lambda(t)$.

2.1.7 Tempo Médio e Vida Média Residual

Duas quantidades básicas que são importantes, é o tempo médio de vida e a vida média residual. O tempo médio de vida é definida pela área sob a função de sobrevivência,

$$t_m = \int_0^\infty S(t)dt. \tag{11}$$

Em seguida, a vida média residual, o tempo médio de vida restante de um paciente, que é então, a área sob a curva de sobrevivência à direita do tempo t divida por $S(t)$,

$$vmr(t) = \frac{\int_t^\infty (u-t)f(u)du}{S(t)} = \frac{\int_t^\infty S(u)du}{S(t)}, \tag{12}$$

sendo que $f(.)$ a função densidade de T.

2.1.8 Técnicas Não-Paramétricas

As técnicas não-paramétricas assumem que não existe vestígios de que os dados de uma amostra tenham uma distribuição de probabilidade conhecida, portando eles são denominados como testes livres de distribuição. Uma técnica muito importante na análise de sobrevivência é o Kaplan-Meier que será mencionado posteriormente.

2.1.8.1 Estimador de Kaplan-Meier

O estimador não-paramétrico de Kaplan-Meier é muito utilizado em estudos clínicos, no qual serve para estimar a função de sobrevivência. É também chamada de estimador-produto limite. Desta maneira, para qualquer t, $S(t)$ pode ser expressa em termos de probabilidades condicionais. Tomando $S(t)$ como uma função discreta de probabilidade maior que zero apenas nos tempos de falha $t_j, j = 1, \dots, k$,

$$S(t_j) = (1 - q_j)(1 - q_2) \dots (1 - q_j), \tag{13}$$

tal que q_j é a probabilidade de um paciente morrer no intervalo $[t_{j-1}, t_j)$ sabendo que ele não morreu até t_{j-1}, considerando $t_0 = 0$, podemos denotar por,

$$q_j = P(T \in [t_{j-1}, t_j) | T \geq t_{j-1}), \tag{14}$$

assim o estimador de q_j para estimar $\hat{S}(t_j)$ em termos de probabilidade é expresso

$$\hat{q}_j = \frac{n^\circ \, de \, falhas \, em \, t_j}{n^\circ \, de \, observações \, sob \, risco \, em \, t_{j-1}}, j = 1, \dots k. \tag{15}$$

Para a estimação de uma função de sobrevivência conhecida como o estimador de Kaplan-Meier, algumas considerações devem ser feitas:

- $t_1 < t_2 < \cdots < t_k$, os k tempos distintos e ordenados de falha;
- d_j: Número de falhas em $t_j, j = 1, \dots, k$;
- n_j: Número de indivíduos sob risco em t_j, indivíduos que não falharam e não foram censuradas até o instante imediatamente anterior a t_j.

O estimador de Kaplan-Meier é definido como:

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j}\right). \tag{16}$$

As principais propriedades do estimador de Kaplan-Meier são:

- Não-viciado para amostras grandes;
- Consistente;
- Converge assintoticamente para um processo gaussiano;
- É estimador de máxima verossimilhança.

A variância assintótica de $\hat{S}(t)$ para construir intervalos de confiança e testar a hipótese é dada por:

$$\widehat{Var}\left(\hat{S}(t)\right) = \left[\hat{S}(t)\right]^2 \sum_{j:t_j<t} \frac{d_j}{n_j(n_j-d_j)}. \quad (17)$$

Por outro lado, o intervalo de confiança de $\hat{S}(t)$ para um t fixo, tem distribuição assintóticamente Normal, e portanto, o intervalo com $100(1-\alpha)\%$ de confiança para $S(t)$ é:

$$\hat{S}(t) \pm z_{\alpha/2}\sqrt{\widehat{Var}\left(\hat{S}(t)\right)}, \quad (18)$$

tal que $\alpha/2$ denota o $\alpha/2$-percentil superior da distribuição Normal. No entanto, quando existir valores extremos de t, os intervalos podem ter limite inferior negativo e limite superior maior do que 1, assim deve-se usar a transformação $\widehat{U}(t) = \log\left[-\log\left(\hat{S}(t)\right)\right]$ e sua variância assintótica é dada por

$$\widehat{Var}\left(\widehat{U}(t)\right) = \frac{\sum_{j:t_j<t} \frac{d_j}{n_j(n_j-d_j)}}{\left[\sum_{j:t_j<t} \log\left(\frac{n_j-d_j}{n_j}\right)\right]^2}, \quad (19)$$

Logo, o intervalo aproximado de $100(1-\alpha)\%$ de confiança para $S(t)$ é dado por:

$$\hat{S}(t)^{\exp\left\{\pm z_{\alpha/2}\sqrt{\widehat{Var}\left(\widehat{U}(t)\right)}\right\}}. \quad (20)$$

2.1.9 Estimação de Quantidades Básicas

A estimação de quantidades básicas se dá, por exemplo, pela estimação da probabilidade de um paciente sobreviver a t semanas de tratamento. Como a as probabilidades do estimador de Kaplan-Meier está ao longo dos degraus de sua curva, pode-se utilizar então, a interpolação linear para saber sua probabilidade. De modo semelhante, podemos utilizar desse artificio para obter outras estimativas, como os percentis. Para o cálculo do tempo médio de vida, em que é obtida pelo cálculo da área sob a curva de Kaplan-Meier, é obtida como a soma de áreas dos retângulos,

$$\hat{t}_m = t_1 \sum_{j=1}^{k-1} \hat{S}(t_j)(t_{j+1} - t_j),\ t_1 < \cdots < t_k, \quad (21)$$

são os k tempos distintos. Sua variância assintótica é,

$$\widehat{Var}(\hat{t}_m) = \frac{r}{r-1}\left[\sum_{j=1}^{r-1} \frac{(A_j)^2}{n_j(n_j - d_j)}\right], A_j = \hat{S}(t_j)(t_{j+1} - t_j) + \cdots + \hat{S}(t_j)(t_{r+1} - t_r), \quad (22)$$

e r é número de observações não censuradas, ou seja, r é o número de falhas.

E por fim, a vida média residual, no qual se dá pelo tempo médio restante de vida após um determinado tempo t,

$$vmr = \frac{\text{área sob a curva } S(t) \text{à direita de } t}{S(t)}. \quad (23)$$

2.1.10 Comparação de Curvas de Sobrevivência

Em situações em que vários tratamentos são realizados, para a verificação da existência de diferença entre elas, pode ser feita pela comparação de curvas de sobrevivência. Um procedimento usual para essa comparação, é a utilização do teste Log-Rank. Este teste consiste na diferença entre número observado de falhas de cada grupo e o seu número esperado de falhas, no qual é apenas apropriado quando as razões das funções de riscos, de cada grupo, são constantes.

Tabela 1: Tabela de Contingência gerada no tempo t_j

	1	2	
Falha	d_{1j}	d_{2j}	d_j
Não Falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	n_{1j}	n_{2j}	n_j

Fonte: (COLOSIMO; GIOLO, 2006, p. 43).

Pode-se denotar, por exemplo, a média e variância, respectivamente, de d_{2j}:

$$w_{2j} = n_{2j} d_j n_j^{-1}, \tag{24}$$

$$(V_j)_2 = n_{2j}(n_j - n_{2j}) d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}. \tag{25}$$

Portanto, se k tabelas forem independentes, o teste para comparação de dois grupos distintos, em que a hipótese testada $H_0: \hat{S}_1(t) = \hat{S}_2(t)$ versus $H_1: \hat{S}_1(t) \neq \hat{S}_2(t)$ é baseado na estatística,

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2}, \tag{26}$$

no qual tem distribuição qui-quadrado com 1 grau de liberdade para grandes amostras.

Se houver mais de dois grupos, $r > 2$, o teste de Log-Rank será generalizado para $\hat{S}_1(t), \ldots, \hat{S}_r(t)$. Assim, sua média e variância serão, respectivamente,

$$w_{ij} = n_{ij} d_j n_j^{-1}, \tag{27}$$

$$(V_j)_{ii} = -n_{ij} n_{lj} d_j (n_j - d_j) n_j^{-2} (n_j - 1)^{-1}. \tag{28}$$

Há também, outros testes como, de Wilcoxon e Tarone-Ware, onde deve-se apenas adicionar pesos u_i. Portanto, a generalização do teste T se dá,

$$S = \frac{\left[\sum_{j=1}^{k} u_j (d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k} u_j^2 (V_j)_2}. \tag{29}$$

Para o teste de Wilcoxon, o peso será $u_j = n_j$ e para Tarone-Ware, $u_j = \sqrt{n_j}$. Existe diferença entre os pesos de cada teste, como para o teste de Wilcoxon, no qual o mesmo direciona a diferença a ser encontrada, os pesos são

distribuídos igual ao número de indivíduos sob risco. Por outro lado, no teste de Log-Rank, os pesos são arranjados de forma uniforme. O teste de Tarone-Ware está entre o teste de Log-Rank e Wilcoxon.

2.1.11 Modelos Probabilísticos em Análise de Sobrevivência

2.1.11.1 Distribuição Exponencial

O modelo exponencial é a distribuição mais simples dentre as distribuições que modelam o tempo. Este modelo contém apenas um parâmetro e é caracterizada por uma função de taxa de risco constante. Sua aplicação se dá no meio industrial na modelagem o tempo de vida de materiais. A função densidade de probabilidade para a variável aleatória tempo de falha T é:

$$f(t) = \frac{1}{\theta}\exp\left\{-\left(\frac{t}{\theta}\right)\right\} \quad t \geq 0, \tag{30}$$

no qual o parâmetro $\theta \geq 0$ é o tempo médio de vida, a unidade de tempo de θ é o mesmo que o tempo de falha t.

Sua taxa de risco $\lambda(t)$ e função de sobrevivência $S(t)$ são dadas, respectivamente,

$$S(t) = \exp\left\{-\left(\frac{t}{\theta}\right)\right\}, \tag{31}$$

$$\lambda(t) = \frac{1}{\theta}, t \geq 0. \tag{32}$$

O seu percentil $100p\%$ pode ser obtido através da fórmula de t_p, que é dado por:

$$t_p = -\theta\log(1-p). \tag{33}$$

2.1.11.2 Distribuição Weibull

O modelo Weibull para o estudo do tempo de sobrevivência pode ser realizada nas mais diversas áreas, entre elas, na área biomédica ou na área industrial, especialmente na fadiga de materiais. A forma de sua função de taxa de falha é monótona, isto é, crescente, decrescente ou constante.

Seja T uma variável aleatória com distribuição Weibull, a função de densidade de probabilidade é:

$$f(t) = \frac{\gamma}{\theta^{\gamma}} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\theta}\right)^{\gamma}\right\}, t \geq 0. \quad (34)$$

em que γ é parâmetro de forma e θ é parâmetro de escala.

Para a distribuição Weibull, sua função de sobrevivência e risco, respectivamente são dadas por,

$$S(t) = \exp\left\{-\left(\frac{t}{\theta}\right)^{\gamma}\right\}, \quad (35)$$

$$\lambda(t) = \frac{\gamma}{\theta^{\gamma}} t^{\gamma-1}, \quad (36)$$

com t, θ e $\gamma \geq 0$. A distribuição exponencial é um caso particular da distribuição Weibull quando se tem $\gamma = 1$. A esperança e variância, respectivamente, da Weibull são,

$$E[T] = \theta\Gamma[1 + (1/\gamma)], \quad (37)$$

$$Var[T] = \theta^2[\Gamma[1 + (2/\gamma) - \Gamma[1 + (1/\gamma)]^2], \quad (38)$$

no qual $\Gamma(r) = (r-1)!$, para r inteiro. Para calcular os percentis t_p é denotado por

$$t_p = \theta[-\log(1-p)]^{1/\gamma}. \quad (39)$$

Uma função que tem relação com a distribuição Weibull é a função valor extremo ou Gumbel. A mesma é definida a partir do logaritmo da variável aleatória, $Y = \log(T)$. Sua função é denotada por,

$$f(y) = \frac{1}{\sigma} \exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}. \quad (40)$$

Os parâmetros apresentados são, μ, parâmetro de locação e σ, parâmetro de escala.

A função de sobrevivência e função de risco são,

$$S(y) = \exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}, \quad (41)$$

$$\lambda(y) = \frac{1}{\sigma}\exp\left\{\frac{y-\mu}{\sigma}\right\}. \quad (42)$$

A média e variância, respectivamente, da função valor extremo são expressas, tal que $v = 0{,}5772$ ... é conhecida como constante de Euler,

$$E[Y] = \mu - v\sigma, \text{ e} \quad (43)$$

$$Var[Y] = (\pi^2/6)\sigma^2. \quad (44)$$

Por fim, o percentil $100p\%$ é denotado por,

$$t_p = \mu + \sigma\log[-\log(1-p)]. \quad (45)$$

2.1.11.3 Distribuição Log-Normal

Como as distribuições já mencionadas, a distribuição log-normal também é utilizada para estudar o tempo de vida de matérias, pacientes e etc. A função densidade de probabilidade da variável aleatória T modelada a partir da log-normal é

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma}\exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\}, \ t \geq 0, \quad (46)$$

em μ é média do logaritmo de falha, como o desvio padrão σ.

A função de sobrevivência e taxa de falha não tem uma forma explicita, são escritas como

$$S(t) = \Phi\left(\frac{-\log(t)+\mu}{\sigma}\right), \quad (47)$$

no qual $\Phi(.)$ é a função de distribuição acumulada de uma normal padrão. A função risco é dada por:

$$\lambda(t) = \frac{f(t)}{S(t)}. \quad (48)$$

A média, variância e percentil $100p\%$ de $S(t)$ são, respectivamente, denotados por

$$E[T] = \exp\{\mu + \sigma^2/2\}, \quad (49)$$

$$Var[T] = \exp\{2\mu + \sigma^2\}(\exp\{\sigma^2\} - 1) \text{ e} \quad (50)$$

$$t_p = \exp\{z_p\sigma + \mu\}. \quad (51)$$

2.1.11.4 Distribuições Gama e Gama Generalizada

A distribuição gama é muito utilizada na área de confiabilidade, pois ela pode-se adequar a vários problemas. Os parâmetros de sua distribuição são k, no qual se caracteriza como parâmetro de forma, e α que é um parâmetro de escala. Sua função densidade de probabilidade é dada por,

$$f(t) = \frac{1}{\Gamma(k)\alpha^k} t^{k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, t > 0. \quad (52)$$

O seu pico acontece se $k > 1$, assim $t = (k-1)/\alpha$. Logo, sua função de sobrevivência e taxa de falha são

$$S(t) = \int_t^{\infty} \frac{1}{\Gamma(k)\alpha^k} u^{k-1} \exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du \text{ e} \quad (53)$$

$$\lambda(t) = \frac{f(t)}{S(t)}. \quad (54)$$

Assim, a sua média e variância são dadas, respectivamente, por

$$E[T] = k\alpha, \text{ e} \quad (55)$$

$$Var[T] = k\alpha^2. \quad (56)$$

Por outro lado, a distribuição gama generalizada, existe três parâmetrosγ, k e α, em que todos são positivos. Sua função é expressa por,

$$f(t) = \frac{\gamma}{\Gamma(\mathrm{k})\alpha^{\gamma k}} t^{\gamma^{k-1}} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, \quad (57)$$

em que $\Gamma(k)$ é a função gama, ou seja, $\Gamma(k) = \int_0^{\infty} x^{k-1} \exp\{-x\} dx$. A descrição dos parâmetros são, α, parâmetro de escala, γ e k são de forma.

Há diversas funções que são casos particulares da gama generalizada:

- Para $\gamma = k = 1$ tem-se $\mathrm{T} \sim \mathrm{Exp}(\alpha)$;
- Para $\mathrm{k} = 1$ tem-se $\mathrm{T} \sim \mathrm{Weibull}(\gamma, \alpha)$;
- Para $\gamma = 1$ tem-se $\mathrm{T} \sim \mathrm{Gama}(\mathrm{k}, \alpha)$.

2.1.12 Estimação de Parâmetros

2.1.12.1 Método da Máxima Verossimilhança

O método da máxima verossimilhança tem como o objetivo encontrar o melhor estimador que tenha a máxima informação dos dados. De forma generalizada, a função de verossimilhança para um θ é expressa por:

$$L(\theta) = \prod_{i=1}^{n} f(t_i, \theta). \tag{58}$$

Como o estudo será para dados contendo censura, a expressão (59) não é a ideal, assim, a função de verossimilhança para dados de sobrevivência se dá por

$$L(\theta) = \prod_{i=1}^{r} f(t_i, \theta) \prod_{i=r+1}^{n} S(t_i, \theta), \tag{59}$$

no qual, seria o mesmo que,

$$L(\theta) = \prod_{i=1}^{n} [f(t_i; \theta)]^{\delta_i} [S(t_i; \theta)]^{1-\delta_i}, \tag{60}$$

em que δ_i indica a existência de falha ou censura.

2.1.12.2 Precisão das Estimativas e Intervalos de Confiança

Para a construção de intervalos de confiança para as estimativas dos parâmetros, a normalidade assintótica dos estimadores de máxima verossimilhança para as grandes amostras é usada. A matriz de Informação de Fisher é dada por:

$$Var(\hat{\theta}) \approx -\left[E\left(\frac{\partial^2 log L(\theta)}{\partial \alpha^2}\right)\right]^{-1}, \tag{61}$$

ou seja, a matriz de variância-covariância. Para a construção do intervalo de confiança, utiliza-se o erro-padrão de $\hat{\theta}$, ou seja, $\sqrt{Var(\hat{\theta})}$. Desta forma, o intervalo aproximado de $(1-\alpha)100\%$ de confiança para θ é dado por

$$\hat{\theta} \pm z_{\alpha/2}\sqrt{\widehat{Var}(\hat{\theta})}. \quad (62)$$

Para uma distribuição multiparamétrica, onde há mais de um parâmetro, o método delta deve ser usado para encontrar a variância,

$$Var(\hat{\phi}) = Var(\hat{\alpha})\left(\frac{\partial\phi}{\partial\alpha}\right)^2 + 2cov(\hat{\alpha},\hat{\gamma})\left(\frac{\partial\phi}{\partial\alpha}\right)\left(\frac{\partial\phi}{\partial\gamma}\right) + Var(\hat{\gamma})\left(\frac{\partial\phi}{\partial\gamma}\right)^2. \quad (63)$$

2.1.12.3 Escolha do Melhor Modelo

A escolha do modelo paramétrico é a parte mais importante na análise paramétrica de dados de sobrevivência. É recomendável a utilização do método da máxima verossimilhança quando o modelo paramétrico se ajusta bem aos dados. A seguir será descrito algumas técnicas para seleção dos modelos.

2.1.12.4 Métodos Gráficos

O primeiro método se caracteriza pela comparação entre a função de sobrevivência dos modelos paramétricos e o estimador de Kaplan-Meier. Pode-se citar, como exemplo, $x = \hat{S}_{km(t)}$ versus $y = \hat{S}_{e(t)}$, $x = \hat{S}_{km(t)}$ versus $y = \hat{S}_{w(t)}$ e $x = \hat{S}_{km(t)}$ versus $y = \hat{S}_{ln(t)}$. O modelo adequado será aquele que se aproxima mais do estimador de Kaplan-Meier.

O segundo método consiste na linearização das funções de sobrevivência de cada modelo, a linearização para o modelo exponencial seria,

$$-\log[S(t)] = \frac{t}{\alpha}. \quad (64)$$

O gráfico com a linearização do modelo exponencial será $-\log[\hat{S}(t)]$ versus t, em que deve ser aproximadamente linear, onde $\hat{S}(t)$ é o estimador de Kaplan-Meier.

Para a distribuição Weibull de parâmetros (γ, α) a sua linearização será expressa por,

$$log[-\log[S(t)]] = -\gamma \log(\alpha) + \gamma \log(t), \quad (65)$$

a linearização $log\left[-\log\left[\hat{S}(t)\right]\right]$ versus $log(t)$, para $\hat{S}(t)$ sendo o estimador de Kaplan-Meier.

E para o modelo Log-Normal, sua forma linearizada será da forma,

$$\Phi^{-1}\left(S(t)\right) = \frac{-\log(t)+\mu}{\sigma}, \quad (66)$$

onde, $\Phi^{-1}(.)$ é percentil da distribuição normal. Logo, $\Phi^{-1}(\hat{S}(t))$ versus $\log(t)$ deve ser linear.

2.1.12.5 Teste da Razão de Verossimilhança

Os métodos gráficos mencionados anteriormente serão bastantes úteis para selecionar alguns modelos específicos a ser estudado. Há também outras formas de fazer a seleção do melhor modelo, usando alguns testes de hipóteses. Neste caso, utilizamos o teste da razão da verossimilhança com as hipóteses H_0: $O\ modelo$ é $adequado$ versus a hipótese alternativa, modelo não é adequado. Assim, a estatística da razão da verossimilhança é

$$TRV = -2\log\left[\frac{L(\hat{\theta}_M)}{L(\hat{\theta}_G)}\right] = 2\left[logL\left(\hat{\theta}_G\right) - logL\left(\hat{\theta}_M\right)\right] \quad (67)$$

Para este teste, o modelo generalizado e seu logaritmo de sua função de verossimilhança $L\left(\hat{\theta}_G\right)$ e $L\left(\hat{\theta}_M\right)$, modelo de interesse com logaritmo de sua função de verossimilhança deve ser usado. O teste da razão de verossimilhança tem uma distribuição aproximada qui-quadrado com a diferença entre o número de parâmetros de $\hat{\theta}_G$ e $\hat{\theta}_M$.

2.1.13 Modelos de Regressão em Análise de Sobrevivência

Em estudos que envolvem covariáveis, os modelos de regressão para dados censurados podem auxiliar nesta pesquisa. Há duas classes de modelos de regressão, paramétricos e semi - paramétricos. Os modelos paramétricos, podem ser chamados de modelo de tempo de vida acelerado, no qual as

variáveis explicam o tempo até a ocorrência de um evento, porém, são menos flexíveis do que os modelos semi - paramétricos. A segunda classe é caracterizada pelo modelo de regressão de Cox, em que, a sua principal vantagem é que ela pode ser modelada para diferentes situações e permite fácil interpretação dos parâmetros de interesse.

No caso dos modelos de regressão linear, a variável resposta é explicada pelas variáveis explicativas, no qual ambas as variáveis devem ter uma relação linear. Logo, o modelo é representado da seguinte forma,

$$Y = \beta_0 + \beta_1 x + \epsilon, \tag{68}$$

sendo que Y é a variável resposta, x é a variável explicativa, β_0 e $\beta_1$são os parâmetros estimados e por fim ϵ são os erros, com a suposição de que segue uma distribuição normal.

Para podermos aplicar o modelo de regressão nos dados de sobrevivência deve-se considerar estes pontos:

- Transformar a variável resposta;
- Componente sistemático não-linear nos parâmetros e uma distribuição assimétrica.

Considerando esses dois tópicos, a transformação logarítmica da variável resposta é dada por:

$$\exp\{\beta_0 + \beta_1 x\}. \tag{69}$$

2.1.13.1 Modelo de Regressão Exponencial

O modelo de regressão da distribuição exponencial é o mais simples e mais utilizado na análise de sobrevivência no qual,

$$T = \exp\{\beta_0 + \beta_1 x\}\epsilon \tag{70}$$

no qual $(f(\epsilon) = \exp\{-\epsilon\})$. O modelo tem uma relação não-linear entre T e a variável explicativa x. Com a linearização da expressão (78) obtém-se

$$Y = \log(T) = \beta_0 + \beta_1 x + v, em\ que\ v = \log(\epsilon). \tag{71}$$

Pode-se observar que o modelo de regressão exponencial é bem semelhante ao linear, no entanto os erros não seguem distribuição normal, mas segue a distribuição valor extremo,

$$f(v) = \exp\{v - \exp\{v\}\}. \tag{72}$$

A função de sobrevivência para o modelo de regressão exponencial de Y é expresso por,

$$S(y|x) = \exp\{-\exp\{y - (\beta_0 + \beta_1 x)\}\}, \tag{73}$$

e para $T = \exp\{Y\}$ dado x,

$$S(t|x) = \exp\left\{-\left(\frac{1}{\exp\{\beta_0 + \beta_1 x\}}\right)\right\} \tag{74}$$

Dessa forma, a estimação dos parâmetros dos modelos através do método da máxima verossimilhança para dados com censura,

$$L(\beta) = \prod_{i=1}^{n}[f(y_i, \beta|x_i)]^{\delta_i}[S(y_i, \beta|x_i)]^{(1-\delta_i)}, \tag{75}$$

como feito anteriormente para o modelo de regressão exponencial, para $y_i = \log(t_i)$ a expressão é

$$L(\beta) = \prod_{i=1}^{n}[f(t_i, \beta|x_i)]^{\delta_i}[S(t_i, \beta|x_i)]^{(1-\delta_i)}. \tag{76}$$

Para se obter os estimadores de máxima verossimilhança deve-se substituir a função densidade pela função de valor extremo (75) ou exponencial (76). Assim, aplicando o logaritmo na função (75) temos que,

$$l(\beta) = \sum_{i=1}^{n}[\delta_i(y_i - \beta_0 - \beta_1 x_i) - \exp\{y_i - \beta_0 - \beta_1 x_i\}]. \tag{77}$$

2.1.13.2 Modelo de Regressão Weibull

Para a construção do modelo de regressão Weibull em uma situação que contêm p variáveis, podemos escrever,

$$Y = \log(T) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \boldsymbol{X\beta} + \sigma v. \tag{78}$$

Assim, temos o modelo de regressão Weibull, se T tiver uma distribuição weibull com $\log(T)$ e σ, parâmetro de escala com distribuição valor extremo.
Sendo assim, sua função de sobrevivência para Y condicionado a X pode ser denotada por

$$S(y|x) = \exp\left\{-\exp\left\{\frac{y-\boldsymbol{X\beta}}{\sigma}\right\}\right\}, \tag{79}$$

para T,

$$S(t|x) = \exp\left\{-\left(\frac{t}{\exp\{\boldsymbol{X\beta}\}}\right)^{\frac{1}{\sigma}}\right\}. \tag{80}$$

2.1.14 Modelo de Tempo de Vida Acelerado

Os modelos de regressão exponencial e Weibull descritos na seção anterior, utiliza as variáveis como forma de acelerar ou desacelerar o tempo até a ocorrência do evento, dessa forma, o modelo é descrito por,

$$T = \exp\{\boldsymbol{X\beta}\}\exp\{\sigma v\}. \tag{81}$$

2.1.15 Adequação do Modelo Ajustado

Antes de finalizar a pesquisa é necessário saber se o modelo é adequado ou não. Dessa forma, existem técnicas gráficas com o uso dos resíduos para observar como os erros se comportam, como por exemplo, resíduo de Cox-Snell. No entanto, deve-se tomar cuidado ao escolher os modelos, pois os resíduos não são utilizados para mostrar qual é o modelo correto e sim, o mais conveniente para os dados.

2.1.15.1 Resíduos de Cox-Snell

Este resíduo auxilia na avaliação do modelo. A expressão generalizada do resíduo se dá,

$$\hat{e}_i = \widehat{\Lambda}(t_i|X_i). \tag{82}$$

Como mostrado anteriormente, $\widehat{\Lambda}(.)$ é a função de risco acumulada do modelo apropriado. Para cada distribuição mencionada, os resíduos de Cox-Snell serão,

$$Exponencial: \hat{e}_i = \left[t_i exp\{-\boldsymbol{X_i}\hat{\beta}\}\right], \tag{83}$$

$$Weibull: \hat{e}_i = [t_i exp\{-\boldsymbol{X_i}\hat{\beta}\}]^{\frac{1}{\hat{\sigma}}}, \tag{84}$$

$$log - normal: \hat{e}_i = -\log\left[1 - \Phi\left(\frac{\log(t_i) - X_i\widehat{\beta}}{\hat{\sigma}}\right)\right]. \tag{85}$$

Desta maneira, o gráfico que deve ser construído entre $\hat{e}_i$ versus $\widehat{\Lambda}(\hat{e}_i)$, em que $\widehat{\Lambda}(\hat{e}_i)$, é o risco acumulado dos erros e a reta deve ser de inclinação 1. Podemos também aproveitar a função de sobrevivência para avaliar os erros dos modelos, como $\hat{e}_i$ versus $-\log(\hat{S}(\hat{e}_i))$, desta forma, sua reta deve ter inclinação 1.

Contudo, este tipo de técnica de análise de resíduos não detecta o tipo de falha, outro, com esse mesmo objetivo seria o resíduo martingale.

2.1.15.2 Resíduos Padronizados

O resíduo padronizado segue o mesmo princípio que os resíduos Cox-Snell. A estatística dos erros padronizados é expressa por,

$$\hat{v}_i = \frac{(y_i - x_i\widehat{\beta})}{\hat{\sigma}}. \tag{86}$$

Se estes erros forem homocedástico, e estiverem próximos de uma reta, independentemente de seu modelo, o modelo é ideal aos dados.

2.1.15.3 Resíduos Martingale

Os resíduos mantingale estuda o número de falhas nos dados não predito no modelo. Sua expressão é,

$$\hat{m}_i = \delta_i - \hat{e}_i, \tag{87}$$

no qual δ_i indica censuras e $\hat{e}_i$ os erros.

2.1.15.4 Resíduos Deviance

O objetivo deste resíduo é de tornar o resíduo mantingale mais simples de detectar observações atípicas. Para detectar se os erros são aleatórios, a estimativas dos resíduos deviance versus o tempo deve ser construído, e espera-se que sejam aleatórios e em torno de zero. Portanto, o resíduo deviance é expresso por,

$$\hat{d}_i = sinal(\hat{m}_i)[-2(\hat{m}_i + \delta_i \log(\delta_i - \hat{m}_i))]^{1/2}. \quad (88)$$

3 Análise Exploratória de Dados

Antes de apresentar os resultados, é necessário mostrar como foi feita toda estruturação dos dados para utilizar nos modelos do estudo. Para organização dos dados foi necessário fazer a junção entre tabelas transacionais de pedidos, tabela de clientes e produtos. Nessa junção foram trazidas as informações do status do pedido, data do pedido, data da aprovação do pedido, data limite de envio do pedido, data da entrega na transportadora, data de entrega estimada para o cliente, data de entrega no cliente, informações de localização do cliente, categoria do produto comprado, valor do produto, valor do frete, tipo de pagamento, se o pagamento foi feito em parcelas, quantidade de parcelas e o valor do pagamento. Para o estudo, as variáveis que foram consideradas importantes foram Tempo de permanência, ticket médio, preço médio do frete, tipo de pagamento, se a compra foi parcelada, a categoria do produto e a quantidade de parcelas. A variável resposta de interesse, é o tempo de permanência, ou seja, o tempo entre a última compra e a data de aquisição dos dados. Foi considerado como censura, se o consumidor fez apenas uma compra no período que consta na base de dados.

Tabela 2: Variáveis escolhidas.

Variável	**Tipo da Variável**	**Exemplos**
Tempo	Quantitativa Contínua	331 Dias
Ticket Médio	Quantitativa Contínua	R$ 20,70
Preço Médio do Frete	Quantitativa Contínua	R$ 8,72
Tipo de Pagamento	Qualitativa Nominal	Cartão de Crédito
Parcelado	Qualitativa Nominal	Sim
Região	Qualitativa Nominal	Sudeste
Categoria	Qualitativa Nominal	Diversão

Fonte: Autoria Própria.

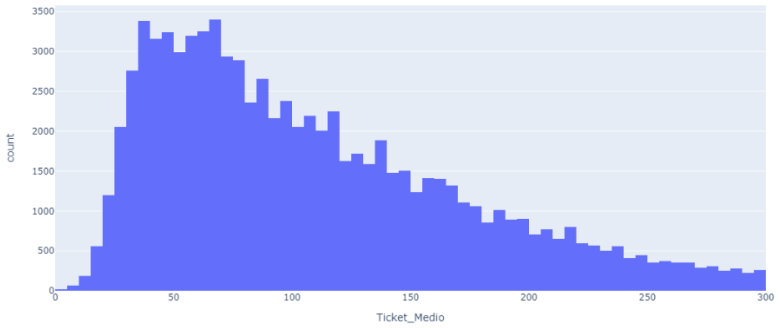
A variável ticket médio foi calculado pela média dos valores de pagamento das compras feito pelos clientes. Além disso, como havia muitas categorias de produtos foi resumido em apenas quatro subcategorias, casa, eletrônico, diversão e outros.

Por meio da análise exploratória de dados podemos identificar características nos dados que fornecem importantes informações, para que assim possamos aprofundar em uma análise mais completa.

Foi feito ainda a junção das tabelas transacionais para ter as informações completas de cada pedido, como, a data do pedido, data de aprovação, localização de onde foi feito o pedido, categoria do produto, preço, valor do frete, tipo de pagamento e entre outras. Nos dados, cada cliente tem uma identificação diferente para diferentes pedidos, logo um mesmo cliente pode aparecer mais de uma vez, assim, existe uma chave com uma identificação única para cada cliente. Com esta chave única de cada cliente é possível ter informações como o ticket médio de cada *customer.*

Observando a figura 7, o ticket médio se concentra entre 50 e 200 reais, a partir deste valor, o ticket médio vai decaindo. Além disso, é possível observar a distribuição do ticket médio pelo Brasil.

Figura 7: Histograma do Ticket Médio.



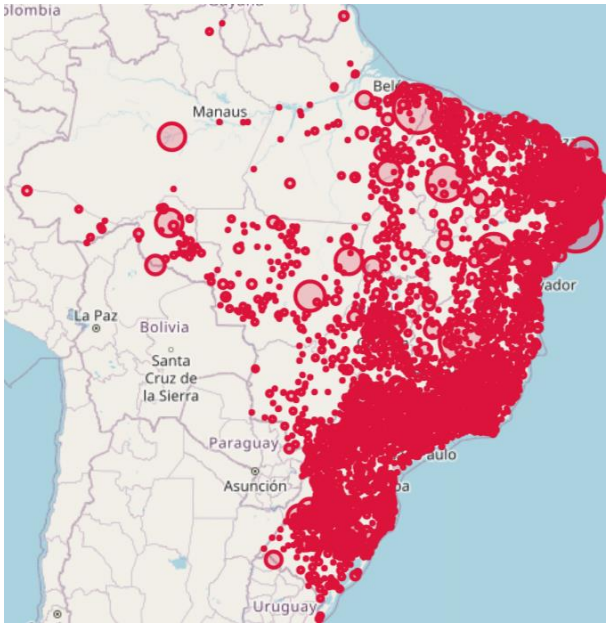
Fonte: Autoria Própria.

Tabela 3: Medidas resumo do ticket médio.

Count	Média	Desvio Padrão	Min	25%	75%	Max
92691	157,75	217,23	1,86	60,70	174,61	13664,08

Fonte: Autoria Própria.

Figura 8: Distribuição do ticket médio no Brasil.

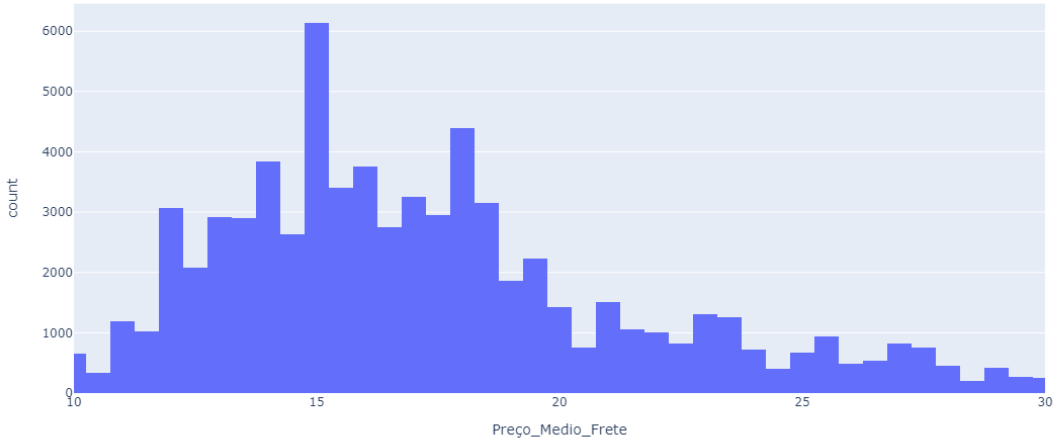


Fonte: Autoria Própria.

Na figura 8 o tamanho dos círculos representa o quão grande é o valor do ticket médio. Assim, a partir da distribuição do ticket médio no Brasil é possível observar que muito das compras está na região de São Paulo, Rio de Janeiro,

Minas Gerais, ou seja, na região sudeste. Outro critério que o consumidor leva em consideração no momento da compra do produto é o frete.

Figura 9: Distribuição do preço médio do frete.



Fonte: Autoria Própria.

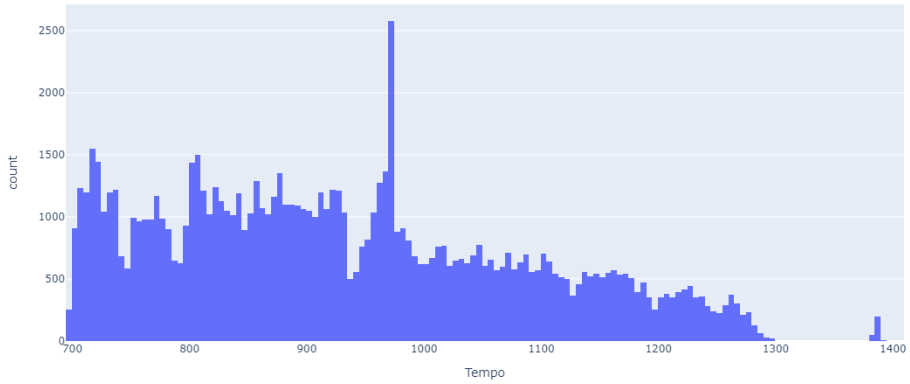
Tabela 4: Medidas resumo do preço médio do frete.

Count	Média	Desvio Padrão	Min	25%	50%	75%	Max
92691	20,162	15,69	0,00	13,37	16,39	21,18	409

Fonte: Autoria Própria.

Como o foco do estudo é saber quando o cliente deixará de comprar no *e-commerce* é necessário calcular o tempo de sua permanência na base de dados. Logo, a partir da última compra de cada cliente e a data do download da base, foi calculado a diferença entre essas duas datas para que assim seja possível estudar o tempo até o *churn.*

Figura 10: Distribuição do tempo de permanência.



Fonte: Autoria Própria.

Na figura 10, o cálculo do tempo foi feito da diferença entre a data da última compra do cliente, que definimos como a entrada do cliente no estudo e a data limite do estudo, que seria a data do download dos dados (24/07/2020). É possível observar uma pequena queda próximo de 900 dias e logo após isso apresenta um crescimento da quantidade de cliente que permaneceram entre 900 e 1000 dias, e logo após esse pico repentino, o número de consumidores vai decaindo.

Tabela 5: Medidas resumo do tempo de permanência.

Count	**Média**	**Desvio Padrão**	**Min**	**25%**	**50%**	**75%**	**Max**
92691	932,33 Dias	152,52 Dias	695 Dias	809 Dias	913 Dias	1041 Dias	1408 Dias

Fonte: Autoria Própria.

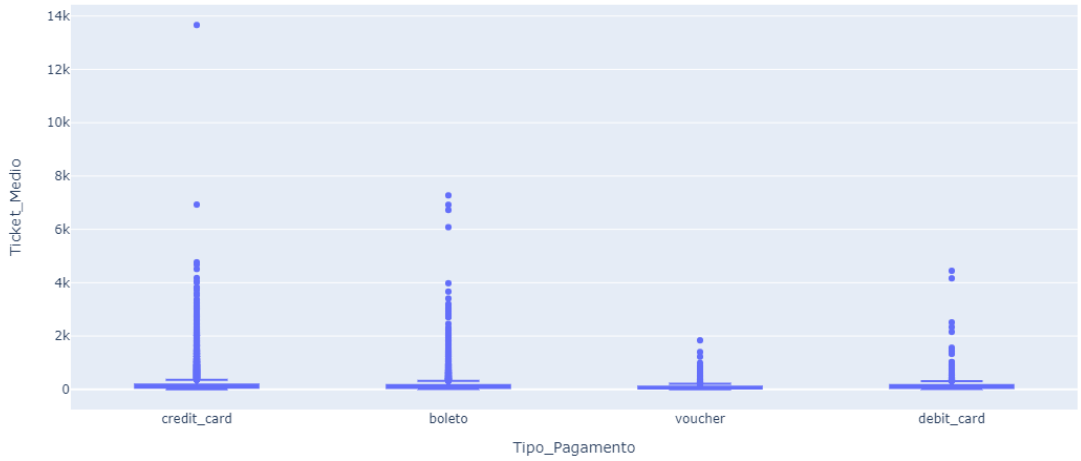
Tabela 6: Média do tempo por tipo de pagamento.

Tipo de Pagamento	**Tempo**
Boleto	941,8 Dias
Cartão de Crédito	930,9 Dias
Cartão de Débito	862,6 Dias
Voucher	947,9 Dias

Fonte: Autoria Própria.

Sobre a média do tempo por tipo de pagamento, é notável que os clientes que utilizaram cartão de débito permaneceram menos tempo comprando, em comparação aos outros métodos de pagamentos.

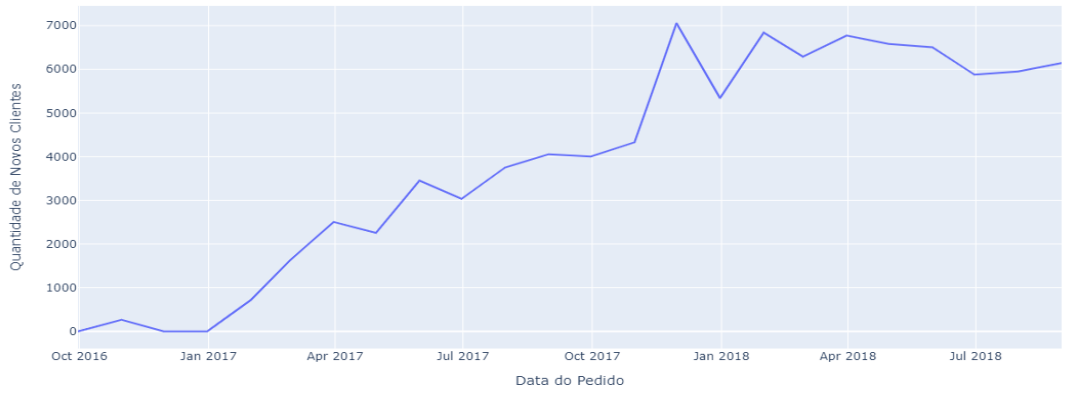
Figura 11: Box-plot do ticket médio pelo método de pagamento.



Fonte: Autoria Própria.

De acordo com o box plot, aqueles clientes que compram por meio do cartão de crédito e boleto possuem um ticket médio maior, do que aqueles utilizaram cartão de débito ou voucher. A justificativa para os clientes que compraram seus produtos com o cartão de crédito é que o cartão pode possuir um limite maior que a sua renda e assim poder parcelar suas compras. Observa-se ainda que nos dados existem muitas compras com valores pequenos, como R$50,00 reais e R$100,00 reais, e poucas compras com valores grandes, outliers, que ficaram mais evidentes no blox-plot.

Figura 12: Crescimento no número de novos clientes por mês.



Fonte: Autoria Própria.

Na figura 12 é possível ver que houve um crescimento na chegada de novos clientes durante o período da base, até o final de 2017, mantendo-se estável em 2018.

4 Aplicação da Análise de Sobrevivência

4.1 Construção dos Dados de Sobrevivência

Como mencionado anteriormente, o cálculo do tempo foi feito da diferença entre a data da última compra do cliente, que definimos como a entrada do cliente no estudo e a data limite do estudo, que seria a data do download dos dados (24/07/2020), logo a escala de medida do estudo é o tempo em dias. Foi estabelecido como censura aleatória. E por fim, as variáveis mencionadas acima. Como a base de dados é muito grande, para efeito didático, no estudo usamos uma amostra de tamanho 5000, escolhida de forma aleatória. Nessa amostra de tamanho 5000, 4254 observações são consideradas como falha e 746 são consideradas como censura, pois são de clientes com somente uma compra. Para cada variável a contagem de censuras e falhas são apresentadas nas Tabelas 7 a 13.

Na tabela 7 é possível observar que 12,75% dos casos de fizeram o pagamento com boleto, são censurados, 15,41% dos que compraram com cartão de crédito foram censurados.

Tabela 7: Contagem de censuras e falhas para a variável tipo de pagamento.

Censura	**Tipo Pagamento**	**Contagem**
Censura	Boleto	126
Censura	Cartão de Crédito	596
Censura	Cartão de Débito	9
Censura	Voucher	15
Falha	Boleto	862
Falha	Cartão de Crédito	3272
Falha	Cartão de Débito	75
Falha	Voucher	45

Fonte: Autoria Própria.

Na tabela 8 mostra como é a distribuição de censura e falha para compras que foram ou não parceladas. É possível observar que 7,88% das compras parceladas foram censuras.

Tabela 8: Contagem de censuras e falhas para a variável parcelado.

Censura	**Parcelado**	**Contagem**
Censura	Não	352
Censura	Sim	394
Falha	Não	2027
Falha	Sim	2227

Fonte: Autoria Própria.

Já na tabela 9, a maioria dos consumidores são da região sudeste, no qual, 10,68% foram censuras e 58,56% dos consumidores da região sul foram tomados como falha.

Tabela 9: Contagem de censuras e falhas para a variável região.

Censura	**Região**	**Contagem**
Censura	Centro – Oeste	53
Censura	Nordeste	55
Censura	Norte	10
Censura	Sudeste	534
Censura	Sul	94
Falha	Centro – Oeste	229
Falha	Nordeste	397
Falha	Norte	78
Falha	Sudeste	2928
Falha	Sul	622

Fonte: Autoria Própria.

Para outras categorias de produtos, 44,18% foram tomados como censura.

Tabela 10: Contagem de censuras e falhas para a variável categoria.

Censura	**Categoria**	**Contagem**
Falha	Casa	162
Falha	Diversão	102
Falha	Eletrônico	111
Falha	Outros	371
Censura	Casa	684
Censura	Diversão	653
Censura	Eletrônico	708
Censura	Outros	2209

Fonte: Autoria Própria.

Na tabela 11, a censura teve o maior ticket médio.

Tabela 11: Média do ticket médio para censuras e falhas.

Censura	**Ticket Médio**
Censura	R$ 205,00
Falha	R$ 148,00

Fonte: Autoria Própria.

Já para o preço médio do frete, a falha teve o maior preço médio de frete.

Tabela 12: Média do preço médio frete para censuras e falhas.

Censura	**Preço Médio Frete**
Censura	R$ 19,60
Falha	R$ 20,70

Fonte: Autoria Própria.

Na tabela 13 mostra uma semelhança na média da falha e censura.

Tabela 13: Média do tempo para censuras e falhas.

Censura	**Tempo**
Censura	939 Dias
Falha	933 Dias

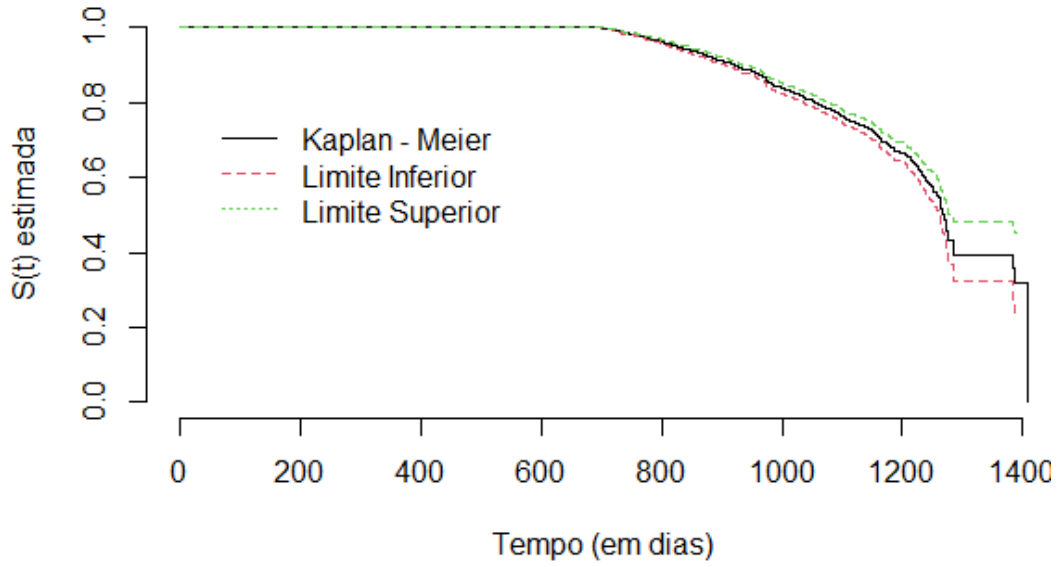
Fonte: Autoria Própria.

4.2 Aplicação das Técnicas Não – Paramétricas nos dados

4.2.1 Kaplan-Meier

Por meio do estimador de Kaplan - Meier foram obtidas as probabilidades de sobrevivência para cada tempo e os resultados podem ser apresentados por meio do gráfico, apresentado na Figura 13.

Figura 13: Sobrevivência estimada por Kaplan-Meier para os dados de *e-commerces* brasileiros.



Fonte: Autoria Própria.

O tempo médio de sobrevivência de clientes que compraram em *e-commerces* brasileiros é de $\hat{t}_m = 609{,}9243$ dias. Por meio da interpolação linear temos que o tempo mediano de dias de é

$$\frac{1275-1274}{0{,}4926929-0{,}5141143} = \frac{1274-MED}{0{,}4926929-0{,}50}, \tag{89}$$

tal que a solução é $MED = 1274{,}6588$ dias, ou seja, a estimativa de tempo em que 50% dos indivíduos que compraram em *e-commerces* brasileiros foi de 1274,6588 dias.

4.3 Aplicação dos Modelos Probabilísticos em Análise de Sobrevivência

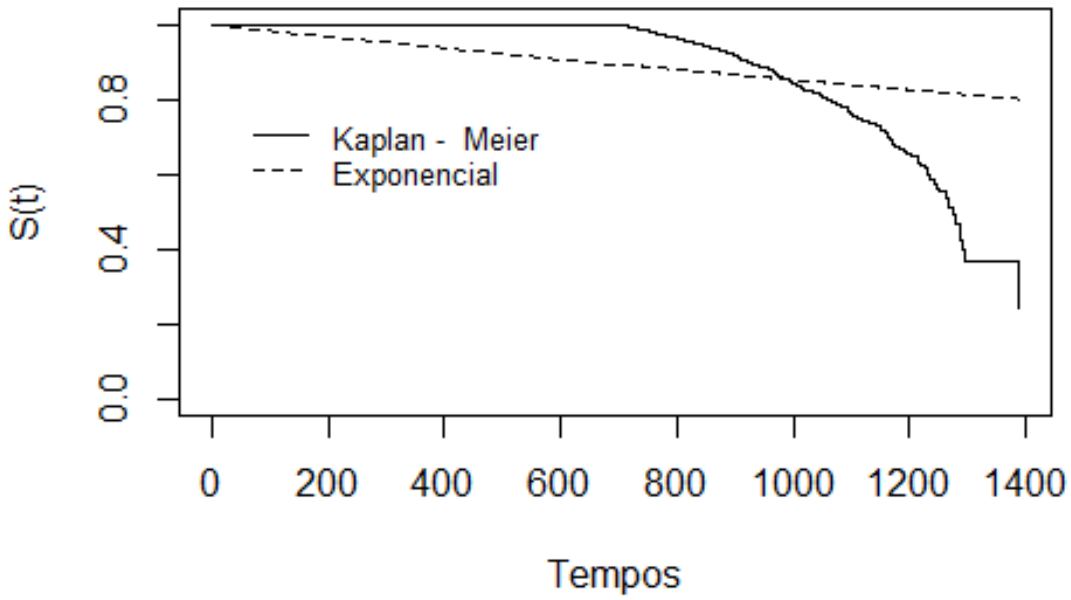
Para saber em qual modelo probabilístico de sobrevivência os dados mais se adequam, foi calculada as estimativas de cada modelo.

4.3.1 Distribuição Exponencial

A função de sobrevivência da distribuição construída a partir dos dados foi definida por,

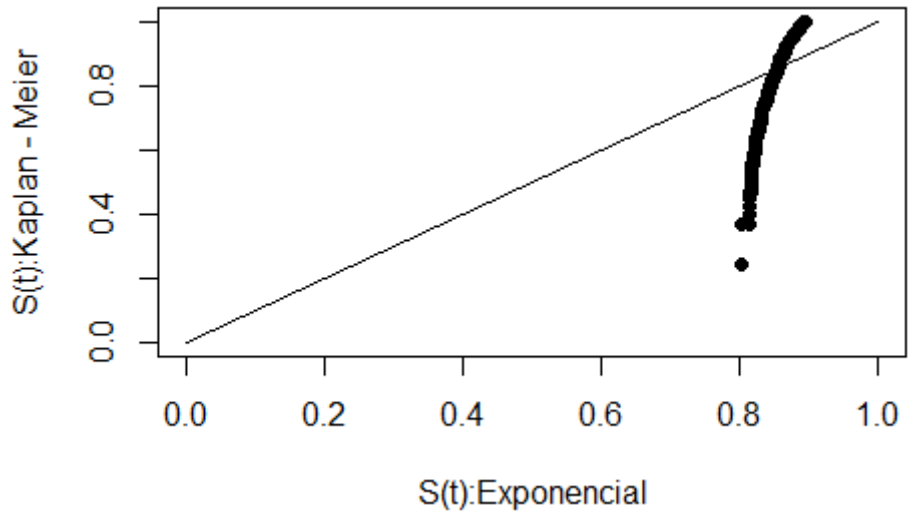
$$S(t) = \exp\left\{-\left(\frac{t}{6760{,}501}\right)\right\}. \tag{90}$$

Figura 14: Curvas de sobrevivência estimadas pelo modelo exponencial
Versus a curva de sobrevivência estimada por Kaplan-Meier.



Fonte: Autoria Própria.

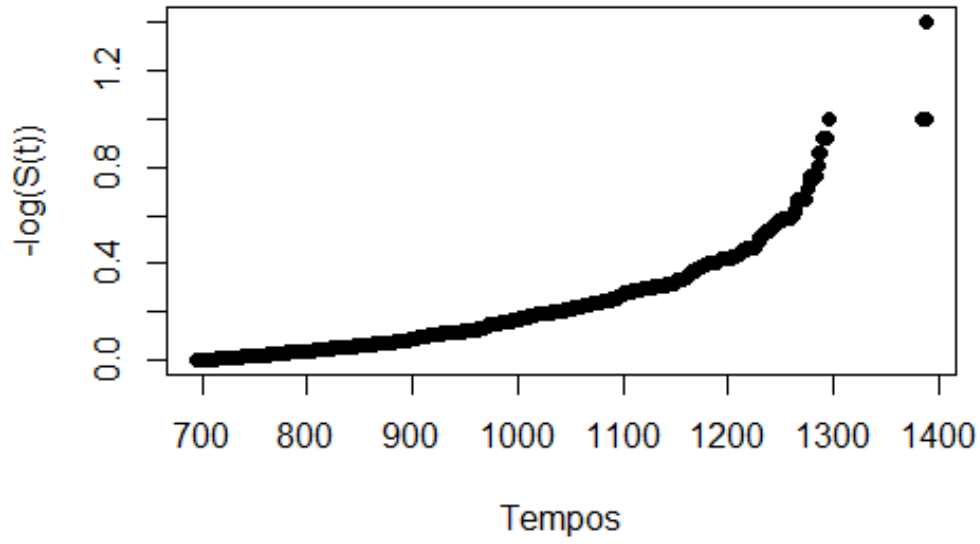
Figura 15: Gráfico das sobrevivências estimadas por Kaplan-Meier versus as sobrevivências estimadas pelo modelo exponencial.



Fonte: Autoria Própria.

É possível observar que na figura 15, o modelo exponencial não se adequa bem aos dados, pois a curva apresenta um afastamento grande da reta $y = x$.

Figura 16: Gráfico de t vs $-\log(\hat{S}(t))$.



Fonte: Autoria Própria.

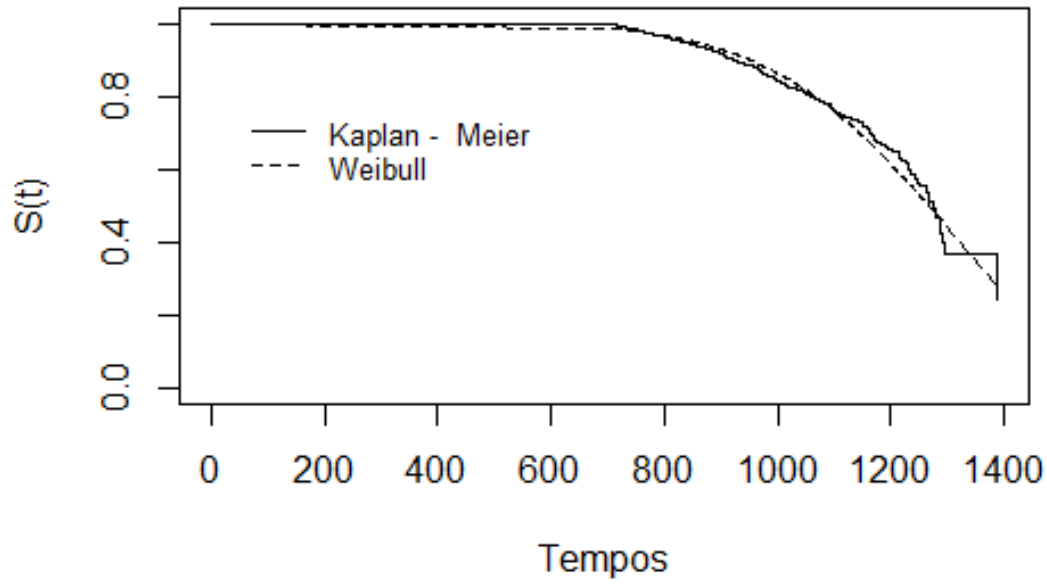
Aplicando o modelo exponencial nos dados do estudo, é possível perceber que existe um desvio marcante na reta do gráfico de linearização, assim o modelo exponencial não é adequado.

4.3.2 Distribuição Weibull

A função de sobrevivência dada pelo modelo weibull construída com os dados de sobrevivência de clientes que compraram em *e-commerces* brasileiros é,

$$S(t) = \exp\left\{-\left(\frac{t}{7{,}54222}\right)^{6{,}54222}\right\}. \quad (91)$$

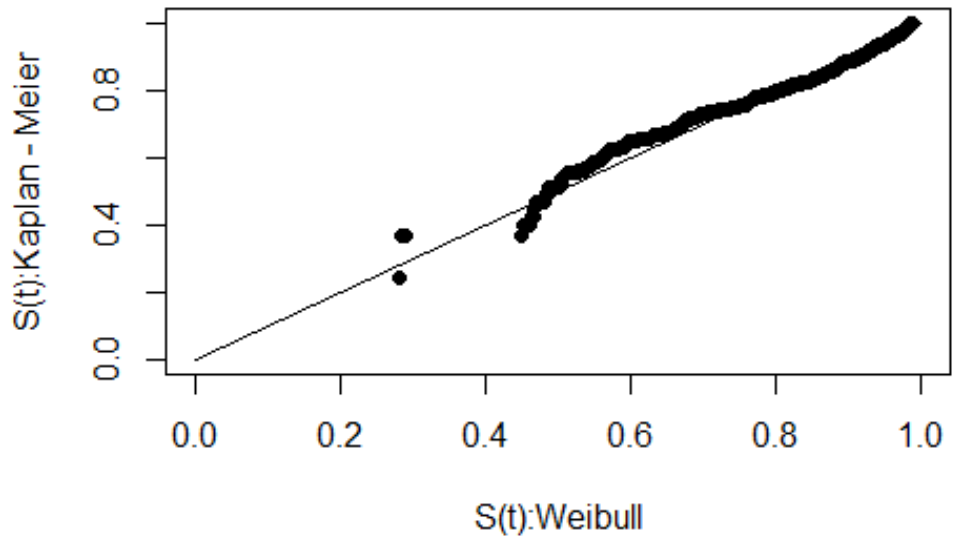
Figura 17: Curvas de sobrevivência estimadas pelo modelo weibull
Versus a curva de sobrevivência estimada por Kaplan-Meier.



Fonte: Autoria Própria.

Comparando graficamente as duas curvas de sobrevivência pode-se observar que, aparentemente a curva do modelo de Weibull está bem próxima da curva de sobrevivência obtida pelo Método de Kaplan-Meier.

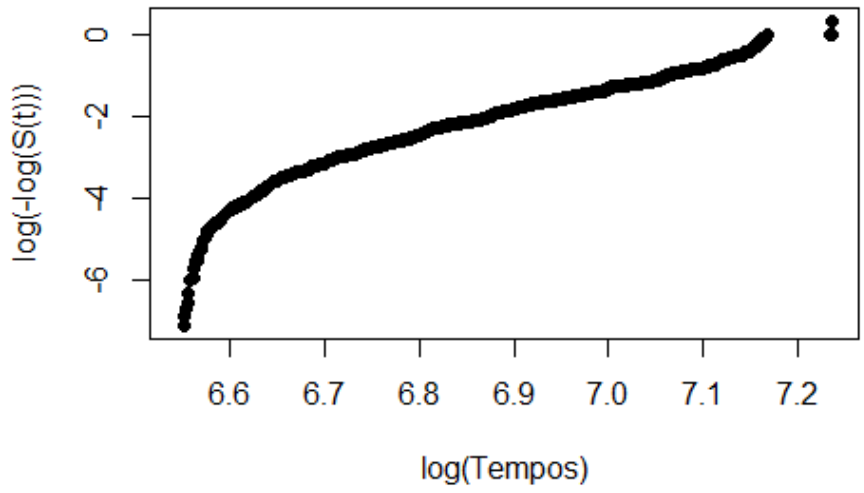
Figura 18:Gráfico das sobrevivências estimadas por Kaplan-Meier versus as sobrevivências estimadas pelo modelo weibull.



Fonte: Autoria Própria.

Diferente do modelo exponencial, o modelo weibull construído parece se adequar aos dados de consumidores que compraram em *e-commerce* brasileiros. A curva mostra - se bem próxima da reta $y = x$.

Figura 19: $\log(t)$ vs $\log(-\log(\hat{S}(t)))$.



Fonte: Autoria Própria.

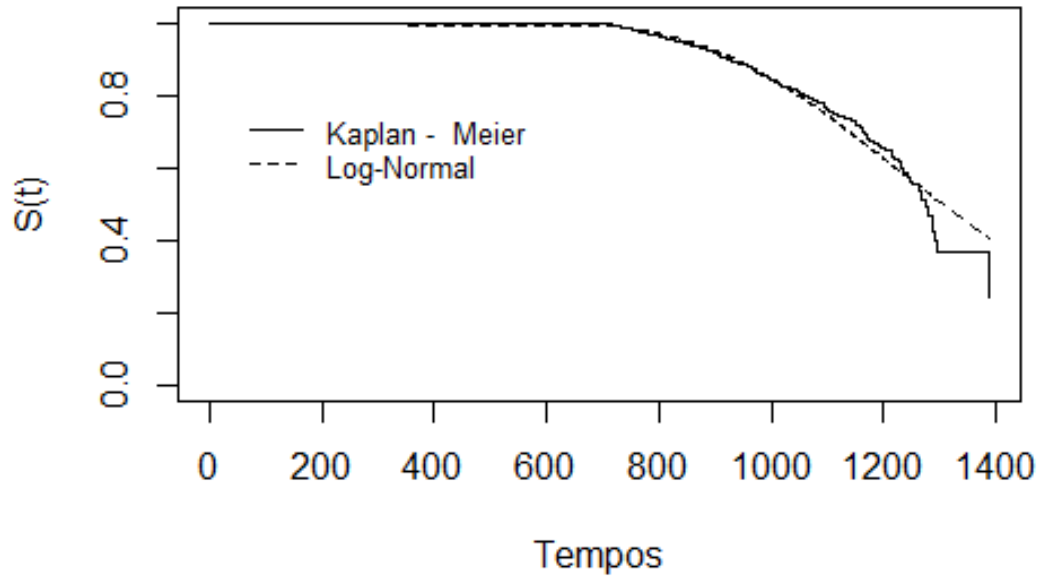
Assim como no gráfico de linearização do modelo exponencial, a Figura 19, para o modelo Weibull também apresenta um desvio menos acentuado do que o modelo exponencial, no entanto, ainda não se aproxima de uma reta

4.3.3 Distribuição Log-Normal

A expressão do modelo log – normal construída a partir dos dados de sobrevivência do estudo é denominada por,

$$S(t) = \Phi\left(\frac{-\log(t) + 7{,}17384}{0{,}2601503}\right). \quad (92)$$

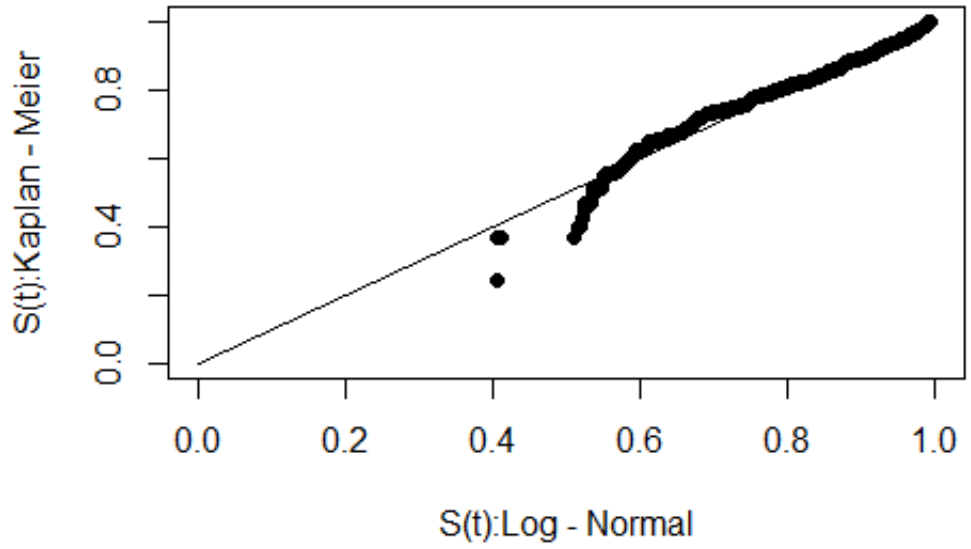
Figura 20: Curvas de sobrevivência estimadas pelo modelo log-normal
Versus a curva de sobrevivência estimada por Kaplan-Meier.



Fonte: Autoria Própria.

A figura 20 exibe uma diferença entre os modelos Exponencial e Weibull, pois suas curvas de sobrevivência se aproximam da curva de Kaplan-Meier. Assim, com o gráfico de sobrevivência estimada podemos saber qual modelo se adequa mais aos dados.

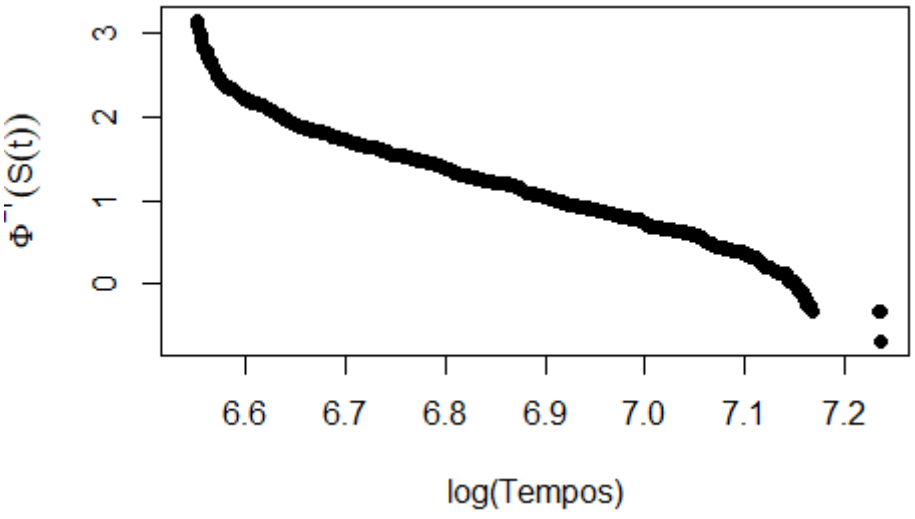
Figura 21: Gráfico das sobrevivências estimadas por Kaplan-Meier versus as sobrevivências estimadas pelo modelo log-normal.



Fonte: Autoria Própria.

Com o modelo log-normal ajustado aos dados, a curva se mostra muito próxima da reta $y = x$, indicando novamente ser um modelo adequado ao estudo em questão.

Figura 22: $\log(t)$ vs $\Phi^{-1}(\hat{S}(t))$.

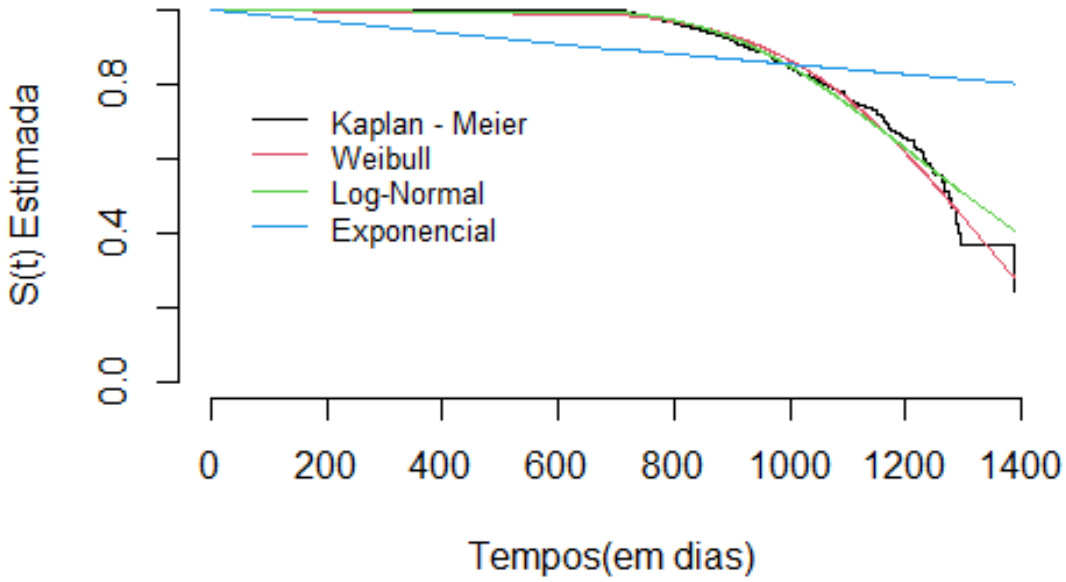


Fonte: Autoria Própria.

No gráfico de linearização da função de sobrevivência do modelo log-normal construída a partir dos dados não apresenta uma reta no gráfico.

Na figura 23 tem-se uma melhor visualização sobre qual modelo é mais adequado aos dados. Na curva verde, que se refere ao modelo Log-Normal, é a curva que está mais próxima a curva de sobrevivência não-paramétrico de Kaplan-Meier.

Figura 23: Gráfico das sobrevivências estimadas por Kaplan-Meier versus todos os modelos estudados.



Fonte: Autoria Própria.

Além dos métodos gráficos, utilizamos o critério de informação de akaike (AIC), que é utilizada para mensurar a qualidade de um modelo estatístico. O objetivo do método é ajustar um modelo com um número mínimo de parâmetros. Assim, o melhor modelo é aquele que obtém um menor valor do critério de Akaike.

$$AIC = 2[p - l(\hat{\theta}_M)], \tag{93}$$

tal que $l(\hat{\theta}_M)$ é o Log - verossimilhança do modelo.

Tabela 14: Log da verossimilhança e critério de akaike de todos os modelos ajustados.

Modelo	$\boldsymbol{l(\hat{\theta}_M)}$	**AIC**
Exponencial	-7267,544	14537,09
Weibull	-6239,116	12482,23
Log – Normal	-6202,008	12408.02

Fonte: Autoria Própria.

Portanto, a partir da aplicação dos três modelos nos dados do estudo, no qual foram os modelos exponencial, weibull e log-normal, podemos dizer que, por meio da análise gráfica, aplicação do critério de Akaike e o log da verossimilhança o melhor modelo que se adequa aos dados de clientes que compraram em *e-commerces* brasileiros é o modelo Log – Normal, cuja função de sobrevivência é dada por:

$$S(t) = \Phi\left(\frac{-\log(t)+ 7{,}17384}{0{,}2601503}\right). \quad (94)$$

O tempo médio de sobrevivência de consumidores que compraram em *e-commerces* brasileiros é de

$$\hat{E}(T) = \exp\left\{7{,}17384 + \frac{0{,}2601503^2}{2}\right\} = 1349{,}56. \quad (95)$$

ou seja, $1349{,}56$ dias. A variância do tempo médio de sobrevivência é,

$$\widehat{Var}\left(\hat{E}(T)\right) = (167{,}4935) + (0{,}7671767) + (65782{,}02) = 65950{,}28. \quad (96)$$

O intervalo de confiança de 95% para $\hat{E}(T)$ é dado por,

$$\hat{E}(T) \pm z_{\alpha/2}\sqrt{\widehat{Var}\left(\hat{E}(T)\right)} = 1349{,}56 \pm 1{,}96\sqrt{65950{,}28} = [846{,}4126; 1853{,}099], \quad (97)$$

O tempo mediano de sobrevivência dos consumidores que se mantém na base é,

$$\widehat{t_{0,5}} = \exp\{0 * 7{,}17384 + 0{,}2601503\} = 1304{,}846, \quad (98)$$

ou seja, 50% dos consumidores permanecem na base por $1304{,}846$ dias. Sua variância é dada por:

$$\widehat{Var}\left(\widehat{t_{0,5}}\right) = Var(\mu) * [\exp\{\hat{\mu}\}]^2 = (9{,}193641e-05) * [\exp(7{,}17384)]^2 = 156{,}533. \quad (99)$$

já o intervalo de confiança de 95% para o tempo mediano é

$$\widehat{t_{0,5}} \pm z_{\alpha/2}\sqrt{\widehat{Var}\left(\widehat{t_{0,5}}\right)} = 1304{,}846 \pm 1.96\sqrt{156{,}533} = [1280{,}323; 1329{,}368]. \quad (100)$$

4.4 Aplicação do Modelo de Regressão Log – Normal

4.4.1 Seleção de Covariáveis

Para construir os modelos de regressão serão utilizados 14 covariáveis no estudo que explica o comportamento da variável resposta, logo será necessário construir a seleção de variáveis que sejam que explicam os dados do estudo. Além disso, para as variáveis parcelado, tipo de pagamento, região e categoria foram transformadas em variáveis *dummies,* que o valor zero ou um indica a ausência ou presença do atributo.

Para a seleção destas covariáveis existem algoritmos implementados em pacotes estatísticos que fazem este trabalho de forma automática, como *forward*, *backward* e *stepwise.* No entanto, os mesmos algoritmos identificam apenas uma covariável que se mostra relevante para o modelo, não um conjunto de variáveis. Com o método proposto por Collet(1994) o pesquisador tem uma postura livre no momento da seleção de covariáveis. Os passos para a seleção são da seguinte forma:

i. Ajustar o modelo apenas com uma variável. Incluir todas as covariáveis que forem significativas ao nível de 0,10.

Tabela 15: Seleção de variáveis com modelos univariados.

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
Passo 1	Nulo	$l_1 = -6202$	12404.02	$\beta_0 = 7.17384$	$< 2e^{-16}$
	Ticket Médio	$l_2 = -6181.6$	12363.22	$\beta_0 = 7.20e^{+00}$ $\beta_1 = -1.68e^{-04}$	$5.5e^{-11}$
	Preço Médio Frete	$l_3 = -6201.9$	12403.83	$\beta_0 = 7.170705$ $\beta_1 = 0.000159$	0.66
	Parcelado	$l_4 = -6201.6$	12403.25	$\beta_0 = 7.1683$ $\beta_1 = 0.0103$	0.38
	Pagamento:	$l_5 = -6199.9$	12399.77	$\beta_0 = 7.1965$	0.041

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
	Cartão de Crédito			$\beta_1 = -0.0294$	
	Pagamento: Cartão de Débito	$l_6 = -6201.9$	12403.84	$\beta_0 = 7.1740$ $\beta_1 = -0.0212$	0.67
	Pagamento: Voucher	$l_7 = -6200.1$	12400.16	$\beta_0 = 7.17507$ $\beta_1 = -0.09487$	0.045
	Região: Sudeste	$l_8 = -6199.2$	12398.35	$\beta_0 = 7.1945$ $\beta_1 = -0.0305$	0.018
	Região: Sul	$l_9 = -6200.2$	12400.35	$\beta_0 = 7.16899$ $\beta_1 = 0.03257$	0.058
	Região: Norte	$l_{10} = -6200.8$	12401.57	$\beta_0 = 7.1724$ $\beta_1 = 0.0728$	0.13
	Região: Nordeste	$l_{11} = -6197.7$	12395.46	$\beta_0 = 7.16768$ $\beta_1 = 0.01305$	0.0041
	Categoria: Diversão	$l_{12} = -6199$	12398.09	$\beta_0 = 7.16714$ $\beta_1 = 0.04003$	0.016
	Categoria: Eletrônico	$l_{13} = -6201.2$	12402.49	$\beta_0 = 7.17063$ $\beta_1 = 0.02000$	0.22
	Categoria: Outros	$l_{14} = -6201.6$	12403.30	$\beta_0 = 7.16880$ $\beta_1 = 0.00997$	0.4

Fonte: Autoria Própria.

ii. As covariáveis que foram significativas no passo i são ajustadas conjuntamente. Com a inclusão de outras covariáveis outras podem deixarem de ser significativas, assim tem – se modelos reduzidos. Somente aquelas variáveis que possuem significância permanecem no modelo.

Tabela 16: Seleção de variáveis com modelo multivariado o passo 2.

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
Passo 2	Intercepto, Ticket Médio, Cartão de Crédito,	$l_1 = -6163.3$	12326.56	$\beta_0 = 7.20e^{+00}$ $\beta_1 = -1.77e^{-04}$ $\beta_2 = -3.65e^{-02}$ $\beta_3 = -1.43e^{-01}$ $\beta_4 = -1.52e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 2.9e^{-12}$ $p_2 = 0.01355$ $p_3 = 0.00276$ $p_4 = 0.47761$

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
	Voucher, Sudeste, Nordeste, Sul e Diversão.			$\beta_5 = 9.67e^{-02}$ $\beta_6 = 4.99e^{-02}$ $\beta_7 = 4.35e^{-02}$	$p_5 = 0.00096$ $p_7 = 0.05226$ $p_6 = 0.00843$
	Intercepto, Cartão de Crédito, Voucher, Sudeste, Sul, Nordeste e Diversão.	$l_2 = -6186.4$	12372.70	$\beta_0 = 7.1648$ $\beta_1 = -0.0367$ $\beta_2 = -0.1244$ $\beta_3 = 0.0218$ $\beta_4 = 0.0910$ $\beta_5 = 0.0561$ $\beta_6 = 0.0410$	$p_0 < 2e^{-16}$ $p_1 = 0.0130$ $p_2 = 0.0100$ $p_3 = 0.3093$ $p_4 = 0.0019$ $p_5 = 0.0297$ $p_6 = 0.0131$
	Intercepto, Ticket Médio, Voucher, Sudeste, Nordeste, Sul e Diversão.	$l_3 = -6168.4$	12332.76	$\beta_0 = 7.17e^{+00}$ $\beta_1 = -1.78e^{-04}$ $\beta_2 = -1.14e^{-01}$ $\beta_3 = -1.47e^{-02}$ $\beta_4 = 9.57e^{-02}$ $\beta_5 = 5.57e^{-02}$ $\beta_6 = 4.41e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 1.6e^{-12}$ $p_2 = 0.0137$ $p_3 = 0.2062$ $p_4 = 0.0102$ $p_5 = 0.0089$ $p_6 = 0.0456$
	Intercepto, Ticket Médio, Cartão de Crédito, Sudeste, Nordeste, Sul e Diversão.	$l_4 = -6167.5$	12335.04	$\beta_0 = 7.19e^{+00}$ $\beta_1 = -1.73e^{-04}$ $\beta_2 = -2.65e^{-02}$ $\beta_3 = 1.39e^{-01}$ $\beta_4 = 9.35e^{-02}$ $\beta_5 = 5.06e^{-02}$ $\beta_6 = 4.38e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 9.5e^{-12}$ $p_2 = 0.0637$ $p_3 = 0.5184$ $p_4 = 0.0014$ $p_5 = 0.0497$ $p_6 = 0.0080$
	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Nordeste, Sul e Diversão.	$l_5 = -6163.5$	12327.06	$\beta_0 = 7.21e^{+00}$ $\beta_1 = -1.78e^{-04}$ $\beta_2 = -3.64e^{-02}$ $\beta_3 = -1.43e^{-01}$ $\beta_4 = 8.31e^{-02}$ $\beta_5 = 3.63e^{-02}$ $\beta_6 = 4.32e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 2.2e^{-12}$ $p_2 = 01384$ $p_3 = 0.00290$ $p_4 = 0.00018$ $p_5 = 0.03411$ $p_6 = 0.00885$
	Intercepto, Ticket Médio, Cartão de Crédito,	$l_6 = -6168.8$	12337.50	$\beta_0 = 7.25e^{+00}$ $\beta_1 = -1.76e^{-04}$ $\beta_2 = -3.59e^{-02}$ $\beta_3 = -1.38e^{-01}$ $\beta_4 = -3.37e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 5.7e^{-12}$ $p_2 = 0.0155$ $p_3 = 0.0042$ $p_4 = 0.0375$

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
	Voucher, Sudeste, Sul e Diversão.			$\beta_5 = 1.11e^{-03}$ $\beta_6 = 4.17e^{-02}$	$p_5 = 0.9590$ $p_6 = 0.0117$
	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Sudeste, Nordeste e Diversão.	$l_7 = -6165.1$	12330.30	$\beta_0 = 7.23e^{+00}$ $\beta_1 = -1.79e^{-04}$ $\beta_2 = -3.72e^{-02}$ $\beta_3 = -1.44e^{-01}$ $\beta_4 = -1.65e^{-02}$ $\beta_5 = 6.65e^{-02}$ $\beta_6 = 4.27e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 1.7e^{-12}$ $p_2 = 0.0120$ $p_3 = 0.0026$ $p_4 = 0.2492$ $p_5 = 0.0080$ $p_6 = 0.0097$
	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Sudeste, Nordeste e Sul.	$l_8 = -6166.8$	12333.65	$\beta_0 = 7.21e^{+00}$ $\beta_1 = -1.76e^{-04}$ $\beta_2 = -3.71e^{-02}$ $\beta_3 = -1.45e^{-01}$ $\beta_4 = 1.38e^{-02}$ $\beta_5 = 9.41e^{-02}$ $\beta_6 = 4.82e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 4.6e^{-12}$ $p_2 = 0.0123$ $p_3 = 0.0026$ $p_4 = 0.5201$ $p_5 = 0.0013$ $p_6 = 0.0613$

Fonte: Autoria Própria.

iii. Ajusta-se um novo modelo com as covariáveis retidas no passo ii. Neste passo as covariáveis no passo ii retornam ao modelo para confirmar que elas não são estatisticamente significativas.

Tabela 17: Seleção de variáveis com modelo multivariado o passo 3.

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
Passo 3	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Nordeste, Sul e Diversão.	$l_1 = -6163.5$	12327.06	$\beta_0 = 7.21e^{+00}$ $\beta_1 = -1.78e^{-04}$ $\beta_2 = -3.64e^{-02}$ $\beta_3 = -1.43e^{-01}$ $\beta_4 = 8.31e^{-02}$ $\beta_5 = 3.63e^{-02}$ $\beta_6 = 4.32e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 2.2e^{-12}$ $p_2 = 01384$ $p_3 = 0.00290$ $p_4 = 0.00018$ $p_5 = 0.03411$ $p_6 = 0.00885$
	Intercepto,	$l_2 = -6163.3$	12326.56	$\beta_0 = 7.20e^{+00}$	$p_0 < 2e^{-16}$

	Ticket Médio, Cartão de Crédito, Voucher, Nordeste, Sul, Diversão e Sudeste			$\beta_1 = -1.77e^{-04}$ $\beta_2 = -3.65e^{-02}$ $\beta_3 = -1.43e^{-01}$ $\beta_4 = 9.67e^{-02}$ $\beta_5 = 4.99e^{-02}$ $\beta_6 = 4.35e^{-02}$ $\beta_7 = 1.52e^{-02}$	$p_1 = 2.9e^{-12}$ $p_2 = 0.01355$ $p_3 = 0.00276$ $p_4 = 0.00096$ $p_5 = 0.05226$ $p_6 = 0.00843$ $p_7 = 0.47761$

Fonte: Autoria Própria.

iv. As eventuais covariáveis significativas no passo iii são incluídas no modelo juntamente aquelas do passo ii. Neste processo retorna-se com as covariáveis excluídas no passo i para confirmar que elas não são estatisticamente significativas.

Tabela 18: Seleção de variáveis com modelo multivariado o passo 4.

Passos	**Modelo**	**Log - Verossimilhança**	**-2log L**	**Estimativa**	**Valor p**
Passo 4	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Nordeste, Sul e Diversão.	$l_1 = -6163.5$	12320.40	$\beta_0 = 7.21e^{+00}$ $\beta_1 = -1.78e^{-04}$ $\beta_2 = -3.64e^{-02}$ $\beta_3 = -1.43e^{-01}$ $\beta_4 = 8.31e^{-02}$ $\beta_5 = 3.63e^{-02}$ $\beta_6 = 4.32e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 2.2e^{-12}$ $p_2 = 01384$ $p_3 = 0.00290$ $p_4 = 0.00018$ $p_5 = 0.03411$ $p_6 = 0.00885$
	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Nordeste, Sul, Diversão e Preço Médio Frete	$l_2 = -6160.2$	12315.26	$\beta_0 = 7.20e^{+00}$ $\beta_1 = -2.09e^{-04}$ $\beta_2 = -3.71e^{-02}$ $\beta_3 = -1.43e^{-01}$ $\beta_4 = 6.86e^{-02}$ $\beta_5 = 3.42e^{-02}$ $\beta_6 = 4.37e^{-02}$ $\beta_7 = 1.05e^{-03}$	$p_0 < 2e^{-16}$ $p_1 = 1.2e^{-14}$ $p_2 = 0.0123$ $p_3 = 0.0028$ $p_4 = 0.0027$ $p_5 = 0.0469$ $p_6 = 0.0082$ $p_7 = 0.0126$
	Intercepto, Ticket Médio, Cartão de	$l_3 = -6157.6$	12326.03	$\beta_0 = 7.21e^{+00}$ $\beta_1 = -1.63e^{-04}$ $\beta_2 = -6.92e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 4.8e^{-14}$ $p_2 = 7.6e^{-05}$

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
	Crédito, Voucher, Nordeste, Sul, Diversão e Parcelado.			$\beta_3 = -1.44e^{-01}$ $\beta_4 = 7.88e^{-02}$ $\beta_5 = 3.41e^{-02}$ $\beta_6 = 4.33e^{-02}$ $\beta_7 = 4.87e^{-03}$	$p_3 = 0.00255$ $p_4 = 0.00037$ $p_5 = 0.04592$ $p_6 = 0.00842$ $p_7 = 0.00056$
	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Nordeste, Sul, Diversão e Cartão de Débito.	$l_4 = -6161.4$	12324.06	$\beta_0 = 7.22e^{+00}$ $\beta_1 = -1.78e^{-04}$ $\beta_2 = -3.99e^{-02}$ $\beta_3 = -1.46e^{-01}$ $\beta_4 = 8.28e^{-02}$ $\beta_5 = 3.57e^{-02}$ $\beta_6 = 4.30e^{-02}$ $\beta_7 = -5.34e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 2.1e^{-12}$ $p_2 = 0.00868$ $p_3 = 0.00233$ $p_4 = 0.00018$ $p_5 = 0.03712$ $p_6 = 0.00904$ $p_7 = 0.30232$
	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Nordeste, Sul, Diversão e Eletrônico.	$l_6 = -6152.7$	12322.74	$\beta_0 = 7.21e^{+00}$ $\beta_1 = -1.79e^{-04}$ $\beta_2 = -3.51e^{-02}$ $\beta_3 = -1.40e^{-01}$ $\beta_4 = 8.31e^{-02}$ $\beta_5 = 3.59e^{-02}$ $\beta_6 = 4.84e^{-02}$ $\beta_7 = 2.82e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 1.5e^{-12}$ $p_2 = 0.01774$ $p_3 = 0.00353$ $p_4 = 0.00018$ $p_5 = 0.03588$ $p_6 = 0.00386$ $p_7 = 0.08582$
	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Nordeste, Sul, Diversão e Outros.	$l_7 = -6126.8$	12253.53	$\beta_0 = 7.21e^{+00}$ $\beta_1 = -1.78e^{-04}$ $\beta_2 = -3.68e^{-02}$ $\beta_3 = -1.45e^{-01}$ $\beta_4 = 8.10e^{-02}$ $\beta_5 = 3.60e^{-02}$ $\beta_6 = 5.88e^{-02}$ $\beta_7 = 2.66e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 1.9e^{-12}$ $p_2 = 0.01267$ $p_3 = 0.00255$ $p_4 = 0.00025$ $p_5 = 0.03556$ $p_6 = 0.00114$ $p_7 = 0.03747$

Fonte: Autoria Própria.

v. Ajusta-se um modelo incluindo as covariáveis significativas no passo 4. Neste passo é testado se alguma delas pode ser retirada do modelo.

Tabela 19: Seleção de variáveis com modelo multivariado o passo 5.

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
Passo 5	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Nordeste, Sul, Diversão, Preço Médio Frete, Parcelado e Outros.	$l_1 = -6152.7$		$\beta_0 = 7.19e^{+00}$ $\beta_1 = -2.22e^{-04}$ $\beta_2 = -6.89e^{-02}$ $\beta_3 = -1.46e^{-01}$ $\beta_4 = 6.35e^{-02}$ $\beta_5 = 3.18e^{-02}$ $\beta_6 = 5.84e^{-03}$ $\beta_7 = 9.73e^{-02}$ $\beta_8 = 4.68e^{-02}$ $\beta_9 = 2.49e^{-02}$	$p_0 < 2e^{-16}$ $p_1 = 4.8e^{-15}$ $p_2 = 8.1e^{-05}$ $p_3 = 0.00223$ $p_4 = 0.00538$ $p_5 = 0.06310$ $p_6 = 0.00122$ $p_7 = 0.01965$ $p_8 = 0.00091$ $p_9 = 0.05133$

Fonte: Autoria Própria.

vi. Utilizando as covariáveis que restaram ao passo v ajusta-se o modelo final para os efeitos principais. Para completar a modelagem deve-se verificar a probabilidade de inclusão de termos de interação. Deve se testar se cada uma das interações duas a duas possíveis entre as covariáveis incluídas no modelo. O modelo fica determinado pelos efeitos principais identificados no passo v e os termos de interação significativos neste passo.

Tabela 20: Seleção de covariáveis com a inclusão da interação no passo 6.

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
Passo 6	Intercepto, Ticket Médio, Cartão de Crédito, Voucher, Nordeste, Sul, Diversão, Preço Médio Frete, Parcelado, Outros, Ticket Médio * Cartão de Crédito, Ticket Médio * Voucher, Ticket Médio * Nordeste,	$l_1 = -6126.8$		$\beta_0 = 7.24e^{+00}$ $\beta_1 = -5.25e^{-04}$ $\beta_2 = -9.19e^{-02}$ $\beta_3 = -2.13e^{-01}$ $\beta_4 = 2.61e^{-01}$ $\beta_5 = 3.13e^{-02}$ $\beta_6 = -3.20e^{-02}$ $\beta_7 = -5.26e^{-04}$ $\beta_8 = 6.96e^{-02}$ $\beta_9 = -1.68e^{-02}$ $\beta_{10} = 9.57e^{-05}$ $\beta_{11} = 2.83e^{-03}$	$p_0 < 2e^{-16}$ $p_1 = 3.1e^{-09}$ $p_2 = 0.0242$ $p_3 = 0.1585$ $p_4 = 0.0022$ $p_5 = 0.5637$ $p_6 = 0.5395$ $p_7 = 0.7389$ $p_8 = 0.0263$ $p_9 = 0.6317$ $p_{10} = 0.3958$ $p_{11} = 0.0338$

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
	Ticket Médio * Sul, Ticket Médio * Diversão, Ticket Médio * Preço Médio do Frete, Ticket Médio * Parcelado, Ticket Médio * Outros, Cartão de Crédito * Voucher, Cartão de Crédito * Nordeste, Cartão de Crédito * Voucher, Cartão de Crédito * Nordeste, Cartão de Crédito * Sul, Cartão de Crédito * Diversão, Cartão de Crédito * Preço Médio do Frete, Cartão de Crédito * Parcelado, Cartão de Crédito * Outros, Voucher * Nordeste, Voucher * Sul, Voucher * Diversão, Voucher * Preço Médio do Frete, Voucher * Parcelado, Voucher * Outros, Nordeste * Sul, Nordeste * Diversão, Nordeste * Preço Médio do Frete, Nordeste * Parcelado, Nordeste * Outros, Sul * Diversão, Sul * Diversão, Sul * Preço Médio do Frete, Sul * Parcelado, Sul * Outros, Diversão * Preço Médio do Frete, Diversão * Parcelado, Diversão			$\beta_{12} = 9.00e^{-05}$ $\beta_{13} = -4.80e^{-05}$ $\beta_{14} = 1.64e^{-04}$ $\beta_{15} = 4.50e^{-06}$ $\beta_{16} = 2.35e^{-05}$ $\beta_{17} = 2.35e^{-05}$ $\beta_{18} = 4.54e^{-05}$ $\beta_{19} = 0.00e^{+00}$ $\beta_{20} = -1.45e^{-01}$ $\beta_{21} = -1.05e^{-01}$ $\beta_{22} = 9.03e^{-02}$ $\beta_{23} = -5.51e^{-04}$ $\beta_{24} = 00e^{+00}$ $\beta_{25} = 4.93e^{-02}$ $\beta_{26} = -2.13e^{-01}$ $\beta_{27} = -1.96e^{-01}$ $\beta_{28} = 3.13e^{-02}$ $\beta_{29} = -8.64e^{-03}$ $\beta_{30} = 0.00e^{+00}$ $\beta_{31} = 1.13e^{-01}$ $\beta_{32} = 0.00e^{+00}$ $\beta_{33} = -6.37e^{-02}$ $\beta_{34} = -1.41e^{-03}$ $\beta_{35} = -9.22e^{-03}$ $\beta_{36} = -2.45e^{-02}$ $\beta_{37} = 9.98e^{+00}$ $\beta_{38} = 2.05e^{-03}$ $\beta_{39} = 3.39e^{-02}$ $\beta_{40} = 4.73e^{-02}$ $\beta_{41} = 1.35e^{-03}$ $\beta_{42} = -7.84e^{-02}$ $\beta_{43} = 0.00e^{+00}$ $\beta_{44} = -1.19e^{-06}$ $\beta_{45} = 3.16e^{-04}$ $\beta_{46} = -2.67e^{-02}$	$p_{12} = 0.2609$ $p_{13} = 0.6440$ $p_{14} = 0.0973$ $p_{15} = 0.0015$ $p_{16} = 0.8084$ $p_{17} = 0.8084$ $p_{18} = 0.5147$ $p_{19} = Nan$ $p_{20} = 0.1032$ $p_{21} = 0.0415$ $p_{22} = 0.0952$ $p_{23} = 0.7503$ $p_{24} = Nan$ $p_{25} = 0.1982$ $p_{26} = 0.2476$ $p_{27} = 0.3752$ $p_{28} = 0.8478$ $p_{29} = 0.3190$ $p_{30} = Nan$ $p_{31} = 0.3536$ $p_{32} = Nan$ $p_{33} = 0.4063$ $p_{34} = 0.1972$ $p_{35} = 0.8863$ $p_{36} = 0.6585$ $p_{37} = 0.0790$ $p_{38} = 0.2692$ $p_{39} = 0.4313$ $p_{40} = 0.2055$ $p_{41} = 0.4611$ $p_{42} = 0.0757$ $p_{43} = Nan$ $p_{44} = 0.9992$ $p_{45} = 0.7812$ $p_{46} = 0.3867$

Passos	Modelo	Log - Verossimilhança	-2log L	Estimativa	Valor p
	* Outros, Preço Médio do Frete * Parcelado, Preço Médio do Frete * Outros, Parcelado * Outros.				

Fonte: Autoria Própria.

4.4.2 Adequação do Modelo

Para identificar qual é o mais adequado aos dados do estudo também foi utilizado é o critério de Akaike. Como mencionado anteriormente, o melhor modelo é aquele que tiver o menor valor de AIC.

$$AIC = 2\left[p - l\left(\hat{\theta}_M\right)\right], \tag{101}$$

tal que $l(\hat{\theta}_M)$ é o Log - verossimilhança do modelo.

Tabela 21: Critério de *Akaike* dos todos os Modelos Ajustados.

Modelos	AIC
Modelo 1	12408.02
Modelo 2	12369.22
Modelo 3	12409.83
Modelo 4	12409.25
Modelo 5	12405.77
Modelo 6	12409.84
Modelo 7	12406.16
Modelo 8	12404.35
Modelo 9	12406.35
Modelo 10	12407.57
Modelo 11	12401.46

Modelos	**AIC**
Modelo 12	12404.09
Modelo 13	12408.49
Modelo 14	12409.30
Modelo 15	12344.56
Modelo 16	12388.70
Modelo 17	55978.68
Modelo 18	12351.04
Modelo 19	12343.06
Modelo 20	12353.50
Modelo 21	12346.30
Modelo 22	12349.65
Modelo 23	12343.06
Modelo 24	12344.56
Modelo 25	12338.40
Modelo 26	12333.26
Modelo 27	12344.03
Modelo 28	12342.06
Modelo 29	12340.74
Modelo 30	12327.46
Modelo 31	12347.53

Fonte: Autoria Própria.

De acordo com o critério AIC (Critério de Informação de Akaike) o modelo regressão que se mostra como o melhor modelo é o 30, com as variáveis ticket médio, cartão de crédito, voucher, nordeste, sul, diversão, preço médio do frete, parcelado e outros, que são produtos de outras categorias.

Tabela 22: Coeficientes estimados, erro padrão, valor da estatística e valor p para o Modelo Final Ajustado.

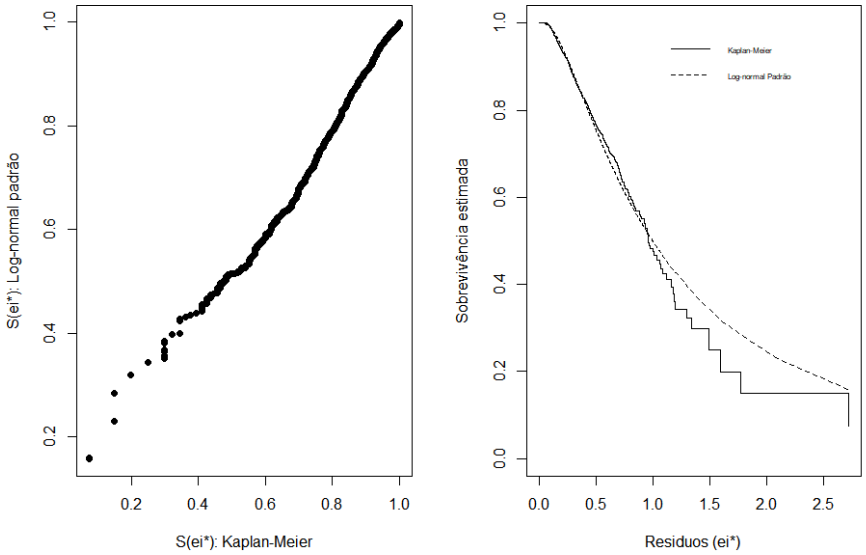
Modelo	Parâmetros	Erro - Padrão	Z	P - valor
Intercepto	$\beta_0 = 7.19ee^{+00}$	1.87e-02	384.47	$p_0 < 2e^{-16}$
Ticket Médio	$\beta_1 = -2.22e^{-04}$	2.83e-05	-7.83	$p_1 = 4.8e^{-15}$
Cartão de Crédito	$\beta_2 = -6.89e^{-02}$	1.75e-02	-3.94	$p_2 = 8.1e^{-05}$
Voucher	$\beta_3 = -1.46e^{-01}$	4.77e-02	-3.06	$p_3 = 0.00223$
Nordeste	$\beta_4 = 6.35e^{-02}$	2.28e-02	2.78	$p_4 = 0.00538$
Sul	$\beta_5 = 3.18e^{-02}$	1.71e-02	1.86	$p_5 = 0.06310$
Diversão	$\beta_6 = 5.84e^{-02}$	1.81e-02	3.23	$p_4 = 0.00122$
Preço Médio do Frete	$\beta_7 = 9.73e^{-04}$	4.17e-04	2.33	$p_5 = 0.01965$
Parcelado	$\beta_8 = 4.68e^{-02}$	1.41e-02	3.32	$p_5 = 0.00091$
Outros	$\beta_8 = 2.49e^{-02}$	1.28e-02	1.95	$p_6 = 0.05133$

Fonte: Autoria Própria.

Para confirmar se o modelo regressão é adequado aos dados, a distribuição dos resíduos na escala logarítmica $(\hat{v}_i)$ deve estar próxima de distribuição padrão. Assim, como os resíduos são censurados, o estimador de Kaplan – Meier deve ser utilizada para estimar a função acumulada dos resíduos.

Logo, aplica-se a transformação exponencial nos resíduos $\hat{v}_i$, isto é, $\hat{e}_i^* = \exp\{\hat{v}_i\}$ que produz resíduos do tipo normal – padrão.

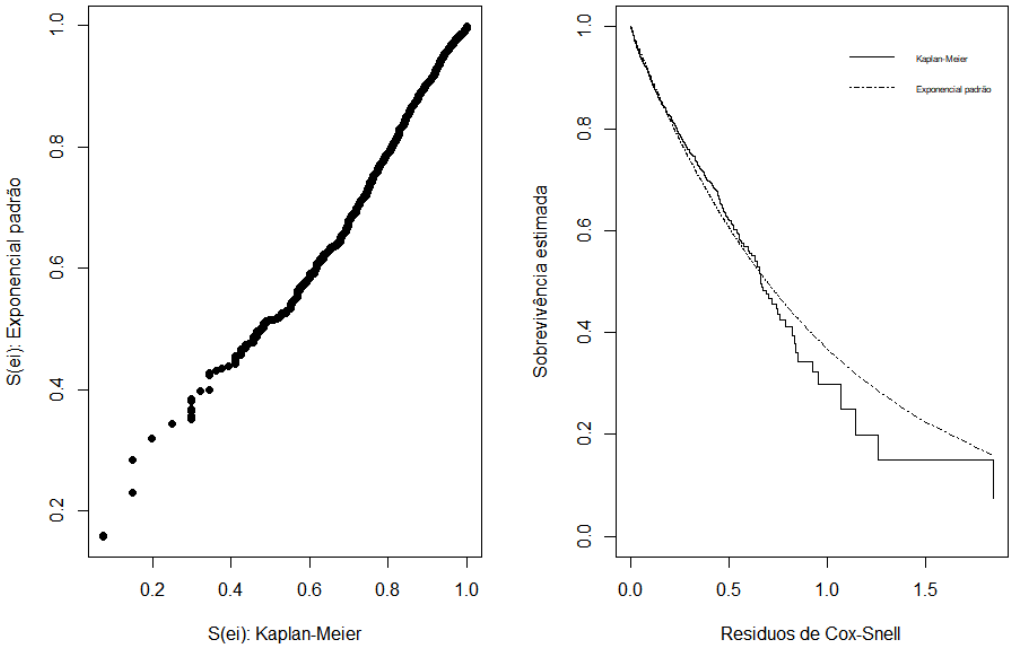
Figura 24: Resíduos $\hat{e}_i^*$ estimada pelo método de Kaplan – Meier e pelo modelo Log-Normal padrão (gráfico à esquerda) e as curvas de sobrevivência estimada (gráfico à direita).



Fonte: Autoria Própria.

Do mesmo modo, os Resíduos de Cox – Snell devem seguir uma distribuição exponencial padrão para que o modelo de regressão Log – Normal possa ser considerado adequado.

Figura 25: Resíduos de Cox-Snell estimadas pelos métodos de Kaplan – Meier e pelo modelo exponencial padrão (gráfico à esquerda) e as curvas de sobrevivência estimadas (gráfico à direita).



Fonte: Autoria Própria.

4.4.3 Interpretação dos Coeficientes

Na tabela 22 são apresentados os coeficientes do modelo que mais representa os dados, ou seja, as variáveis ticket médio, cartão de crédito, voucher, nordeste, sul, diversão, preço médio do frete, parcelado e produtos de outras categorias. Assim, de acordo com o modelo final, para o coeficiente do ticket médio, quando positivo, indica que aqueles consumidores que têm um poder de compra maior tem uma tendência de ficar mais tempo na base de dados, ou seja, de não apresentar o *churn*, para o coeficiente negativo do cartão de crédito, segundo o modelo, indica que para aqueles consumidores que fazem compras com o cartão de crédito ficam menos tempo na base de dados, isto é, até não ocorrer o *churn* e para o coeficiente negativo da variável do tipo de pagamento com voucher mostra que consumidores que fizeram compras com voucher representa que os consumidores ficam menos tempo até não apresentar

o *churn*. Na Tabela 23, é apresentada o tempo médio de cada região, em que se observa que nas regiões norte e nordeste, os tempos médios na base são levemente menores que nas demais regiões.

Tabela 23: Média do Tempo Médio para cada região do Brasil.

Região	**Média do Tempo**
Sul	142, 17 Dias
Sudeste	142, 35 Dias
Norte	129, 67 Dias
Nordeste	130, 53 Dias
Centro - Oeste	144, 17 Dias

Fonte: Autoria Própria.

Na Tabela 24 são apresentados os tickets médios por regiões. Observa-se que nas regiões Norte e Nordeste os tickets médios são maiores do que aquelas regiões onde estão os grandes centros, como a região sudeste.

Tabela 24: Média do ticket médio para cada região do Brasil.

Região	**Média do Ticket Médio**
Sul	R$ 159,55
Sudeste	R$ 148,51
Norte	R$ 222,84
Nordeste	R$ 197,99
Centro - Oeste	R$ 176,34

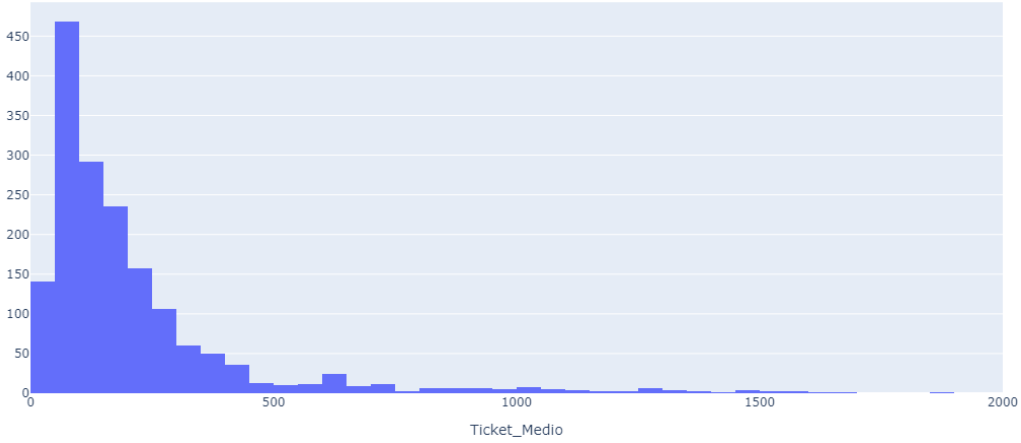
Fonte: Autoria Própria.

O coeficiente positivo da categoria diversão indica que consumidores que compram produtos da categoria diversão ficam mais tempo na base até não apresentar o *churn*. Para o coeficiente positivo do preço médio do frete, indica que para aqueles consumidores que fazem compras com o frete muito caro ficam mais tempo na base de dados, ou, o não *churn* e para o coeficiente positivo da variável parcelado indica que os consumidores que parcelam suas compras demonstram que os clientes ficam mais tempo não apresentar o *churn* e, por fim, o coeficiente dos produtos de outras categorias também ficam mais tempo até não ocorrer o *churn*.

4.5 Discussão dos resultados

A partir da fundamentação teórica apresentada e modelada através dos dados, foi possível estudar o tempo em que os clientes não cheguem ao *churn,* de clientes que compraram no *marketplace* da Olist entre os anos de 2016 a 2018. Por meio das técnicas mencionadas anteriormente foi possível observar que consumidores que possuem um ticket médio alto, tem menos chance de não apresentar o *churn* e ficam mais tempo na base de dados da empresa, além disso, os clientes que fazem compras de produtos com cartão de crédito tendem a ficar mais tempo com a empresa, isto é, de não ocorrer o *churn* e aqueles que utilizam voucher como forma de pagamento nas compras, também ficam mais tempo de não apresentar o *churn* com a companhia. Ademais, de acordo com o coeficiente do modelo final, clientes que são da região nordeste ficam mais tempo sem apresentar o *churn*, e os consumidores que são da região sul também permanecem mais tempo sem apresentar o *churn*, uma justificativa para isso seria o ticket médio da região nordeste é de R$ 197, 51 e da região sul é de R$ 159, 55. Essa diferença entre as duas regiões seria a distribuição do ticket médio nas duas regiões. Nas regiões onde existe os grandes centros como São Paulo e Minas Gerais tem uma maior variabilidade no ticket médio do que nas regiões do Norte e Nordeste do Brasil, onde os valores do ticket médio está muito mais agrupado.

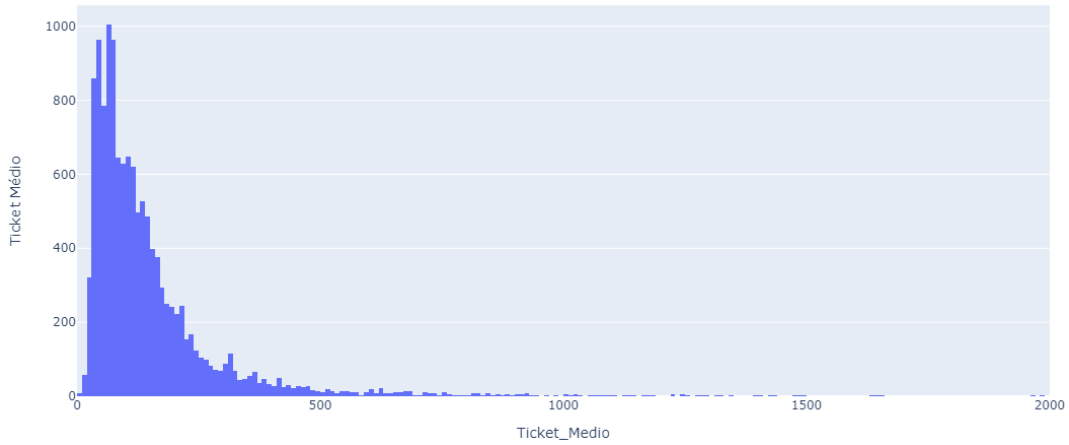
Figura 26: Histograma do Ticket Médio da Região Nordeste.



Fonte: Autoria Própria.

Os valores do ticket médio para a região Nordeste estão concentrados entre R$ 50,00 e $200,00.

Figura 27: Histograma do Ticket Médio da Região Sul.



Fonte: Autoria Própria.

Por outro lado,os valores do ticket médio para a região Sul concentra-se entre R$ 30,00 e R$ 200,00. Aqueles consumidores que compram produtos da categoria diversão ficam mais tempo no conjunto de dados da empresa e de não apresentar o *churn.* Ademais, os clientes que fazem compras de produtos que tem o preço médio do frete com um valor alto tendem a ficar mais tempo com a empresa, ou melhor, não acontecer o *churn*, já os clientes que parcelam suas compras ficam mais tempo na base de dados da empresa, em outros termos, de não acontecer o *churn*. E por fim, consumidores que compraram produtos de outros tipos de categoria como papelaria, beleza e saúde ficam mais tempo com a empresa, ou, de não ocorrer o *churn.*

Sobre a definição do que é censura na pesquisa, isto é, aqueles consumidores que fizeram apenas uma compra no período que está na base de dados, foi feita somente por questões didáticas. Para uma perspectiva do negócio a melhor forma seria truncar o tempo em um certo período aqueles consumidores que estão acima deste período seria censura e os que estão abaixo seriam definidos como falha. Para um estudo futuro este critério poderia ser utilizado sobre a ótica do negócio.

4.5.1 Considerações Finais

É importante citar que a cultura direcionada para o *customer sucess* (sucesso do cliente) pode levar grandes benefícios para às empresas. A visão focada ao consumidor propõe as empresas a se conscientizarem no combate ao *churn,* no qual as leva a ter um grande conhecimento dos perfis dos seus consumidores. Com a crescente competitividade das empresas, a retenção de clientes se torna prioridade nas companhias, pois o investimento na angariação de novos consumidores, onde o sucesso não é garantindo, torna muito mais cara do que manter os clientes cativos.

Por meio do modelo de previsão de quando os consumidores estarão mais propensos ao *churn,* as companhias poderão usufruir do modelo para retirar *insights* e reduzir a taxa de cancelamento e propor planos de ações. Pode possibilitar o entendimento do comportamento do cliente que abandona a empresa, pode possibilitar o conhecimento do seu cliente usando a escuta ativa, investindo na comunicação com os seus consumidores, ou seja, usar o *feedback* como uma oportunidade de melhoria. Uma forma de valorização do cliente que é já utilizada atualmente é o clube de benefícios, que são programas que as empresas utilizam para engajar funcionários e clientes através de promoções e descontos em estabelecimentos. Outro plano de ação que as empresas devem tomar como estratégia é o *onboarding* direcionado para o cliente, que é basicamente apresentar para o consumidor tudo o que o produto comprado pode oferecer, como funciona e como chegar ao seu objetivo final. Outra forma que as companhias já utilizam para reter seus clientes são os cupons de descontos que são oferecidos por meio eletrônico com divulgação de promoções de produtos ou serviços que são de interesse de cada cliente de acordo com o seu perfil. O investimento nas notificações dos aplicativos de compras mostrando os benefícios e valores dos seus produtos e serviços e deixar os consumidores a vontade ao comprar os produtos das companhias.

Dessa forma, será possível que as companhias consigam reter e evitar uma pré – fuga de seus consumidores utilizando o modelo proposto em conjunto com as estratégias mencionadas acima. O modelo de regressão em análise de sobrevivência, permitiram identificar as principais características que podem

fazer com que o tempo de *churn* pode aumentar ou diminuir. Esta informação pode ser extremamente importante para que a área de marketing possa trabalhar diretamente na retenção destes clientes.

Referências

BUSSAB, Wilton; MORETTIN**,** Pedro**. Estatística Básica.** 9 ed. Editora Saraiva. 2017.

LAWLESS, J.F. **Statistical models and methods for lifetime data**, John Wiley & Sons. 1982

NELSON, W. **Applied Life Data Analysis**, Wiley, New York, 2003.

ELANDT-JOHNSON, R.C.; JOHNSON, N.L. **Survival Models and Data Analysis**. Wiley, 2014.

LEE, E.T.; WANG, J. **Statistical Methods for Survival Data Analysis**. 2003.

ALBON, C. **Python Machine Learning Cookbook.** California: O'Reilly, 2018.

Brazilian E-Commerce Public Dataset by Olist. **Kaggle,** 2020. Disponível em: <https://www.kaggle.com/olistbr/brazilian-ecommerce>. Acesso em: 29 dez. 2020

CHALLET, F. **Deep Learning With Python.** New York: Manning, 2018.
COLOSIMO, A. E; GIOLO, R. S. **Análise de Sobrevivência Aplicada.** São Paulo: blucher, 2006.

Ebit. **WebShoopers**. 39 ed, 2019. Disponível em: < https://company.ebit.com.br/webshoppers/webshoppersfree >. Acesso em: 02 jan 2021

Ebit. **WebShoopers**. 40 ed, 2019. Disponível em: < https://company.ebit.com.br/webshoppers/webshoppersfree >. Acesso em: 02 jan. 2021

Ebit. **WebShoopers**. 41 ed, 2020. Disponível em: < https://company.ebit.com.br/webshoppers/webshoppersfree >. Acesso em: 02 jan. 2021

Ebit. **WebShoopers**. 42 ed, 2020. Disponível em: <https://company.ebit.com.br/webshoppers/webshoppersfree>Acesso em: 02 jan. 2021.

Ebit. **WebShoopers**. 43 ed, 2021. Disponível em: < https://company.ebit.com.br/webshoppers/webshoppersfree >. Acesso em: 10 abr. 2022.

Ebit. **WebShoopers**. 44 ed, 2021. Disponível em: < https://company.ebit.com.br/webshoppers/webshoppersfree >. Acesso em: 10 abr. 2022.

Ebit. **WebShoopers**. 45 ed, 2022. Disponível em: < https://company.ebit.com.br/webshoppers/webshoppersfree >. Acesso em: 10 abr. 2022.

Faturamento do e-commerce cresce 56,8% neste ano e chega a R$ 41,92 bilhões. **ABCOMM (Associação Brasileira de Comércio Eletrônico).** Disponível em: https://abcomm.org/noticias/faturamento-do-e-commerce-cresce-568-neste-ano-e-chega-a-r-4192-bilhoes/. Acesso em: 18 dez 2020.

Pandemia do COVID - 19 levou ao fechamento de mais de 75,5 mil lojas no país. **Estados de Minas.** Disponível em: <https://www.em.com.br/app/noticia/economia/2021/03/01/internas_economia,1242013/pandemia-de-covid-19-levou-ao-fechamento-de-mais-de-75-mil-lojas-no-pais.shtml>. Acesso em:17 jul 2022.

GÉRON, A. **Mãos à Obra: Aprendizado de Máquina com Scikit-Learn &TensorFlow.** Rio de Janeiro: Alta Books, 2019.

HARRISON, MATT. **Machine Learning: Guia de Referência Rápida.** São Paulo: Novatec, 2020.

Intuition of Adam Optimizer. **Geeks for Geeks**. Disponivel em: https://www.geeksforgeeks.org/intuition-of-adam-optimizer/. Acesso em: 23 out 2022.

Intro to optimization in deep learning: Momentum, RMSProp and Adam. **PaperspaceBlog**. Disponivel em: https://blog.paperspace.com/intro-to-optimization-momentum-rmsprop-adam/. Acesso em: 23 out 2022.

Gentle Introduction to the Adam Optimization Algorithm for Deep Learning. **Machine Lerning Mistery.** Disponível em: https://machinelearningmastery.com/adam-optimization-algorithm-for-deep-learning/. Acesso em: 23 out 2022.

A Gentle Introduction to the Rectified Linear Unit (ReLU). **Machine Lerning Mistery**. Disponível em: https://machinelearningmastery.com/rectified-linear-activation-function-for-deep-learning-neural-networks/. Acesso em: 23 out 2022.

A Gentle Introduction To Sigmoid Function. **Machine Lerning Mistery**. Disponível em: https://machinelearningmastery.com/a-gentle-introduction-to-sigmoid-function/. Acesso em: 23 out 2022.

Binary crossentropy.**ML Plataform – Knowledge Center**. Disponível em: < https://peltarion.com/knowledge-center/modeling-view/build-an-ai-model/loss-functions/binary-crossentropy>. Acesso em: 23 out 2022.

Understanding binary cross-entropy / log loss: a visual explanation. **Towards Data Science**. Disponível em: <

https://towardsdatascience.com/understanding-binary-cross-entropy-log-loss-a-visual-explanation-a3ac6025181a>. Acesso em: 23 out 2022.
Nielse, M.A. **Neural Network and Deep Lerning.** Determination Press, 2015.

Moretin, P.A; Singer. J.M. **Estatística e Ciência de Dados**.LTC | Livros Técnicos e Científicos Editora Ltda, jun/2022.

A história do e-commerce no Brasil: entenda como o modelo cresceu e se transformou no país. **Com School.** Disponível em: https://news.comschool.com.br/a-historia-do-e-commerce-no-brasil/. Acesso em: 14 jan 2023.

Glossário

Churn - Casos em que um cliente para de comprar ou utilizar um produto ou serviço de uma empresa.

Customer Churn - *Rotatividade* de Clientes.

Gross Merchandise Volume (GMV) - Refere-se ao valor total de mercadorias vendidas em um e-commerce durante um período.

Marketplace - É um espaço de compra e venda de produtos.