UNIVERSIDADE DE SÃO PAULO

ESCOLA DE ENGENHARIA DE LORENA

TAYRO STRINGARI DE TOLEDO

Utilização de análises estatísticas para estimar a propensão de compra de clientes do segmento de automóveis

Lorena - SP

2020

TAYRO STRINGARI DE TOLEDO

Utilização de análises estatísticas para estimar a propensão de compra de clientes do segmento de automóveis

Trabalho de conclusão de curso apresentado à Escola de Engenharia de Lorena - Universidade de São Paulo como requisito parcial para conclusão da Graduação do curso de Engenharia Química.

Orientadora: Prof.ª Dr.ª Mariana Pereira de Melo

Lorena - SP

2020

Ficha catalográfica elaborada pelo Sistema Automatizado
da Escola de Engenharia de Lorena,
com os dados fornecidos pelo(a) autor(a)

Toledo, Tayro Stringari de
Utilização de análises estatísticas para estimar a propensão de compra de clientes do segmento de automóveis / Tayro Stringari de Toledo; orientadora Mariana Pereira de Melo. – Lorena, 2020.
102 p.

Monografia apresentada como requisito parcial para a conclusão de Graduação do Curso de Engenharia Química – Escola de Engenharia de Lorena da Universidade de São Paulo. 2020

1. Modelos de propensão. 2. Regressão logística múltipla. 3. Análise discriminante. 4. Propensão de compra. 5. Modelagem de dados. I. Título. II. Melo, Mariana Pereira de, orient.

AGRADECIMENTOS

Primeiramente, gostaria de agradecer à minha família, em especial ao meu pai, Paulo, por sempre ter me dado apoio e ter me ensinado os caminhos corretos a serem seguidos na vida e à minha avó, Izolda, por todo carinho e suporte que me permitiu estudar e concluir uma graduação. Ambos nunca duvidaram da minha capacidade e sou muito grato por tudo que me ensinaram e me proporcionaram.

Aos meus irmãos, por todos os bons momentos que passamos, vocês me mostraram que é possível ser feliz, apesar de todas as adversidades e que é através da dedicação que podemos alcançar nossos objetivos.

À Thainara, que me acompanhou durante uma boa parte da graduação e durante todo o desenvolvimento desse trabalho, sempre me apoiando e ajudando quando precisei.

Aos meus amigos e colegas de trabalho, em especial Rafhael, Thaís e Beatriz por me permitirem desenvolver esse trabalho incrível e me guiarem pelos caminhos corretos, não só na execução do trabalho, mas no meu desenvolvimento dentro da empresa. Vocês são incríveis!

Aos meus colegas e professores da Escola de Engenharia de Lorena e de toda Universidade de São Paulo por todos os conhecimentos que me proporcionaram tanto na área acadêmica, quanto no meu desenvolvimento pessoal e profissional.

À minha orientadora, Mariana Pereira de Melo, por todo suporte no desenvolvimento desse trabalho, pela paciência, sabedoria, liderança, pela compreensão que demonstrou durante todas as nossas conversas e pela prontidão em me ajudar sempre que necessário.

“No fim tudo dá certo, e se não deu certo
é porque ainda não chegou ao fim.”

Fernando Sabino

RESUMO

TOLEDO, T. S. **Utilização de análises estatísticas para estimar a propensão de compra de clientes do segmento de automóveis.** 2019. 49 f. Monografia (Graduação em Engenharia Química) – Escola de Engenharia de Lorena, Universidade de São Paulo, Lorena – SP, 2019.

Este trabalho foi desenvolvido com o objetivo de utilizar técnicas de modelagem estatística para calcular a probabilidade de compra de automóveis e, consequentemente, identificar clientes potenciais, através da análise de regressão logística múltipla e da análise discriminante. As informações utilizadas para as análises foram cedidas por uma empresa do segmento de varejo automotivo e financeiro, e os dados são referentes às informações reais de cadastro e histórico de compras dos clientes da empresa. O tema se justifica pela dificuldade da empresa em selecionar em seu banco de dados quais clientes estão mais propensos a adquirirem um novo veículo, a fim de direcionar os esforços das equipes de vendas e obter uma melhor conversão nas campanhas. Por conta da quantidade de informações para análise, foram utilizadas também técnicas de *Data Mining* para identificação e preparação das variáveis para modelagem, auxiliando em todo processo de manuseio de dados. A qualidade dos ajustes e os resultados obtidos pelos modelos foram comparados, concluindo que ambas as metodologias se mostraram eficazes e capazes de atingir os objetivos do trabalho, ainda que o método logístico apresente vantagens em relação ao discriminante para aplicação nessa situação. Através destes resultados, acredita-se no potencial avanço a ser obtido pela empresa utilizando análises estatísticas para direcionamento de suas campanhas, quando comparado ao status quo.

Palavras Chave: modelos de propensão, regressão logística múltipla, análise discriminante, propensão de compra, segmento de automóveis, modelagem de dados.

ABSTRACT

TOLEDO, T. S. **Use of statistical analysis to estimate the propensity of buying clients on automobile field.** 2019. 49 f. Monograph (Graduation in Chemical Engineering) – Escola de Engenharia de Lorena, Universidade de São Paulo, Lorena - SP, 2019.

This work was made using statistical modeling techniques as means to calculate the probability of cars being bought and, consequently, to identify potential customers through multiple logistic regression analysis and discriminant analysis. The information used for the analyzes was provided by a company in the automotive and financial retail segment, and the data refers to real information of customers' registration and purchase history. The theme is justified by the company's difficulty on selecting in its database which customers are more likely to acquire a new vehicle in order to direct the efforts of the sales teams, and to get a better conversion rate on its campaigns. Due to the amount of information to be analyzed, Data Mining techniques were also used to identify and prepare the variables for modeling, assisting in all data handling processes. The quality of the adjustments and the results obtained by the models were compared, concluding that as the methodologies are allowed and able to achieve the objectives of the work, even though the logistic method represents advantages in relation to the discrimination by application in this case. Through these results, we believe in a significant growth potential, using statistical analyzes to target their campaigns, as to when compared to the status quo.

Keywords: propensity models, regression, logistics, discriminant analysis, buying prospection, automobile segment, data modeling.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

IoT	Internet das coisas
ROC	Reciever Operating Characteristic
AUC	Area under the ROC curve
MDA	Análise discriminante múltipla
TEE	Taxa Estimada de Erro
ECM	Expected cost off missclassification
IV	Information Value
WOE	Weight of Evidence
PF	Pessoa Física
PJ	Pessoa Jurídica

SUMÁRIO

1 – INTRODUÇÃO

A era da revolução digital, Internet das coisas (IoT), mídias sociais e da *Big Data* tem gerado um recorde de geração e acúmulo de dados estruturados e não estruturados (HASHEM, 2014), o que despertou o interesse de pesquisadores, governos e também das entidades privadas por identificaram nesse grande volume de informações disponíveis e armazenados novas oportunidades de comércio, inovação e engenharia social (EKBIA, 2014). A importância e o potencial dos dados para as instituições pode ser percebida quando, em 2013, o *World Economic Forum* (Fórum Econômico Mundial) definiu os dados pessoais como uma nova classe de ativos, chegando a ser apelidado como "O novo petróleo" por comentaristas (por exemplo, KUNEVA, 2009; ROTELLA, 2012). Inúmeros debates sobre o tema foram realizados pelo mundo, levantando questões a respeito da privacidade das informações, segmentação, segurança e como utilizar seu potencial para fins comerciais, saúde, infraestrutura, alimentação, política e proteção, por exemplo.

Em seu estudo, CASTELLANO, FORTUNATO e LORETO (2009) apresentaram uma série de modelos e simulações utilizando dados e com base em técnicas de física estatística que seriam aplicados em diferentes áreas, como na política, para previsão de comportamento e intenção de voto em eleições, na área social, para previsão de crimes, tumultos, mudanças de gosto e/ou comportamento, identificar doenças e epidemias, e na área econômica, para previsão de tendências, acertos nos mercados, direcionamento de *marketing*, por exemplo.

Dentro das oportunidades para o setor econômico mais voltadas para empresas, podemos citar a utilização de dados para direcionar, determinar e prever comportamentos, de forma a antecipar ações, oportunidades e mudanças no mercado visando atender as necessidades e desejos dos consumidores. Para se ter uma ideia do potencial, as companhias de internet acumulam um grande volume de informações, como o Google, que processa todos os dias mais de 24 petabytes de dados. O Facebook, empresa fundada em 2004, recebe mais de 10 milhões de fotos novas carregadas a cada hora e seus membros clicam no botão de "*like"* ou comentam uma publicação mais de 3 bilhões de vezes por dia, dados esses que podem ser utilizados para minerar e entender melhor sobre os interesses e

comportamento dos usuários (VIKTOR MAYER-SCHÖNBERGER, 2013). Esse tipo de informação tem sido analisada e utilizada nas organizações para construir informações relevantes e que poderão ser aproveitadas pelos seus negócios (UN GLOBAL PULSE, 2012) para design de novos produtos e serviços, entender o mercado e definir suas estratégias quase que em tempo real frente às mudanças de padrões e novas oportunidades (DAVENPORT; BARTH; BEAN, 2012).

Destacam-se, nesse momento atual, as empresas que conseguem fazer uso dessas informações relevantes para seus negócios, tomando decisões assertivas, criando experiências únicas para seus usuários e acrescentando valor aos seus negócios. Entretanto, para a maioria das entidades esse ainda é um processo novo, mas que é necessário para se adequar às necessidades do mercado e tirar maior proveito das informações que possuem. Muitas instituições vêm tentando se adequar a essa nova realidade, tomando medidas para melhorar seus sistemas de coleta e armazenamento de dados, segurança, qualidade e confiabilidade das informações obtidas e também para explorar os dados que possuem a fim de trazer receita para os negócios.

Com as novas tecnologias disponíveis, as negociações se tornaram mais dinâmicas, trazendo oportunidades a todo momento e, as entidades que não entendem a necessidade de se reinventar, colocando o consumidor como foco, muitas vezes à frente do lucro, a fim de solucionar suas necessidades, estão perdendo espaço no mercado ou sucumbiram à revolução digital e a era da *Big Data*. A Kodak, por exemplo, empresa do ramo da fotografia, fundada em 1880 e registrada como marca em 1888, inventora da tecnologia digital para as máquinas fotográficas, líder no segmento na década de 80 e 90 e, responsável por mais de 80% das vendas nesse período (Terra, 2012) foi perdendo mercado até decretar falência em 2012, por não apostar no potencial da fotografia digital, produto desenvolvido por eles que tornou muito mais simples o processo de captura de imagens para os usuários, acreditando que seria ruim para seus negócios acabar com o uso de filmes fotográficos e máquinas tradicionais. No outro extremo, inúmeros exemplos de sucesso apareceram do início do milênio até os tempos de hoje, empresas estas denominadas num primeiro momento como *Startups* (empresas emergentes, inovadoras, que utilizam muita tecnologia em seus processos e, por conta disso, conseguem escalar facilmente seus negócios) e que

hoje valem milhões, ou até bilhões adotando como foco principal de seu modelo de negócios solucionar uma necessidade dos consumidores. *Cases* de empresas como *Uber, Netflix, Airbnb, Nubank*, etc., que até 10 anos atrás eram pouco conhecidas, mas que mudaram totalmente a forma como o mercado atua no segmento em que estão inseridos, são exemplos da importância de dar voz ao cliente, entender e acompanhar seu comportamento, valorizar suas necessidades e buscar sua satisfação. Nem mesmo as grandes entidades, consolidadas no mercado e com um modelo de negócios lucrativo estão livres das mudanças e passam a temer não apenas a concorrência existente, mas também a ameaça de novas empresas que podem vir a surgir de uma oportunidade não explorada e revolucionar o segmento.

Por conta disso, até as grandes entidades estão se reinventando. Recentemente, o Grupo Pão de Açúcar (GPA), maior grupo varejista e de distribuição do Brasil viu uma oportunidade de recompensar e fidelizar seus clientes, além de economizar com campanhas de *Marketing* para seus produtos sem ter que descontar da sua margem os descontos oferecidos. A empresa conseguiu reestruturar seu programa de relacionamento e fidelidade, lançado nos anos 2000, utilizando tecnologia, o método de *Data-Driven Marketing* (Marketing orientado por dados), realizando análises de perfis, segmentação e análise de comportamento para entender seus consumidores e apresentar descontos e ofertas 100% personalizadas no aplicativo “Meu Desconto” para estes, utilizando como base o seu histórico de compras. E não foi só isso, a empresa viu no volume gasto com *Marketing* pelas indústrias e fornecedores de seus produtos a oportunidade de oferecer a plataforma para que o próprio fornecedor a utilizasse para criar suas ofertas e segmentar os grupos de clientes que teriam acesso à ela, de forma a permitir que os consumidores fossem impactados por ofertas de produtos relevantes à eles. Com isso as indústrias vendem seus produtos, apresentam outros produtos do seu portfólio, conseguem maior penetração nas campanhas, economizam com campanhas de *Marketing* e fidelizam os clientes da marca. Com a estratégia, o GPA trouxe atratividade para seu programa de fidelidade e conseguiu oferecer para seus clientes melhores condições de compra dos produtos que costumam comprar e, como resultado, a empresa alavancou suas vendas e reduziu gastos com campanhas de *Marketing.* (MANZINI, 2017)

Inseridas nesse contexto a utilização de inteligência de dados para segmentar e focar determinadas ações em um grupo específico de consumidores têm se mostrado efetiva tanto para redução de custos, quanto para aumento de vendas. Ganham destaque ferramentas, técnicas e modelos estatísticos que surgem como forma de traduzir os dados disponíveis em informações relevantes a fim de identificar, segmentar e caracterizar, por exemplo, qual público é mais aderente a um produto específico e deverá ser impactado por uma determinada campanha, direcionando então os esforços da equipe de vendas para criação de conteúdo e condições baseadas nessas informações. Para isso, é fundamental a utilização das informações disponíveis sobre os consumidores, sejam armazenadas na forma de dados de pessoais (cadastro) e registros das negociações (transacionais) da empresa quanto os disponíveis no *Word Wide Web* (Web), a fim de possuir o máximo de informações sobre os clientes. Entretanto, por conta do aumento significativo no volume de dados, apenas técnicas estatísticas são insuficientes ou incapazes de analisar a grande quantidade de informações em um prazo de tempo razoável, surgindo como resposta a estes problemas as ferramentas e técnicas de *data mining* (Mineração de dados). De acordo com Weis e Indurkhya (1998, p. 1, tradução nossa)

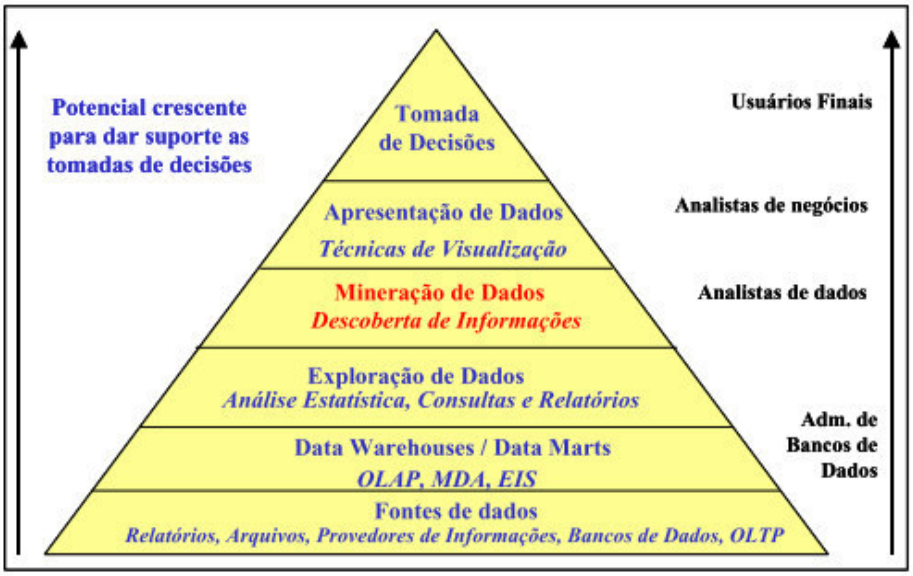
> Data mining é a busca por informações valiosas em grandes volumes de dados. É um esforço cooperativo entre homem e computadores. Homens projetam os bancos de dados, descrevem os problemas e selecionam os objetivos. Os computadores analisam os dados e procuram por padrões que combinem com as metas estabelecidas.

Dessa forma, as técnicas de *data mining* são de muita utilidade no processo de extração e manuseio de dados em meio ao grande volume disponível, resultando na apresentação de informações valiosas para a tomada de decisão dentro das empresas. De acordo com Berry e Linoff (2004), podemos utilizar os conhecimentos e técnicas de *data mining* para responder inúmeros problemas de interesse intelectual, econômico e comercial, formulados a partir de seis finalidades ou resultados esperados: Classificação, Estimação, Predição, Agrupamento ou Associação, *Clustering* e Descrição.

A Figura 1, retirada de Côrtes, Porcaro e Sérgio (2002), apresenta as etapas do processo de manuseio de dados, desde a etapa de coleta de informações, sistemas

utilizados para essas funcionalidades, até exploração, mineração, apresentação e, por fim, tomada de decisão.

Figura 1- Mineração de dados no contexto da inteligência de negócios.



Fonte: Côrtes, Porcaro e Sérgio, 2002.

No estudo objeto do presente trabalho, foram utilizados os dados fornecidos por uma empresa do setor financeiro e de varejo automotivo referentes às informações de cadastro (nome, sexo, data de nascimento etc.) e de compras de produtos de seus clientes. Por conta da variedade de produtos disponíveis na empresa, os estudos foram focados nos dados de clientes de automóveis, e foram utilizadas técnicas estatísticas e de *data mining* com o objetivo de identificar quais clientes eram mais propensos a realizar a compra do automóvel num período de tempo definido, fornecendo um direcionamento para a equipe de vendas.

Os dados foram selecionados e preparados, utilizando softwares de manuseio de dados, antes da realização das análises, buscando minimizar inconsistências (*noise data)* nas informações. Para a análise de dados, foram utilizadas duas técnicas estatísticas amplamente conhecidas: a análise de discriminante e o modelo de regressão logística múltipla.

1.1 Objetivo

Este trabalho foi um estudo de caso em que a proposta era utilizar técnicas estatísticas para elaboração de um modelo para definir a probabilidade dos clientes realizarem uma nova compra de automóvel a partir de fatores demográficos (sexo, estado e local de venda, por exemplo) e do histórico financeiro desses clientes, avaliando e mensurando a importância de cada uma das variáveis, além de identificar quais clientes são mais propensos a adquirirem o novo produto.

1.2 Objetivos Específicos

Como objetivos específicos, foram traçadas as seguintes metas:

- Entender e aplicar técnicas de *data mining* para seleção, limpeza e transformação dos dados a serem utilizados;
- Elaborar um modelo de propensão de compra baseado em análise de regressão logística múltipla para determinado produto de automóveis;
- Elaborar um modelo de propensão de compra baseado em análise discriminante para determinado produto de automóveis;
- Comparar a qualidade de ajuste dos modelos obtidos pela análise de regressão logística múltipla e análise discriminante, a fim de identificar qual o mais adequado;
- Utilizar o modelo mais adequado para compreender o perfil dos clientes, avaliando quais as características tornam o cliente mais propenso ou menos propenso a adquirir o produto em questão.

Para alcançar os objetivos do estudo, foram utilizados dados reais de uma empresa do segmento financeiro e de varejo automotivo, que serviram como base para realização da Análise de Discriminante e também para construção do modelo de Regressão Logística Múltipla. Ambas as ferramentas são metodologias adequadas para predição de variáveis categóricas/binárias.

1.3 Justificativa

Por conta do porte, dos segmentos e do tempo de atuação no mercado, a empresa em questão possui muitos clientes em sua carteira. Apesar de haver conhecimento prévio de que grande parte desses consumidores adquirirem o produto determinada recorrência, atualmente não é claro para a equipe de *marketing* e vendas quais clientes estão no momento ideal de compra. Diante disso, surge a necessidade de priorizar e orientar os esforços desta equipe a fim de aumentar os ganhos obtidos.

A tarefa de gerenciamento da carteira de clientes em uma empresa, muitas vezes, não é feita de forma eficaz. Na equipe comercial, por conta do volume de trabalho, quantidade de clientes ou tempo disponível, poucos buscam identificar e abordar oportunidades de vendas dentro dos seus antigos clientes, e, quando o fazem, nem sempre são assertivos nas suas ações. Se por um lado, deixar de abordar o cliente desperdiça oportunidades valiosas de rendimento para a empresa, por outro lado, a abordagem indiscriminada causa desconforto dos consumidores e desperdiça tempo e recursos. Sendo assim, nota-se a importância de priorizar as abordagens, direcionando à equipe comercial, um grupo selecionado de clientes que possuem o perfil e que estão propensos a adquirirem um novo produto naquele momento.

Dessa forma, os modelos e técnicas estatísticas de predição surgem como opção para identificar quais os clientes, dentre o grande número disponível na carteira, são mais propensos a adquirirem um novo produto. Consequentemente, um maior direcionamento das equipes de vendas resultará em um aumento das vendas e redução de tempo e recursos da equipe.

2 – REVISÃO DA LITERATURA

2.1 Análise de Regressão

Podemos recorrer à estatística para compreender problemas nas áreas de ciências naturais, engenharia, economia e finanças etc., utilizando a relação entre as variáveis envolvidas no fenômeno para extrair análises valiosas sobre comportamento, variação e até prever uma determinada resposta futura a partir dos dados fornecidos. Uma das técnicas mais utilizadas para tais objetivos são as análises de regressão, pois permitem criar modelos empíricos que relacionem uma variável de interesse com uma ou mais variáveis explicativas obtidas a partir de observações, coleta de dados e experimentos do fenômeno em questão.

O relacionamento entre as variáveis analisadas e a resposta pode ocorrer de várias maneiras, entretanto, o que vai definir qual análise de regressão será utilizada é o tipo da variável resposta, que pode ser quantitativa (medida em valores) ou qualitativa (respostas binárias, como “sucesso ou fracasso, “compra ou não compra”) (MONTOGOMERY, D. C.; RUNGER, 2011). Essa diferença no tipo de variável resposta influencia tanto na escolha do modelo quanto nas premissas usadas para a elaboração do mesmo, sendo assim, é importante entender os conceitos relacionados a esses tipos de análise para definir qual será o método adotado.

Nos casos em que a variável resposta é quantitativa, geralmente, é utilizada a análise de regressão linear, onde considera-se que o valor da resposta seja uma função linear das demais variáveis relevantes envolvidas no fenômeno, mais um termo de erro aleatório (MONTOGOMERY, D. C.; RUNGER, 2011). Nesta análise, o método dos mínimos quadrados é utilizado para se obter uma equação linear que relacione a variável resposta ou dependente à(s) variável(eis) independente(s) ou regressor(es). Essa equação permite aproximar ou prever os valores da variável dependente/resposta a partir dos valores da(s) variável(eis) dependente(s) dentro de um determinado intervalo de confiança especificado durante a análise. Como nesse trabalho, a resposta a ser determinada é qualitativa (compra ou não compra),

e uma vez que tal análise não é eficiente para atingir os objetivos esperados, não serão apresentados conceitos mais aprofundados relacionados à regressão linear.

Para os casos em que a variável resposta é qualitativa, geralmente, é utilizada a análise de regressão logística, onde considera-se a variável resposta (dependente) seja uma função probabilística não linear das variáveis relevantes envolvidas no fenômeno, mais um termo de erro aleatório (MONTOGOMERY, D. C.; RUNGER, 2011). Nesta análise, o método da máxima verossimilhança é utilizado para estimar os parâmetros do modelo que permite prever a probabilidade da variável dependente ou resposta ser sucesso ou fracasso (0 ou 1) a partir das variável(eis) independente(s). Neste trabalho, faremos o uso deste método para determinar a probabilidade de compra de automóveis pelos clientes.

No item seguinte, serão apresentados conceitos relacionados à construção do modelo de regressão logística.

2.2 Regressão Logística

Modelos de regressão logística é conhecido como o método tradicional ou padrão utilizado para relacionar uma variável resposta binária (qualitativa) à uma ou mais variáveis relevantes ao fenômeno. Tal método é amplamente utilizado em estudos nas mais diversas áreas, pois permite explicar determinados fenômenos e, muitas vezes, até prever eventos futuros, servindo como suporte para a tomada de decisão.

Como já discutido no item anterior, a utilização desse método faz-se necessária quando o modelo de regressão linear não é apropriado para o fenômeno a ser estudado. Pela própria característica da variável resposta, na regressão logística a resposta fica restrita a apenas duas possibilidades (sucesso e fracasso), o que implica que os erros associados ao modelo também ficam restritos a apenas dois valores. Isto acaba inviabilizando a utilização da análise de regressão linear, uma vez que esta possui como suposição que os erros do modelo seguem uma distribuição normal.

Segundo Montogomey e Runger (2011), os modelos de regressão logística simples são baseados na Equação 1:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (1)$$

onde i=1, 2, ..., n, e a variável resposta Y_i assumindo os valores de 0 ou 1. Considerando-se que a variável resposta Y_i seja uma variável aleatória de Bernoulli, a probabilidade de acontecer tais possibilidades pode ser vista na Tabela 1:

Tabela 1 - Distribuição de probabilidades

Y_i	**Probabilidade**
1	$P(Y_i = 1) = \pi_i$
0	$P(Y_i = 0) = 1 - \pi_i$

Fonte: Montogomey e Runger, 2011.

Conforme Montogomey e Runger (2011), uma vez que o valor do erro seja nulo, ou seja, $E(\epsilon_i) = 0$, o valor esperado da variável resposta Y_i é apresentado na Equação 2:

$$E(Y_i) = 1(\pi_i) + 0(1 - \pi_i) \quad (2)$$

resultando em (Equação 3):

$$Y_i = \beta_0 + \beta_1 x_i = \pi_i \quad (3)$$

Sendo assim, o modelo apresenta como resultado a probabilidade de a variável resposta Y_i ter valor 1.

Como já foi apresentado, dado o fato de a variável resposta do modelo não ser linear, mas sim binária (sucesso ou fracasso), os erros associados ao modelo também se comportam dessa forma, pois existem apenas duas possibilidades de resposta, conforme mostrado a seguir nas Equações 4 e 5:

$$\epsilon_i = 1 - (\beta_0 + \beta_1 x_i), \text{ quando } Y_i = 1 \quad (4)$$

$$\epsilon_i = -(\beta_0 + \beta_1 x_i), \text{ quando } Y_i = 0 \quad (5)$$

Seguindo o mesmo princípio, a variância dos erros também não é constante, sendo uma média do valor esperado, como segue abaixo, nas Equações 6 e 7:

$$\sigma^2 = E\{Y_i - E(Y_i)\}^2 = (1 - \pi_i)^2\pi_i + (0 + \pi_i)^2(1 - \pi_i) = \pi_i(1 - \pi_i) \quad (6)$$

ou apenas:

$$\sigma^2 = E(Y_i)[1 - E(Y_i)] \quad (7)$$

Este é o segundo argumento que demonstra a inviabilidade do uso da regressão linear para modelagem de uma variável resposta binária uma vez que, na análise de regressão linear, supõe-se que a distribuição condicional dos erros segue uma Normal, com média 0 e variância constante.

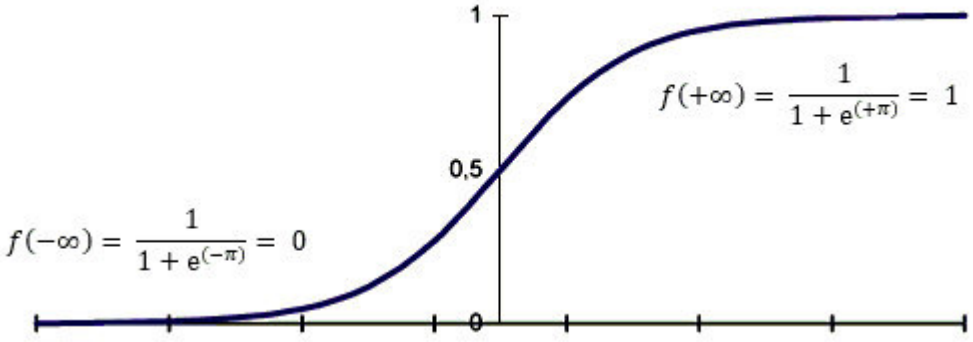
Por fim, a variável resposta de resposta fica restrita ao intervalo de 0 à 1, como apresentado abaixo, na Equação 8:

$$0 \leq E(Y_i) = \pi_i \leq 1 \quad (8)$$

diferentemente da análise de regressão linear, cuja variável resposta pode assumir qualquer valor.

Ainda segundo Montogomey e Runger (2011), geralmente, quando a variável resposta é binária, existem evidências indicando que a forma da função da resposta deve ser suposta como não linear. Para esses casos, normalmente emprega-se a função de resposta logit, uma função monotonicamente crescente ou decrescente, em forma de S (Figura 2). Para utilização da função, deve-se definir a quantidade de variáveis independentes utilizadas, como apresentado a seguir.

Figura 2 - Função Logística



Fonte: Arquivo pessoal.

2.2.1 Regressão Logística Simples

A regressão logística simples é utilizada para modelar uma variável resposta binária (qualitativa) em função de apenas uma variável independente ou explicativa. Segundo Montogomey e Runger (2011), a função de resposta *logit* que representa esse tipo de modelo é apresentada na Equação 9:

$$E(Y_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)} = \frac{1}{1+\exp[-(\beta_0 + \beta_1 x_i)]} \quad (9)$$

Sendo assim, o modelo que relaciona a variável dependente à variável independente é dado por (Equação 10):

$$\frac{E(Y_i)}{1-E(Y_i)} = \exp(\beta_0 + \beta_1 x_i) \quad (10)$$

Para o caso da regressão logística simples, a grandeza $\exp(\beta_0 + \beta_1 x_i)$ apresentada na Equação 10, é definida como razão de chances (*odds ratio*), que representa a probabilidade de sucesso dividido pela probabilidade de fracasso, como apresentado na Equação 11:

$$\exp(\beta_0 + \beta_1 x_i) = \frac{E(Y_i)}{1 - E(Y_i)} = \frac{\pi_i}{1 - \pi_i} = \frac{P(Y_i=1)}{P(Y_i=0)} \quad (11)$$

Como pode ser observado, o logaritmo natural da razão de chances é uma função linear da variável regressora, com inclinação β_1. Na prática, isso significa que a razão de chances varia $\exp(\beta_1)$ para cada aumento na unidade da variável independente x_i.

Nesse modelo de regressão logística, os parâmetros são geralmente estimados pelo método de máxima verossimilhança. Para mais detalhes, consultar Montogomey, Peck e Vining (2006).

2.2.2 Regressão Logística Multivariada

A regressão logística multivariada é utilizada para modelar uma variável resposta binária (qualitativa) em função de outras variáveis independentes ou explicativas. Nesse caso, as funções empregadas permanecem as mesmas da regressão logística simples, com a adição das kn variáveis independentes relevantes no modelo, conforme apresentado nas Equação 12:

$$E(Y_i) = \frac{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)}{1 + \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)} = \frac{1}{1 + \exp[-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)]} \quad (12)$$

A Equação 12 pode ser reescrita, conforme apresentado em Kleinbaum e Klein (2002):

$$E(Y_i) = \frac{\exp(\beta_0 + \sum \beta_i x_i)}{1 + \exp(\beta_0 + \sum \beta_i x_i)} = \frac{1}{1 + \exp[-(\beta_0 + \sum \beta_i x_i)]} \quad (13)$$

Sendo assim, o modelo que relaciona a variável dependente às variáveis independentes na regressão logística com função de resposta *logit* é apresentada na Equação 14:

$$\frac{E(Y_i)}{1-E(Y_i)} = \exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k) \quad (14)$$

onde k representa o número de variáveis independentes relevantes ao modelo.

Análogo à regressão logística simples, a grandeza $\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$ apresentada na Equação 14, é definida como razão de chances (*odds ratio*), que representa a probabilidade de sucesso dividido pela probabilidade de fracasso, como apresentado na Equação 15:

$$\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k) = \frac{E(Y_i)}{1-E(Y_i)} = \frac{\pi_i}{1-\pi_i} = \frac{P(Y_i=1)}{P(Y_i=0)} \quad (15)$$

Nesse caso, o logaritmo natural da razão de chances é uma função linear do somatório das k variáveis regressoras, com inclinação β_k. Na prática, isso significa que a razão de chances varia $\exp(\beta_1)$ para cada aumento na unidade da variável independente x_1, mantendo-se constante as demais variáveis; a razão de chances varia $\exp(\beta_2)$ para cada aumento de uma unidade na variável independente x_2, mantendo-se constante as demais variáveis, e assim por diante, ou ainda, que a razão de chances varia $\exp(\sum \beta_k)$ para cada aumento na unidade da variável independente $\sum x_k$.

Como no modelo de regressão logística simples, os parâmetros são geralmente estimados pelo método de máxima verossimilhança. Para mais detalhes, consultar Hosmer e Lemeshow (2000) e/ou Kleinbaum e Klein (2002).

2.2.3 Interpretação dos resultados e avaliação da qualidade do ajuste

Como já apresentado nos itens anteriores, os modelos de regressão logística (simples e múltipla) fornecem como resposta valores de $E(Y_i)$ que variam de 0 a 1. Esses valores permitem distinguir dois grupos de interesse, sucesso e fracasso, baseados nas probabilidades π_i, calculadas a partir de valores específicos das variáveis explicativas. Para tal, é necessário estabelecer um ponto de corte (*classification cutoff*), que define a partir da resposta o limite entre o sucesso (valores iguais ou acima do ponto) e fracasso (valores abaixo do ponto). A seguir, é apresentado um exemplo desta regra de decisão onde estabeleceu-se como ponto de corte o valor de 0,5:

Tabela 2 – Análise dos resultados

Y_i	**Probabilidade**	**Resultado**
1	$\pi_i \geq 0{,}5$	Sucesso
0	$\pi_i < 0{,}5$	Fracasso

Fonte: Arquivo pessoal.

O ponto de corte é um ajuste definido pelo desenvolvedor do modelo e varia de acordo com os dados utilizados e a finalidade da previsão, levando em conta a relação risco x retorno, determinado pelo utilizador.

A qualidade do ajuste é um item importante quando se espera que os resultados fornecidos pelo modelo sejam os mais próximos possíveis dos resultados reais. Para isso, alguns métodos podem ser utilizados para avaliar a qualidade do modelo, permitindo determinar se ele se encaixa nos objetivos para os quais foram construídos. É importante salientar que essa etapa é realizada após a construção do modelo, servindo para avaliar os resultados obtidos.

Segundo Hosmer e Lemeshow (2000), as abordagens para avaliação do ajuste incluem basicamente (1) cálculo e avaliação das medidas gerais do ajuste, (2) análise dos componentes estatísticos individuais de resumo, geralmente

graficamente, e (3) exame de outras medidas de diferença entre os valores reais e obtidos pelo modelo.

O primeiro indicador que pode ser observado são as medidas de resumo fornecidas na saída de qualquer modelo ajustável, que apresentam um painel geral sobre o ajuste realizado. Essas informações não fornecem dados sobre os componentes individuais do modelo, e mesmo que valores pequenos de uma dessas estatísticas não garantam com precisão a qualidade do ajuste, valores muito elevados fornecem um importante indicador da presença de problemas no modelo. Nessa etapa, pode-se realizar o teste qui-quadrado de Pearson, que fornece indicadores para avaliar não só a qualidade do ajuste, mas a homogeneidade e independência das variáveis. Basicamente, o teste do qui-quadrado de Pearson analisa a existência de diferenças entre o resultado observado (real) e o resultado classificado pelo modelo. Para maiores informações sobre teste do qui-quadrado de Person, consultar Hosmer e Lemeshow (2000). Há também, o teste de Hosmer-Lemeshow, conhecido como *Goodness-of-fit Test* que também analisa a existência de diferenças entre o resultado observado (real) e o resultado classificado pelo modelo, mas este método estratifica as observações em classes, definidas como *decis*, e aplica à cada classe um teste do qui-quadrado.

Além destes, um método simples, mas muito eficiente é a comparação dos resultados reais com os resultados previstos pelo modelo, informações que permitem construir a chamada Matriz de Confusão do modelo, apresentada no Quadro 1.

Quadro 1 - Matriz de Confusão padrão

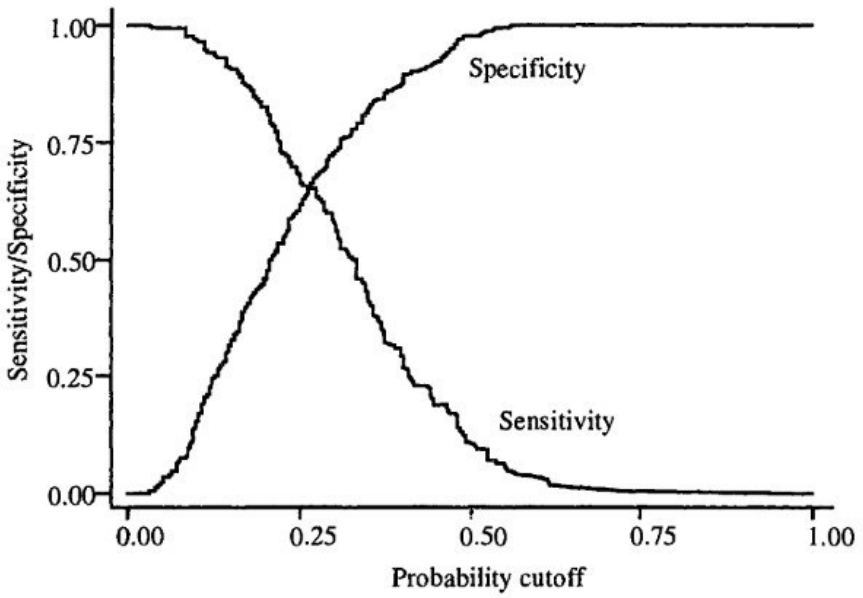
		Resultados classificados pelo modelo	
		Sucesso	Fracasso
Resultados reais observados	Sucesso	Verdadeiro positivo	Falso negativo
	Fracasso	Falso positivo	Verdadeiro negativo

Fonte: Hosmer e Lemeshow, 2000.

A matriz de confusão permite observar de forma simples informações muito importantes sobre a qualidade do ajuste realizado, como a quantidade de resultados reais apresentados incorretamente pelo modelo e também a quantidade de os casos em que o modelo de fato acertou na previsão. Com essas informações é possível calcular os valores de Sensibilidade (taxa de acertos positivos) e Especificidade (taxa de acertos negativos) do modelo em questão.

Valores de sensibilidade e especificidade utilizam um único ponto de corte para avaliar a qualidade do modelo. Para descrições mais completas sobre a precisão da classificação, pode-se utilizar a análise da área abaixo da curva ROC (*Reciever Operating Characteristic*). Essa curva se origina da probabilidade do modelo detectar um resultado verdadeiro (sensibilidade), relacionada à probabilidade de se obter um resultado falso (1 – especificidade) para todos os pontos possíveis no intervalo de dados do modelo. A figura abaixo (Figura 3) apresenta o relacionamento entre a sensibilidade e a especificidade:

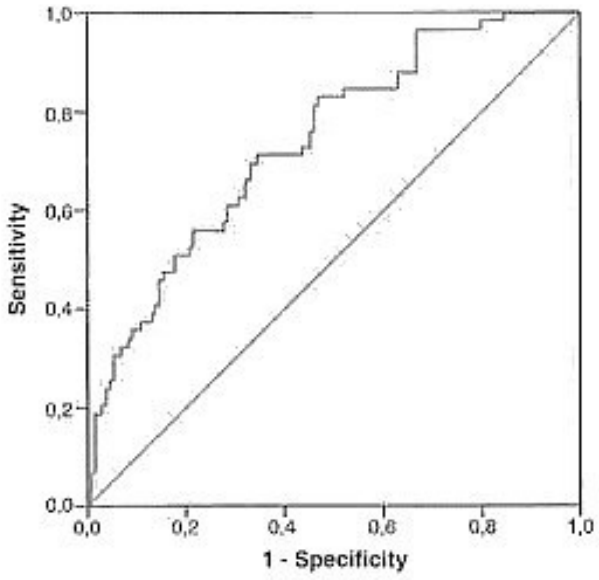
Figura 3 - Plotagem da sensibilidade e especificidade



Fonte: Hosmer e Lemeshow, 2000.

A imagem seguinte (Figura 4) apresenta a plotagem da curva ROC, relacionando sensibilidade *versus* (1 – especificidade):

Figura 4 - Curva ROC



Fonte: : FÁVERO *et al*., 2009.

A área abaixo da curva ROC (*AUC - Area under the ROC curve*) apresenta a probabilidade de acerto sobre a probabilidade de erro, em cada par de sucesso/fracasso limitado de 0 (fracasso) à 1 (sucesso). Desse modo, quanto maior a área abaixo da curva, e como consequência, maior aproximação da curva com o eixo esquerdo do gráfico, maior é a taxa de acertos e menor é a taxa de erros do modelo (maior número de previsões corretas) servindo como ferramenta para identificação da qualidade do ajuste obtido (HOSMER; LEMESHOW, 2000). Em outras palavras, a curva ROC define a diferença entre o sucesso de abordagens aleatórias (reta diagonal) e o sucesso das predições utilizando o modelo, de modo que, quanto mais próxima a curva ROC da reta diagonal, piores são os resultados obtidos pelo modelo.

Os modelos obtidos com as técnicas apresentadas permitem atribuir aos indivíduos da amostra um *score* que está relacionada à propensão de sucesso e fracasso em relação ao objetivo definido. Para um modelo corretamente ajustado, quanto maior o valor do *score*, maiores são chances de sucesso. Sendo assim, uma das formas de fazer uso dos modelos é selecionar apenas os indivíduos com os *scores* mais altos. Nesse sentido, a medida mais utilizada para modelos empregados em ações de *marketing* e negócios é o *Lift*, que tem como objetivo

mostrar a performance do modelo em cada decil (divisão ordenada dos dados de uma variável em dez partes iguais), de acordo com a Equação 16:

$$Lift\ (S,d) = \frac{\%Targets(S,d)}{d} \tag{16}$$

onde o numerador representa o percentual de clientes que se encontra no decil d com a previsão de resposta afirmativa pelo modelo step S, e d é a densidade da amostra. Tal densidade é representada pela divisão do número de eventos raros pelo número total de eventos do conjunto de dados em estudo.

Um ponto que deve ser considerado no momento de definir o melhor modelo para uso são os custos causados pelos erros de previsão e classificação. Esses custos estão relacionados à quantidade de falsos positivos e falsos negativos apresentados na matriz de confusão, impactando diretamente no retorno obtido com a utilização do modelo. Se, por um lado, o custo para abordar um indivíduo que responderá à ação negativamente (falso positivo) está relacionado com o esforço e valor de envio da comunicação, podendo ser relacionado à divulgação do produto e da marca e que poderá ocasionar uma oportunidade futura, por outro lado, deixar de abordar um indivíduo que responderia afirmativamente à ação (falso negativo) tem impacto direto nos lucros da empresa causados pela não realização do negócio e a empresa ainda corre o risco perder o cliente para a concorrência. Sendo assim, é necessário analisar os custos causados por cada um dos erros mencionados para que se possa concluir sobre os custos decorrentes da má classificação do modelo. A matriz de custos poderia ser representada como (Equação 17):

$$\begin{bmatrix} 0 & k_2 \\ k_1 & 0 \end{bmatrix} \tag{17}$$

onde k_1 representa o custo unitário para realização da comunicação e k_2 a perda de lucro causada pela não adesão ao produto. Os valores nulos se referem às previsões corretas.

2.2.4 Exemplos de aplicação

Como já discutido anteriormente, modelos de regressão logística são aplicados nas mais diversas áreas como forma de estudar, explicar e até prever eventos futuros. A previsão de eventos futuros se dá a partir da determinação da probabilidade de ocorrência de um determinado evento de interesse (*target*), e tem muita utilidade para a área de *Marketing*, vendas e relacionamento com o cliente.

Como exemplo de aplicação do modelo de regressão logística, em SILVA (2000), a técnica foi utilizada como opção para comparação dos resultados obtidos com os de um modelo de redes neurais, uma técnica muito mais complexa de previsão de eventos futuros e que exige um grande esforço computacional para construção e aplicação. O objetivo do trabalho era determinar a propensão dos clientes de uma determinada empresa do segmento financeiro à aquisição de crédito a fim de direcionar as campanhas ao grupo com perfil mais aderente ao produto, reduzindo gastos com envio de *mailing* e aumentando a conversão. Foram construídos diversos modelos a partir de ambas as técnicas e pode-se constatar que, nesse caso, modelos sem expurgo de variáveis (sendo significantes ou não) apresentaram melhor performance e, mesmo que optassem pelo uso de um modelo com pior performance, os resultados obtidos já apresentariam um ganho significativo em relação ao *status quo*.

Em outro exemplo, ADORNO (2011) também fez a utilização de modelos estatísticos para determinação da propensão de clientes de uma instituição bancária ao crédito pessoal. Nesse estudo, foram utilizadas as técnicas de árvore de decisão, regressão logística e redes neurais e, assim como o trabalho apresentado anteriormente, a aplicação de qualquer uma das técnicas já representaria um ganho para a instituição, uma vez que todas se mostraram eficientes para o objetivo proposto. Entretanto, técnica de árvore de decisão se mostrou menos eficiente em relação às demais e, dentre as duas técnicas restantes, o modelo de redes neurais se mostrou mais trabalhoso e ainda

apresentou problemas para aplicação fora da ferramenta de desenvolvimento, o que inviabilizaria a seleção de clientes em outro software de consulta de dados.

Em CABRAL (2013), foi realizado um estudo de mercado para avaliar a posição de uma marca em relação à concorrência. No trabalho, o modelo de regressão logística foi aplicado para definir a propensão de compra ou não dos produtos dessa marca partindo de informações importantes como o perfil dos consumidores, motivo de compra e se foi recomendação médica para verificar a propensão de compra dos clientes dentro de cada grupo.

Em outro caso que pode ser citado, BOJANOWSKI e LOLATTO (2018) utilizaram o modelo de regressão logística para identificar empresas que possuem baixa propensão de pagamento (risco de inadimplência) entre os clientes de uma grande instituição financeira a fim de antecipar ações para evitar a deterioração do portifólio e também identificar os mais propensos ao pagamento, possibilitando ações de cobranças diferenciadas. Além da técnica de regressão logística, foi utilizado também nesse trabalho a análise de sobrevivência (modelo de mistura e o modelo de tempo de promoção). No resultado final, o modelo de análise de sobrevivência mostrou um ganho em relação à determinação da propensão dos clientes por contar com a variável tempo embutida na resposta o que permite prever o período de inadimplência, o que, em relação ao modelo logístico é limitado à janela de performance analisada.

Como apresentado, a técnica de regressão fornece um ganho para direcionamento e tomada de decisão das instituições em relação ao *status quo*, sendo uma técnica que traz bons resultados quando comparada às demais, ainda que possua uma limitação em relação à janela de performance, e ainda não apresenta tanto esforço no desenvolvimento da técnica.

2.3 Análise Discriminante

A análise discriminante, assim como a regressão logística, é uma técnica estatística apropriada para os casos em que a variável dependente é qualitativa (categórica), o que impede o uso do método de regressão linear, conforme discutido

anteriormente. Tal análise é utilizada para classificar e discriminar objetos, além de permitir prever o resultado de eventos futuros. Porém, diferente da regressão logística que é limitada em sua forma básica à apenas dois grupos de respostas (sucesso e fracasso, por exemplo), essa técnica é capaz de lidar com os casos em que a resposta possui três ou mais classificações (faixas de idade ou risco alto, médio e baixo, por exemplo). Sendo assim, divide-se a técnica em análise discriminante de dois grupos, quando apenas duas classificações estão envolvidas, e análise discriminante múltipla (MDA), quando três ou mais classificações são utilizadas para a variável resposta.

2.3.1 Análise discriminante linear

Nesse método, na primeira etapa, conhecida como análise exploratória, é feita a definição de como serão agrupados ou separados os objetos. Para tal, deve-se identificar características que possam ser utilizadas para diferenciar os objetos em grupos, definindo regras que permitem tanto alocar, quanto para separar tais objetos. Em seguida determina-se uma função dada por uma combinação linear de duas ou mais variáveis independentes que melhor irão discriminar os objetos nos grupos definidos inicialmente. Esta função, também conhecida como função discriminante, assume uma forma semelhante à da regressão múltipla, conforme apresentado na Equação 18:

$$Z_{jk} = a + W_1 X_{1k} + W_2 X_{2k} + \cdots + W_n X_{nk} \tag{18}$$

onde

Z_{jk} representa o escore Z discriminante da função discriminante j para o objeto k

a representa o intercepto

W_i representa o peso discriminante para a i-ésima variável independente

X_{ik} representa a variável independente i para o objeto k

Dessa forma, a função discriminante permite alocar os objetos em grupos semelhantes, buscando minimizar as probabilidades de classificações incorretas ou sobreposições. Na Equação 16, também conhecida como função discriminante

linear de Fisher, o escore discriminante é a soma dos valores obtidos pela multiplicação de cada variável independente por seu peso discriminante. Segundo Hair *et al*. (2009), o que torna tal análise única é a possibilidade de utilizar-se mais de uma função discriminante, resultando na probabilidade de cada objeto tenha mais de um escore discriminante.

O método de Fisher parte da suposição de que a covariância de observações multivariadas das n populações são iguais, ou seja, supondo duas populações dadas por τ_1 e τ_2, a covariância é dada por (Equação 19):

$$\sum_1 = Cov(X|\tau_1) = \sum_2 = Cov(X|\tau_2) = \sum \quad (19)$$

onde X representa um vetor de variáveis aleatórias proveniente de uma das populações τ_1 e τ_2, cujos vetores de médias são dados por:

$\mu_1 = E(X|\tau_1)$: vetor de médias de observação multivariada de τ_1

$\mu_2 = E(X|\tau_2)$: vetor de médias de observação multivariada de τ_2

A ideia inicial apresentada pelo método de Fisher era encontrar uma combinação linear das variáveis originais, permitindo uma melhor diferenciação das populações. Considerando uma combinação linear das variáveis em estudo, as médias de $Y = l^T X$, para as duas populações definidas acima são dadas por:

$$\mu_{1Y} = E(Y|\tau_1) = E(l^T X|\tau_1) = l^T \mu_1 \quad (20)$$

$$\mu_{2Y} = E(Y|\tau_2) = E(l^T X|\tau_2) = l^T \mu_2 \quad (21)$$

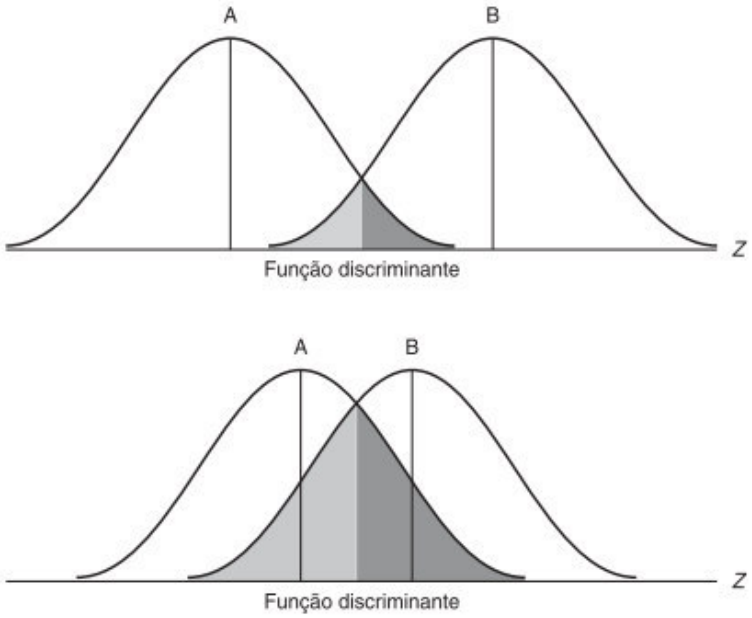
A variância, como consequência da suposição inicial, é igual para qualquer uma das populações, e se dá por:

$$\sigma^2{}_Y = Var(l^T X) = l^T Cov(X) = l^T \sum l \quad (22)$$

Segundo Fisher (1936), a função discriminante deve maximizar a diferença entre as médias específicas dentro de cada grupo. A média de um grupo em particular pode ser obtida calculando-se a média dos escores discriminantes para todos os indivíduos do grupo. Tal média, chamada de centroide, está presente em cada grupo da análise e representa o local típico de um indivíduo dentro do mesmo. Logo, quando a análise envolve dois grupos, possuem dois centroides, quando envolvem três grupos, possuem três centroides e assim por diante. Após a determinação a primeira função discriminante, as demais serão obtidas pela restrição de que os escores das funções não sejam correlacionados.

Segundo Hair *et al*. (2009), o teste de significância estatística da função discriminante é uma medida da distância entre os centroides dos grupos, obtida comparando as distribuições dos escores discriminantes para os grupos, conforme ilustrado na Figura 5:

Figura 5- Representação dos escores Z discriminantes entre dois grupos



Fonte: Hair *et al*., 2009.

Conforme observado na Figura 5, o diagrama na parte superior fornece uma melhor separação entre os grupos, uma vez que a sobreposição da distribuição dos escores é menor do que a verificada no diagrama apresentado na parte inferior da Figura 5, que possui uma função discriminante relativamente pobre entre os grupos A e B. As áreas sombreadas na intersecção entre os gráficos representam os casos em que podem ocorrer classificações ruins de objetos dos grupos A no grupo B e vice-versa.

A definição dos grupos da variável dependente deve ser feita de modo que esses sejam distintos, mutuamente excludentes e possam abranger todas as hipóteses, ou seja, cada observação pode ser inserida em apenas um grupo. Teoricamente, a análise discriminante permite a utilização de um número ilimitado de categorias na variável dependente, entretanto, quanto maior o número de categorias, mais complexas se tornam as análises, uma vez que são estimadas NG – 1 (número de grupos menos um) funções discriminantes a fim de se obter melhores resultados. Portanto, deve-se equilibrar o número de categorias a fim de manter a exclusividade dos grupos e a efetividade de uma análise com um número de categorias reduzida.

Segundo Hair *et al*. (2009), para a utilização da análise discriminante linear, algumas premissas e suposições devem ser adotadas, tais como: normalidade multivariada das variáveis independentes, de estruturas (matrizes) de dispersão e covariância desconhecidas (mas iguais) para os grupos definidos pela variável dependente; linearidade e ausência de multicolinearidade entre as variáveis. Tais premissas devem ser adotadas para evitar impactos sobre a estimação da função discriminante, bem como na classificação e interpretação dos resultados obtidos.

A avaliação do nível de significância para o poder discriminatório das funções de cada grupo coletivamente (significância geral) e separadamente (significância individual), caso o número de grupos seja superior à dois, permite identificar a necessidade de reformular o modelo definido. Caso o modelo geral seja significante, a avaliação das funções separadamente identifica quais podem ser mantidas e interpretadas posteriormente.

A avaliação da significância geral se dá a partir de testes estatísticos que permitem avaliar a habilidade das funções obtidas de definirem escores Z capazes

de discriminar significativamente os grupos, tais como a avaliação simultânea das medidas de lambda de Wilks, traço de Hotelling e o critério de Pillai. Para os casos em que o número de grupos é superior ou igual a três, se uma das funções obtidas é considerada não significante, o modelo discriminante deve ser reestruturado, obtendo apenas funções significantes.

O cálculo do *score* de corte entre os grupos é o critério utilizado para determinar à qual grupo um determinado objeto deverá ser classificado. Tal análise é feita utilizando apenas as funções consideradas significantes, permitindo então a construção de matrizes de classificação e avaliações mais precisas sobre o poder discriminatório das funções. Segundo Hair *et al*. (2009), o cálculo do escore de corte entre dois grupos é baseado nos centroides e no tamanho relativo dos grupos. Assumindo que as distribuições são normais e as estruturas de dispersão dos grupos são conhecidas, podemos utilizar a Equação 23 para cálculo do corte ótimo entre os grupos, conforme apresentado:

$$Z_{cs} = \frac{N_A Z_B + N_B Z_A}{N_A + N_B} \tag{23}$$

onde

Z_{cs} representa o escore de corte ótimo entre os grupos A e B

N_A representa o número de observações no grupo A

N_B representa o número de observações no grupo B

Z_A representa o centroide do grupo A

Z_B representa o centroide do grupo B

Se os grupos são especificados como sendo de tamanhos iguais, então o escore de corte ótimo será a média dos dois centroides, como apresentado na Equação 24:

$$Z_{cs} = \frac{Z_A + Z_B}{2} \tag{24}$$

onde

Z_{cs} representa o escore de corte ótimo entre os grupos A e B;

Z_A representa o centroide do grupo A;

Z_B representa o centroide do grupo B.

Deve-se levar em conta os custos de má classificação gerada pela classificação de objetos em grupos errados utilizando o escore de corte. Para os casos em que os custos de má classificação são aproximadamente iguais para todos os grupos, o escore de corte ótimo será aquele que classificar erroneamente o menor número de objetos em todos os grupos, já para os casos em que os custos de má classificação são desiguais, o escore de corte ótimo será o que minimizar os custos de má classificação. Para maiores informações sobre abordagens mais sofisticadas para determinar escores de corte, consultar Dillion e Goldstein (1984) e Huberty *et al*. (1987).

Segundo Johnson e Wichern (2007), o custo esperado de má classificação (ECM – *expected cost off missclassification*) pode ser calculado multiplicando o número de classificações observadas na população i incorretamente classificadas na população j pela probabilidade de tal classificação ocorrer, ou seja, considerando duas populações τ_1 e τ_2, temos o custo de má classificação apresentado (Equação 25):

$$ECM = c(2|1)P(2|1)p_1 + c(1|2)P(1|2)p_2 \quad (25)$$

onde:

$c(2|1)$ representa o número de classificações observadas na população τ_1 incorretamente classificadas na população τ_2;

$c(1|2)$ representa o número de classificações observadas na população τ_2 incorretamente classificadas na população τ_1;

$P(2|1)$ representa a probabilidade de classificações observadas na população τ_1 serem incorretamente classificadas na população τ_2;

$P(1|2)$ representa a probabilidade de classificações observadas na população τ_2 serem incorretamente classificadas na população τ_1;

p_1 e p_2 representam a probabilidade de a observação ser da população τ_1 ou τ_2, respectivamente, sendo $p_1 + p_2 = 1$.

Ainda segundo Johnson e Wichern (2007), para classificar duas populações multivariadas normais, deve-se buscar atingir o menor valor possível de ECM, ou seja, supondo que as densidade normais multivariadas das populações τ_1 e τ_2 são dadas por (Equação 26):

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\sum|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu_i)'\sum^{-1}(x - \mu_i)\right], \text{ sendo } i = 1 \text{ e } 2 \qquad (26)$$

Supondo ainda que os parâmetros μ_1, μ_2 e $\sum$ são conhecidos e cancelando os termos $(2\pi)^{\frac{p}{2}}|\sum|^{1/2}$, as regiões ($R_1$ e R_2) para os valores mínimos de ECM (Equação 25) são obtidas conforme apresentado nas Equações 27 e 28:

$$R_1: \exp\left[-\frac{1}{2}(x - \mu_1)'\sum^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_2)'\sum^{-1}(x - \mu_2)\right] \geq \left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (27)$$

$$R_2: \exp\left[-\frac{1}{2}(x - \mu_1)'\sum^{-1}(x - \mu_1) + \frac{1}{2}(x - \mu_2)'\sum^{-1}(x - \mu_2)\right] < \left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (28)$$

A partir das Equações 27 e 28, pode-se construir as regras de classificação de uma observação x_0 nas duas populações τ_1 e τ_2, conforme as Equações 29 e 30:

- Alocar x_0 em τ_1 se:

$$(\mu_1 - \mu_2)'\sum^{-1}x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\sum^{-1}(\mu_1 + \mu_2)] \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \qquad (29)$$

- Alocar x_0 em τ_2 se:

$$(\mu_1 - \mu_2)'\sum^{-1}x_0 - \frac{1}{2}(\mu_1 - \mu_2)'\sum^{-1}(\mu_1 + \mu_2)] < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (30)$$

2.3.2 Análise discriminante quadrática

Para os casos em que as matrizes de covariância entre as populações são diferentes ($\sum_1 \neq \sum_2$), assim como os vetores médios, a função discriminante linear de não é adequada para discriminação entre os grupos. Nestes casos, deve-se fazer uso da função discriminante quadrática. As funções de densidades normais multivariadas das populações τ_1 e τ_2 são dadas por (Equação 31):

$$f_i(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\sum_i|^{1/2}} \exp\left[-\frac{1}{2}(x - \mu_i)'\sum_i^{-1}(x - \mu_i)\right], \text{ sendo } i = 1 \text{ e } 2 \quad (31)$$

Nesse caso, os termos $(2\pi)^{\frac{p}{2}}|\sum_i|^{1/2}$ não são cancelados por se tratar de covariâncias diferentes entre as populações e a regra de classificação se torna um pouco mais complicada. Segundo Johnson e Wichern (2007), como no caso da análise discriminante linear, as regiões que minimizam os valores de ECM dependem da razão das densidades $f_1(x)/f_2(x)$, ou equivalente, ao logaritmo natural da razão das densidades, $\ln\left[\frac{f_1(x)}{f_2(x)}\right] = \ln[f_1(x)] - \ln[f_2(x)]$. Sendo assim, as regiões (R_1 e R_2) para os valores mínimos de ECM (Equação 25) são obtidas conforme apresentado nas Equações 32 e 33:

$$R_1: -\frac{1}{2}x'(\sum_1^{-1} - \sum_2^{-1})x + (\mu_1'\sum_1^{-1} - \mu_2'\sum_2^{-1})x - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (32)$$

$$R_2: -\frac{1}{2}x'(\sum_1^{-1} - \sum_2^{-1})x + (\mu_1'\sum_1^{-1} - \mu_2'\sum_2^{-1})x - k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (33)$$

onde:

$$k = \frac{1}{2}\ln\left(\frac{|\sum_1|}{\sum_2|}\right) + \frac{1}{2}(\mu_1'\sum_1^{-1}\mu_1 - \mu_2'\sum_2^{-1}\mu_2) \quad (34)$$

A partir das Equações 32 e 33, pode-se construir as regras de classificação de uma observação x_0 nas duas populações τ_1 e τ_2, conforme as Equações 35 e 36:

- Alocar x_0 em τ_1 se:

$$-\frac{1}{2}x_0'(\sum_1^{-1} - \sum_2^{-1})x_0 + (\mu_1'\sum_1^{-1} - \mu_2'\sum_2^{-1})x_0 - k \geq \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (35)$$

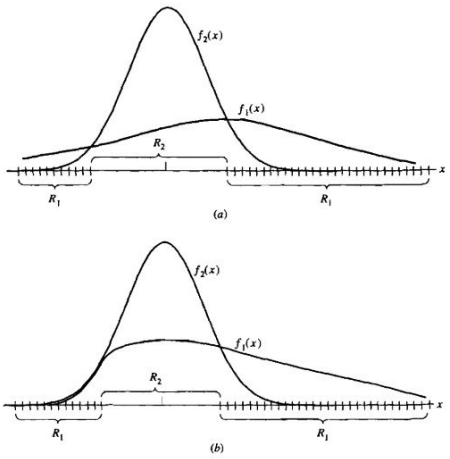
- Alocar x_0 em τ_2 se:

$$-\frac{1}{2}x_0'(\sum_1^{-1} - \sum_2^{-1})x_0 + (\mu_1'\sum_1^{-1} - \mu_2'\sum_2^{-1})x_0 - k < \ln\left[\left(\frac{c(1|2)}{c(2|1)}\right)\left(\frac{p_2}{p_1}\right)\right] \quad (36)$$

As regiões de classificação são definidas por funções quadráticas de x. Nas regiões em que $\sum_1 = \sum_2$, o termo quadrático, $-\frac{1}{2}x'(\sum_1^{-1} - \sum_2^{-1})x$, desaparece e a equação é reduzida às Equações 29 e 30.

Ainda segundo Johnson e Wichern (2007), a classificação com funções quadráticas é desajeitada para mais de duas dimensões e pode gerar resultados um pouco estranhos. Essa particularidade é verdadeira quando o dado não é essencialmente normal multivariado. A Figura 6, apresentada abaixo, mostra funções quadráticas para classificação de, no caso (a) duas populações normais com variâncias diferentes e, no caso (b) duas populações, sendo uma delas não normal, resultando em uma regra não apropriada. A regra quadrática no caso da Figura 6 gerou uma região R_1 constituída de dois pontos disjuntos.

Figura 6 - Função discriminante quadrática de (a) duas populações normais e (b) uma população não normal – regra não apropriada



Fonte: Johnson e Wichern, 2007.

Segundo Johnson e Wichern (2007), como apresentado na Figura 6, e visto em muitas outras aplicações, a cauda da primeira região R_1 para a distribuição τ_1 é menor no caso de distribuições não normais classificadas pelo procedimento quadrático e não se alinha bem com as distribuições populacionais, podendo levar a grandes taxas de erros. Um grave problema com o procedimento quadrático é a sensibilidade do método à distúrbios de normalidade.

Para os casos em que os dados não são normais multivariados, pode-se transformá-los os mais próximos possíveis do normal e testar a igualdade das matrizes de covariância para avaliar qual das regras de discriminação é apropriada. Mais detalhes, ver Johnson e Wichern (2007). Tais resultados mostram a importância de checar a performance de qualquer procedimento de classificação, tal qual será apresentado no item 2.3.3.

2.3.3 Interpretação dos resultados e avaliação da qualidade do ajuste

A classificação dos objetos é feita utilizando os valores definidos no escore de corte. Para as τ_g populações, a classificação é feita conforme apresentado na Tabela 3:

Tabela 3 – Classificação dos resultados

Probabilidade	**Resultado da Classificação**
$Z_n < Z_{cs1}$	Grupo 1
$Z_{cs1} < Z_n < Z_{cs2}$	Grupo 2
.	.
.	.
.	.
$Z_n > Z_{csg}$	Grupo g

Fonte: Arquivo pessoal.

onde

Z_n representa o escore discriminante para o n-ésimo indivíduo;

Z_{cs1} representa o escore de corte crítico entre os grupos 1 e 2;

Z_{cs2} representa o escore de corte entre para os grupos 2 e 3;

$Z_{cs(g-1)}$ representa o escore de corte crítico entre os grupos $(g-1)$ e g.

Após a realização da discriminação do conjunto de dados entre as τ_g populações $(g = 1, 2, \ldots)$ pelo modelo, pode-se construir a matriz de confusão, conforme apresentado no Quadro 2. A matriz fornece então uma perspectiva sobre a efetividade do modelo, permitindo o cálculo do percentual de acertos e a taxa estimada de erro (TEE), uma estimativa da taxa de erro verdadeira.

Quadro 2 - Matriz de Confusão da análise discriminante

População verdadeira	População classificada pelo modelo				Total
	τ_1	τ_2	...	τ_g	
τ_1	n_{11}	n_{12}	...	n_{1g}	n_1
τ_2	n_{21}	n_{22}	...	n_{2g}	n_2
.	.	.	.	.	.
.	.	.	.	.	.
.	.	.	.	.	.
τ_g	n_{g1}	n_{g2}	...	n_{gg}	n_g
Total	n'_1	n'_2	...	n'_g	n

Fonte: Arquivo pessoal.

Onde:

n_{ij} é o número de observações que foram observadas em τ_i e classificadas em τ_j;

τ_i e τ_j são, respectivamente, as populações observadas e classificadas pelo modelo;

n'_i é o número de observações classificadas em τ_i;

n_i é o número de observações em τ_i;

n é o número de observações da amostra.

A partir da matriz de confusão, a TEE pode ser calculada, de acordo com a Equação 37 (abaixo), permitindo então identificar quão bem a função classificou os objetos em cada grupo. Uma boa discriminação deve fornecer baixos valores de TEE.

$$TEE = \frac{n - \sum_{i=1}^{g} n_{ij}}{n} \tag{37}$$

A taxa estimada de erro (TEE) de se classificar uma observação da população τ_i na população τ_j é então:

$$TEE(j|i) = \frac{n_{ij}}{n_i} \tag{38}$$

Tal parâmetro fornece informações sobre a precisão da classificação realizada ao utilizar o modelo. Segundo Hair *et al*. (2009), o critério sugerido é de que a precisão de classificação deva ser pelo menos um quarto maior do que a obtida por chances (aleatoriamente). Também deve ser levado em conta os percentuais de acertos específicos dos grupos, a fim de evitar que alguns grupos

possuam percentuais de acertos inaceitáveis, ainda que o percentual de acertos geral do modelo seja aceitável.

Ainda segundo Hair *et al*. (2009), um teste estatístico do poder discriminatório da matriz de classificação quando comparada com um modelo de chances é a estatística Q de Press. Essa medida compara o número de classificações corretas com o tamanho da amostra total e o número de grupos, conforme apresentado na Equação 39:

$$Q\ de\ Press = \frac{[N - (nK)]^2}{N(K-1)} \tag{39}$$

onde

N representa o tamanho da amostra total

n representa o número de observações corretamente classificadas

K representa o número de grupos

O valor calculado é comparado com um valor crítico (o valor qui-quadrado para o grau de liberdade no nível de confiança desejado) e, se ele exceder o valor crítico, então a matriz de classificação pode ser considerada estatisticamente melhor do que os resultados obtidos aleatoriamente.

Um método que pode ser utilizado para a validação dos resultados e da verificação da efetividade do modelo obtido é a utilização de uma amostra de testes. Esse método consiste em dividir a amostra total aleatoriamente em dois grupos, de análise e de teste, para então utilizar a função discriminante para classificar uma amostra de teste que não foi utilizada para obtenção da mesma, eliminando assim o viés de que o resultado obtido seja subestimado por utilizar a mesma amostra para construção do modelo e verificação do ajuste. Segundo Hair *et al*. (2009), alguns pesquisadores sugerem que esse procedimento deveria ser seguido diversas vezes, a fim de aumentar a confiança da validade da função obtida.

A validação cruzada (ou *cross-validation*) é o método mais comum utilizado para avaliar a validade externa da função discriminante e é realizada utilizando

múltiplos subconjuntos da amostra total. Segundo Hair *et al*. (2009), a abordagem mais amplamente utilizada é o método *jackknife*, baseada no princípio do "deixa um de fora", ou seja, estima-se a função discriminante para $k-1$ amostras, eliminando-se uma observação por vez a partir de uma amostra de k casos. Em seguida, após o cálculo da função discriminante para a subamostra, é verificada a classificação da amostra eliminada utilizando a função discriminante estimada. Novamente, outra amostra é retirada e o mesmo procedimento é feito. Depois que todas as previsões de classificação do grupo foram realizadas, uma a uma, uma matriz de classificação é construída e a razão de sucessos é calculada a fim de determinar a qualidade do ajuste.

Ainda segundo Hair *et al*. (2009), a validação cruzada é muito sensível a amostras pequenas, alguns pesquisadores sugerem a utilização desse método apenas quando o tamanho do grupo menor seja pelo menos três vezes o número de variáveis preditoras, outros ainda, sugerem uma proporção de cinco para um. Esse método vem se tornando amplamente utilizado à medida que os softwares computacionais disponibilizam tal análise como opção de cálculo.

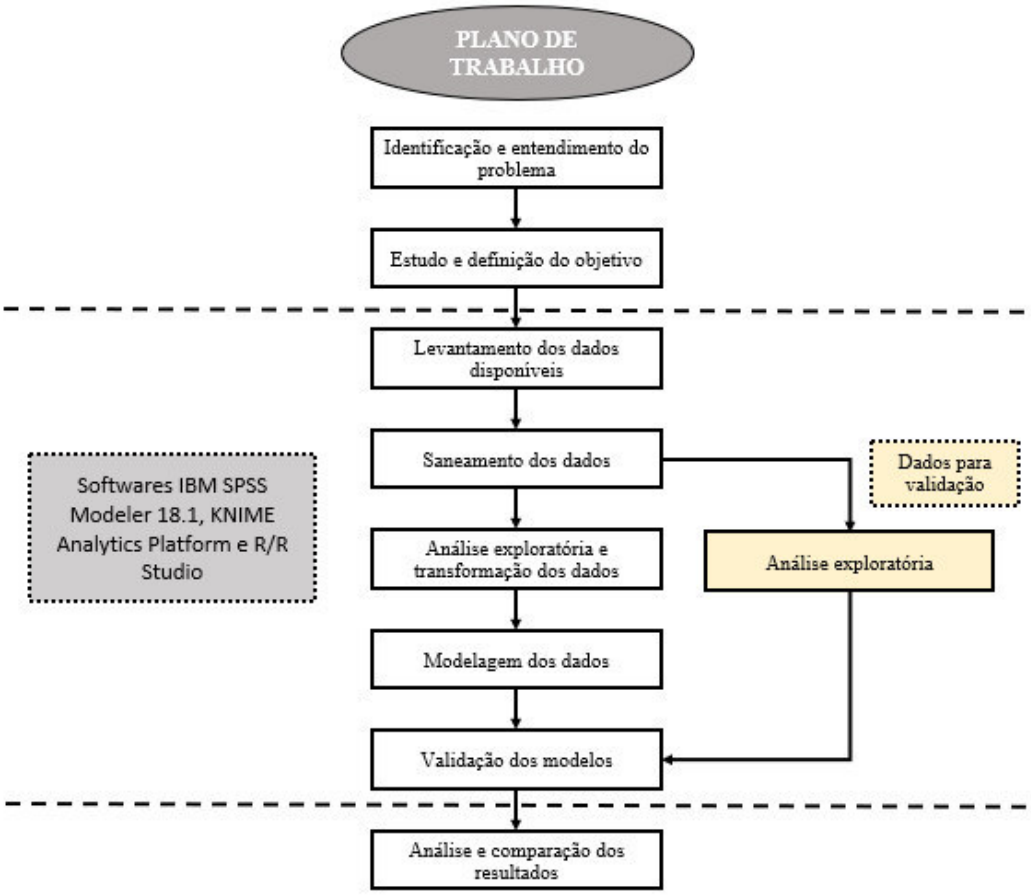
3 – MATERIAIS E MÉTODOS

No presente trabalho, realizou-se uma pesquisa explicativa, quantitativa e descritiva utilizando métodos e modelos estatísticos para estimar a probabilidade de compra de clientes do segmento de automóveis. A amostra utilizada para construção do modelo foi fornecida por uma empresa que atua no setor de varejo automotivo e serviços financeiros. Os dados referem-se às informações de perfil dos clientes, tanto pessoa física como pessoa jurídica, que adquiriram pelo menos um automóvel e os serviços fornecidos pela empresa para esses clientes, desde janeiro de 2010 até novembro de 2019. As informações sobre o produto e as operações de compra e venda de automóveis na empresa, coletados em seus pontos de venda e armazenados nos bancos de dados, também foram considerados na análise. A análise e modelagem dos dados foi realizada utilizando os softwares IBM SPSS Modeler 18.1, KNIME Analytics Platform e R/R Studio.

A fim de comparar os resultados obtidos e, consequentemente, buscar os melhores resultados, foram utilizadas as técnicas de regressão logística múltipla e análise de discriminante para construção dos modelos de propensão de compra.

Para coordenar e estruturar as etapas de trabalho, ao longo do projeto serão seguidas as seguintes etapas apresentadas na Figura 7:

Figura 7- Fluxograma de trabalho



Fonte: Arquivo pessoal.

3.1 Identificação e entendimento do problema

Para identificar e entender mais detalhadamente o problema, deve-se analisar primeiramente o quanto as instituições gastam todos os anos com estratégias para impactar seus clientes, a fim de convencê-los a realizar uma nova compra. Muitas dessas estratégias se baseiam apenas em quantidade, ou seja, oferecer tudo para todos os clientes, o que aumenta significativamente o custo com marketing e nem sempre acaba sendo efetivo para a empresa. Olhando do ponto de vista do cliente que recebe uma enorme quantidade de mensagens, ligações e comunicações, pode ocorrer um desgaste no relacionamento dele com a empresa/marca.

Sendo assim, identificar quais clientes são mais propensos e que, possuem perfil mais aderente ou teriam mais interesse em um determinado produto é objeto

de interesse em estudos em áreas como marketing, estatística e ciências de dados, a fim de fornecer estratégias que tragam mais resultados aos negócios.

Para este estudo, a empresa em questão conta com mais de 60 anos de tradição no segmento de serviços financeiros e de varejo automotivo e possui uma ampla base de clientes com um diversificado portfólio de produtos, o que justifica a necessidade de buscar estratégias para melhorar a assertividade e reduzir os custos das suas estratégias de vendas.

3.2 Levantamento dos dados disponíveis

A construção do subconjunto de dados foi feita utilizando o software IBM SPSS Modeler 18.1, que permitiu consultar e integrar as informações dos diversos sistemas utilizados pela instituição, possibilitando o levantamento dos dados de vendas do produto de automóveis, bem como referentes a outros produtos que também foram adquiridos por esses clientes, além de suas informações demográficas. Vale ressaltar que, nem sempre a escolha dos dados que serão utilizados para o processo de modelagem é feita considerando aquilo que o analista considera como as informações mais relevantes, mas sim, levando em conta quais os dados disponíveis, bem como a qualidade e volume desses dados.

No processo inicial de levantamento de informações, não foi feito nenhum tipo de filtro nos dados disponíveis, uma vez que o objetivo era acumular o máximo de informações possíveis para realizar uma análise inicial de toda massa histórica e a partir daí preparar os dados para os procedimentos de modelagem.

Após as análises iniciais, foram consideradas questões que permitissem atingir os objetivos do modelo, como quais seriam os dados e o período relevante para a amostra de dados, o tamanho necessário e como seria definida a marcação do *target* (sucesso ou fracasso dentro da amostra de dados). Por se tratar de um trabalho que utiliza dados históricos para “prever” um comportamento futuro, deve-se partir do princípio de que os comportamentos e experiências passados são aplicáveis ao futuro. Sendo assim, foram consideradas variáveis que permitissem descrever e discriminar os clientes, como informações pessoais (idade, sexo,

endereço, por exemplo), seguindo para dados que representassem o relacionamento do cliente com a instituição (como quantidade de compras, valor gasto, produtos adquiridos, dias desde compra, por exemplo), para serem utilizados como *input* do modelo. Da mesma forma, é necessário apresentar o *output* do modelo, que nesse caso será definido na forma de sucesso ou fracasso para a recompra de automóvel em um determinado tempo.

Nessa etapa de definição das variáveis e coleta dos dados, é muito importante que as marcações temporais sejam coerentes e respeitem o momento em que os fatos ocorreram. Não se deve, por exemplo, apresentar no *input* do modelo informações que sejam posteriores ou que estejam diretamente relacionadas ao *output*, pois dessa forma, pode se chegar a um modelo que fornece previsões incorretas sobre seu conjunto de dados.

Neste trabalho, a fim de reduzir os impactos das causados pelos diferentes cenários econômicos em um longo período, optou-se por utilizar o histórico de vendas de um intervalo de 5 anos.

Para reduzir a variabilidade e sazonalidade no evento modelado ao longo do tempo, a amostra foi construída no formato de safras, onde os clientes podem ser selecionados como sucesso em vários pontos (safras diferentes), indicando novas compras em datas diferentes. Uma safra pode ser entendida como uma fotografia do conjunto de dados no momento da safra em questão, apresentando as informações de um determinado registro até o momento que define a safra. Para este trabalho, optou-se pela utilização de safras mensais, então, por exemplo, se um cliente já realizou uma compra anteriormente, dentro do período histórico levantado (5 anos), mas não comprou novamente, esse registro estará em todas as safras, com as informações de dias desde compra, idade e outros produtos que ele possuía variando de acordo com a evolução das safras em relação às datas de compra desse cliente, mas marcado como fracasso, uma vez que não realizou uma recompra. Diferente desse caso, se um cliente que estava no histórico recomprou na terceira safra, esse registro estará na primeira e na segunda, com suas informações referentes àquelas safras, mas marcados como fracasso e, na terceira safra ele seria considerado um sucesso, mas, a partir da quarta safra, esse registro já seria considerado um fracasso novamente, até o momento em que ele realizar uma nova compra em outra safra. Isso ocorre pois, todos os clientes que já

realizaram uma compra possuem potencial para serem um sucesso e essa visão nos permite analisar quais produtos ele possuía enquanto era um fracasso e depois no momento de em que se tornou um sucesso, quantos dias ele demorou para isso, dentre outras análises que podem ser feitas olhando o comportamento daquele registro mês a mês.

O conjunto de dados foi então composto por 14 safras (14 meses), divididas posteriormente em 12 safras para formação do conjunto de treinamento e testes do modelo, um período que minimiza as instabilidades causadas pelo tempo, e 2 safras para validação do modelo. Mais à frente, na sessão 3.4, serão abordados os motivos da divisão da amostra dessa forma.

Segue no Quadro 3 as informações sobre o conjunto de dados construído nessa etapa de levantamento de dados:

Quadro 3 – Conjunto de dados inicialmente construído.

ATRIBUTO	DESCRIÇÃO
Tipo pessoa	Tipo de pessoa do cliente (física ou jurídica)
Idade | Tempo de fundação	Idade (PF) ou tempo desde a fundação (PJ) do cliente
Sexo	Sexo do cliente (apenas PF)
Valor da venda automóvel	Valor do veículo adquirido pelo cliente
Ano fabricação | Ano modelo	Ano de fabricação e modelo do veículo adquirido
Local da venda	Revenda onde foi realizada a venda do veículo
Estado do cliente	Estado de residência do cliente
Dias desde a compra	Quantidade de dias desde a aquisição do último veículo
Sinergia pós-vendas automóveis	Quantidade de serviços realizados no pós-vendas de automóveis
Sinergia pós-vendas caminhões	Quantidade de serviços realizados no pós-vendas de caminhões
Sinergia cota automóveis	Quantidade de cotas de consócio de automóveis adquiridas pelo cliente
Sinergia cota caminhões	Quantidade de cotas de consócio de caminhões adquiridas pelo cliente
Sinergia cota imóveis	Quantidade de cotas de consócio de imóveis adquiridas pelo cliente
Sinergia outras cotas	Quantidade de cotas de consócio de outros produtos adquiridas pelo cliente
Sinergia banco	Quantidade de produtos do banco da empresa adquiridos
Sinergia corretora	Quantidade de produtos da corretora adquiridas pelo cliente
Sinergia vendas caminhões	Quantidade de veículos comerciais adquiridas pelo cliente

Fonte: Arquivo pessoal.

3.3 Saneamento dos dados

Na etapa anterior, de levantamento de dados, foram descartados do conjunto de dados atributos que apresentavam grandes quantidades de valores ausentes/nulos (*missing values*) ou fora de padrão (*noise data*), desde que não representassem informações muito relevantes sobre o fenômeno analisado. Nessa etapa, o conjunto de dados construído foi analisado de forma minuciosa a fim de identificar e tratar inconsistências nas informações consideradas relevantes.

Para os dados considerados relevantes para realização das análises, os valores ausentes foram preenchidos da seguinte forma:

Tabela 4 – Tratamento dos atributos do conjunto de dados.

ATRIBUTO	TIPO	%NULOS	TRATAMENTO
Idade | Tempo de fundação	Numérico	0,021%	Média dos valores do atributo em questão
Sexo	Categórico	0,008%	Categoria mais frequente apresentada no atributo
Estado do cliente	Categórico	0,015%	Categoria mais frequente apresentada no atributo

Fonte: Arquivo pessoal.

Apesar das técnicas utilizadas para tratamento de valores ausentes inferirem na adição de dados que podem não estar corretos, seu uso se faz necessário no mundo real, onde os bancos de dados são suscetíveis a armazenar registros inconsistentes, incoerentes com a realidade ou ainda não registrarem determinada informação, o que, sem técnicas como essas, inviabilizariam todo um grupo de informações para procedimentos de análise e modelagem.

3.4 Divisão do conjunto de dados: Treino, teste e validação.

Conforme discutido na seção 3.2, neste trabalho foram utilizados dois conjuntos de dados, que denominamos como conjunto de modelagem e conjunto de teste. O conjunto de modelagem foi subdividido em duas partes: o conjunto de treino e o conjunto de validação. Portanto, neste trabalho temos três conjuntos diferentes de dados: de treino, de teste e de validação.

De acordo com BERRY e LINOFF (2004), o conjunto de treino é utilizado para explicar a variável dependente (*target*) em termos das variáveis independentes (*input*). A partir do ajuste obtido por esse conjunto é que se encontram os padrões apresentados nos dados. Entretanto, o modelo construído levando em conta apenas os dados de treino tendem a se ajustar apenas a esse conjunto, enviesando as previsões de outros conjuntos de dados.

Ao aplicar o modelo obtido a partir do conjunto de treino para com um novo conjunto de dados, o conjunto de teste, pode-se verificar a capacidade do modelo desenvolvido em classificar novos conjuntos de dados. Portanto, como o objetivo do modelo é fazer previsões em novos conjuntos de dados, representado nesse caso pelo conjunto de teste, seu objetivo passa a ser medir a performance do modelo desenvolvido na etapa de treino em relação à sua capacidade de generalização, obtendo uma estimação coerente sem grandes vieses de erros.

O conjunto de validação, por sua vez, funciona como um terceiro conjunto de dados, a parte dos anteriores, com informações que não foram alvo das análises até o momento, utilizado então para medir o comportamento do ajuste frente a essa nova amostra de dados. Esse conjunto permite verificar qual seria o comportamento das classificações realizadas pelo modelo caso esse fosse utilizado em uma situação real a partir daquele momento.

Após a construção do modelo, caso a qualidade do ajuste seja comprovada na etapa de testes, esse estará apto para ser utilizado em uma nova base de dados, isto é, um novo conjunto de clientes com o objetivo de identificar a probabilidade desse grupo realizar a ação definida pelo objetivo do modelo (recompra de automóveis).

Neste trabalho, optou-se por segmentar o conjunto de dados inicial pelo tipo de registro da amostra (pessoa física e pessoa jurídica), pelos motivos apresentados na seção seguinte (seção 3.5) e, utilizou-se 80% para a etapa de construção e treinamento do modelo e 20% para a etapa de testes. Na etapa de validação, utilizou-se um novo conjunto de dados, composto pelas vendas realizadas no período de 2 meses (2 safras) após o período do primeiro conjunto de dados. Nas Tabelas 6 e 7 estão apresentadas a quantidade de amostras em cada conjunto de dados, em relação aos sucessos e fracassos:

Tabela 5 – Quantidade de registros que compõem cada conjunto de dados do grupo pessoa física

	Sucesso	Fracasso
Conjunto de dados de treinamento	1.534	287.554
Conjunto de dados de testes	285	71.987
Conjunto de dados de validação	157	36.143

Fonte: Arquivo pessoal.

Tabela 6 – Quantidade de registros que compõem cada conjunto de dados do grupo pessoa jurídica

	Sucesso	Fracasso
Conjunto de dados de treinamento	216	38.876
Conjunto de dados de testes	47	9.726
Conjunto de dados de validação	23	3.647

Fonte: Arquivo pessoal.

3.5 Análise exploratória e transformação dos dados

Nessa etapa, foram realizadas análises no conjunto de dados utilizando o software KNIME Analytics Platform a fim de verificar o relacionamento da variável independente (*target* – sucesso ou fracasso na recompra do veículo) com as variáveis dependentes, uma a uma, de forma independente. Nessa etapa, com as análises realizadas obteve-se uma maior familiarização com os dados, o que permitiu observar um comportamento muito diferente entre os públicos (pessoa física e pessoa jurídica) considerados no conjunto de dados, diferenciados pela variável “Tipo pessoa”, de tal forma que optou-se por trabalhar com eles separadamente para os processos de modelagem a fim de melhorar os resultados obtidos. Portanto, foram construídos dois modelos distintos, de acordo com os públicos pessoa física e pessoa jurídica.

Nesse processo, foram realizadas a categorização ou recategorização das variáveis preditoras a fim de trabalhar com 2 a 4 grupos por variável, além de identificar quais dessas variáveis não tinham grande significância para o processo de predição com o uso da técnica *Information Value*.

Segundo HENRIQUE, FRANCISCO e NETO (2008), é muito comum no desenvolvimento de modelos tratar as variáveis como categóricas, independente

da sua natureza discreta ou contínua, buscando sempre a simplicidade na interpretação dos resultados obtidos. Tal procedimento pode também trazer ganhos no poder preditivo do modelo e deve ser realizado tanto para originalmente contínuas, quanto para as que já são categóricas. Para o processo de categorização das variáveis, foi utilizada a técnica de árvore de decisão, analisando a relação de forma independente entre cada preditor em par com a variável resposta. Esse processo permitiu classificar, por exemplo, a variável "Valor da venda automóvel" em 3 categorias de acordo com a faixa de valor do veículo, transformando uma variável numérica em categórica. O procedimento foi aplicado em todas as variáveis, a fim de que sejam utilizadas para a modelagem apenas preditores categóricos.

A técnica de *Information Value (IV)* é uma das mais utilizadas para selecionar variáveis importantes em um modelo preditivo. E ela permite, por exemplo, ranquear as variáveis de acordo com sua importância, segundo a Equação 40:

$$IV = \sum(\%\ não\ eventos - \%\ eventos) * WOE \tag{40}$$

Onde, WOE (*Weight of Evidence*) representa a medida de separação entre os clientes que atingem e os que não atingem o objetivo definido, calculado da seguinte forma:

$$WOE = \ln\left(\frac{\%\ não\ eventos}{\%\ eventos}\right) \tag{41}$$

Sendo assim, os valores de *IV* permitem classificar as variáveis de acordo com seu poder preditivo, conforme a o Quadro 4:

Quadro 4 – Regras de classificação do *Information Value (IV)*.

Information Value	Poder preditivo da variável
Menor do que 0,02	Não preditivo
Entre 0,02 e 0,1	Pouco preditivo
Entre 0,1 e 0,3	Moderadamente preditivo
Entre 0,3 e 0,5	Fortemente preditivo
Maior do que 0,5	Poder preditivo suspeito - verificar

Fonte: Arquivo pessoal.

Com a utilização das técnicas de categorização e de *Information Value* (IV), foram retiradas do conjunto de dados algumas variáveis que não possuem poder preditivo para o evento modelado (IV menor do que 0,02), minimizando a entrada de preditores sem efeito prático no modelo. As categorias foram definidas de modo a maximizar o valor do IV, levando em conta que, valores de IV acima de 0,5 podem indicar variáveis que estavam tão relacionadas à resposta desejada que podem ser parte dela ou caracterizam um erro no levantamento do conjunto de dados (apresentar uma informação referente ao momento futuro do registro analisado). No nosso conjunto de dados, nenhum dos valores de IV foram superiores à 0,5.

Outro critério considerado para seleção das variáveis que seriam utilizadas para modelagem foi a correlação linear entre elas. Variáveis com elevado grau de correlação (valores mais próximos de 1, para correlação positivas, ou -1, para correlações negativas) foram analisadas a fim de utilizar apenas a que possuísse o maior IV dentre as correlacionadas. Nas seções 3.5.1 e 3.5.2 serão apresentadas as categorizações obtidas pela árvore de decisão, as informações de IV e correlação linear para os conjuntos de dados PF e PJ:

3.5.1 Conjunto de dados Pessoa Física (PF):

As variáveis do conjunto de dados PF foram categorizadas, de acordo com a Tabela 7:

Tabela 7 – Categorias das variáveis do conjunto de dados PF.

VARIAVEIS	CATEGORIAS
DIAS DESDE A ÚLTIMA COMPRA	ATÉ 430 DIAS ENTRE 430 E 700 DIAS ENTRE 700 E 1250 DIAS MAIS DE 1250 DIAS
IDADE	MENOS DE 40,5 ANOS IGUAL OU MAIOR DO QUE 40,5 ANOS
LOCAL DA VENDA	REVENDA DE SINOP REVENDA DE BELÉM REVENDA DE RONDONÓPOLIS DEVENDAS DO RIO DE JANEIRO REVENDAS DE MINAS GERAIS E SÃO PAULO OUTRAS REVENDAS
PASSAGENS POS-VENDA AUTO	NÃO POSSUI ENTRE 1 E 4 MAIS DE 4
PASSAGENS BANCO	NÃO POSSUI UMA PASSAGEM MAIS DE UMA PASSAGEM
POSSUI CONSÓRCIO DE AUTOMÓVEIS	NÃO SIM
POSSUI CONSÓRCIO DE CAMINHÕES	NÃO SIM
PASSAGEM POS-VENDA VEIC. COMERCIAIS	NÃO SIM
VALOR DA VENDA	ATÉ R$ 50.000,00 ENTRE R$ 50.000,00 E R$ 140.000,00 MAIOR QUE R$ 140.000,00
POSSUI CONSÓRCIO DE IMÓVEIS	SIM NÃO
POSSUI PRODUTOS NA CORRETORA	SIM NÃO
POSSUI OUTRAS COTAS DE CONSÓRCIO	SIM NÃO
POSSUI VENDAS DE CAMINHÕES	SIM NÃO
SEXO	M F
ESTADO DO CLIENTE	PA SP, RJ E MG MT OUTROS ESTADOS

Fonte: Arquivo pessoal.

Os valores de IV para as variáveis do conjunto de dados PF podem ser visualizados na Tabela 8:

Tabela 8 –Classificação do IV para as variáveis do conjunto de dados PF.

ATRIBUTO	IV	Poder preditivo da variável
Idade	0,000000	Não preditivo
Sexo	0,047346	Pouco preditivo
Valor da venda	0,154297	Moderadamente preditivo
Local da venda	0,139776	Moderadamente preditivo
Estado do cliente	0,147083	Moderadamente preditivo
Dias desde a compra	0,043733	Pouco preditivo
Sinergia pós-vendas automóveis	0,402085	Fortemente preditivo
Sinergia pós-vendas caminhões	0,052435	Pouco preditivo
Sinergia consórcio automóveis	0,052357	Pouco preditivo
Sinergia consórcio caminhões	0,012934	Não preditivo
Sinergia consórcio imóveis	0,002472	Não preditivo
Sinergia outras cotas de consórcio	0,000089	Não preditivo
Sinergia banco	0,288020	Moderadamente preditivo
Sinergia corretora	0,005358	Não preditivo
Sinergia vendas caminhões	0,007312	Não preditivo

Fonte: Arquivo pessoal.

A partir da Tabela 8, é possível observar o poder preditivo de cada variável, separadamente. Na Tabela 9, serão apresentados os dados de correlação linear entre as variáveis do conjunto de dados PF:

Tabela 9 – Correlação linear das variáveis do conjunto de dados PF.

	Sexo	Idade	Local da venda	Dias desde a compra	Sinergia banco	Sinergia corretora	Sinergia cota automóveis	Sinergia cota caminhões	Sinergia cota imóveis	Sinergia outras cotas	Sinergia pós-vendas automóveis	Sinergia pós-vendas caminhões	Sinergia vendas caminhões	Valor da venda	Estado do cliente
Sexo	1,000														
Idade	0,000	1,000													
Local da venda	0,082	0,000	1,000												
Dias desde a compra	0,007	0,000	0,057	1,000											
Sinergia banco	0,026	0,000	0,090	0,086	1,000										
Sinergia corretora	0,007	0,000	0,046	0,070	0,073	1,000									
Sinergia cota automóveis	0,048	0,000	0,045	0,023	0,020	0,035	1,000								
Sinergia cota caminhões	0,028	0,000	0,044	0,018	0,006	0,011	0,061	1,000							
Sinergia cota imóveis	0,011	0,000	0,019	0,009	0,014	0,013	0,085	0,024	1,000						
Sinergia outras cotas	0,005	0,000	0,014	0,001	0,019	0,018	0,049	0,002	0,055	1,000					
Sinergia pós-vendas automóveis	0,114	0,000	0,122	0,129	0,064	0,042	0,085	0,031	0,016	0,022	1,000				
Sinergia pós-vendas caminhões	0,068	0,000	0,161	0,025	0,035	0,028	0,033	0,075	0,001	0,002	0,076	1,000			
Sinergia vendas caminhões	0,023	0,000	0,057	0,009	0,032	0,034	0,013	0,066	0,003	0,001	0,018	0,250	1,000		
Valor da venda	0,211	0,000	0,196	0,219	0,064	0,011	0,107	0,081	0,018	0,008	0,104	0,132	0,049	1,000	
Estado do cliente	0,081	0,000	0,808	0,067	0,087	0,043	0,041	0,045	0,009	0,007	0,198	0,149	0,057	0,197	1,000

Fonte: Arquivo pessoal.

No conjunto de dados PF, observa-se que a maioria das variáveis apresentam baixa ou nenhuma correlação entre si, exceto as informações de “Estado do Cliente” e “Local da venda”, que apresentaram forte correlação linear positiva e, tal correlação é esperada ao analisar o significado dessas variáveis. Como a empresa possui pontos de vendas em algumas cidades dos estados de São Paulo, Minas Gerais, Mato Grosso, Pará e Pernambuco e, revendem seus produtos localmente, o estado dos clientes que compram os produtos tem forte relacionamento com a cidade do ponto de venda que realizou a venda. Dessa

forma, optou-se por utilizar apenas a variável “Estado do Cliente” nos processos de modelagem, uma vez que essa possui maior poder preditivo (IV).

3.5.2 Conjunto de dados Pessoa Jurídica (PJ):

As variáveis do conjunto de dados PJ foram categorizadas, de acordo com a Tabela 10:

Tabela 10 – Categorias das variáveis do conjunto de dados PJ.

VARIAVEIS*	CATEGORIAS
DIAS DESDE A ÚLTIMA COMPRA	ATÉ 320 DIAS ENTRE 320 E 1200 DIAS MAIS DE 1200 DIAS
LOCAL DA VENDA	REVENDAS DO MATO GROSSO REVENDA DE BELÉM DEVENDA DO RIO DE JANEIRO REVENDA DE MINAS GERAIS E SÃO PAULO OUTRAS REVENDAS
TEMPO DE FUNDAÇÃO	ATÉ 5,5 ANOS ENTRE 5,5 E 26,5 ANOS MAIS DE 26,5 ANOS
PASSAGENS POS-VENDA AUTO	NÃO POSSUI ENTRE 1 E 10 MAIS DE 10
PASSAGENS BANCO	NÃO SIM
POSSUI CONSÓRCIO DE AUTOMÓVEIS	NÃO SIM
POSSUI CONSÓRCIO DE CAMINHÕES	NÃO SIM
PASSAGEM POS-VENDA VEIC. COMERCIAIS	NÃO SIM
POSSUI COTA DE IMÓVEIS	NÃO SIM
POSSUI PRODUTOS NA CORRETORA	SIM NÃO
POSSUI VENDAS DE CAMINHÕES	SIM NÃO
POSSUI OUTRAS COTAS DE CONSÓRCIO	SIM NÃO
VALOR DO PRODUTO	ATÉ R$ 55.000,00 ENTRE R$ 55.000,00 E R$ 96.000,00 MAIOR QUE R$ 96.000,00
ESTADO DO CLIENTE	PA, MT E RJ SP E MG OUTROS ESTADOS

Fonte: Arquivo pessoal.

Na Tabela 11 serão apresentados os valores de IV para as variáveis do conjunto de dados PJ:

Tabela 11 –Classificação do IV para as variáveis do conjunto de dados PJ.

ATRIBUTO	IV	Poder preditivo da variável
Tempo de fundação	0,037257	Pouco preditivo
Valor da venda	0,031980	Pouco preditivo
Local da venda	0,032026	Pouco preditivo
Estado do cliente	0,040255	Pouco preditivo
Dias desde a compra	0,215484	Moderadamente preditivo
Sinergia pós-vendas automóveis	0,206666	Moderadamente preditivo
Sinergia pós-vendas caminhões	0,026246	Pouco preditivo
Sinergia consórcio automóveis	0,160161	Moderadamente preditivo
Sinergia consórcio caminhões	0,086825	Pouco preditivo
Sinergia consórcio imóveis	0,038620	Pouco preditivo
Sinergia outras cotas de consórcio	0,007546	Não preditivo
Sinergia banco	0,111289	Moderadamente preditivo
Sinergia corretora	0,000784	Não preditivo
Sinergia vendas caminhões	0,010145	Não preditivo

Fonte: Arquivo pessoal.

Comparando as tabelas de categorização e IV dos conjuntos PF (Tabelas 7 e 8) e PJ (Tabelas 10 e 11), é possível evidenciar as diferenças entre os públicos. Tais diferenças podem ser percebidas tanto pelas categorizações realizadas ou pelas variáveis com maior poder preditivo em cada conjunto de dados.

Tabela 12 — Correlação linear das variáveis do conjunto de dados PJ.

	Tempo de fundação	Local da venda	Dias desde a compra	Sinergia banco	Sinergia corretora	Sinergia cota automóveis	Sinergia cota caminhões	Sinergia cota imóveis	Sinergia outras cotas	Sinergia pós-vendas automóveis	Sinergia pós-vendas caminhões	Sinergia vendas caminhões	Valor da venda	Estado do cliente
Tempo de fundação	1,000													
Local da venda	0,099	1,000												
Dias desde a compra	0,126	0,079	1,000											
Sinergia banco	0,077	0,113	0,024	1,000										
Sinergia corretora	0,014	0,052	0,014	0,060	1,000									
Sinergia cota automóveis	0,013	0,122	0,081	0,005	0,043	1,000								
Sinergia cota caminhões	0,016	0,182	0,055	0,099	0,078	0,151	1,000							
Sinergia cota imóveis	0,014	0,033	0,013	0,016	0,031	0,074	0,029	1,000						
Sinergia outras cotas	0,015	0,033	0,040	0,037	0,028	0,058	0,061	0,001	1,000					
Sinergia pós-vendas automóveis	0,114	0,098	0,079	0,138	0,028	0,107	0,055	0,022	0,027	1,000				
Sinergia pós-vendas caminhões	0,037	0,269	0,042	0,178	0,098	0,031	0,204	0,004	0,006	0,158	1,000			
Sinergia vendas caminhões	0,011	0,127	0,050	0,138	0,076	0,024	0,184	0,008	0,003	0,083	0,370	1,000		
Valor da venda	0,059	0,154	0,177	0,051	0,010	0,073	0,067	0,012	0,011	0,033	0,101	0,050	1,000	
Estado do cliente	0,067	0,700	0,038	0,061	0,039	0,024	0,095	0,019	0,004	0,044	0,102	0,077	0,064	1,000

Fonte: Arquivo pessoal.

No conjunto de dados PJ, de forma idêntica ao conjunto PF, as únicas variáveis que apresentaram forte correlação linear positiva entre si foram “Estado do Cliente” e “Local da venda”. Também se optou por utilizar apenas a variável “Estado do Cliente” nos processos de modelagem por conta do seu maior poder preditivo (IV).

3.6 Oversampling

Oversampling é um processo de criação de um conjunto de dados com proporções maiores do evento considerado raro e, um menor número de eventos comuns, quando comparado ao conjunto de dados original. Tal método também é conhecido como Amostra Aleatória Estratificada, uma vez que utiliza proporções diferentes de cada categoria da variável binária. BERRY e LINOFF (2004) trabalham com a teoria de que, em um problema com uma variável resposta binária, a ideia é ter uma amostra de desenvolvimento com 10% a 40% dos indivíduos menos frequente, e que, valores entre 20% e 30% normalmente produzem resultados satisfatórios em modelos no contexto de *Data Mining* de forma geral.

Como o objetivo do trabalho é definir a propensão dos clientes recomprarem um novo veículo, tal evento (sucesso) se torna raro se comparado com a grande quantidade de clientes na amostra que não realizaram a aquisição do produto no intervalo de tempo. A utilização da amostra dessa forma levaria a um modelo com alta capacidade de prever os fracassos dentro do conjunto de dados, o que não é de fato o objetivo final que se quer atingir. Nesse contexto, a técnica de *Oversampling* se apresenta como uma possível solução para esse problema, reequilibrando dentro do conjunto de dados a proporção de sucessos e fracassos para o processo de modelagem. Nesse trabalho, a proporção utilizada na etapa de construção foi de 1:1, selecionadas aleatoriamente, a fim de maximizar a capacidade preditiva do modelo elaborado.

3.7 Modelagem de dados

A construção dos modelos foi realizada utilizando o *Software* estatístico R versão 3.6.2 dentro do ambiente do *Software* KNIME Analytics Platform que possui extensões para tal finalidade.

Como apresentado e fundamentado na Seção 2, as técnicas de modelagem aplicadas para análise de dados foram regressão logística e análise discriminante,

cujos resultados serão analisados e comparados a fim de determinar qual o melhor modelo para o objetivo desejado.

3.7.1 Regressão logística

O modelo logístico múltiplo foi construído utilizando o método de seleção de variáveis *Stepwise*. Esse método seleciona as variáveis com maior significância para compor o modelo final e, consequentemente, minimiza a entrada e permanência de variáveis com pouco poder preditivo para essa técnica.

Como citado anteriormente, foram construídos dois modelos separados de acordo com o tipo pessoa (PF ou PJ) do registro, uma vez que foram identificados comportamentos muito diferentes para esses públicos.

Todos os modelos construídos passaram pelas etapas de treinamento (utilizando 80% do conjunto de dados), testes (utilizando 20% do conjunto de dados) e validação (utilizando um novo conjunto de dados). Na Seção 4, a seguir, serão apresentados os resultados obtidos.

3.7.2 Análise discriminante

De forma análoga ao modelo logístico, a construção do modelo discriminante também se deu em 3 etapas: construção, teste e validação. Para simular as mesmas condições de desenvolvimento e permitir a comparação os resultados obtidos por ambas as técnicas, também foram elaborados dois modelos separados de acordo com o tipo pessoa (PF ou PJ) do registro e foram utilizadas as mesmas composições do conjunto de dados treinamento, teste e validação.

Neste trabalho, optou-se pelo uso da técnica de análise discriminante linear (comparativamente à análise discriminante quadrática), uma vez que rejeitou-se a hipótese de normalidade dos dados. Tal resultado já era esperado, já que a maioria dos dados utilizados nesse trabalho são categóricos ou passaram por uma categorização na etapa de preparação do conjunto e, portanto, não seguem uma

distribuição normal. Apesar do método ser preferencialmente desenvolvido para dados normais, tal requisito tem menor impacto no ajuste final quando comparada à análise discriminante quadrática.

4 – RESULTADOS E DISCUSSÃO

Nessa seção serão apresentados os resultados obtidos para os modelos construídos ao longo deste trabalho, realizando então discussões sobre a qualidade dos ajustes e definindo qual o modelo mais adequado para o caso em questão.

4.1 Pessoa Física (PF)

4.1.1 Regressão logística múltipla PF

Os parâmetros do melhor ajuste obtido para o modelo logístico utilizando o conjunto de dados PF estão apresentados na Tabela 1:

Tabela 13 – Parâmetros obtidos para o modelo de Regressão Logística do público PF (pessoa física).

VARIAVEIS	CATEGORIAS	ESTIMATIVA	ERRO PADRÃO	p-VALOR	ODDS RATIO
INTERCEPT	-	0,85738	0,13585	<0,001	2,357
DIAS DESDE A ÚLTIMA COMPRA	ATÉ 430 DIAS	-0,2528	0,11036	0,023	0,777
	ENTRE 430 E 700 DIAS	-0,32182	0,12871	0,012	0,725
	ENTRE 700 E 1250 DIAS	-	-	-	-
	MAIS DE 1250 DIAS	-0,32678	0,10096	0,001	0,721
SINERGIA PÓS-VENDAS AUTOMÓVEIS	NÃO POSSUI	-1,58496	0,24123	<0,001	0,205
	ENTRE 1 E 4	-0,95546	0,08611	<0,001	0,385
	MAIS DE 4	-	-	-	-
SINERGIA BANCO	NÃO POSSUI	-	-	-	-
	APENAS UM	0,23568	0,09526	0,013	1,266
	MAIS DE UM	3,13083	0,46378	<0,001	22,893
SINERGIA COTA DE AUTOMÓVEIS	NÃO	-	-	-	-
	SIM	0,54628	0,15297	<0,001	1,727
SINERGIA POS-VENDA VEIC. COMERCIAIS	NÃO	-	-	-	-
	SIM	1,28315	0,3399	<0,001	3,608
VALOR DO PRODUTO	ATÉ R$ 50.000,00	-0,62391	0,13456	<0,001	0,536
	ENTRE R$ 50.000,00 E R$ 140.000,00	-0,38369	0,09943	<0,001	0,681
	MAIOR QUE R$ 140.000,00	-	-	-	-
GÊNERO	M	0,18768	0,08665	0,030	1,206
	F	-	-	-	-
ESTADO DO CLIENTE	PA	-0,1715	0,12906	0,184	0,824
	SP, RJ E MG	-0,37465	0,09729	<0,001	0,687
	MT	-	-	-	-
	OUTROS ESTADOS	-0,88906	0,45905	0,052	0,411

Fonte: Arquivo pessoal.

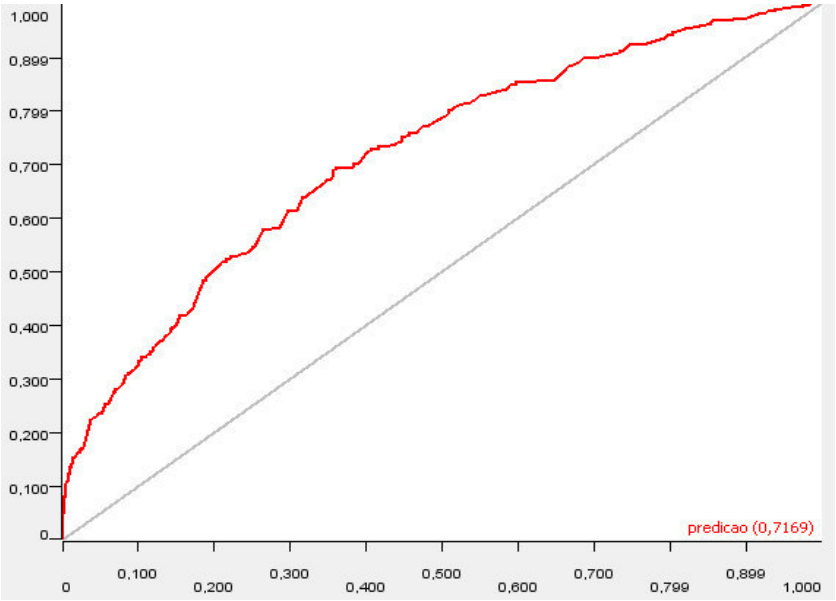
Pode-se observar que duas variáveis têm alta influência nas chances de um cliente ser considerado sucesso ou fracasso (*odds ratio*). São elas, a sinergia com o banco (22,893 vezes mais chance) e sinergia com pós vendas de veículos comerciais (3,608 vezes mais chance). Tais resultados fazem sentido quando analisamos os dados de negócios por trás dessas variáveis, sendo assim, os registros que possuem sinergia com o banco são clientes que conseguiram financiamento de crédito pela empresa anteriormente, logo, muito provavelmente, conseguiriam mais crédito para realizar uma recompra. Registros que possuem sinergia com pós vendas de veículos comerciais são clientes que possuem

caminhões, ônibus, vans, dentre outros, e, portanto, possuem maior poder aquisitivo e propensão de conseguir crédito para comprar um automóvel.

O modelo obtido foi então submetido à segunda etapa, de testes. Nessa etapa, o modelo é utilizado para classificar o conjunto de dados separado anteriormente para realização dos testes. Tal etapa é importante para verificar se o ajuste obtido é satisfatório.

A seguir, serão apresentadas as medidas a respeito da qualidade do modelo na etapa de testes, sendo essas, curva ROC, AUC, matriz de confusão e *Lift*.

Figura 8 - Curva ROC (AUC = 0,7169) obtida na etapa de testes do modelo logístico múltiplo PF.



Fonte: Arquivo pessoal.

Tabela 14 - Matriz de Confusão obtida na etapa de testes do modelo logístico múltiplo PF – Conjunto teste.

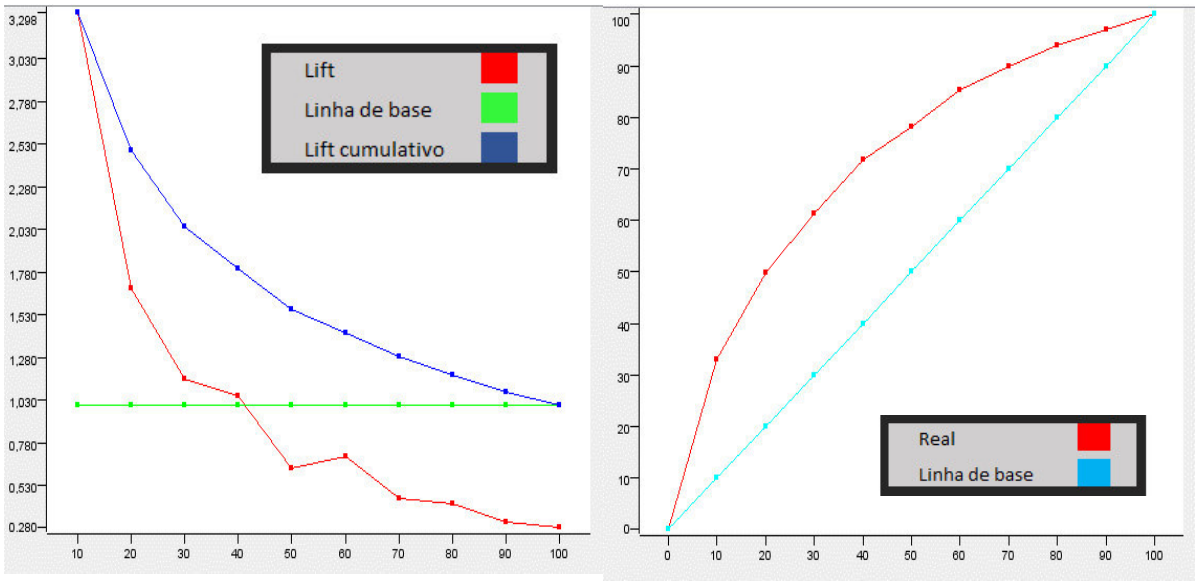
		Resultados classificados pelo modelo	
		Sucesso	Fracasso
Resultados reais observados	Sucesso	201	84
	Fracasso	21.854	50.129

Fonte: Arquivo pessoal.

Com base na matriz de confusão (Tabela 14), foram calculados os valores de Especificidade (taxa de acertos negativos) e a Sensibilidade (taxa de acertos positivos) para o ajuste obtido. Os valores são 0,696 de especificidade e 0,705 de sensibilidade. A taxa de acertos global para o modelo PF nessa etapa foi de 69,64%.

Outra medida que pode ser avaliada para modelos logísticos é o *Lift*, que é capaz de mensurar o poder de classificação e a vantagem de utilização do modelo em relação a ações aleatórias em cada decil. Segue abaixo os gráficos obtidos para o *lift* para o modelo logístico PF:

Figura 9 – Gráficos do *Lift* e dos ganhos cumulativos obtidos na etapa de testes do modelo logístico múltiplo PF.



Fonte: Arquivo pessoal.

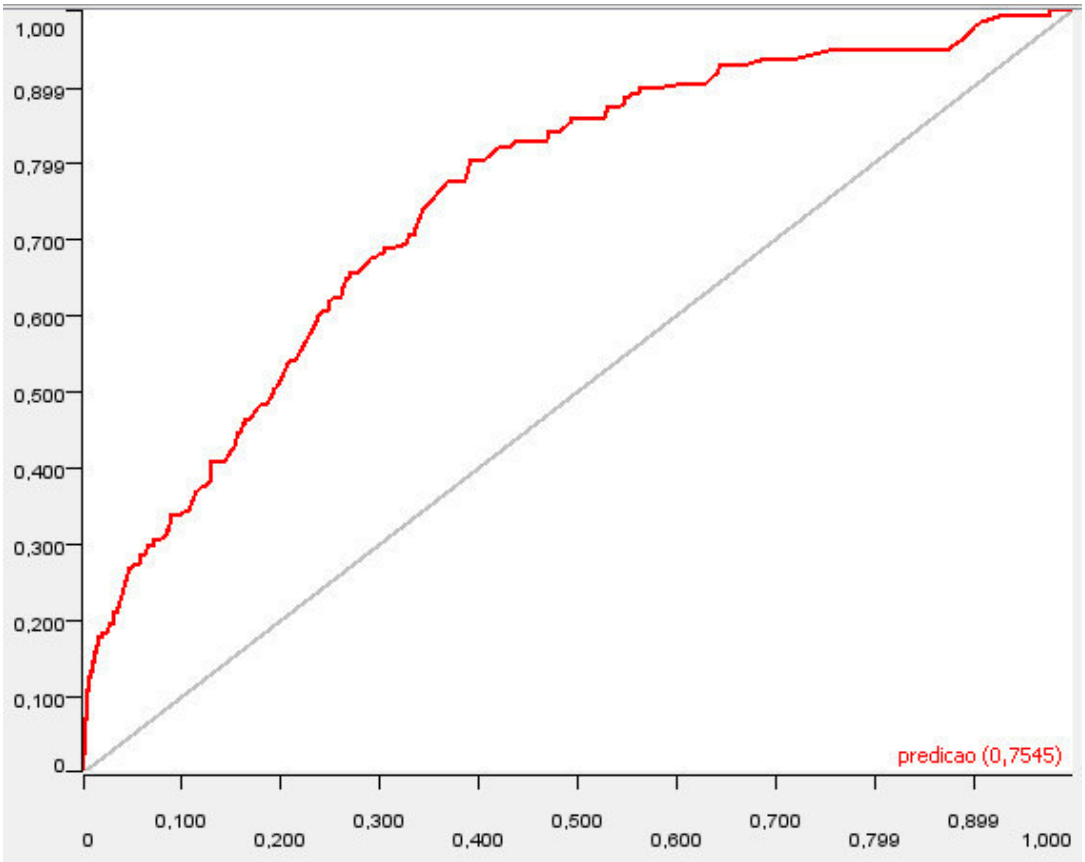
Analisando o resultado obtido pelo *Lift*, apresentado na imagem anterior, observa-se o grande poder preditivo do modelo obtido em cada decil. O poder máximo do modelo pode ser obtido no primeiro decil, onde os resultados chegam a ser 3,2 vezes melhores do que os resultados obtidos caso o modelo não fosse aplicado. Esse poder vai decaindo até o quarto decil, onde o valor ainda é um pouco superior ao aleatório.

Com os dados apresentados da etapa de testes, pode-se verificar que o modelo obtido teve um bom desempenho na previsão dos eventos do conjunto de testes, o que atesta a sua qualidade e permite seguir para a etapa seguinte.

A última etapa do desenvolvimento do modelo é a etapa de validação, que consiste em utilizar o modelo obtido para classificar um conjunto de dados diferente do utilizado para construção e teste e, a partir daí, avaliar se o *score* obtido consegue distinguir os eventos sucesso e fracasso. Essa etapa é muito importante para constatar o poder preditivo do modelo em outros conjuntos de dados, uma vez que o objetivo do modelo é ser utilizado para prever outras amostras de dados e gerar informações que permitam abordagens distintas para cada cliente.

Segue abaixo a curva ROC, medida AUC, matriz de confusão e *Lift* para ambos o modelo logístico PF na etapa de validação:

Figura 10 - Curva ROC (AUC = 0,7454) obtida na etapa de validação do modelo logístico PF.



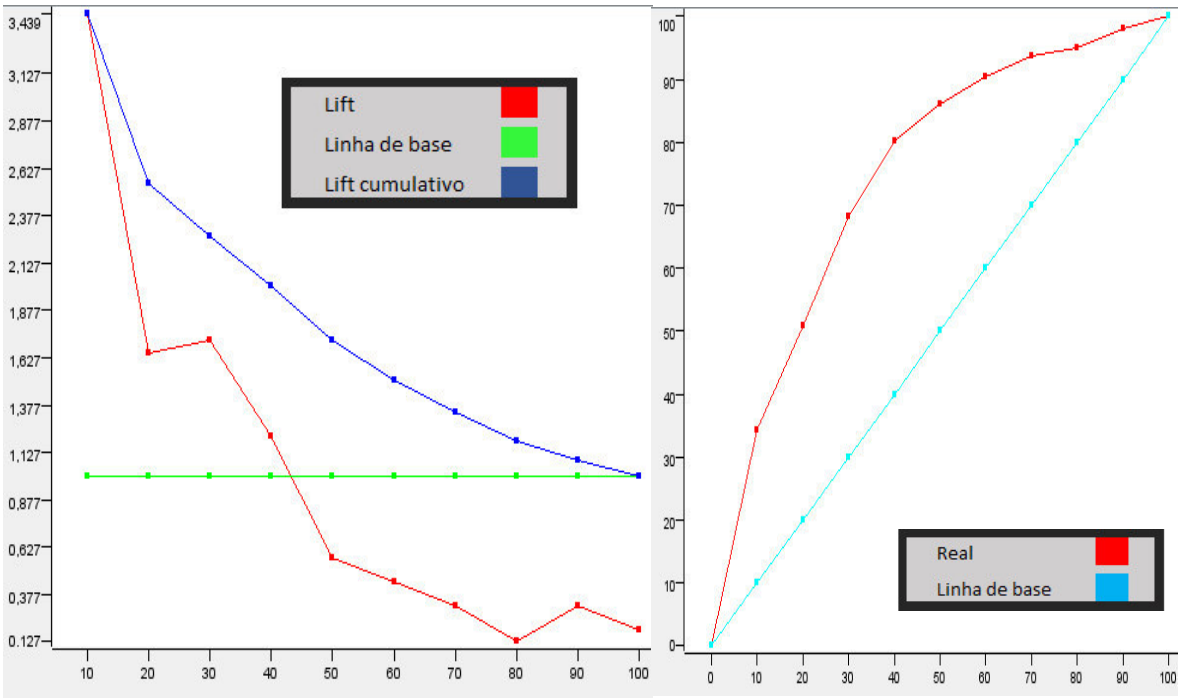
Fonte: Arquivo pessoal.

Tabela 15 - Matriz de Confusão obtida na etapa de validação do modelo logístico PF com um novo conjunto de dados.

		Resultados classificados pelo modelo	
		Sucesso	Fracasso
Resultados reais observados	Sucesso	111	46
	Fracasso	11.936	24.207

Fonte: Arquivo pessoal.

Figura 11 – Gráficos do *Lift* e dos ganhos cumulativos obtidos na etapa de validação do modelo logístico PF.



Fonte: Arquivo pessoal.

Os valores obtidos na etapa de validação para o modelo logístico PF na foram: 0,700 para especificidade e 0,707 para sensibilidade. A taxa de acertos global para o modelo PF nessa etapa foi de 66,99%.

Tais resultados comprovam a eficácia do modelo na predição da propensão de compra do grupo analisado, obtendo bons resultados quando submetido a um novo conjunto de dados. É possível observar uma queda na taxa de acertos global do modelo em relação à etapa de testes, o que já era esperado pelo fato de

utilizarmos um novo conjunto de dados, diferente daquele utilizado para construção e testes.

Analisando o *Lift* das etapas de testes e validação, observa-se uma certa constância nos resultados obtidos para o modelo logístico PF, que manteve o poder preditivo no modelo acima de 3 vezes no primeiro decil (3,29 na etapa de testes e 3,43 na etapa de validação).

Segue, na Tabela 16, os valores obtidos de sensibilidade, especificidade e taxa global de acertos do modelo logístico PF nas duas etapas de desenvolvimento do modelo: etapa de testes e etapa de validação.

Tabela 16 – Valores de sensibilidade, especificidade e taxa global de acertos obtidos nas etapas de testes e validação do modelo logístico PF.

	Etapa de testes	Etapa de Validação
Especificidade	0,696	0,700
Sensibilidade	0,705	0,707
Taxa global de acertos	69,65%	66,99%

Fonte: Arquivo pessoal.

4.1.2 Análise discriminante linear PF

Os parâmetros obtidos pelo modelo de análise discriminante linear para o público PF construído a partir do mesmo conjunto de dados que foi utilizado para construir o modelo de regressão logística podem ser observados na Tabela 17:

Tabela 17 – Parâmetros obtidos para o modelo de Análise Discriminante linear do público PF.

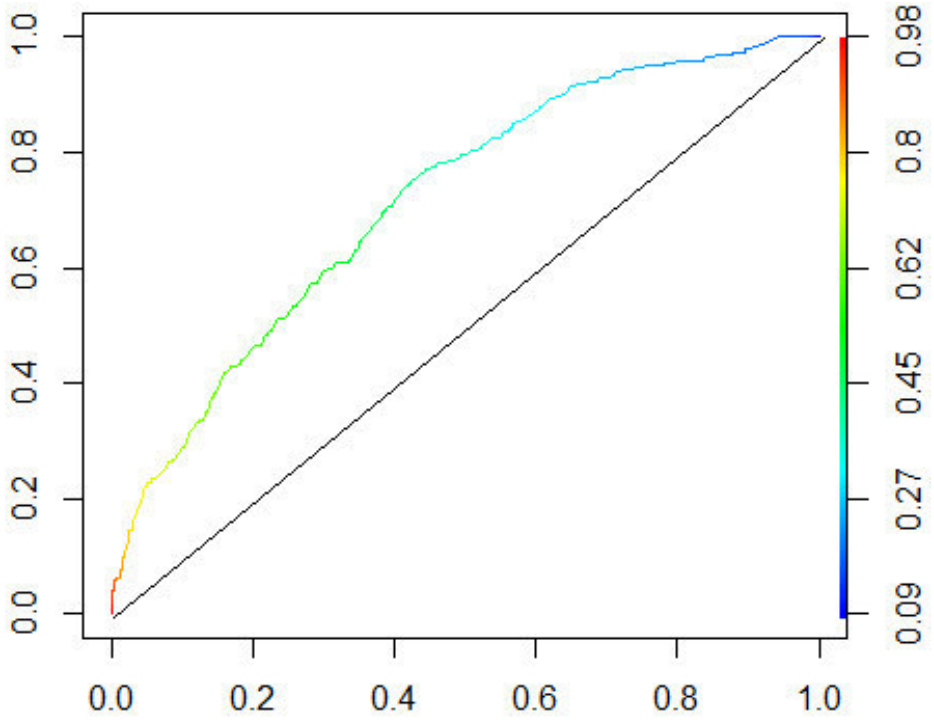
VARIAVEIS*	CATEGORIAS	COEFICIENTES (LD1)	MEANS	
			NÃO	SIM
DIAS DESDE A ÚLTIMA COMPRA	ATÉ 430 DIAS	0,235860	0,236636	0,235333
	ENTRE 430 E 700 DIAS	0,037082	0,147327	0,129074
	ENTRE 700 E 1250 DIAS	0,509866	0,279009	0,377445
	MAIS DE 1250 DIAS	-	-	-
PASSAGENS POS-VENDA AUTO	NÃO POSSUI	-	-	-
	ENTRE 1 E 4	0,558700	0,528031	0,258801
	MAIS DE 4	1,822893	0,413299	0,724902
PASSAGENS BANCO	NÃO POSSUI	-	-	-
	UMA PASSAGEM	0,243747	0,185789	0,234681
	MAIS DE UMA PASSAGEM	2,024325	0,004563	0,083442
POSSUI COTA DE AUTOMÓVEIS	NÃO	-	-	-
	SIM	1,036781	0,031943	0,108866
POSSUI COTA DE CAMINHÕES	NÃO	-	-	-
	SIM	0,404185	0,003259	0,013690
PASSAGEM POS-VENDA VEIC. COMERCIAIS	NÃO	-	-	-
	SIM	0,530429	0,013038	0,048240
VALOR DO PRODUTO	ATÉ R$ 50.000,00	-0,660153	0,225554	0,135593
	ENTRE R$ 50.000,00 E R$ 140.000,00	-0,425560	0,587353	0,501304
	MAIOR QUE R$ 140.000,00	-	-	-
GÊNERO	M	0,143961	0,614733	0,713820
	F	-	-	-
ESTADO DO CLIENTE	PA	-0,026034	0,132334	0,145372
	SP, RJ E MG	-0,420068	0,640808	0,473924
	MT	-	-	-
	OUTROS ESTADOS	-0,551516	0,018253	0,006519

Fonte: Arquivo pessoal.

É possível notar que, como esse método não faz a seleção das variáveis com maior poder preditivo e descarte das demais para a construção do ajuste, todas foram utilizadas como preditoras no modelo.

Para a etapa de testes do modelo discriminante linear foi utilizado o mesmo conjunto de dados segmentado anteriormente para teste do modelo logístico. Segue abaixo a curva ROC, a medida AUC e matriz de confusão para o modelo discriminante:

Figura 12 - Curva ROC (AUC = 0,7157) obtida na etapa de testes do modelo discriminante linear PF.



Fonte: Arquivo pessoal.

Tabela 18 - Matriz de Confusão obtida na etapa de testes do modelo discriminante linear PF.

		Resultados classificados pelo modelo	
		Sucesso	Fracasso
Resultados reais observados	Sucesso	172	113
	Fracasso	23.150	48.837

Fonte: Arquivo pessoal.

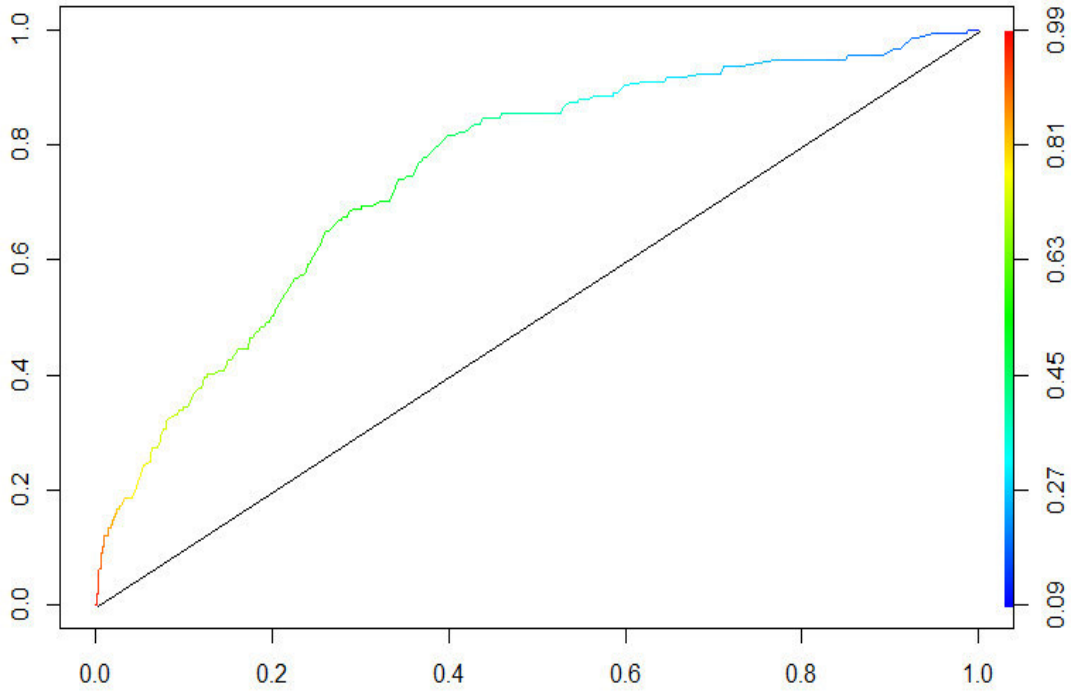
Com base na matriz de confusão, foram calculados os valores de Especificidade (taxa de acertos negativos) e a Sensibilidade (taxa de acertos positivos), e os valores obtidos para o modelo PF são: 0,678 e 0,603. A taxa de acertos para o modelo PF foi de 67,82%.

Com base nos dados apresentados, pode-se verificar uma qualidade satisfatória do ajuste para esse modelo e seguir para a última etapa de desenvolvimento, a etapa de validação.

Nesta etapa de validação, o modelo discriminante foi submetido ao mesmo conjunto de dados utilizado nessa etapa do modelo logístico. Os resultados obtidos pelo ajuste também se mostraram consistentes e evidenciaram a eficácia da técnica em prever o evento de interesse (recompra de automóvel).

Segue abaixo a curva ROC, medida AUC e matriz de confusão para o modelo discriminante linear PF:

Figura 13 - Curva ROC (AUC = 0,7511) obtida na etapa de validação do modelo discriminante linear PF.



Fonte: Arquivo pessoal.

Tabela 19 - Matriz de Confusão obtida na etapa de validação do modelo discriminante linear PF com um novo conjunto de dados.

		Resultados classificados pelo modelo	
		Sucesso	Fracasso
Resultados reais observados	Sucesso	116	41
	Fracasso	12.404	23.739

Fonte: Arquivo pessoal.

A partir da Tabela 19, foram calculados os valores de Especificidade (taxa de acertos negativos) e a Sensibilidade (taxa de acertos positivos), e os valores obtidos na etapa de validação para o modelo discriminante PF são: 0,657 e 0,739. A taxa de acertos para o modelo discriminante PF foi de 65,72%.

Segue, na Tabela 20, os valores obtidos de sensibilidade, especificidade e taxa global de acertos do modelo discriminante PF nas etapas de desenvolvimento:

Tabela 20 – Valores de sensibilidade, especificidade e taxa global de acertos obtidos nas etapas de testes e validação do modelo discriminante PF.

	Etapa de testes	Etapa de Validação
Especificidade	0,678	0,657
Sensibilidade	0,603	0,739
Taxa global de acertos	67,81%	65,72%

Fonte: Arquivo pessoal.

4.1.1 Comparação dos resultados obtidos pelos modelos de Regressão Logística Múltipla e Análise discriminante linear do conjunto de dados PF

De fato, ambas as técnicas se mostraram eficazes para o cumprimento dos objetivos desse trabalho para os clientes PF. Os modelos apresentaram bons resultados nos seus ajustes e superiores às abordagens aleatórias, o que já seria um ganho significativo e justificaria sua utilização no processo em questão. Na Tabela 21 estão apresentados os resultados de especificidade, sensibilidade e a taxa de acertos obtida para cada ajuste obtido no conjunto de dados PF:

Tabela 21 – Dados de especificidade, sensibilidade e taxa de acerto global do modelo logístico PF e discriminante PF.

	Etapa de testes				Etapa de Validação			
	AUC	Especificidade	Sensibilidade	Taxa global de acertos	AUC	Especificidade	Sensibilidade	Taxa global de acertos
Modelo Logístico PF	0,7169	0,696	0,705	69,64%	0,7454	0,700	0,707	66,99%
Modelo Discriminante PF	0,7157	0,678	0,603	67,81%	0,7511	0,657	0,739	65,72%

Fonte: Arquivo pessoal.

Comparando os valores da Tabela 23, nota-se que ambos os modelos apresentaram resultados constantes e pouca variação nos indicadores de eficiência entre as etapas, com grande parte dos indicadores ligeiramente melhores na etapa de validação. Isso poderia indicar um *underfitting* (quando o ajuste obtido não aprendeu muito bem sobre o conjunto de dados de treinamento e, por isso, não se adapta tão bem ao conjunto de dados de testes). Ainda, para o modelo discriminante, talvez pelo fato de utilizar no seu desenvolvimento todas as variáveis disponíveis no conjunto de dados permitiu mais informações e uma adaptabilidade maior para classificar um novo conjunto de dados, conforme desenvolvido na etapa de validação.

Analisando a taxa global de acertos do modelo, pode-se supor que os modelos obtidos apresentam resultados similares, com uma pequena vantagem para o modelo logístico, entretanto, por se tratar de um trabalho com o objetivo de classificar potenciais clientes entre compradores e não compradores, deve-se levar em conta os custos de má-classificação atrelado ao processo. Dessa forma, quando levamos em conta os custos de má classificação para modelos voltados para aplicações em negócios, a maior perda é representada pela classificação incorreta de um sucesso em fracasso (falso negativo), uma vez que, um indivíduo incorretamente classificado como sucesso (falso positivo) gera apenas o custo/esforço de realizar a abordagem, que poderia inclusive gerar uma oportunidade futura, entretanto, um indivíduo incorretamente classificado como fracasso ocasionaria a perda de um cliente e impactaria diretamente no resultado

da empresa. Dito isso, pode-se observar as taxas de falsos negativos em cada modelo analisando a Tabela 22:

Tabela 22 – Taxa de sucessos incorretamente classificadas como fracassos (valores falso negativos) em cada modelo PF

	Etapa de testes	Etapa de Validação
Modelo Logístico PF	29,47%	29,30%
Modelo Discriminante PF	39,65%	26,11%

Fonte: Arquivo pessoal.

Levando em conta os resultados da Tabela 22, nota-se que o modelo logístico PF, apesar de ter uma taxa de falso negativo um pouco maior na etapa de validação comparado ao modelo discriminante, ele apresentou um indicador quase constante entre as etapas de testes e validação. Já o modelo discriminante PF teve uma queda brusca na taxa entre as etapas o que, novamente, poderia indicar um *underfitting*.

Pode-se destacar ainda que a técnica de regressão logística tem uma vantagem em relação à discriminante para esse tipo de aplicação. Como o *output* do modelo logístico é um *score* (entre 0 e 1), que nesse trabalho pode ser definido como a probabilidade de um determinado cliente realizar a compra de um produto, cabe ao utilizador definir a faixa de corte para realização das abordagens, ponderando a relação risco e retorno decidido pelo negócio. Ou seja, se determinado utilizador decidir que quer atuar apenas com registros de *score* iguais ou superiores a 0,8, o volume de potenciais clientes para realização da ação provavelmente seria menor, mas o poder preditivo e, consequentemente as taxas de acertos das predições nessas faixas de *score* tendem a ser muito maiores. Tal afirmação pode ser confirmada pelo *Lift*, apresentado nas Figuras 9 e 11, onde vemos que no primeiro decil, o modelo logístico PF chega a seu poder máximo, sendo mais de mais 3 vezes melhor do que abordagens aleatórias e vai decaindo para os demais decis. Tal análise e possibilidade não é possível para os modelos discriminantes, uma vez que a saída dessa técnica é uma classificação apenas entre sim e não para o evento modelado, impedindo outros cortes que auxiliassem a tirar maior proveito o resultado. Outra vantagem do modelo logístico múltiplo frente ao modelo discriminante linear é a possibilidade de avaliar o efeito de cada

variável explicativa na variável resposta, verificando sua importância, bem como mensurando o efeito de cada uma delas na razão de chances de o cliente efetuar a compra.

4.2 Pessoa Jurídica (PJ)

4.2.1 Regressão logística múltipla PJ

Os parâmetros do melhor ajuste obtido para o modelo logístico utilizando o conjunto de dados PJ estão apresentados na Tabela 23:

Tabela 23 – Parâmetros obtidos pelo método de Regressão Logística Múltipla do público PJ (pessoa jurídica).

VARIAVEIS*	CATEGORIAS	ESTIMATIVA	ERRO PADRÃO	p-VALOR	ODDS RATIO
INTERCEPT	-	-0,2658	0,2815	0,344	0,767
DIAS DESDE A ÚLTIMA COMPRA	ATÉ 320 DIAS	1,1717	0,3031	<0,001	3,228
	ENTRE 320 E 1200 DIAS	0,6121	0,2612	0,019	1,844
	MAIS DE 1200 DIAS	-	-	-	-
SINERGIA PÓS-VENDAS AUTO	NÃO POSSUI	-1,3911	0,4339	<0,001	0,249
	ENTRE 1 E 10	-0,9027	0,2381	0,001	0,405
	MAIS DE 10	-	-	-	-
SINERGIA BANCO	NÃO	-	-	-	-
	SIM	0,6371	0,2483	0,010	1,891
SINERGIA COTA DE AUTOMÓVEIS	NÃO	-	-	-	-
	SIM	1,4009	0,3426	<0,001	4,059

Fonte: Arquivo pessoal.

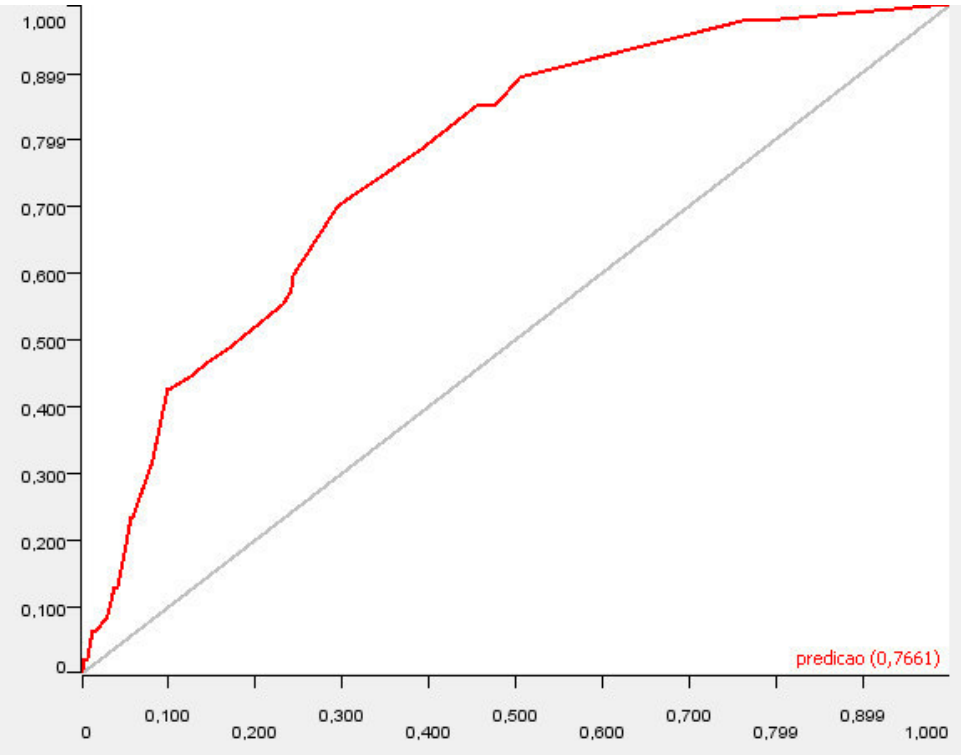
Observa-se que, para esse público, o método *Stepwise* selecionou apenas 4 variáveis para construção do melhor ajuste, dentre todas as disponíveis no conjunto de dados PJ. Uma análise preliminar dos modelos logísticos obtidos permite evidenciar os comportamentos diferentes dos públicos PF e PJ. Se forem levadas em conta as variáveis dependentes, por exemplo, podemos identificar que para o modelo PJ são mais relevantes informações de frequência de compra e histórico de produtos, enquanto o PF também considera informações de gênero e endereço (dados referentes ao indivíduo). Sabendo que *odds ratio* representa a

probabilidade de observar-se o sucesso (nesse caso, recompra de um veículo novo) para um indivíduo classificado em uma determinada categoria da variável dependente em relação à categoria utilizada como referência, pode se observar que as chances de sucesso para clientes PJ aumentam quanto menor for o período desde a última compra realizada, enquanto para o público PF, as chances de sucesso são maiores no período entre 700 e 1250 dias desde a última compra. Sendo assim, é possível evidenciar a diferença comportamental entre os públicos PF e PJ, o que justifica a escolha de trabalhar com esses públicos de forma separada.

Analisando os parâmetros obtidos pelo modelo, observam-se duas variáveis que têm alta influência nas chances de um cliente ser considerado sucesso ou fracasso (odds ratio). São elas, a sinergia cotas de automóveis (4,059 vezes mais chance) e dias desde a última compra (3,228 vezes mais chance para a categoria até 320 dias e 1,844 vezes mais chances na categoria entre 320 e 1200 dias). Esses resultados podem ser explicados pela estratégia do público PJ com o produto de estudo. Geralmente, clientes PJ adquirem várias cartas de consórcio de automóveis para trocarem suas frotas com um custo de crédito menor do que financiamento. Outro fator é que, essas empresas geralmente adquirem muitos veículos, às vezes até mais de um em um mesmo mês, o que aumenta as chances de uma recompra quanto menor o tempo da última compra em dias.

Segue abaixo a curva ROC, a medida AUC e matriz de confusão e *Lift* para o modelo logístico PJ na etapa de testes:

Figura 14 - Curva ROC (AUC = 0,7661) obtida na etapa de testes do modelo logístico múltiplo PJ.



Fonte: Arquivo pessoal.

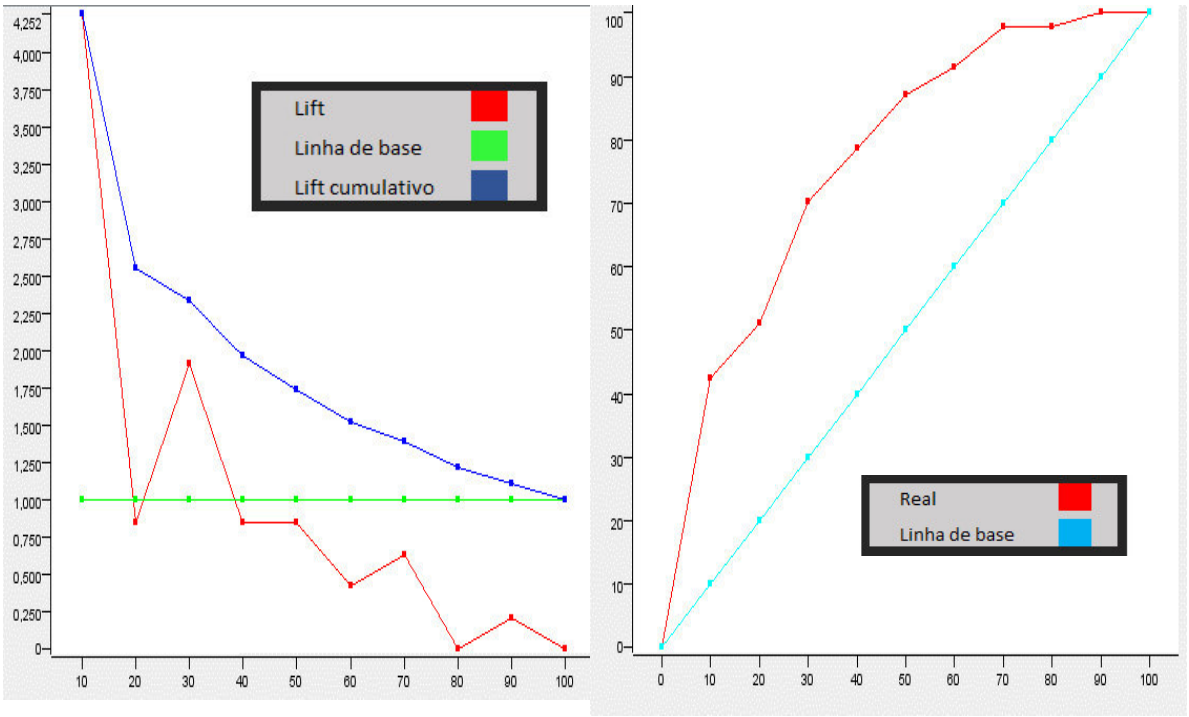
Tabela 24 - Matriz de Confusão obtida na etapa de testes do modelo logístico múltiplo PJ com 20% do conjunto de dados.

		Resultados classificados pelo modelo	
		Sucesso	Fracasso
Resultados reais observados	Sucesso	39	8
	Fracasso	3.816	5.910

Fonte: Arquivo pessoal.

Com base na matriz de confusão, foram calculados os valores de Especificidade (taxa de acertos negativos) e a Sensibilidade (taxa de acertos positivos), e os valores obtidos para o modelo logístico PJ são de: 0,607 e 0,829, respectivamente. A taxa de acertos global para o modelo logístico PJ foi de 60,87%. Também foram realizadas as análises de *Lift* apresentada na imagem abaixo:

Figura 15 – Gráficos do *Lift* e dos ganhos cumulativos obtidos na etapa de testes do modelo PJ.

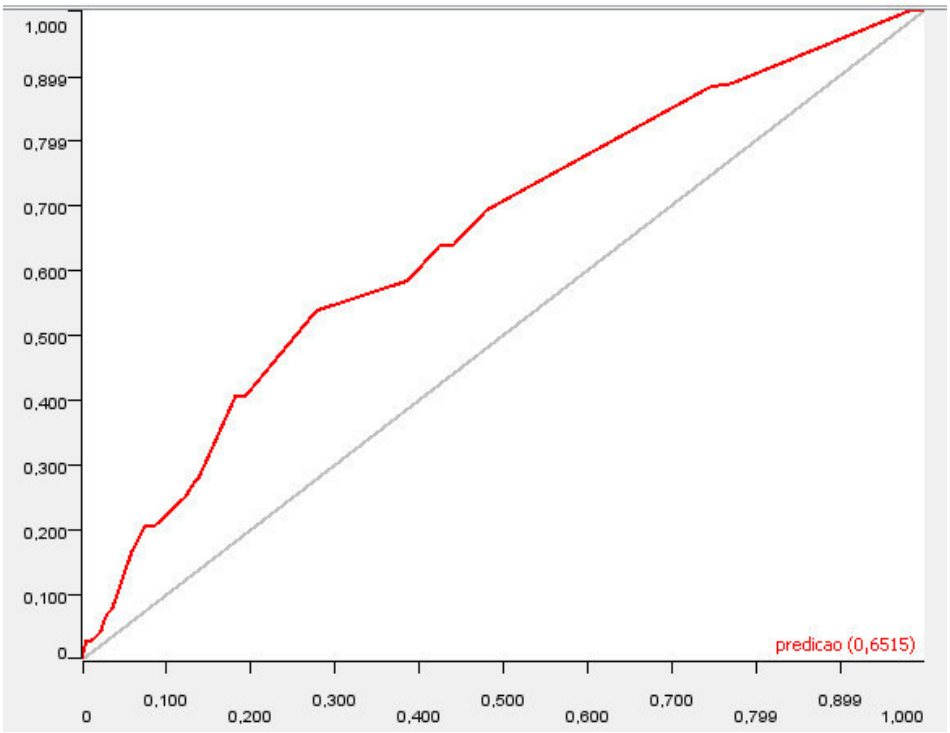


Fonte: Arquivo pessoal.

O *lift* obtido para o modelo logístico múltiplo PJ na etapa de teste se mostrou instável, com um bom resultado no primeiro decil, sendo 4,20 vezes melhor do que os resultados obtidos sem a utilização do modelo, mas decaiu no segundo decil, sendo pior do que o aleatório, e subiu novamente no terceiro decil, para 2 vezes melhor. Esse resultado pode estar ligado ao baixo volume de dados para construção do modelo PJ, com uma pequena quantidade de eventos sucesso ou pode indicar que as variáveis utilizadas não são suficientes para fornecer um ajuste que permita diferenciar com alta precisão os eventos sucesso e fracasso. Apesar disso, o resultado obtido pelo modelo na etapa de testes como um todo se mostrou satisfatório, o que possibilita partir para a etapa de validação do modelo.

Segue abaixo a curva ROC, a medida AUC, matriz de confusão e *Lift* para o modelo logístico múltiplo PJ na etapa de validação:

Figura 16 - Curva ROC (AUC = 0,6515) obtida na etapa de validação do modelo logístico múltiplo PJ.



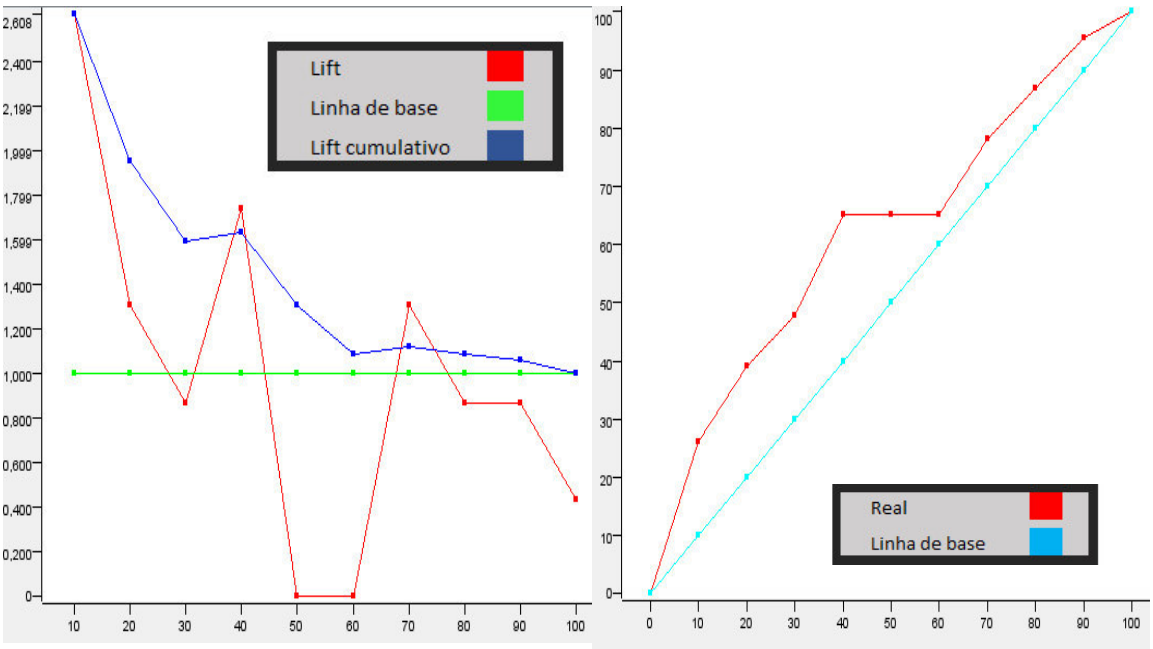
Fonte: Arquivo pessoal.

Tabela 25 - Matriz de Confusão obtida na etapa de validação do modelo logístico múltiplo PJ.

		Resultados classificados pelo modelo	
		Sucesso	Fracasso
Resultados reais observados	Sucesso	15	8
	Fracasso	1.534	2.103

Fonte: Arquivo pessoal.

Figura 17 – Gráficos do *Lift* e dos ganhos cumulativos obtidos na etapa de validação do modelo PJ.



Fonte: Arquivo pessoal.

Novamente, o *Lift* apresentou resultados instáveis ao longo de cada decil, mas isso já era esperado, uma vez que o número de sucessos no conjunto de dados de validação era muito pequeno.

O ajuste obtido se mostrou satisfatório e coerente com os objetivos do trabalho. Na Tabela 26 estão apresentadas as métricas obtidas pelo modelo logístico PJ nas etapas de testes e validação:

Tabela 26 – Valores de sensibilidade, especificidade e taxa global de acertos obtidos nas etapas de testes e validação do modelo logístico múltiplo PJ.

	Etapa de testes	Etapa de Validação
Especificidade	0,607	0,578
Sensibilidade	0,829	0,652
Taxa global de acertos	60,87%	57,87%

Fonte: Arquivo pessoal.

4.2.2 Análise discriminante linear PJ

Os parâmetros obtidos pelo modelo de análise discriminante podem ser observados na Tabela 27:

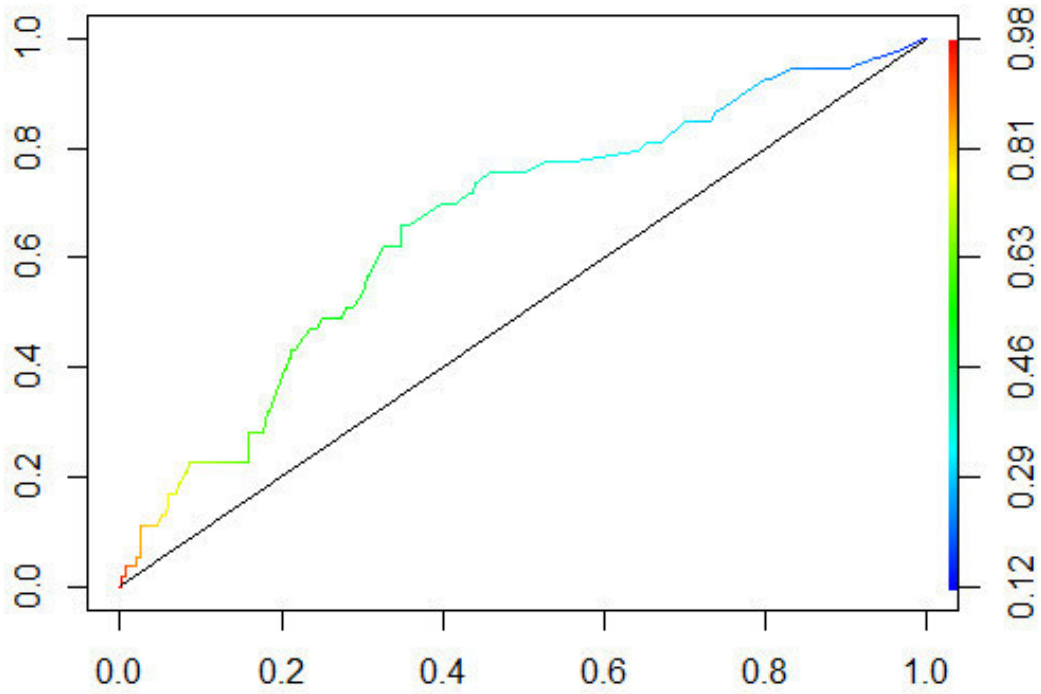
Tabela 27 – Parâmetros obtidos pelo modelo de Análise Discriminante linear do público PJ (pessoa jurídica) utilizando 80% do conjunto de dados.

VARIAVEIS*	CATEGORIAS	COEFICIENTES (LD1)	MEANS	
			NÃO	SIM
DIAS DESDE A ÚLTIMA COMPRA	ATÉ 320 DIAS	1,11103	0,19535	0,35349
	ENTRE 320 E 1200 DIAS	0,65201	0,46977	0,46977
	MAIS DE 1200 DIAS	-	-	-
PASSAGENS POS-VENDA AUTO	NÃO POSSUI	-	-	-
	ENTRE 1 E 10	-0,45703	0,75349	0,53023
	MAIS DE 10	0,51500	0,20000	0,41860
PASSAGENS BANCO	NÃO	-	-	-
	SIM	1,06437	0,13953	0,33953
POSSUI COTA DE AUTOMÓVEIS	NÃO	-	-	-
	SIM	1,15179	0,07442	0,25116
POSSUI COTA DE CAMINHÕES	NÃO	-	-	-
	SIM	0,93864	0,01860	0,14419
PASSAGEM POS-VENDA VEIC. COMERCIAIS	NÃO	-	-	-
	SIM	-0,27378	0,08372	0,12093
POSSUI COTA DE IMÓVEIS	NÃO	-	-	-
	SIM	0,53733	0,00000	0,02326
VALOR DO PRODUTO	ATÉ R$ 55.000,00	0,39424	0,14884	0,13953
	ENTRE R$ 55.000,00 E R$ 96.000,00	-0,15945	0,22326	0,18605
	MAIOR QUE R$ 96.000,00	-	-	-
ESTADO DO CLIENTE	PA, MT E RJ	-1,83725	0,53488	0,63721
	SP E MG	-2,17245	0,46512	0,35814
	OUTROS ESTADOS	-	-	-

Fonte: Arquivo pessoal.

Segue abaixo a curva ROC, a medida AUC e matriz de confusão para o modelo discriminante linear PJ:

Figura 18 - Curva ROC (AUC = 0,6602) obtida na etapa de testes do modelo discriminante linear PJ.



Fonte: Arquivo pessoal.

Tabela 28 - Matriz de Confusão obtida na etapa de testes do modelo discriminante linear PJ.

		Resultados classificados pelo modelo	
		Sucesso	Fracasso
Resultados reais observados	Sucesso	27	20
	Fracasso	2.987	6.739

Fonte: Arquivo pessoal.

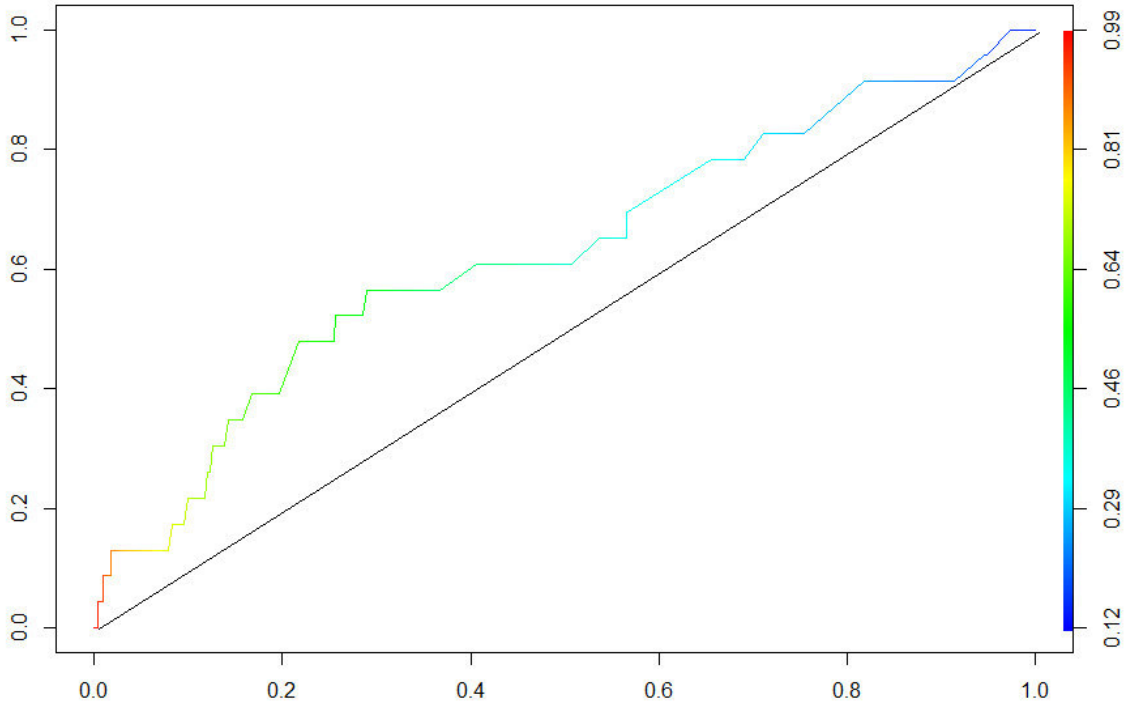
Com base na matriz de confusão (Tabela 28), foram calculados os valores de Especificidade (taxa de acertos negativos) e a Sensibilidade (taxa de acertos positivos), e os valores obtidos para o modelo discriminante PJ são: 0,693 e 0,574, respectivamente. A taxa de acertos foi 69,23%.

Com base nos dados apresentados, pode-se verificar uma qualidade de ajuste mediana para esse modelo e, então, podemos seguir para a última etapa de desenvolvimento, a etapa de validação.

Nesta etapa de validação, o modelo discriminante foi submetido ao mesmo conjunto de dados utilizado nesta etapa do modelo logístico.

Segue abaixo a curva ROC, medida AUC e matriz de confusão para o modelo discriminante PJ na etapa de validação:

Figura 19 - Curva ROC (AUC = 0,6272) obtida na etapa de validação do modelo discriminante linear PJ.



Fonte: Arquivo pessoal.

Tabela 29 - Matriz de Confusão obtida na etapa de validação do modelo discriminante linear PJ com um novo conjunto de dados.

		Resultados classificados pelo modelo	
		Sucesso	Fracasso
Resultados reais observados	Sucesso	13	10
	Fracasso	1.160	2.477

Fonte: Arquivo pessoal.

A partir da Tabela 29, foram calculados os valores de Especificidade (taxa de acertos negativos) e a Sensibilidade (taxa de acertos positivos), e os valores obtidos na etapa de validação para o modelo discriminante PJ: 0,681 e 0,565, respectivamente. A taxa de acertos global foi de 68,03%. Abaixo estão apresentados os resultados obtidos pelo modelo discriminante linear PJ nas etapas de testes e validação:

Tabela 30 – Valores de sensibilidade, especificidade e taxa global de acertos obtidos nas etapas de testes e validação do modelo discriminante PJ.

	Etapa de testes	Etapa de Validação
Especificidade	0,693	0,681
Sensibilidade	0,574	0,565
Taxa global de acertos	69,23%	68,03%

Fonte: Arquivo pessoal.

4.2.3 Comparação dos resultados obtidos pelos modelos de Regressão Logística e Análise discriminante do conjunto de dados PJ

No caso do conjunto de dados PJ, os modelos logístico de discriminante apresentaram qualidades de predição menos parecidas, com resultados razoáveis, permitindo assim a discussão de qual o melhor modelo para o objetivo final do trabalho.

Tabela 31 – Dados de especificidade, sensibilidade e taxa de acerto global do modelo logístico PJ e discriminante PJ.

	Etapa de testes				Etapa de Validação			
	AUC	Especificidade	Sensibilidade	Taxa global de acertos	AUC	Especificidade	Sensibilidade	Taxa global de acertos
Modelo Logístico PJ	0,7661	0,60765	0,82979	60,87%	0,6515	0,57822	0,65217	57,87%
Modelo Discriminante PJ	0,6602	0,69289	0,57447	69,23%	0,6272	0,68106	0,56522	68,03%

Fonte: Arquivo pessoal.

Analisando as informações contidas na Tabela 31, pode-se notar uma queda acentuada nos valores AUC e sensibilidade no modelo logístico PJ, o que poderia indicar um *overfitting*. Analisando também a taxa global de acertos dos modelos, pode-se supor que o modelo discriminante PJ é muito superior ao modelo logístico PJ, mas, novamente levando em conta os custos de má classificação atrelados ao processo, pode-se analisar as valores de sucessos que foram incorretamente classificados como fracassos na Tabela 32:

Tabela 32 – Taxa de sucessos incorretamente classificadas como fracassos (valores falso negativos) em cada modelo PJ

	Etapa de testes	Etapa de Validação
Modelo Logístico PJ	17,02%	34,78%
Modelo Discriminante PJ	42,55%	43,48%

Fonte: Arquivo pessoal.

Analisando a Tabela 32, nota-se uma grande alteração na taxa de sucessos incorretamente classificadas como fracassos (falso negativo) no modelo logístico PJ entre as etapas de testes e validação, entretanto, o ajuste obtido se mostrou bem superior ao modelo discriminante PJ, quando levamos em conta a finalidade do modelo e ainda, conforme dito anteriormente, o modelo logístico ainda tem vantagens em relação ao modelo discriminante por permitir trabalhar com cortes nos *scores* obtidos e avaliar o efeito de cada variável explicativa na variável resposta, verificando sua importância, bem como mensurando o efeito de cada uma delas na razão de chances do cliente efetuar a compra.

Ainda que o modelo logístico PJ apresente uma queda acentuada no *Lift* para tal ajuste de 4,25 na etapa de testes para 2,61 na etapa de validação, esse modelo ainda mostra bons resultados para atuação com a segmentação do score, o que é inviável para o modelo discriminante.

Pode-se observar que, ainda que os modelos obtidos para o público PJ sejam melhores do que o *status quo* e, portanto, satisfatórios, observa-se uma qualidade inferior à obtida para o público PF. A qualidade dos modelos apresentados obtidos reflete, em sua maioria, na qualidade dos dados utilizados e,

no caso do conjunto de dados PJ, a quantidade de registros e variáveis disponíveis para construção dos modelos foram aquém do desejado. Variáveis como renda, score de crédito, bens, faturamento médio, dentre outras que permitissem estimar o poder aquisitivo dos clientes, poderiam ser de grande aproveitamento para os modelos desse trabalho.

5 – CONCLUSÃO

Neste trabalho, foram analisados dados de perfil e histórico de compras de clientes do segmento de automóveis de uma empresa em um intervalo de 5 anos, a fim de prever o comportamento de compra desses clientes. Tal tema surgiu a partir do interesse da empresa em obter informações mais assertivas sobre seus atuais clientes a fim de realizar ações de vendas com direcionamentos melhores e conversões maiores. Partindo do fato de que tal empresa possuía uma base de mais de 300.000 clientes, tão ação é justificável, uma vez que permitiria aumento das vendas e redução dos custos de marketing.

Para chegar ao objetivo do trabalho, os dados coletados foram então analisados, manuseados e transformados a fim de remover inconsistências e construir um conjunto de dados em que pudessem ser aplicadas técnicas estatísticas. Tais análises identificaram dois públicos com características distintas, de acordo com o tipo pessoa do registro (pessoa física e pessoa jurídica), o que exigiu a separação do conjunto original entre esses públicos. O conjunto ainda foi separado para construção e treinamento do modelo, da seguinte forma: 80% para construção e treinamento do modelo e 20% para realização da etapa de testes.

Os modelos foram então construídos utilizando duas técnicas distintas, a regressão logística múltipla e a análise discriminante linear, cujos resultados foram comparados ao longo do desenvolvimento deste trabalho. Para validar a qualidade do ajuste, os modelos obtidos foram utilizados para classificar um outro conjunto de dados, diferente do que foi utilizado nas etapas de treinamento e testes, composto pelos clientes que realizaram uma compra até 2 meses após a última data do primeiro conjunto.

Com os resultados obtidos, pode-se concluir que os modelos obtidos para os públicos PF e PJ apresentaram bons resultados, ainda que para o público PJ tenham apresentado uma qualidade mediana, a performance obtida com sua utilização ainda representaria ganhos para o processo em relação *status quo*, o que por si só já justificaria a utilização de qualquer um dos modelos.

Para escolha do melhor modelo, foram apresentados os indicadores que representavam a qualidade dos ajustes obtidos, curva ROC, medida AUC,

especificidade, sensibilidade e *Lift*. Quando analisados esses resultados, poder-se-ia inferir que ambos os modelos obtiveram performances bem semelhantes, com uma pequena vantagem para a técnica logística no modelo PF com uma taxa de acertos global de 66,99% contra 65,72% para o modelo discriminante PF na etapa de validação, e uma grande superioridade do modelo discriminante PJ, com uma taxa de 68,03% contra 57,87% do modelo logístico PJ na etapa de validação.

Entretanto, por se tratar de modelos voltados para aplicações em negócios, devem ser analisados os custos de má classificação nesse tipo de utilidade. Nesse caso, a maior perda representa os clientes compradores (sucesso), mas que foram incorretamente classificados como não compradores (fracasso), denominados valores falso-negativos. Analisando esse indicador, o modelo discriminante PF obteve um resultado melhor na etapa de validação, com uma taxa de 26,11% contra 29,30% do modelo logístico PF na etapa de validação, apesar de apresentar uma oscilação nesse indicador entre a etapa de testes (39,65%) e validação, o que poderia indicar que, coincidentemente, esse modelo "adaptou-se" melhor ao conjunto ao conjunto de validação do que ao de teste (*underfitting*). Já para os modelos PJ, o modelo de regressão logística se mostrou bem superior ao discriminante, com resultados de 34,78%, contra 43,48%.

Comparando as particularidades e diferenças de cada uma das técnicas, nota-se que para a análise discriminante, foram utilizadas todas as informações disponíveis no conjunto de dados enquanto para regressão logística, a técnica *stepwise* seleciona apenas as variáveis mais significantes, com maior poder preditivo, para construir um ajuste parcimonioso para o conjunto de dados submetido, o que representa um ganho para o processo, uma vez que exige um esforço computacional menor. Analisando também o *output* de cada modelo, a técnica logística fornece um *score*, que pode ser definido neste trabalho como a propensão de compra de um determinado cliente. Tal *output* permite que sejam realizados cortes e selecionados apenas uma determinada faixa de utilização dos resultados, de acordo com a relação risco e retorno definidos pelo usuário, o que tende a maximizar os resultados obtidos pelo modelo. No caso do modelo discriminante, o *output* é apenas a classificação entre grupos "sim" e "não" para o evento modelado, impedindo um melhor aproveitamento. Outra vantagem do método logístico frente ao discriminante é a possibilidade de avaliar o efeito de cada

variável explicativa na variável resposta, verificando sua importância, bem como mensurando o efeito de cada uma delas na razão de chances de o cliente efetuar a compra, o que permitiu conclusões importantes a respeito das variáveis utilizadas.

Um ponto importante que pode ser discutido em todos os trabalhos de manuseio e modelagem de dados é que o ajuste obtido reflete muito nas informações disponíveis no seu conjunto de dados e nem sempre pode-se trabalhar com o conjunto de dados ideal para determinada finalidade, com todas as informações necessárias, quantidade de registros suficiente ou estruturados da forma correta e isso impacta diretamente no ajuste obtido. Neste trabalho, pode-se notar que o volume de dados para registros PJ é bem menor do que os registros PF e, decorrente disso, a quantidade de sucessos desse conjunto também são escassos. Nota-se ainda que, no conjunto de variáveis utilizadas para esse trabalho, não foram levadas em conta informações que pudessem inferir a saúde financeira ou o poder aquisitivo dos clientes de maneira mais direta, como pagamentos de parcelas feitos em dia, atrasados, score de crédito, renda familiar, classe socioeconômica, profissão/área de atuação da empresa, dentre outros dados que provavelmente seriam relevantes para um modelo dessa categoria, mas que não estavam disponíveis no momento da realização desse trabalho. Analisando esse cenário, quando comparamos os ajustes obtidos por ambas as técnicas, os modelos PJ não obtiveram resultados tão bons quanto os modelos PF, sofrendo muitas oscilações entre as etapas de testes e validação e com poucos dados para construção das análises. Tal condição pode ser decorrente do perfil desse tipo de registro, uma vez que se tratam de pessoas jurídicas, ou seja, empresas, que possuem comportamentos totalmente distintos dependendo do setor de atuação da mesma, porte e quantidade de funcionários, perfil de investimentos, concessão de crédito, resultados obtidos em negociações e outras informações que poderiam ser úteis para obtenção de um ajuste de maior qualidade.

Levando em conta todos os pontos citados anteriormente e os objetivos definidos nesse trabalho, a técnica de regressão logística parece ser a mais adequada, seja por utilizar um volume informações menor, selecionando apenas variáveis com maior poder preditivo, por apresentar uma taxa de falso negativos melhor (em consistência ou valor absoluto) ou ainda pelo fato de permitir cortes

para a utilização das faixas de *score* dos registros classificados de acordo com as estratégias do negócio.

Vale ainda ressaltar que uma técnica não exclui a outra e, como neste trabalho os resultados foram consistentes para ambas as técnicas, poderiam ser utilizados ambos os modelos juntos, desde que fosse computacionalmente viável, para classificação dos clientes da empresa, construindo um terceiro indicador com base na comparação das classificações obtidas por cada ajuste.

Por fim, as técnicas de regressão logística e análise discriminante mostraram bastante potencial para aplicações econômicas, o que justifica os grandes investimentos feitos pelas corporações nos últimos anos nas áreas de engenharia de dados, ciência de dados e estatística. Embora ainda existam melhorias possíveis nesse trabalho, como adição de informações mais consistentes com o tipo de uso do modelo ou até utilização de outras técnicas (como redes neurais, por exemplo), as ferramentas utilizadas nesse trabalho mostraram consistência nos seus resultados e já representam ganhos ao processo, quando comparadas ao *status quo*. Tais melhorias e técnicas poderiam ser considerados para futuros trabalhos nessa área.

REFERÊNCIAS

ADORNO, C. F. **MODELOS DE PROPENSÃO: OFERTA DE CRÉDITO PESSOAL**, 2011. .

BERRY, M. J. A.; LINOFF, G. S. **Data mining techniques: For Marketing, Sales and Customer Relationship Management**. [s.l: s.n.]v. 25

BOJANOWSKI, D. Z.; LOLATTO, G. A. APLICAÇÃO DE REGRESSÃO LOGÍSTICA E MODELOS COM FRAÇÃO DE CURA EM UM ESTUDO SOBRE CLIENTES INADIMPLENTES DE UMA INSTITUIÇÃO FINANCEIRA. 2018.

CABRAL, C. I. S. **Aplicação do Modelo de Regressão Logística num Estudo de Mercado**. [s.l: s.n.].

CASTELLANO, C.; FORTUNATO, S.; LORETO, V. Statistical physics of social dynamics. **Reviews of Modern Physics**, v. 81, n. 2, p. 591–646, 2009.

CÔRTES, S. da C.; PORCARO, R. M.; SÉRGIO, L. Mineração de Dados – Funcionalidades , Técnicas e Abordagens. **02/05/2002**, p. 34, 2002.

DAVENPORT, T. H.; BARTH, P.; BEAN, R. How “Big Data” Is Different. **MIT Sloan Management Review**, v. 54, n. 1, p. 22–24, 2012.

DILLON, W. R.; GOLDSTEIN, M. Multivariate Analysis: Methods and Applications. 1984.

EKBIA, H. et al. Big Data, Bigger Dilemmas: A Critical Review. **Journal of the American Society for Information Science**, v. 1, n. 6, p. 2581–2583, 2014.

FÁVERO, L. P.; BELFIORE, P.; SILVA, F. L. da; CHAN, B. L. **Análise de dados - Modelagem multivariada para tomada de decisões**, 2009. .

FISHER, R. A. THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS. **Annals of Eugenics**, v. v.7, p. p.179-188, 1936.

HAIR, J. F.; BLACK, W. C.; BABIN, B. J.; ANDERSON, R. E.; TATHAM, R. L. **Análise Multivariada de Dados**. 6ª Ediçã ed. [s.l: s.n.]

HASHEM, I. A. T. et al. The rise of “big data” on cloud computing: Review and open research issues. **Information Systems 47**, p. 98–115, 2014.

HENRIQUE, P.; FRANCISCO, O. P.; NETO, L. Medidas do Valor Preditivo de Modelos de Classificação Aplicados a Dados de Crédito Sumário. p. 1–41, 2008.

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. Second Edi ed. [s.l.] Jhon Wiley & Sons, Inc., 2000.

HUBERTY, C. J.; WISENBAKER, J. W.; SMITH, J. C. Assessing Predictive Accuracy in Discriminant Analysis. **Multivariate Behavioral Research**, p. 307, 1987.

JOHNSON, R. A.; WICHERN, D. W. **APPLIED MULTIVARIATE STATISTICAL ANALYSIS**. Sixth edit ed. [s.l: s.n.]

KLEINBAUM, D. G.; KLEIN, M. **Logistic Regression A Self-Learning Text Second Edition**. [s.l: s.n.]

KUNEVA, M. **Roundtable on Online Data Collection, Targeting and Profiling**. [s.l: s.n.].

MANZINI, G. **CASE GPA: COMO MELHORAR A EXPERIÊNCIA DO CONSUMIDOR USANDO DADOS E MOBILE**. Disponível em: <https://digitalks.com.br/noticias/case-gpa-como-melhorar-a-experiencia-do-consumidor-usando-dados-e-mobile/>. Acesso em: 14 abr. 2019.

MONTOGOMERY, D. C.; RUNGER, G. C. **Estatística aplicada e probabilidade para engenheiros**, 2011. .

ROTELLA, P. **Is Data The New Oil?** Disponível em: <https://www.forbes.com/sites/perryrotella/2012/04/02/is-data-the-new-oil/#346a62c67db3>. Acesso em: 14 abr. 2019.

SHOLOM M. WEISS, N. I. **Predictive Data Mining: A Practical Guide**. [s.l: s.n.]

SILVA, V. F. do S. **Modelos de Propensão ao Consumo baseados em Redes Neuronais Artificiais, o caso particular do Crédito Pessoal**. 2000. 2000.

TERRA. **Kodak: como a era digital se voltou contra um de seus criadores**. Disponível em: <https://www.terra.com.br/noticias/tecnologia/negocios-e-ti/kodak-como-a-era-digital-se-voltou-contra-um-de-seus-criadores,19382feb711ea310VgnCLD200000bbcceb0aRCRD.html>. Acesso em:

14 abr. 2019.

UN GLOBAL PULSE. **Big Data for Development: Challenges & Opportunities**. May. Disponível em: <http://www.unglobalpulse.org/sites/default/files/BigDataforDevelopment-UNGlobalPulseJune2012.pdf>. Acesso em: 14 abr. 2019.

VIKTOR MAYER-SCHÖNBERGER, K. C. **Big Data: A Revolution That Will Transform How We Live, Work, and ThinkHoughton Mifflin Harcourt**, 2013. .

APÊNDICE A – Regressão Linear Múltipla e Análise Discriminante Linear

Segue apresentado abaixo o código utilizado para construção dos modelos de regressão logística nesse trabalho:

```
my_data <- knime.in


modelo <-glm(TARGET ~ ., data = my_data, family = "binomial")
modelo_step  <- step(modelo,method = "forward")
exp(modelo_step$coefficient)
summary(modelo_step)
```

Para os modelos de análise discriminante, o código utilizado foi:

```
my_data <- knime.in


amostras.lda <- lda(SUCESSO~,data=my_data)
amostras.lda.values <- predict(amostras.lda)
View(amostras.lda)
Summary(amostras.lda)
```

A diferença entre a variável TARGET e SUCESSO é que, uma está expressa na forma de “0” e “1” (TARGET), enquanto a outra na forma de “SIM” e “NÃO” (SUCESSO).

TERMO DE PERMISSÃO DE USO DE INFORMAÇÕES

Através deste termo, nós da Rodobens Administradora de Consórcios, declaramos que estamos de acordo com a utilização das informações desta empresa no Trabalho de Conclusão de Curso intitulado "Utilização de análises estatísticas para estimar a propensão de compra de clientes do segmento de automóveis" desenvolvido pelo aluno Tayro Stringari de Toledo, a ser apresentado à Escola de Engenharia de Lorena no primeiro semestre de 2020.

São José do Rio Preto, 01 de janeiro de 2020.

NOME: Beatriz Rezende Lanjoni
CARGO: Gerente de CRM

Assinatura