UNIVERSIDADE DE SÃO PAULO

Instituto de Ciências Matemáticas e de Computação

Modelagem de propensão ao atrito no setor de telecomunicações

Rodolfo Augusto da Silva Arruda

Dissertação de Mestrado do Programa de Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria (MECAI)

SERVIÇO DE PÓS-GRADUAÇÃO DO ICMC-USP

Data de Depósito:

Assinatura: _____

Rodolfo Augusto da Silva Arruda

Modelagem de propensão ao atrito no setor de telecomunicações

Dissertação apresentada ao Instituto de Ciências Matemáticas e de Computação – ICMC-USP, como parte dos requisitos para obtenção do título de Mestre – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria.
EXEMPLAR DE DEFESA

Área de Concentração: Matemática, Estatística e Computação

Orientador: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
Fevereiro de 2019

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi
e Seção Técnica de Informática, ICMC/USP,
com os dados inseridos pelo(a) autor(a)

A779m Arruda, Rodolfo Augusto da Silva
Modelagem de propensão ao atrito no setor de
telecomunicações / Rodolfo Augusto da Silva Arruda;
orientador Francisco Louzada Neto. -- São Carlos,
2018.
49 p.

Dissertação (Mestrado - Programa de Pós-Graduação
em Mestrado Profissional em Matemática, Estatística
e Computação Aplicadas à Indústria) -- Instituto de
Ciências Matemáticas e de Computação, Universidade
de São Paulo, 2018.

1. Attrition. 2. Regressão Logística. 3. Random
Forest. 4. Algoritmos Genéticos. I. Louzada Neto,
Francisco, orient. II. Título.

Rodolfo Augusto da Silva Arruda

Modeling attrition propensity in the telecommunication sector

Master dissertation submitted to the Institute of Mathematics and Computer Sciences – ICMC-USP, in partial fulfillment of the requirements for the degree of the Master – Professional Masters in Mathematics, Statistics and Computing Applied to Industry. *EXAMINATION BOARD PRESENTATION COPY*

Concentration Area: Mathematics, Statistics and Computing

Advisor: Prof. Dr. Francisco Louzada Neto

USP – São Carlos
February 2019

Dedico este projeto a todos os amigos e familiares!

AGRADECIMENTOS

Agradeço à equipe de Inteligência da Qualidade pela convivência, ensinamentos e viabilidade deste projeto.

Aos companheiros de turma do MECAI, pela parceria e troca de conhecimentos.

Aos professores do ICMC São Carlos, pela dedicação.

Ao orientador, Prof. Dr. Francisco Louzada Neto, pelos ensinamentos e revisão do projeto.

Aos amigos do Max-House, pelo acolhimento e companheirismo.

Aos familiares, pelo apoio e compreensão.

RESUMO

ARRUDA, R. A. S. **Modelagem de propensão ao atrito no setor de telecomunicações**. 2019. 49 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

A satisfação dos clientes é fundamental para a manutenção do relacionamento com a empresa. Quando eles precisam resolver algum problema, a empresa necessita proporcionar bom atendimento e ter capacidade de resolutividade. No entanto, o atendimento massificado, muitas vezes, impossibilita soluções sensíveis às necessidades dos clientes.

A metodologia estatística pode ajudar a empresa na priorização de clientes com perfil a reclamar em um órgão de defesa ao consumidor (ODC), evitando assim uma situação de atrito. Neste projeto, foi realizada a modelagem do comportamento do cliente com relação à propensão ao atrito. Foram testadas as técnicas de Regressão Logística, *Random Forest* e Algoritmos Genéticos.

Os resultados mostraram que os Algoritmos Genéticos são uma boa opção para tornar o modelo mais simples (parcimonioso), sem perda de performance, e que o *Random Forest* possibilitou ganho de performance, porém torna o modelo mais complexo, tanto do ponto de vista computacional quanto prático no que tange à implantação em sistemas de produção da empresa.

Palavras-chave: *Attrition*, Regressão Logística, *Random Forest*, Algoritmos Genéticos.

ABSTRACT

ARRUDA, R. A. S. **Modeling attrition propensity in the telecommunication sector**. 2019. 49 p. Dissertação (Mestrado – Mestrado Profissional em Matemática, Estatística e Computação Aplicadas à Indústria) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos – SP, 2019.

Customer satisfaction is key to maintaining the relationship with the company. When these need to solve some problem, the company needs to provide good service and have resolving capacity. However, the mass service often makes it impossible.

The statistical methodology can help the company in the prioritization of clients with profile to complain in ODC, thus avoiding a situation of attrition. In this project was carried out the modeling of the behavior of the client in relation to the propensity to attrition. Logistic Regression, Random Forest and Genetic Algorithms were tested.

The results showed that the Genetic Algorithms are a good option to make the model simpler (parsimonious) without loss of performance and that Random Forest allowed performance gain, but it makes the model more complex, both from the point of view computational and practical in relation to the implantation in production systems of the company.

Keywords: textit Attrition, Logistic Regression, textit Random Forest, Genetic Algorithms.

LISTA DE TABELAS

Tabela 1 – Classificação do IV com relação ao poder preditivo	20
Tabela 2 – Matriz de Confusão.	28
Tabela 3 – Relação de Variáveis de chamadas e seus respectivos IV	31
Tabela 4 – Ponderação do Primeiro Componente Principal.	32
Tabela 5 – Formação do Componente - TT.	32
Tabela 6 – Componente Massiva/Programada.	33
Tabela 7 – Relação das variáveis de Reparo e seus respectivos IV	33
Tabela 8 – Relação de Variáveis de Motivo da chamada de Primeiro Nível e seus respectivos IV	34
Tabela 9 – Relação de Variáveis de Motivo da Chamada Comercial e seus respectivos IV	34
Tabela 10 – Relação de Variáveis de Histórico ODC e seus respectivos IV	35
Tabela 11 – Modelo de Chamadas	37
Tabela 12 – Modelo de Reparos	37
Tabela 13 – Modelo de Motivo Principal da Chamada	38
Tabela 14 – Modelo de Motivos Comerciais	38
Tabela 15 – Modelo de Passagem anterior ao ODC	39
Tabela 16 – Métricas de Predição para Modelos de Primeiro Nível - Treinamento	39
Tabela 17 – Métricas de Predição para Modelos de Primeiro Nível - Validação	40
Tabela 18 – Modelo de Segundo Nível - Regressão Logística	40
Tabela 19 – Métricas do Modelo de Segundo Nível - Regressão Logística	40
Tabela 20 – Modelo de Regressão Logística Parcimoniosa.	41
Tabela 21 – Métricas do Modelo Parcimonioso - Regressão Logística.	41
Tabela 22 – Probabilidade de Seleção do <i>Score</i> da população inicial	43
Tabela 23 – População Inicial com dez indivíduos - seed(123)	43
Tabela 24 – Métricas do Modelo Selecionado	44
Tabela 25 – Métricas de Predição para Modelos de Primeiro Nível - Treinamento	45
Tabela 26 – Métricas de Predição para Modelos de Primeiro Nível - Validação	45
Tabela 27 – Métricas do Modelo de Segundo Nível - RF	46

SUMÁRIO

1	INTRODUÇÃO	12
1.1	Apresentação do Problema	12
1.2	Regulação no Setor de Telecomunicações	13
1.3	Organização do trabalho	14
2	REVISÃO BIBLIOGRÁFICA	15
2.1	Introdução à Modelagem	15
2.2	Planejamento e Definições	15
2.3	Variáveis Potenciais	16
2.4	Planejamento Amostral	17
2.5	Construção do <i>Score</i>	18
2.5.1	<i>Pré-Processamento - Análise Exploratória</i>	18
2.5.2	<i>Pré-Processamento - Seleção de Variáveis</i>	19
2.5.2.1	<i>Filter vs Wrapper</i>	19
2.5.2.2	<i>Filter: Information Value (IV)</i>	20
2.5.2.3	<i>Parcimônia x Visão de Negócio</i>	21
2.6	Análise de Componentes Principais (PCA)	21
2.7	Algoritmos de Classificação	22
2.7.1	<i>Regressão Logística</i>	22
2.7.2	<i>Árvores de Decisão</i>	23
2.7.3	<i>Combinação de Classificadores</i>	24
2.7.4	<i>Random Forest</i>	25
2.8	Algoritmos Genéticos	26
2.9	Performance dos Algoritmos	27
2.9.1	<i>Métricas da Matriz de Confusão</i>	28
3	MATERIAIS E MÉTODOS	30
3.1	Desenho do Estudo	30
3.2	Bases de Dados e Seleção de Variáveis	30
3.2.1	<i>Dados de Chamadas</i>	31
3.2.2	<i>Dados de Reparos</i>	32
3.2.3	<i>Dados de Motivo das Chamadas</i>	33
3.2.4	<i>Passagem anterior pelo ODC</i>	35

4	RESULTADOS	36
4.1	Modelagem	36
4.1.1	<i>Modelos de Primeiro Nível</i>	37
4.1.2	<i>Poder Preditivo dos Modelos de Primeiro Nível</i>	39
4.1.3	<i>Modelo de Segundo Nível</i>	40
4.1.4	<i>Poder Preditivo dos Modelos de Segundo Nível</i>	40
4.2	Regressão Logística Parcimoniosa	41
4.3	Metodologias Alternativas de Modelagem	41
4.3.1	<i>Algoritmos Genéticos</i>	42
4.3.1.1	<i>Codificação, População inicial e fitness do GA</i>	42
4.3.1.2	<i>Crossover, Mutação e Seleção</i>	43
4.3.1.3	<i>Simulações com GA</i>	44
4.3.2	<i>Stacking Random Forest</i>	45
4.4	Conclusão	46
	REFERÊNCIAS	48

INTRODUÇÃO

1.1 Apresentação do Problema

Manter os clientes é tão importante quanto a conquista de novos. Dentro do ciclo de vida (atração, aquisição, conversão, retenção e fidelização) (CUTLER; STERNE, 2000), a empresa precisa entender e solucionar os motivos de insatisfação para garantir a manutenção do relacionamento. Diante de uma realidade em que a quantidade de clientes é muito grande, a tratativa individualizada é inviável, sobretudo devido ao alto custo operacional envolvido. Assim, as empresas adotam táticas de atuação que se iniciam pela segmentação dos clientes em diferentes perfis, quase sempre priorizando o retorno financeiro como métrica e, assim, oferecem aos diferentes perfis tratamentos diferenciados.

Por outro lado, todos os clientes, independente do segmento ao qual foram alocados, esperam um bom tratamento/atendimento. No entanto, quando precisam entrar em contato com uma empresa, os perfis aos quais foram enquadrados são determinantes na experiência vivenciada no atendimento, pois o direcionamento das chamadas às ilhas são planejadas segundo a segmentação. Basicamente, existem duas categorias de serviços oferecidos no atendimento: o personalizado, que é uma vertente especializada em serviços com maior valor agregado e o massificado, que trata o grande volume das chamadas de forma padronizada. As diferenças entre essas duas formas de atendimento refletem, sobretudo, no tempo de espera para falar com um atendente, na eficácia da solução de problemas e na eficiência.

A experiência com o *Call Center* é determinante para a satisfação do cliente, principalmente, em situações de atrito, nas quais ele precisa ter seus problemas ou dúvidas sanadas de forma efetiva. Se o atendimento não for adequado, o cliente pode buscar órgãos de defesa do consumidor (ODC) para intermediarem as relações com as empresas. Essa situação é ruim para o cliente que se desgasta na busca dessa alternativa, mormente em termos do tempo despendido e também para a empresa que pode ter sua imagem e reputação prejudicadas, além de ser suscetível

de punção pelo órgão regulador.

Diante desse cenário, faz muito sentido o desenvolvimento de uma sistemática capaz de direcionar os clientes, em momentos críticos, para o atendimento personalizado. Ou seja, ao invés de o direcionamento ser feito simplesmente pela segmentação de valor, podem ser utilizadas técnicas de modelagem estatística, baseadas no histórico de comportamento do cliente, para direcionar para ilhas especiais parte do público que esteja passando por uma situação delicada que precisa de tratamento diferenciado.

Esta dissertação de Mestrado buscou a criação de uma inteligência capaz de direcionar para um atendimento personalizado clientes propensos a recorrer a um ODC. Para isso, foram selecionadas bases de dados comportamentais dos clientes e, a partir delas, foram criadas variáveis que serviram de *input* para os modelos estatísticos. Como resultado da modelagem estatística foi obtida uma lista ordenada, de forma que os clientes mais propensos ao atrito ficaram no topo da lista.

1.2 Regulação no Setor de Telecomunicações

As Telecomunicações no Brasil ganharam relevância a partir da década de 50. Esta época foi marcada pelo avanço tecnológico, que permitiu a ampliação da oferta do serviço e o aumento da base de clientes. Nesse período a exploração do serviço era desordenada, com pouca abrangência territorial e de baixa qualidade. Em 1962, o setor teve a sua primeira ação governamental com a Lei 4.117, que instituía o Código Brasileiro de Telecomunicações e disciplinava a prestação de serviço, colocando-o sob o controle do governo.

Em 1967, foi criado o Ministério das Comunicações, o qual passou a fiscalizar as diversas concessionárias do serviço telefônico. Em 1972, o Ministério das Comunicações, por meio da Lei 5.792, propôs uma nova estrutura para o setor, que até então contava com cerca de mil prestadoras de serviço. Foi a criação da TELEBRÁS, a qual incorporou as companhias telefônicas existentes e que, na concepção, seria a grande prestadora de serviços de telecomunicações nacional (NEVES, 2002).

Nos anos 80, as modificações no cenário político e a piora na situação econômica do país reverteram o ritmo de crescimento do setor no País e apontavam para o esgotamento do modelo. Na década de 90, ficou evidente o esgotamento do modelo e as dificuldades em financiar o setor levaram à promulgação da Lei Mínima (a qual permitiu que houvesse a prestação privada de serviços de telefonia móvel celular e de satélite) e da Lei Geral das Telecomunicações que, dentre outros tópicos, tratava do direito do consumidor e da criação da ANATEL e da definição de seu papel como órgão regulador.

Concomitantemente a esse processo, a economia mundial se transformou na década de 90, sobretudo pelo processo de fusões e aquisições e a liberalização dos mercados. O setor

de telecomunicações foi alvo de um grande dinamismo nas operações desse tipo, que exigiu do órgão regulador grande atenção com objetivo de evitar situações que prejudicassem o livre mercado e o direito dos consumidores (PIRES; DORES, 2000).

A ANATEL foi encarregada de construir o arcabouço regulatório, destacando-se: a) Plano Geral de Outorgas que dividiu o Brasil em áreas de prestação de serviço telefônico; b) Plano de Metas de Universalização que visou garantir que todo cidadão tivesse acesso ao serviço; c) Plano de Metas de Qualidade que estabeleceu regras claras para prestação do serviço, estabelecendo condições para reparo, atendimento em uma loja, atendimento por meio de telefone e todo o conjunto de informações e definições necessárias à prestação do serviço (SILVA, 2000).

1.3 Organização do trabalho

Para a elaboração desta dissertação foi utilizado um conjunto de dados reais de uma empresa de Telecomunicações. Um modelo inicial foi elaborado dentro da empresa com restrição de tempo. A metodologia seguiu o fluxo tradicional de modelagem de *Score* pelo mercado, como seleção de variáveis via filtragem por *Information Value* e construção dos *Scores* por meio de Regressão Logística. Uma inovação proposta consistiu na construção de *Scores* de primeiro nível para cada tema de variáveis explicativas, por exemplo, comercial, técnico etc. e na posterior combinação deles via *Stacking* com outra Regressão Logística. Com o projeto de pesquisa foi possível testar outras técnicas de modelagem como *Random Forest* (RF) e Algoritmos Genéticos (GA).

A dissertação foi organizada da seguinte forma. No Capítulo 2, foi apresentada a revisão bibliográfica sobre as etapas utilizadas para a realização da modelagem estatística e a avaliação do modelo. No Capítulo 3, foi apresentado o planejamento amostral e as variáveis disponíveis para a construção dos modelos e seus respectivos potenciais de discriminação. Por fim, no Capítulo 4, as performances dos modelos foram avaliadas sob diversas métricas e foi finalizada com a discussão dos resultados obtidos.

REVISÃO BIBLIOGRÁFICA

2.1 Introdução à Modelagem

O desenvolvimento de um modelo consiste, de forma geral, em determinar a relação entre as diversas características dos clientes, com um fenômeno de interesse. Especificamente, quando o produto final da modelagem é um *Score*, as principais etapas dessa atividade são (DINIZ; LOUZADA, 2013):

1. Planejamento e definições;
2. Identificação de variáveis potenciais;
3. Planejamento amostral;
4. Determinação do *score*: aplicação da metodologia estatística;
5. Validação e verificação da performance do modelo estatístico;
6. Determinação do ponto de corte ou faixas de escore;
7. Determinação de regra de decisão.

Essas etapas foram detalhadas neste Capítulo para o tipo de modelo e problema deste trabalho, sendo que a sequência dos passos foi adaptada.

2.2 Planejamento e Definições

Antes de executar a construção do modelo é preciso ter bem claro algumas características como: o público ou o produto de interesse, a forma de representar o problema em questão na

linguagem matemática, a periodicidade de atualização do modelo e, principalmente, de qual etapa do processo decisório o modelo fará parte.

A definição do público ou produto de interesse é importante devido ao fato de esse recorte tornar o modelo sensível às suas determinadas características quanto à seleção de atributos preditivos. Caso contrário, o modelo pode tornar-se genérico e perder poder preditivo quando ajustado para todos os clientes sem a discriminação de suas características.

A representação matemática do problema traduz a necessidade do negócio na equação do modelo. Nesse momento, o pesquisador define qual o tipo de variável resposta (representação do fenômeno aleatório) ele irá modelar. Geralmente, existe uma simplificação muito comum nessa definição que é a utilização de uma variável binária em que o número um é utilizado para apontar os clientes com uma determinada característica e o número zero, para apontar os clientes que não apresentam tal característica. Por exemplo, se o interesse do estudo é dizer quais clientes são propensos a reclamar no ODC, então, pode-se criar a variável binária com a marcação 0 e 1.

A periodicidade de atualização do modelo é importante, pois a partir dela é que é definida a recorrência de atualização do banco de dados com as variáveis explicativas. Por exemplo, se o modelo terá atualização mensal, então, será necessário atualizar os atributos explicativos uma vez por mês. No entanto, se o modelo for pensado para trabalhar *online*, toda nova informação que faz parte do modelo deve ser imediatamente atualizada.

Por fim, deve-se ter bem claro como será realizada a implantação do modelo em produção e como ele fará parte do processo decisório da empresa. Nessa etapa, são definidas as estratégias ou os planos de negócio que farão uso do modelo.

2.3 Variáveis Potenciais

As variáveis utilizadas para a construção do modelo podem ser divididas basicamente em dois tipos: cadastrais e comportamentais. As variáveis cadastrais dizem respeito tanto ao indivíduo, como idade, sexo, escolaridade, renda, como também ao seu relacionamento com a empresa, como produto, segmento, tempo como cliente, etc. As variáveis comportamentais são referentes ao histórico de interações entre o cliente e a empresa, como por exemplo, a quantidade de ligações efetuadas, a quantidade de problemas técnicos relatados, a receita gerada, etc.

Essas informações, na estrutura tradicional, são armazenadas em tabelas relacionais de bancos de dados e são identificadas por uma ou mais chaves, tais como CPF ou número do telefone. Portanto, na fase de identificação das variáveis potenciais, o pesquisador deve verificar quais são as tabelas disponíveis que contêm informações cadastrais ou comportamentais sobre o público ou produto de interesse. A partir daí, verificar quais os possíveis atributos serão construídos a fim de servir de *input* para o modelo estatístico.

2.4 Planejamento Amostral

O desenho amostral deve ser iniciado pela definição do público alvo, ou seja, qual população a amostra vai representar. Para essa delimitação, geralmente, são utilizados os produtos que o cliente possui ou o segmento ao qual pertence. Feito esse recorte, a próxima etapa é avaliar a incidência do fenômeno em estudo ao longo do tempo. Com essa taxa é possível verificar se o banco de dados é balanceado ou se o pesquisador está diante de um evento raro.

Segundo (DINIZ; LOUZADA, 2013), o desbalanceamento do banco de dados pode prejudicar o desenvolvimento do modelo, pois se uma das classes é pouco representativa pode ser difícil estabelecer correlações com as variáveis explicativas e observar diferenças relevantes com a classe predominante. Com isso, uma amostragem aleatória simples não é indicada, sendo necessária uma metodologia denominada *Oversampling* ou *State Dependent*, que consiste em aumentar a proporção do evento raro na amostra. Geralmente, selecionam-se todos os clientes da categoria de baixa frequência e sorteiam-se os clientes da outra categoria. Segundo (SILVERMAN DT, 1992), não existe ganho relevante em manter a proporção entre as categorias superior a 4:1 - “*there is usually little marginal increase in precision from increasing the ratio of controls to case beyond four*”.

A sazonalidade na ocorrência do evento pode comprometer a performance do modelo, sobretudo, se a estrutura de correlações for determinada em um período atípico. Para minimizar esse efeito a amostra é composta de forma que os clientes sejam selecionados em diversos momentos no tempo. Para o caso em que a variável resposta é binária, pode-se selecionar todos os clientes nos quais a incidência do evento foi observada em diversos meses e sortear, em cada um dos meses, os demais clientes. A visão mensal para a coleta dos clientes, cujo fenômeno não foi observado, é importante para garantir que eles estejam expostos aos mesmos fatores de risco, segundo (SILVERMAN DT, 1992) “*Controls are selected randomly from the ‘risk set’ the subjects in the cohort who are at risk at the time of diagnosis of each case*”.

Os clientes da amostra devem ser identificados por alguma chave única, como por exemplo, CPF ou telefone, para a etapa de enriquecimento do banco de dados com informações explicativas. Além disso, deve-se manter a referência do mês ao qual o cliente representa no plano amostral. A última informação é útil para a divisão do conjunto de dados em duas partes: treinamento e validação. Quando a divisão do banco de dados é feita considerando os períodos de tempo mais antigos para o treinamento e os mais recentes para a validação, diz-se que a validação ocorre *out of time*. A separação do banco de dados é importante, pois a performance deve ser avaliada em parte da amostra que não foi utilizada para ajustar o modelo, o que permite saber se o modelo de fato aprendeu as relações relevantes ou se está superajustado aos dados de treinamento (*overfitting*).

2.5 Construção do *Score*

Após determinado o planejamento amostral e obtidas as variáveis potenciais, seguimos alguns procedimentos descritivos a fim de verificar se elas não possuem algum tipo de sujeira ou ruído. Também, nessa etapa, procuramos reduzir nosso banco de dados, com o intuito de eliminar as variáveis redundantes e com pouco poder discriminativo. A esses passos denominamos pré-processamento.

2.5.1 Pré-Processamento - Análise Exploratória

Nessa fase da análise do banco de dados preocupamo-nos em entender quais são os tipos dos atributos disponíveis para a modelagem, que podem ser dos tipos qualitativos (nominal ou ordinal) ou quantitativos (contínuo ou discreto) (BOLFARINE; BUSSAB, 2005). Para cada um deles existe uma maneira de representar as estatísticas descritivas que, no primeiro, caso basicamente se dá pela análise de frequência e, no segundo, por medidas de centralidade e variabilidade.

Para as variáveis qualitativas nominais, por exemplo, sexo, é importante verificar se para ambas as classes (masculino e feminino) existe uma quantidade de pelo menos 5% de preenchimento. Também é importante verificar se não existe alguma classe que represente um erro preenchimento. Ainda é importante contabilizar qual a relevância dos valores faltantes, uma vez que variáveis com muitos *missing values* prejudicam o ajuste do modelo. A esses passos damos o nome de análise univariada. Na sequência, deve-se realizar a análise cruzada ou bivariada em que a variável é confrontada com a variável resposta ou *target*. Dessa análise, é possível extrair se a variável tem potencial discriminativo com relação ao objetivo.

As variáveis qualitativas ordinais, por exemplo, segmento do cliente, são analisadas de forma semelhante às variáveis nominais, com a preocupação adicional de verificação de uma possível ordenação da taxa de risco ou propensão com as classes da variável. Por exemplo, se para o negócio é esperado que clientes do segmento A sejam menos propensos do que clientes do segmento B e, respectivamente, os clientes do segmento C sejam mais propensos do que os do segmento B, é importante verificar se na amostra está presente esse comportamento.

As variáveis quantitativas (contínuas ou discretas), no geral, são analisadas da mesma forma. No caso univariado, é feito via cálculo de média e desvio padrão, gráficos *boxplot* para identificação de *outliers* e verificação do percentual de valores faltantes. No caso bivariado, são criados gráficos *boxplot* ou outros para confrontar a variável contra cada uma das classes da variável resposta. É comum que as variáveis dessa natureza sejam convertidas para o tipo qualitativo ordinal, por exemplo, renda, que, após a categorização, fica com classes denominadas faixas de renda.

É muito rico para o estudo, que o pesquisador faça diversas combinações entre as variáveis para encontrar novos atributos que possuam maior poder preditivo. Por exemplo, a

combinação entre o segmento do cliente e a renda pode resultar em uma nova variável com um bom poder preditivo. Após essas etapas, o banco de dados ficará com todas as variáveis originais e com outras variáveis adicionais construídas pela combinação *Feature Engenheering*. A próxima etapa do trabalho é a eliminação de variáveis irrelevantes e redundantes, que é apresentada na Seção 2.5.2.

2.5.2 Pré-Processamento - Seleção de Variáveis

Seleção de variáveis representa uma etapa importante para melhorar a performance dos algoritmos de aprendizado de máquina em termos de redução de tempo para a construção do modelo e aumento da acurácia. Dados brutos, em geral, possuem muitos ruídos, além de apresentarem grande dimensionalidade, redundância e muitas vezes irrelevância. Basicamente é desejável escolher um conjunto de atributos que é relevante para o problema em questão para maximizar a performance com a utilização do mínimo de métricas.

Com relação à estratégia de busca desse subconjunto ótimo, a complexidade exponencial desse problema $O(2^N)$ impossibilita testar todas as combinações possíveis para um banco de dados com muitos atributos. Portanto, para casos como esse, o objetivo passa a ser encontrar o melhor subconjunto, gastando o mínimo de tempo possível. Ou seja, o problema que antes era tratado como uma busca exaustiva, passa a ser tratado como um problema de busca heurística, a qual evita a força bruta, mas, ao mesmo tempo, não garante o subconjunto ótimo.

2.5.2.1 Filter vs Wrapper

De forma geral, o processo de remoção de variáveis redundantes e irrelevantes do banco de dados pode ser classificado em duas classes, conhecidas como seleção de subconjunto (wrapper) e filtragem (filter). O primeiro seleciona a melhor combinação de variáveis e o segundo ranqueia os atributos de acordo com uma métrica específica como, por exemplo, o ganho de informação.

A capacidade do método *wrapper* em lidar com conjunto de dados de grande dimensão é limitada. Sendo assim, para problemas com muitos atributos, esse método não pode ser diretamente aplicado. O método de filtragem consiste na Seleção de atributos utilizando medidas como informação, distância, dependência ou consistência, sem usar qualquer tipo de classificador nessa etapa. Este método possui algumas características [(LIU; MOTODA, 1998), pg.36]:

1. Ele não é baseado em qualquer viés indutivo dos classificadores, mas sim em propriedades intrínsecas dos dados, o que permite que os atributos selecionados possam ser utilizados por qualquer classificador;
2. Calcular essas medidas é mais barato em termos de complexidade de tempo;

3. Pode lidar com conjunto de dados com grande quantidade de atributos, o que pode ser utilizado para diminuir a dimensão do conjunto de dados.

Do ponto de vista dos resultados do método de seleção por filtragem, é gerado o ranqueamento dos atributos de acordo com o critério de avaliação escolhido. Nesse tipo de abordagem, não é retornado um subconjunto ótimo de atributos, mas sim uma lista ordenada de acordo com a relevância de cada atributo. A ideia é avaliar cada atributo individualmente e ordenar todos os atributos de acordo com os valores obtidos.

2.5.2.2 Filter: Information Value (IV)

Neste trabalho, devido à grande quantidade de variáveis explicativas, decidimos utilizar a filtragem como seleção de variáveis. Para isso, empregamos uma métrica conhecida como *Information Value (IV)* ou *Mutual Information*, a qual é derivada da teoria da informação:

The mutual information (MI) is a measure of the amount of information that one random variable has about another variable. This definition is useful within the context of feature selection because it gives a way to quantify the relevance of a feature subset with respect to the output. (VERGARA; ESTÉVEZ, 2014)

Essa métrica é utilizada para determinar o poder preditivo de cada uma das variáveis existentes. Por do IV, obtém-se um valor que permite o ranqueamento das variáveis e, pela definição de um ponto de corte, mantêm-se para a fase de modelagem apenas variáveis com valor acima de um determinado ponto. De acordo com (SIDDIQI, 2005), o poder preditivo da variável pode ser classificado conforme os valores do IV, seguindo a lógica da Tabela 1.

Tabela 1 – Classificação do IV com relação ao poder preditivo

Valor do IV	Poder Preditivo
Menor do que 0,02	Nulo
De 0,02 até 0,1	Fraco
De 0,1 até 0,3	Médio
Maior do que 0,3	Forte

O IV é calculado pela seguinte relação:

$$IV = \sum_{i=1}^k \left\{ (\%Prop = 1_i - \%Prop = 0_i) \cdot \ln \left(\frac{\%Prop = 1_i}{\%Prop = 0_i} \right) \right\},$$

em que i é a classe da variável categórica explicativa.

2.5.2.3 *Parcimônia x Visão de Negócio*

Parcimônia vem do latim *parcos* e significa frugalidade, moderação, simplicidade. Nas ciências, esse conceito é comumente associado à economia de suposições em teorias. É importante considerar esse termo na construção de modelos, uma vez que bons modelos podem ser construídos a partir de um número reduzido de variáveis se a seleção delas for adequadamente realizada, evitando redundâncias. É comum, no início do projeto, que a visão de negócio (pensamento dedutivo) enumere diversas características/variáveis que devem ser contempladas no modelo. No entanto, pela exploração dos dados (pensamento indutivo) verifica-se que muitas das variáveis sugeridas são correlacionadas entre si. Pelo princípio da parcimônia, as variáveis redundantes devem ser eliminadas e o modelo final precisa do menor número de variáveis suficientes para explicar o fenômeno. No entanto, quando a relação com o negócio pede que a solução não siga por esse caminho ou o analista que desenvolve os modelos não possui subsídios para remover determinadas variáveis, é possível combinar as redundantes e utilizar como *input* para o modelo. Uma das maneiras de realizar a combinação é a técnica de Componentes Principais.

2.6 **Análise de Componentes Principais (PCA)**

A análise de Componentes Principais (Principal Component Analysis, PCA) é um método para expressar os dados multivariados. Ela permite ao pesquisador reorientar os dados de modo que as primeiras poucas dimensões expliquem o maior número possível de informações disponíveis. Se houver presença de redundância substancial no conjunto de dados, pode ser possível explicar a maior parte das informações no conjunto original de dados com um número relativamente pequeno de dimensões (LATTIN, 2011).

A PCA consiste basicamente em um problema de determinação dos autovalores e autovetores de uma matriz de correlação dos dados. O produto dos dados originais pelos autovetores representa os Escores dos componentes principais ($\mathbf{Z}$), sendo que as variâncias dos componentes são determinadas pelos autovalores associados.

Como os componentes são mutuamente não correlacionados, a variância da soma é simplesmente a soma das variâncias individuais, ou seja, a soma dos autovalores. Essa propriedade é particularmente útil quando se trata de expressar a quantidade de variação explicada por algum subconjunto dos componentes principais.

Outro subproduto útil da solução de componentes principais é a matriz de correlação do Score do Componente Principal ($\mathbf{Z}$) com os dados originais ($\mathbf{X}$). Essa ajuda a interpretar $\mathbf{Z}$, se soubermos o padrão de relacionamento com os dados originais $\mathbf{X}$. Esta relação é conhecida como cargas dos componentes principais.

Para obter a matriz padronizada dos componentes principais $\mathbf{Z}_s$ simplesmente multiplicamos depois de $\mathbf{Z} = \mathbf{XU}$ a diagonal dos autovalores de $\mathbf{X}$, representada por $\mathbf{D}$: $\mathbf{Z}_s = \mathbf{XUD}^{-0,5}$.

$$D = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}.$$

Com um pouco de álgebra, podemos reescrever a equação anterior para expressar $\mathbf{X}$ como uma função de Z_s : $X = Z_s D^{0.5} U^t$. O que isso revela é que qualquer matriz $\mathbf{X}$ pode ser expressa como produto de três matrizes mais simples, sendo que Z_s é uma matriz de variáveis não correlacionadas, $D^{0.5}$ é uma matriz diagonal que executa uma transformação extensora (essencialmente “despadronizando” Z_s , multiplicando pelos desvios padrões de $\mathbf{Z}$, e U^t é uma matriz de transformação que realiza rotação ortogonal. Esse modo de representar $\mathbf{X}$ é conhecido como uma decomposição em valores singulares (SVD).

2.7 Algoritmos de Classificação

2.7.1 Regressão Logística

O modelo de Regressão Logística consiste em uma metodologia usualmente utilizada com o propósito de determinar a relação entre uma variável resposta discreta binária com outras variáveis que podem ser discretas ou contínuas, as chamadas variáveis explicativas. Esse modelo tornou-se a metodologia padrão de análise para esse tipo de problema em diversas áreas. Lemeshow ([LEMESHOW, 2005](#)) afirma que muitas funções de distribuição foram propostas para problemas com variável resposta binária. As principais vantagens da função logística são flexibilidade e facilidade de usar a função e a fácil interpretabilidade do modelo ajustado. A representação do modelo de regressão logística é:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_p x_p}},$$

em que $x' = (x_1, x_2, \dots, x_p)$ são as p variáveis explicativas e $\beta = (\beta_0, \beta_1, \dots, \beta_p)$ são os parâmetros que devem ser estimados.

A transformação de $\pi(x)$ é conhecida como transformação *logit* e é definida como:

$$g(x) = \log \left(\frac{\pi(x)}{1 - \pi(x)} \right) = x' \beta = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p.$$

A importância dessa transformação é que $g(x)$ possui muitas das propriedades desejadas de um modelo de regressão linear, como linearidade nos parâmetros, é contínua entre $-\infty$ e ∞ , dependendo do intervalo das covariáveis. Diferentemente dos modelos de regressão linear, a distribuição dos erros não é normal e sim binomial, com média zero e variância $\pi(x)[1 - \pi(x)]$.

O método de estimação dos parâmetros é chamado de máxima verossimilhança, que, de forma geral, obtém as estimativas dos parâmetros do modelo. A função de verossimilhança para

a função logística é dada por:

$$l(\beta) = \prod_{i=1}^n \pi(x_i)^{y_i} [1 - \pi(x_i)]^{(1-y_i)}.$$

As estimativas dos parâmetros β são obtidas pela maximização dessa função. No entanto, é matematicamente mais fácil trabalhar com o logaritmo dessa equação, que resulta na função *log likelihood*, $L(\beta)$. Dessa forma, os parâmetros são obtidos derivando $L(\beta)$ com relação aos parâmetros e igualando as expressões resultantes a zero:

$$\sum_{i=1}^n [y_i - \pi(x_i)] = 0$$

e

$$\sum_{i=1}^n x_{ij} [y_i - \pi(x_i)] = 0.$$

Uma vez que realizamos o ajuste do modelo de regressão logística múltipla, devemos verificar a significância das variáveis no modelo. Para isso, fazemos uso do teste univariado de Wald, que sob a hipótese de que um coeficiente individual é zero, segue a distribuição de uma normal padrão. Se o nosso objetivo é obter o melhor ajuste com o menor número de parâmetros, o próximo passo é obter um modelo reduzido, mantendo na equação apenas as variáveis significativas.

2.7.2 Árvores de Decisão

Esse algoritmo utiliza da estratégia dividir para conquistar, em que um problema complexo é dividido em problemas mais simples, aos quais recursivamente é aplicada a mesma estratégia. As soluções dos subproblemas podem ser combinadas, na forma de uma árvore, para produzir uma solução do problema complexo. Formalmente, uma árvore de decisão é um grafo acíclico direcionado em que em cada nó ou é um nó de divisão, com dois ou mais sucessores, ou um nó folha e final (FACELI KATTI; LORENA, 2011).

Uma regra de divisão é guiada por uma medida de *goodness of split*, que indica quão bem um dado atributo discrimina as classes. Uma regra de divisão tipicamente funciona como uma heurística que olha um passo para frente. Para cada teste possível, o sistema hipoteticamente considera os subconjuntos dos dados obtidos. O sistema escolhe o teste que maximiza ou minimiza algumas funções heurísticas sobre os subconjuntos. Uma medida muito utilizada para isso é o Ganho de Informação, que é fundamentado no conceito de entropia. A entropia é usada para medir a aleatoriedade (dificuldade para prever) do atributo alvo. A cada nó de decisão, o atributo que mais reduz a aleatoriedade da variável alvo será escolhido para dividir os dados. Os valores de um atributo definem partições no conjunto de exemplos. Para cada atributo, o ganho

de informação mede a redução na entropia nas partições obtidas, de acordo com os valores do atributo.

As vantagens desse algoritmo são:

- **Flexibilidade:** por se tratar de um método não-paramétrico, não faz suposições sobre os dados.
- **Robustez:** são invariantes a transformações (estritamente) monótonas de variáveis de entrada. Como consequência dessa invariância, a sensibilidade a distribuições com grande cauda e *outliers* é também reduzida.
- **Seleção de atributos:** modelos tendem a ser robustos contra a adição de atributos irrelevantes e redundantes.
- **Interpretabilidade:** decisões complexas e globais podem ser aproximadas por uma série de decisões mais simples e locais .
- **Eficiência:** possuem complexidade de tempo linear como o número de exemplos.

A principal desvantagem desse algoritmo é a instabilidade. Pequenas variações no conjunto de treinamento podem produzir grandes variações na árvore final. A cada nó, o critério de mérito de divisão classifica os atributos, e o melhor atributo é escolhido para dividir os dados. Se dois ou mais atributos são classificados similarmente, pequenas variações da classificação dos dados podem alterar a classificação. Todas as subárvores abaixo desse nó mudam. Além disso, a estratégia da partição recursiva implica que a cada divisão que é feita, o dado é dividido com base no atributo de teste. Depois de algumas divisões, há usualmente muitos poucos dados nos quais a decisão se baseia. Há uma forte tendência a inferências feitas próximas das folhas serem menos confiáveis que aquelas feitas próximas à raiz.

2.7.3 Combinação de Classificadores

Quando falamos em combinação de preditores, a ideia é, de alguma forma, considerar o trabalho conjunto dos classificadores individuais na predição de novos exemplos, o que tem como premissa que isso promova um melhor desempenho do que a ação independente de cada um.

O primeiro requisito para que isso seja feito de forma sucedida é que combinar classificadores idênticos é inútil. Ou seja, o ideal é que os classificadores cometam erros independentes, ou seja, não sejam correlacionados. O segundo requisito é que os classificadores devem ter um melhor desempenho do que um classificador aleatório.

Quando combinamos as predições dos classificadores, podemos utilizar diversas estratégias, como a votação, a serialização ou *ensemble stacking*. No método de votação, as duas formas

de classificação mais utilizadas são: **a)** uniforme, em que a opinião de todos os classificadores contribui igualmente para a classificação final, e **b)** ponderada, na qual a contribuição de um bom classificador é reforçada e isso pode mudar ao longo do tempo. No método de serialização, uma melhoria na votação uniforme é obtida quando cada classificador pode produzir uma estimativa de probabilidade de o exemplo pertencer a uma classe, em vez de produzir uma única etiqueta.

Dado um exemplo de teste $\mathbf{x}$, cada classificador probabilístico reporta a probabilidade do exemplo pertencer a cada uma das classes binárias 0 (p_0) ou 1 (p_1). Dado um conjunto de m classificadores probabilísticos, as probabilidades de classe de todos os modelos podem ser combinadas. Esse método é conhecido na literatura como soma de distribuição. Kittler, 1998, apud (FACELI KATTI; LORENA, 2011) estudou várias estratégias para fusão de m classificadores probabilísticos em problemas de k classes. Assumindo que representamos por P_{ik} a probabilidade dada pelo classificador i do exemplo ser da classe k , então:

- Regra do Máximo: $S_k = \max_i [P_{ik}]$;
- Regra da Média Simples: $S_k = \sum_{i=1}^m [P_{ik}] / m$;
- Regra da Média Geométrica: $S_k = \sqrt[m]{(\prod_{i=1}^m [P_{ik}])}$;
- Regra da Média Harmônica: $S_k = \frac{m}{\sum_{i=1}^m [1/[P_{ik}]]}$.

No método de *ensemble stacking*, a ideia é combinar diversos modelos preditivos (primeiro nível) para gerar um novo modelo (segundo nível). Em geral, o modelo de segundo nível possui uma performance superior. Os modelos de primeiro nível podem ser gerados de diversas maneiras, uma delas, pela utilização de diversos algoritmos sobre o mesmo conjunto de dados. A outra é utilizar o mesmo algoritmo sobre partes (variáveis) do conjunto de dados. Ou seja, particionar a base de dados em subconjuntos de variáveis temáticas e ajustar o mesmo algoritmo para cada um desses subconjuntos. No final, um algoritmo é utilizado para combinar os modelos de primeiro nível.

2.7.4 Random Forest

Random Forest (RF) é um algoritmo popular e muito eficiente pertencente à família dos métodos de *ensembles*. Pode ser utilizado tanto para regressão quanto para classificação. O princípio da classificação é baseado na combinação de diversas árvores de classificação via um processo de *bagging*, que consiste em sucessivas árvores de decisão independentes construídas por uma amostra via *bootstrap* dos indivíduos. A atribuição de qual classe um novo indivíduo pertence é baseada na votação das predições feitas por cada uma das árvores e a classe eleita é a que recebe a maior quantidade de votos. São dois os parâmetros mais importantes desse modelo: **mtry**, que consiste na quantidade de variáveis selecionadas aleatoriamente para formar

uma árvore de classificação específica e **ntree**, que consiste na quantidade de árvores que serão ajustadas. De forma mais detalhada, as etapas de ajuste são:

- Amostragem aleatória n dos N indivíduos, com $n < N$ com reposição;
- Seleção aleatória de m variáveis preditoras das M existente no banco de dados, com $m < M$ (**mtry**);
- Crescimento da árvore de classificação sem poda;
- Repetição desses passos de acordo com o número de árvores. (**ntree**)

Esse processo se repete para todas as árvores. No final, a classificação de cada uma delas é combinada para formar a classificação final do indivíduo. A boa performance desse algoritmo (por exemplo, taxa de acerto), de forma geral, depende da baixa correlação entre as árvores geradas e do poder preditivo de cada uma das árvores originais. Algumas vantagens desse algoritmo são:

- Apresentar ganho de acurácia com relação a outros algoritmos;
- Poder lidar com uma grande variedade de variáveis preditoras ao mesmo tempo;
- Fornecer uma estimativa de quais variáveis são importantes para classificação.

Nesse processo, a indução da árvore é influenciada ainda pelo hiper parâmetro K (número de *features*), isto é, o número K de variáveis aleatoriamente selecionadas. Segundo (BERNARD LAURENT HEUTTE, 2009), esse o número permite introduzir mais ou menos aleatoriedade no processo de indução. Breiman estudou a performance do algoritmo de acordo com K . Em seus experimentos, diversas RF foram testadas e avaliadas de acordo com erro de classificação. Seus experimentos não permitiram concluir o comportamento do RF de acordo com a variação de K . No entanto, ele decidiu usar como padrão o valor de $\log_2(M) + 1$.

2.8 Algoritmos Genéticos

Pela complexidade em identificar o subconjunto ótimo de variáveis explicativas para a construção de um *Score*, diversas heurísticas podem ser utilizadas para chegar a um subconjunto não-ótimo. Uma abordagem possível (e muito utilizada) é a filtragem na qual as variáveis potenciais são ranqueadas e, a partir da arbitragem de um ponto de corte e da eliminação de redundâncias, o subconjunto não-ótimo é encontrado. Entretanto, realizar essa metodologia em algumas circunstâncias pode não ser viável, como por exemplo, quando existe o interesse de recalibração do modelo on-line ou com recorrência alta.

Uma alternativa técnica seria a construção de *Scores* intermediários (modelos de primeiro nível) pela metodologia tradicional e utilizá-los como variáveis de entrada para uma função que combina tais entradas e produz uma regra final de decisão, com melhor desempenho de predição. Dessa forma, a atualização do modelo final pode ser feita pela atualização dos parâmetros que combinam esses *Scores* de primeiro nível. Uma proposta para definir com realizar essa combinação é a utilização dos Algoritmos Genéticos, que são modelos computacionais inspirados na evolução biológica das populações. Nessa abordagem, cada solução é codificada como um indivíduo e os seus genes são avaliados por uma função denominada *fitness*. Cada geração sucessora é composta pelos melhores indivíduos das gerações antecessoras e por seus filhos, os quais, por sua vez, são gerados por processos de *crossover* e mutação (SYLVESTER; CHAWLA, 2005).

O uso de Algoritmos Genéticos já foi estudado na aplicação em construção de *Scores*. (FOGARTY, 2012), discute os prós e os contras de sua utilização na construção de *Credit Scoring*. Segundo o autor, uma das principais vantagens disso é justamente a possibilidade de recalibração dos *Scores* de forma frequente, uma vez que eles, geralmente, perdem performance com o passar do tempo. O autor ainda discute o problema de seleção de variáveis. Ratifica que, na construção desses modelos, a metodologia de filtragem para redução de *inputs* é muito utilizada, devido à impossibilidade de combinar todas as variáveis possíveis. Por outro lado, existem algumas barreiras como a legislação do segmento bancário que prevê transparência no sistema de *Scoring* e a falta de *experts* nas áreas. Para superar isso, o autor propõe um sistema de modelagem tradicional dos dados e a introdução dos algoritmos genéticos para a manutenção dos *Scores*.

Uma outra vantagem destacada por (FINLAY, 2006), é que, na prática, o critério pelo qual os parâmetros dos modelos são determinados são diferentes dos critérios de sua avaliação. Por exemplo, no caso de regressão logística, a avaliação da performance dos modelos não é realizada via *likelihood ratio* e sim por medidas como coeficientes de Gini, KS ou métricas de erro de classificação *misclassification rate*, quando considerado um ponto de corte no *Score*. Dessa forma, os Algoritmos Genéticos podem selecionar os parâmetros visando maximizar essas métricas.

2.9 Performance dos Algoritmos

Para a avaliação dos algoritmos, de forma geral, a base de dados é dividida nas seguintes proporções: 70% da base constitui a amostra de desenvolvimento/treinamento, a qual é utilizada para fazer toda a etapa de seleção de variáveis e ajuste dos modelos, e os 30% restantes compõem a chamada amostra de teste/validação. Essa metodologia de partição do banco de dados é conhecida como *holdout*, cujo resultado obtido na amostra de treinamento é confrontado com novos dados, não utilizados para treinar o modelo, para verificar a eficácia dos modelos na

previsão de resultados de interesse. Para essa validação, são utilizadas diversas métricas.

2.9.1 Métricas da Matriz de Confusão

Para a apuração das métricas de performance, utilizamos uma tabela de dupla-entrada que relaciona os dados reais aos preditos pelo modelo na amostra de validação. A Tabela 2 obtida é chamada matriz de confusão.

Tabela 2 – Matriz de Confusão.

Modelo/Real	Prop = 0	Prop = 1
Prop = 0	VN	FN
Prop = 1	FP	VP

As principais medidas utilizadas para o balizamento da eficiência do modelo na predição dos resultados são apresentadas a seguir:

- Sensibilidade (S): é a probabilidade de um indivíduo avaliado como propenso pelo algoritmo/modelo ser de fato propenso.

$$S = \frac{VP}{VP + FN}.$$

- Especificidade (E): é a probabilidade de um indivíduo não propenso ser classificado como tal.

$$E = \frac{VN}{VN + FP}.$$

- Valor Preditivo Positivo (VPP): é a proporção de verdadeiros positivos em relação a todas as predições positivas

$$VPP = \frac{VP}{VP + FP}.$$

- Valor Preditivo Negativo (VPN): é a proporção de verdadeiros negativos em relação a todas as predições negativas

$$VPN = \frac{VN}{VN + FN}.$$

- Capacidade Total de Acerto ou Acurácia (CTA): razão entre a soma da quantidade de clientes corretamente identificados pelo modelo em relação ao total de resultados possíveis.

$$CTA = \frac{VP + VN}{VP + FP + FN + VN}.$$

- F1 Score ($F1$): é a média harmônica entre a sensibilidade e a especificidade. Dá uma medida resumo dessas duas métricas, o que pode apontar um modelo equilibrado.

$$F1 = \frac{2 * S * E}{S + E}.$$

- Lift (*lift*): é utilizado quando não estamos interessados na performance do modelo como um todo, mas sim na boa acurácia de apenas 5% ou 10% do banco de dados ordenados por um determinado escore. *Lift* mostra o ganho de captura ao utilizar um escore, fixado um ponto de corte, em relação à busca aleatória. Ou seja, quanto a predição é melhor do que a busca aleatória na fração do banco de dados predito como propensos.

$$lift = \frac{S}{\frac{VP+FP}{VP+VN+FP+FN}}.$$

MATERIAIS E MÉTODOS

Neste Capítulo está descrito o desenho do estudo e planejamento amostral e os conceitos das variáveis disponíveis para a construção dos modelos.

3.1 Desenho do Estudo

O fenômeno de interesse, reclamação no ODC, possui incidência mensal de aproximadamente sete mil clientes (CPFs) dentro do segmento estudado. Diante da quantidade de clientes ativos, isso representa uma taxa de 0,27% ao mês. Pela raridade do evento, um plano amostral foi realizado para compensar o desbalanceamento da base de dados, via um sorteio aleatório na visão mensal. Dentre os clientes propensos à reclamação, foram recuperados registros de 54.221 CPFs distintos nos oito meses de estudo (de agosto de 2016 a março de 2017).

A amostra final consistiu de 279.000 CPFs. Com esse número de clientes, o estudo ficou com uma relação aproximada de um propenso para cada quatro não propensos. Para compor as variáveis explicativas foi observado um histórico retrospectivo de até seis meses. O conteúdo das bases de dados e os conceitos das *features* estão detalhados na Seção [3.2](#).

3.2 Bases de Dados e Seleção de Variáveis

As tabelas utilizadas no estudo refletiam o comportamento do consumidor com relação à quantidade de chamadas realizadas no *Call Center*, os motivos dessas chamadas, a incidência de defeitos/reparos nos produtos, a quantidade de massivas e o histórico do cliente com o ODC. Antes da seleção de variáveis, foi realizada a partição do banco de dados em duas partes. A primeira parte (amostra de treinamento) continha registros dos cinco primeiros meses (ago'16 até dez'16) e a segunda parte (amostra de validação) continha os três meses restantes (jan'17 até mar'17). No total, 947 variáveis estavam disponíveis como potenciais candidatas para predizer o

fenômeno de interesse.

3.2.1 Dados de Chamadas

A base de dados de chamadas continha as ligações de cada cliente, com seus respectivos tempos de duração. A partir dessas informações, foram criadas variáveis relacionadas à quantidade de ligações efetuadas por cada cliente (CPF) e o tempo de duração, considerando diferentes janelas de tempo (últimos 5 dias até 90 dias). A relação dessas variáveis com a propensão do cliente procurar o ODC foi avaliada pelo IV (*Information Value*), ver Tabela 3. Pode-se perceber que tais variáveis não possuem um poder de discriminação muito grande, pois o maior IV obtido foi aproximadamente de 0,1.

Tabela 3 – Relação de Variáveis de chamadas e seus respectivos IV

Variável	IV
Qtd. Chamadas em 05 dias	0,064
Qtd. Chamadas em 15 dias	0,095
Qtd. Chamadas em 30 dias	0,103
Qtd. Chamadas em 60 dias	0,095
Qtd. Chamadas em 90 dias	0,088
Tempo máximo em 05 dias	0,064
Tempo máximo em 15 dias	0,095
Tempo máximo em 30 dias	0,103
Tempo máximo em 60 dias	0,094
Tempo máximo em 90 dias	0,090
Primeiro Componente Principal	0,107
Segundo Componente Principal	0,030

Além das variáveis individuais, foram criadas duas outras, derivadas da combinação das variáveis originais, os componentes principais. A criação destes componentes fez sentido devido à grande correlação existente nos dados originais. O primeiro componente explicou 75,66% e o segundo, 8,51% da variabilidade original dos dados. A ponderação do primeiro componente pode ser conferida na Tabela 4.

Tabela 4 – Ponderação do Primeiro Componente Principal.

Variável	Ponderação
Qtd. Chamadas em 05 dias	0,019
Qtd. Chamadas em 15 dias	0,099
Qtd. Chamadas em 30 dias	0,213
Qtd. Chamadas em 60 dias	0,394
Qtd. Chamadas em 90 dias	0,527
Tempo máximo da Chamada em 05 dias	0,096
Tempo máximo da Chamada em 15 dias	0,149
Tempo máximo da Chamada em 30 dias	0,290
Tempo máximo da Chamada em 60 dias	0,424
Tempo máximo da Chamada em 90 dias	0,476

3.2.2 Dados de Reparos

A base de dados de reparos continha variáveis como a quantidade de defeitos nos produtos dos clientes, a classificação do tipo de defeito (massivo ou pontual), apontava se o defeito era reincidente, se ocorreu logo no início do relacionamento do cliente com a empresa e o prazo que a empresa levou para corrigi-lo.

Ao todo foram construídas 283 variáveis para essa tabela de dados. As variáveis com os maiores poderes discriminativos desse conjunto referiam-se à quantidade de defeitos nos intervalos de tempo (de 05 dias a 180 dias). Devido à forte estrutura de correlação, foram criados componentes principais para ponderar a importância do tempo na ocorrência do evento. Por exemplo, para o subconjunto de variáveis que tratavam de defeitos pontuais, nos quais o cliente liga e abre um bilhete de defeito (*Trouble Ticket*) para resolver um defeito particular, foi realizada a análise de componentes principais que possibilitou obter pesos para ponderar esse fenômeno entre os intervalos de tempo. Foram criadas duas novas variáveis: a primeira, Componente TT, primeiro componente principal de variáveis sobre reparos pontuais, explicou 82,29% do conjunto de dados original, conforme destacado na Tabela 5.

Tabela 5 – Formação do Componente - TT.

Variáveis	Ponderação
Qtd. de Reparos em 05 dias	0,0319
Qtd. de Reparos em 15 dias	0,0926
Qtd. de Reparos em 30 dias	0,1904
Qtd. de Reparos em 60 dias	0,3233
Qtd. de Reparos em 90 dias	0,4347
Qtd. de Reparos em 120 dias	0,5149
Qtd. de Reparos em 180 dias	0,6290

A segunda, Componente Massiva, foi obtida via o primeiro componente principal que considerou as variáveis de reparos massivos, no qual a empresa detecta que ocorreu um evento de grandes proporções que afetou muitos clientes, que explicou 80,76% do conjunto de dados original, conforme destacado na Tabela 6.

Tabela 6 – Componente Massiva/Programada.

Variáveis	Ponderação
Qtd. de Massiva/Programada em 05 dias	0,0326
Qtd. de Massiva/Programada em 15 dias	0,0910
Qtd. de Massiva/Programada em 30 dias	0,1869
Qtd. de Massiva/Programada em 60 dias	0,3281
Qtd. de Massiva/Programada em 90 dias	0,4396
Qtd. de Massiva/Programada em 120 dias	0,5153
Qtd. de Massiva/Programada em 180 dias	0,6240

Após essa etapa de construção dos componentes, partiu-se para a análise de correlações para eliminar variáveis redundantes do conjunto de dados original. Foram eliminadas variáveis que possuíam correlações de *Pearson* acima de 0,7. A lógica para a eliminação das variáveis foi: 1 – Cálculo das correlações entre todas as variáveis; 2 – Eliminação das variáveis com correlação acima de 0,7, mantendo no banco de dados a variável com maior IV. Após essa etapa, foram mantidas, no conjunto de dados, nove variáveis, conforme apresentado na tabela 7.

Tabela 7 – Relação das variáveis de Reparo e seus respectivos IV

Variável	IV
Qtd. reparos (180 dias) Motivo 1	0,141
Primeiro Componente Principal - TT	0,124
Reincidência Reparos (30 dias)	0,117
Qtd. reparos (180 dias) Motivo 4	0,050
SLA reparos (90 dias)	0,042
Qtd. reparos (180 dias) Motivo 2	0,036
Qtd. reparos (180 dias) Motivo 5	0,030
Qtd. reparos (90 dias) Motivo 3	0,025
Primeiro Componente Principal - Massiva/Programada	0,022

3.2.3 Dados de Motivo das Chamadas

Outra base disponível para a modelagem continha um fluxo com diversos níveis que registrava o caminho que o cliente percorreu durante o atendimento no *Call Center*. Por exemplo, se o cliente ligou e disse que queria uma informação, o atendente assinalava em seu programa no primeiro nível que se trata de uma informação. Com a evolução do diálogo, o cliente poderia pedir informação sobre sua fatura. Então, o atendente assinalava no segundo nível que se tratava de fatura. Finalmente, o cliente poderia dizer que desejava esclarecimento sobre o vencimento

da fatura e, assim, o atendente marcava no terceiro nível que se tratou de uma informação sobre o vencimento da fatura. Portanto, três campos foram assinalados durante esse atendimento: no primeiro nível, "Informação", no segundo nível, "Fatura" e no terceiro, "Vencimento".

Diversos outros caminhos podem ocorrer no atendimento e o atendente classifica a solicitação do cliente na melhor opção disponível em seu *menu* de opções (não há criação de novas categorias). Portanto, essa base possuía uma grande quantidade de possibilidades de caminhos que o cliente poderia percorrer durante o atendimento telefônico. Não foi possível analisar todos os caminhos. A estratégia para abordar o problema foi elencar os caminhos mais frequentes.

O IV foi calculado para os 500 caminhos mais frequentes. Com esses IV, foi realizado um alinhamento com a área de negócio de quais caminhos faziam sentido e como melhor utilizar as informações. As Tabelas 8 e 9 apresentam as variáveis selecionadas por IV, agrupadas pelo conceito desejado pela área de negócio.

Tabela 8 – Relação de Variáveis de Motivo da chamada de Primeiro Nível e seus respectivos IV

Variável	IV
Tipo 1	0,2895
Tipo 2	0,2765
Tipo 3	0,2734
Tipo 4	0,2092
Tipo 5	0,0792

A Tabela 8 apresenta os IV das variáveis relacionadas ao motivo principal da chamada. O nome das variáveis está codificado, porém o conceito delas representa o motivo mais geral pelo qual o cliente ligou, por exemplo, informação, reclamação, etc.

Tabela 9 – Relação de Variáveis de Motivo da Chamada Comercial e seus respectivos IV

Variável	IV
Fatura: Tipo 1	0,1431
Fatura: Tipo 2	0,0536
Fatura: Tipo 3	0,0530
Fatura: Tipo 4	0,0529
Fatura: Tipo 5	0,0495
Fatura: Tipo 6	0,0454
Oferta: Tipo 1	0,0721
Oferta: Tipo 2	0,0692
Oferta: Tipo 3	0,0663

A Tabela 9 apresenta os IV dos motivos de chamadas referentes a alguns fluxos que se referiam a assuntos comerciais, como fatura e oferta. Embora a nomenclatura esteja codificada,

seus significados são detalhes (fluxos) sobre os quais os clientes ligaram para falar sobre as suas faturas e as ofertas.

3.2.4 *Passagem anterior pelo ODC*

A última base disponível trazia o histórico de todos os clientes que já reclamaram no ODC. Com essa base foi possível construir variáveis que assinalavam o motivo que levou o cliente a ir ao ODC no passado. Como muitas delas eram correlacionadas, foi realizada a eliminação de muitas delas. Após a filtragem, ficaram pré-selecionadas para a modelagem onze variáveis não redundantes, conforme apresentado na Tabela 10.

Tabela 10 – Relação de Variáveis de Histórico ODC e seus respectivos IV

Variável	IV
Qtd. ODC Geral	0,7036
Qtd. ODC (180 dias) Motivo X	0,3387
Reaberturas	0,1110
Qtd. ODC (180 dias) Motivo Y	0,1006
Reiteraões	0,0695
Qtd. ODC (05 dias) Geral	0,0525
Qtd. ODC (180 dias) Motivo Z	0,0380
Agging do Cliente	0,0329
Qtd. ODC (180 dias) Motivo Q	0,0312
Qtd. ODC (05 dias) Mesmo segmento	0,0308
Qtd. ODC (180 dias) Motivo J	0,0308

Na Tabela 10, estão apresentadas variáveis que apontam se o cliente tem algum histórico de passagem no ODC, em qualquer segmento, não apenas no segmento estudado. Também mostra se o cliente já fez reaberturas ou reiteraões. Além disso, destaca alguns motivos de passagem pela ODC.

Com as variáveis selecionadas nessa Seção, partiu-se para a criação dos modelos estatísticos.

RESULTADOS

4.1 Modelagem

No Capítulo 3, foram descritas as bases de dados e como foi realizada a seleção/filtragem das variáveis. Neste Capítulo está descrita a construção dos modelos de aprendizado de máquina. A abordagem inicial escolhida para modelar os dados foi a construção de modelos de regressão logística individuais para cada uma das bases de dados temáticas. Denominamos esses modelos de primeiro nível, cujo propósito foi gerar *Scores* para cada uma das dimensões de negócio, refletidas pelas diferentes bases de dados. Os *Scores* gerados nesse processo foram:

- Camadas: identifica o perfil do cliente que liga recorrentemente ao *Call Center* (3.2.1);
- Reparos: registro de problemas técnicos (3.2.2);
- Comercial: apontamentos de desvios na fatura ou outros motivos comerciais (3.2.3);
- Motivo Chamada: palitagem do primeiro nível que identifica o motivo principal da chamada (3.2.3);
- Passou ODC: verifica se o cliente já recorreu ao órgão de defesa do consumidor no passado (3.2.4).

Esses *Scores* tiveram como finalidade servir de *input* para o modelo de segundo nível, o qual também foi modelado via Regressão Logística. A essa estratégia de modelagem dá-se o nome de *Stacking*, conforme explicado no Capítulo 2. A amplitude teórica desses *Scores* variava de zero a cem.

4.1.1 Modelos de Primeiro Nível

Para a construção dos modelos de primeiro nível foi utilizada a amostra de treinamento, conforme descrito no Capítulo 2.

Para o modelo de Chamadas, as variáveis de entrada estão descritas na Tabela 3 do Capítulo 3. Devido à pouca variedade de informações e à estrutura de correlações, o modelo final ficou com apenas uma variável, o primeiro componente principal (descrito na Tabela 4). O segundo componente foi descartado em discussão com a área de negócio. A Tabela 11 apresenta o ajuste desse modelo. Nessa tabela, observamos a relação positiva entre o valor do componente e a propensão do cliente procurar o ODC, ou seja, quanto mais chamadas o cliente tiver ao longo de seis meses e quanto mais demorada for a ligação, maior será o *Score* de Propensão.

Tabela 11 – Modelo de Chamadas

Variável	Estimativa	Erro Padrão	P-valor
Intercepto	-1,6701	0,0078	<0,0001
Primeiro Componente Principal	0,0377	0,0005	<0,0001

A segunda regressão ajustada utilizou dados a respeito de reparos, conforme descrito no Capítulo 3. As nove variáveis finais foram descritas na Tabela 7. O modelo ajustado pode ser conferido na Tabela 12. Percebemos que todas as variáveis contribuem positivamente para aumentar a propensão do cliente procurar o ODC. A variável com maior contribuição refere-se à quantidade de vezes que o cliente procurou a empresa para resolver algum problema técnico de motivo 1. Os dois componentes principais, reparos pontuais e reparos massivos, foram significativos para o modelo. Essas variáveis agregam ao modelo o perfil de cliente crônico, que sofre muito com problemas técnicos. Além dessas variáveis, também compõem o modelo variáveis de quantidade de reparos por motivos 2 e 3 e a média do tempo de reparo.

Tabela 12 – Modelo de Reparos

Variável	Estimativa	Erro Padrão	P-valor
Intercepto	-1,7050	0,0079	<0,0001
Primeiro Comp. Principal - TT	0,2199	0,0046	<0,0001
Primeiro Comp. Principal - Massiva/Programada	0,2251	0,0151	<0,0001
Qtd. reparos (180 dias) Motivo 1	1,5546	0,0479	<0,0001
Qtd. reparos (180 dias) Motivo 2	0,1735	0,0242	<0,0001
Qtd. reparos (90 dias) Motivo 3	0,0639	0,0294	0,0299
SLA reparos (90 dias)	0,0185	0,0080	0,0208

O terceiro modelo ajustado (Tabela 13) trata do motivo principal pelo qual o cliente fez a chamada ao *Call Center*. Todas as variáveis foram positivamente relacionadas com o evento ODC. O motivo da ligação do tipo 3 tem um peso superior aos outros motivos, seguido pelo motivo do tipo 4.

Tabela 13 – Modelo de Motivo Principal da Chamada

Variável	Estimativa	Erro Padrão	P-valor
Intercepto	-2,0932	0,0098	<0,0001
Qtd de Tipo 1 (90 dias)	0,0645	0,0015	<0,0001
Qtd de Tipo 2 (90 dias)	0,0055	0,0012	<0,0001
Qtd de Tipo 3 (90 dias)	2,4050	0,0645	<0,0001
Qtd de Tipo 4 (90 dias)	0,1201	0,0026	<0,0001
Qtd de Tipo 5 (90 dias)	0,0380	0,0031	<0,0001
Qtd de Tipo 6 (90 dias)	0,0737	0,0106	<0,0001

O quarto modelo (Tabela 14) reflete ligações por aspectos comerciais, tal como fatura e oferta. Todas as variáveis foram positivamente relacionadas com ODC. As mais importantes estão relacionadas a problemas de oferta do tipo 1 e problemas com fatura do tipo 1.

Tabela 14 – Modelo de Motivos Comerciais

Variável	Estimativa	Erro Padrão	P-valor
Intercepto	-1,8034	0,0084	<0,0001
Fatura: Tipo 1	0,3948	0,0099	<0,0001
Fatura: Tipo 2	0,2335	0,0327	<0,0001
Fatura: Tipo 3	0,1270	0,0050	<0,0001
Fatura: Tipo 4	0,2933	0,0203	<0,0001
Fatura: Tipo 5	0,1658	0,0068	<0,0001
Oferta: Tipo 1	0,7011	0,0202	<0,0001
Oferta: Tipo 2	0,1911	0,0121	<0,0001
Oferta: Tipo 3	0,3236	0,0122	<0,0001

Por fim, a última regressão ajustada (Tabela 15) trata de variáveis relacionadas à passagem anterior no ODC. As variáveis mais fortes na regressão referiam-se ao fato de o cliente ter alguma experiência passada recente com o ODC, no mesmo segmento de negócio, o que aponta a provável não resolutividade do problema. Também foi relevante a variável que aponta se o cliente já possui histórico de passagem pelo ODC, em algum momento de sua vida, em qualquer segmento, o que reflete um conhecimento prévio do cliente sobre o fluxo necessário para acionar o ODC e qual tipo de tratativa/resolutividade o ODC dá ao problema.

Tabela 15 – Modelo de Passagem anterior ao ODC

Variável	Estimativa	Erro Padrão	P-valor
Intercepto	-1,4592	0,0106	<0,0001
Já foi ao ODC (todos segmentos)	2,8550	0,0419	<0,0001
Foi ao ODC (180 dias - Motivo X)	0,7946	0,0813	<0,0001
Foi ao ODC 180 dias - Motivo Y	0,8146	0,1407	<0,0001
Possui Reaberturas	0,4341	0,1316	<0,0001
Possui Reiteraões	0,5115	0,1756	0,0036
Foi ao ODC 5 dias (geral)	0,6783	0,2654	0,0106
Foi ao ODC 5 dias (mesmo Segmento)	2,9314	1,0250	0,0042
Agging do cliente (meses)	-0,0048	0,0002	<0,0001

4.1.2 Poder Preditivo dos Modelos de Primeiro Nível

Para avaliar o desempenho preditivo dos modelos de primeiro nível, foi utilizada a matriz de confusão 2.9.1. Os pontos de corte nos *Scores* foram definidos considerando como propensos os cinquenta mil clientes com maiores *Scores*. Ou seja, pela ordenação dos clientes por essa métrica, atribuiu-se a marcação Propenso = 1 aos cinquenta mil primeiros e Propenso = 0 aos demais. Isso foi feito tanto na amostra de desenvolvimento quanto na amostra de validação. Com essa marcação foram construídas as matrizes de confusão para cada modelo e extraídas as métricas das Tabelas 16 e 17.

Tabela 16 – Métricas de Predição para Modelos de Primeiro Nível - Treinamento

Modelos	Acurácia	S	E	VPP	VPN	F1-Score	Lift
Passou ODC	0,6673	0,4825	0,7117	0,2868	0,8512	0,5751	1,4803
Chamadas	0,6487	0,4345	0,7001	0,2583	0,8374	0,5362	1,3331
Reparos	0,6616	0,4680	0,7082	0,2782	0,8470	0,5636	1,4358
Motivo Chamada	0,7033	0,5754	0,7340	0,3421	0,8779	0,6451	1,7655
Comercial	0,6776	0,5092	0,7181	0,3027	0,8589	0,5958	1,5621

Na Tabela 16, podemos conferir que o modelo de motivo da chamada apresenta o maior *lift*, com um ganho de 76,55% em relação a uma amostra aleatória. A acurácia desse modelo é de 70,33% e o F1-Score de 64,51%. O modelo com o segundo melhor desempenho foi o comercial, com *lift* de 56,21%, acurácia de 67,76% e F1-Score de 59,58%. Em seguida, ficou o modelo de passagem pela ODC com *lift* de 48,03%, acurácia de 66,73% e F1-Score de 57,51%.

Pela tabela 17, verificamos que o desempenho dos modelos na amostra de validação é parecido com o desempenho na etapa de treinamento. O modelo com melhor desempenho novamente foi o que trata do motivo da chamada, com *lift* de 62,19%, acurácia 65,87% e F1-Score de 65,37%. O segundo modelo com melhor desempenho também foi o comercial, com *lift* de 50,01%, acurácia de 63,98% e F1-Score de 62,25%. Em seguida, ficou o modelo de passagem pela ODC com *lift* de 45,95%, acurácia de 63,35% e F1-Score de 61,19%.

Tabela 17 – Métricas de Predição para Modelos de Primeiro Nível - Validação

Modelos	Acurácia	S	E	VPP	VPN	F1-Score	Lift
Passou ODC	0,6335	0,5810	0,6462	0,2846	0,8642	0,6119	1,4595
Chamadas	0,6017	0,4994	0,6265	0,2447	0,8378	0,5558	1,2545
Reparos	0,6212	0,5495	0,6386	0,2692	0,8540	0,5907	1,3804
Motivo Chamada	0,6587	0,6457	0,6619	0,3163	0,8852	0,6537	1,6219
Comercial	0,6398	0,5972	0,6501	0,2926	0,8695	0,6225	1,5001

4.1.3 Modelo de Segundo Nível

O modelo de segundo nível foi construído pelo ajuste da regressão logística com os *Scores* de primeiro nível como covariáveis. A Tabela 18 apresenta os valores das estimativas dos parâmetros. Como esperado, todos os coeficientes associados com os *Scores* de primeiro nível apresentaram valores positivos e foram estatisticamente significantes, ao nível de 5%.

Tabela 18 – Modelo de Segundo Nível - Regressão Logística

Modelos	Estimativa	Erro Padrão	p-valor
Intercepto	-3,8304	0,0267	<0,0001
Score Passou ODC	0,0401	0,0005	<0,0001
Score de Chamadas	0,0176	0,0009	<0,0001
Score de Reparos	0,0154	0,0008	<0,0001
Score Principal da Chamada	0,0257	0,0007	<0,0001
Score Comercial	0,0188	0,0007	<0,0001

4.1.4 Poder Preditivo dos Modelos de Segundo Nível

A Tabela 19 apresenta o desempenho dos modelos de segundo nível nas amostras de treinamento e validação. O modelo apresentou *lift* de 86,83% na amostra de treinamento e 73,34% na amostra de validação. As outras métricas na validação mostraram melhora no desempenho com relação aos modelos individuais de primeiro nível. A acurácia foi de 67,60% e o F1-Score foi de 68,12%.

Tabela 19 – Métricas do Modelo de Segundo Nível - Regressão Logística

Período	Acurácia	S	E	VPP	VPN	F1-Score	Lift
Desenvolvimento	0,7163	0,6089	0,7421	0,3620	0,8876	0,6690	1,8683
Validação	0,6760	0,6900	0,6726	0,3381	0,8996	0,6812	1,7334

Este foi o modelo entregue para a empresa, que por motivo de cronograma, não pôde ser comparado com outras metodologias. No entanto, este projeto de mestrado permitiu a construção de metodologias alternativas que possibilitaram verificar se esta foi a melhor abordagem. Os resultados comparativos podem ser conferidos na Seção 4.3

4.2 Regressão Logística Parcimoniosa

Nesta Seção apresentamos o resultado do modelo de Regressão Logística com apenas uma equação. Denominamos de modelo parcimonioso, pois antes da realização da modelagem em si, realizamos um processo que elimina muitas variáveis redundantes. Portanto, esta abordagem foi construída por meio da realização de filtragem das variáveis brutas, sem a construção de componentes principais ou modelos de primeiro nível. A Tabela 20 apresenta o modelo ajustado.

Tabela 20 – Modelo de Regressão Logística Parcimoniosa.

Variável	Estimativa	Erro Padrão	P-Valor
Intercepto	-1,9646	0,0131	<0,0001
Agging do cliente	-0,0039	0,0002	<0,0001
Foi ao ODC (180 dias) Geral	2,2048	0,0380	<0,0001
Qtd de Chamadas (90 dias) Tipo 1	0,0699	0,0028	<0,0001
Qtd de Chamadas (90 dias) Outros Tipos	0,0246	0,0005	<0,0001
Comercial (90 dias) Motivo X	0,9901	0,0716	<0,0001
Qtd Reparos (90 dias)	0,1378	0,0147	<0,0001
Qtd Massivas (180 dias)	0,1849	0,0343	<0,0001

As variáveis finais do modelo foram o *agging* do cliente, ou seja, quanto tempo o cliente tem relação com a empresa nesse segmento, determinados motivos de chamadas, se teve algum desfecho que gerou reclamação sobre o comercial e a quantidade de reparos e massivas.

Tabela 21 – Métricas do Modelo Parcimonioso - Regressão Logística.

Modelo Parcimonioso	Acurácia	S	E	VPP	VPN	F1-Score	Lift
Desenvolvimento	0,7123	0,5987	0,7396	0,3559	0,8846	0,6618	1,8369
Validação	0,6660	0,6643	0,6664	0,3254	0,8912	0,6653	1,6687

A Tabela 21 apresenta as métricas desse modelo. A acurácia na amostra de validação foi de 66,60%, menor em um ponto percentual do modelo ajustado via *Stacking*. O *lift* também apresentou valor inferior 66,87%. Dessa forma, conclui-se que, para este conjunto de dados, a abordagem de *Stacking* mostrou ganhos em performance.

4.3 Metodologias Alternativas de Modelagem

Nesta sessão, apresentamos os modelos desenvolvidos na etapa experimental deste projeto. O modelo anteriormente apresentado foi construído dentro da empresa em um tempo reduzido, porém, seguindo a metodologia julgada adequada para o problema. A ideia desta sessão foi explorar alternativas de modelagem e comparar sob diversas métricas (Subseção 2.9.1) as performances dos modelos.

4.3.1 Algoritmos Genéticos

Esta abordagem foi desenvolvida com o objetivo de utilizar os Algoritmos Genéticos para explorar outras formas de combinação dos *Scores* de primeiro nível (o valor máximo, a média simples, a média geométrica, a média harmônica e a regressão logística) e ainda ajudar a entender se seria necessário utilizar todos os *Scores* ou apenas alguns deles. O problema modelado possuía um espaço de busca de $(2^5 - 1) * 5 = 155$ possibilidades. Foram realizadas simulações de probabilidade de *crossover* e mutação e a seleção foi feita via elitismo, mantendo na população sempre os indivíduos mais aptos.

4.3.1.1 Codificação, População inicial e fitness do GA

Cada indivíduo foi codificado em um vetor de seis posições, em que as cinco posições iniciais indicavam, de forma binária, se um *Score* de primeiro nível estava presente ou ausente naquele indivíduo. A ordem do vetor foi relevante para o problema, pois o *Score* com melhor desempenho aparecia na primeira posição e o *Score* com pior desempenho, na quinta posição.

Para avaliar o desempenho preditivo dos modelos de primeiro nível foi utilizada a matriz de confusão. Os pontos de corte para cada um dos cinco *Scores* foram definidos considerando como propensos os cinquenta mil clientes com maiores *Scores*. Ou seja, pela ordenação dos clientes por essa métrica, atribuiu-se a marcação *Propenso* = 1 aos cinquenta mil primeiros e *Propenso* = 0 aos demais. Isso foi feito na amostra de desenvolvimento (153.400 clientes). As performances podem ser conferidas na Tabela 16.

O *Score* de Motivo de Chamada apresentou a melhor performance com acurácia de 70,33%, sensibilidade de 50,92%, especificidade de 71,81%, VPP de 30,27% e VPN de 85,89%. A segunda melhor performance foi obtida pelo *Score* Comercial. Em seguida, vieram os *Scores* de Histórico ODC, Reparos e, por fim, Chamadas. É interessante perceber que, independente da métrica escolhida, a ordenação dos *Scores* com relação à performance foi a mesma. A tabela 22 representa a probabilidade de seleção baseada no valor da sensibilidade para a população inicial na amostra de treinamento.

Tabela 22 – Probabilidade de Seleção do *Score* da população inicial

Score	Sensibilidade	Proporção	Acumulada
Motivo Chamada (S1)	0,5754	23,3%	23,3%
Comercial (S2)	0,5092	20,6%	43,9%
Passou ODC (S3)	0,4825	19,5%	63,5%
Reparos (S4)	0,4680	19,0%	82,4%
Chamadas (S5)	0,4345	17,6%	100,0%
Soma	2.4696	100.0%	-

O processo de definir quais *Scores* comporiam os genes dos indivíduos foi realizado via sorteio de um número inteiro aleatório entre zero e cem, via um gerador uniforme. O número sorteado definiu o limite superior da proporção acumulada gerada pela sensibilidade dos *Scores* de primeiro nível, representado na tabela 22. Por exemplo, para um determinado indivíduo *i* se o número 45 for sorteado, farão parte do processo de avaliação desse indivíduo os *Scores* Motivo de Chamada e Comercial, pois ambos acumulam 43,9% da proporção da sensibilidade.

A sexta posição do vetor de codificação foi ocupada pela regra de combinação, atribuída para cada indivíduo de forma aleatória, a saber: Regra 1 - Média Simples, Regra 2 - Média Geométrica, Regra 3 - Média Harmônica, Regra 4 - Máximo e Regra 5 - Regressão logística. A Tabela 23 apresenta um exemplo da codificação para uma população inicial de dez indivíduos.

Tabela 23 – População Inicial com dez indivíduos - seed(123)

Ind	S1	S2	S3	S4	S5	Regras	Fitness
1	1	1	0	0	0	4	16.902
2	1	1	0	0	0	5	17.182
3	1	1	1	1	1	1	18.011
4	1	1	1	0	0	5	18.009
5	1	1	1	0	0	3	17.691
6	1	1	1	1	1	3	17.660
7	1	1	1	1	0	3	17.704
8	1	0	0	0	0	5	17.102
9	1	1	0	0	0	1	17.059
10	1	1	0	0	0	5	17.182

A função de *fitness* foi a contagem de quantos clientes apresentavam marcação Propenso = 1 entre os cinquenta mil clientes com os maiores valores no *Score* final. A tabela 23 mostra que o indivíduo 1, com os genes de *Scores* S1 e S2 e Regra 4 obteve *fitness* de 16.902.

4.3.1.2 Crossover, Mutação e Seleção

O processo de *crossover* baseou-se na troca de regras entre dois indivíduos de forma aleatória. Ou seja, para cada geração, um indivíduo foi avaliado segundo a probabilidade de

realizar *crossover* e caso essa probabilidade fosse maior do que um número aleatório, ele trocava de regra com outro elemento escolhido ao acaso. Senão, ele seguia para a próxima geração com o mesmo gene de regra.

A mutação foi responsável por alterar os genes de *Scores* dos indivíduos. Da mesma forma que no processo de *crossover*, existia uma probabilidade de mutação em que cada indivíduo era confrontado com um número aleatório e, caso a probabilidade de mutação fosse maior, os *Scores* eram embaralhados. Essa foi uma característica boa do GA proposto, pois permitiu explorar melhor o espaço de busca, sobretudo, por fazer alterações significativas no gene da população inicial que seguia a lógica da Tabela 22. Com a mutação, por exemplo, cromossomos com a ausência de S1 também eram viáveis de existir, algo que não ocorria na população inicial. Nesse processo, também foi permitida a extinção de algum gene para determinados indivíduos.

A seleção foi realizada via elitismo, em que os dez pais eram comparados com os dez filhos e sobreviviam para a geração seguinte apenas os dez mais aptos, ou seja, os dez indivíduos com maiores *fitness*. Ao final de *n* gerações, a resposta procurada estava contida nos cromossomos do indivíduo mais apto.

4.3.1.3 Simulações com GA

Diversas simulações foram realizadas, alterando valores da população inicial (10, 20 e 50) e os parâmetros de *Crossover* e mutação (5%, 10% e 50%). Em todos os cenários, foi fixada uma única semente para a aleatorização. Um exemplo de população inicial com dez indivíduos está representada na Tabela 23. A seleção ocorreu via elitismo e, entre cada uma das gerações, foram selecionados os dez indivíduos mais aptos. Nesse processo de seleção, foram descartados indivíduos exatamente iguais aos seus pais.

O algoritmo encontrou que a solução com melhor desempenho foi a combinação dos *Scores* S1 (Motivo Chamada) e S3 (Passou ODC) via regra do Máximo, com *fitness* de 18.550. A tabela 24 apresenta as medidas de desempenho da regra final.

Tabela 24 – Métricas do Modelo Selecionado

Período	Acurácia	S	E	VPP	VPN	F1-Score	Lift
Desenvolvimento	0.7221	0.6241	0.7457	0.3710	0.8919	0.6795	1.9146
Validação	0.6770	0.6926	0.6733	0.3393	0.9004	0.6828	1.7398

O resultado obtido, em termos de desempenho, foi um pouco superior ao obtido pela metodologia de *stacking* de Regressão Logística, Seção 4.1. A acurácia obtida com a heurística de GA foi de 67,70% contra 67,60%. O *lift* foi de 1,7398 contra 1,7334. Em termos de desempenho, não representa um ganho relevante, porém, do ponto de vista de parcimônia, é muito melhor e simples a utilização desse modelo alternativo. Do ponto de vista do negócio, é interessante destacar que esse modelo ficou com os dois *Scores* finais, Passou ODC e Motivo Chamada,

que são as duas dimensões que a área de negócio julgou mais relevantes para captar o perfil de clientes propensos a reclamar no ODC.

4.3.2 Stacking Random Forest

Para a elaboração dos modelos de primeiro nível, foi utilizada a técnica de *Random Forest* com os parâmetros **mtry** = $\log_2(M+1)$ e **ntree** = 200 (ver 2.7.4). O banco de dados utilizado foi aquele com as variáveis brutas, com remoção de variáveis muito correlacionadas. Assim, o parâmetro **mtry** para os modelos foram: Passou ODC: 7, Chamadas: 4, Reparos: 8, Motivo Chamada: 3 e Comercial: 9. Para o modelo de segundo nível, esse parâmetro assumiu o valor de 3.

A Tabela 25 apresenta o desempenho dos modelos de primeiro nível. O modelo de Motivo de Chamada apresentou o melhor desempenho com acurácia de 70,37% e *lift* de 1,7689. O segundo melhor desempenho foi obtido pelo modelo Comercial, seguido por Chamadas, Passou ODC e Reparos.

Tabela 25 – Métricas de Predição para Modelos de Primeiro Nível - Treinamento

Modelos	Acurácia	S	E	VPP	VPN	F1-Score	Lift
Passou ODC	0,6641	0,4743	0,7097	0,2820	0,8489	0,5686	1,4552
Chamadas	0,6697	0,4888	0,7132	0,2906	0,8530	0,5801	1,4997
Reparos	0,6609	0,4661	0,7077	0,2771	0,8465	0,5620	1,4299
Motivo Chamada	0,7037	0,5766	0,7343	0,3428	0,8783	0,6459	1,7689
Comercial	0,7011	0,5698	0,7327	0,3388	0,8763	0,6411	1,7482

A Tabela 26 mostra que a ordem de desempenho na amostra de validação se manteve a mesma com relação à amostra de treinamento, com redução no valor das métricas. Por exemplo, a acurácia do modelo de motivo da chamada foi de 66,04%.

Tabela 26 – Métricas de Predição para Modelos de Primeiro Nível - Validação

Modelos	Acurácia	S	E	VPP	VPN	F1-Score	Lift
Passou ODC	0,6285	0,5682	0,6431	0,2784	0,8601	0,6033	1,4273
Chamadas	0,6195	0,5450	0,6375	0,2670	0,8526	0,5877	1,3691
Reparos	0,6287	0,5686	0,6432	0,2786	0,8602	0,6036	1,4284
Motivo Chamada	0,6604	0,6499	0,6629	0,3184	0,8865	0,6563	1,6325
Comercial	0,6563	0,6393	0,6604	0,3132	0,8831	0,6497	1,6060

A Tabela 27 apresenta as métricas de avaliação desse modelo. A acurácia na amostra de validação foi de 68,72%, mais de um ponto percentual com relação ao melhor modelo obtido anteriormente (GA). O *lift* obtido foi de 1,8052, valor superior ao melhor modelo (GA) (1,7398). Ou seja, essa abordagem de fato trouxe ganhos para a melhor assertividade do modelo.

Tabela 27 – Métricas do Modelo de Segundo Nível - RF

Período	Acurácia	S	E	VPP	VPN	F1-Score	Lift
Desenvolvimento	0,7280	0,6392	0,7493	0,3800	0,8963	0,6899	1,9610
Validação	0,6872	0,7186	0,6796	0,3521	0,9088	0,6986	1,8052

4.4 Conclusão

A modelagem desse problema, iniciada dentro da empresa, foi realizada via regressão logística pela técnica de *stacking*, em que cada modelo de primeiro nível foi construído sobre diferentes bases de dados temáticas. O conhecimento das bases de dados e do negócio da empresa foi adquirido pela interação com as áreas de negócio responsáveis pelo produto que delimitou o público modelado.

A empresa possuía uma diretriz clara e consolidada da metodologia de avaliação da performance do modelo, cuja principal métrica era obtida pelo percentual de verdadeiros positivos (VP), fixado o tamanho da lista de clientes, ou seja, dada uma relação fixa de clientes, ordenadas pelo *Score* com relação à propensão à reclamação ao ODC, quantos daquela lista procuravam de fato o ODC, até 30 dias, após uma data de referência. Portanto, o gráfico de ganhos acumulados e a métrica *lift* eram centrais na avaliação.

Após a finalização da modelagem, a área de negócios julgou que o modelo apresentou boa performance. A utilização da Regressão Logística por meio do *Stacking* possibilitou a fácil implantação em linguagem T-SQL no servidor local que atendia a área, com atualização mensal. Não foi possível avaliar o modelo em produção, uma vez que esse projeto foi despriorizado pela empresa.

Durante o mestrado profissional, foi possível aprender novas técnicas de modelagem, sobretudo nas disciplinas de Introdução aos Sistemas Evolutivos e Introdução ao Aprendizado de Máquina. A primeira disciplina possibilitou aplicar a metodologia de algoritmos genéticos para otimizar a combinação dos *Scores* de primeiro nível via diferentes regras de combinação. Com a aplicação, foi possível obter como resultado prático um modelo mais simples (parcimonioso) com relação ao *Stacking* sem perda de performance.

A segunda disciplina possibilitou o contato com diferentes técnicas de *Machine Learning*. Dentre as técnicas apresentadas, foi escolhida para a aplicação neste projeto o algoritmo *Random Forest*. Essa metodologia, possibilitou ganho de performance com relação à Regressão Logística, porém, foi observado, durante o processo de modelagem, que, muitas vezes, essa abordagem requeria grande esforço computacional, além de maior dificuldade de implantação em produção.

Em síntese, este projeto foi relevante pela aplicação em dados reais e por avaliar diferentes abordagens metodológicas. Apresentou que o teste de diferentes técnicas pode trazer ganhos de assertividade, porém mostrou que a dinâmica empresarial (*trimming to money*), muitas vezes, não

permite que sejam exploradas referidas alternativas. Dessa forma, a interação entre empresa e universidade se mostrou muito importante por proporcionar espaço para esse tipo de reflexão.

REFERÊNCIAS

- BERNARD LAURENT HEUTTE, S. A. S. Influence of Hyperparameters on Random Forest Accuracy. **Lecture Notes in Computer Science**, v. 5519, n. 6, p. 171–180, 2009. Citado na página 26.
- BOLFARINE, H.; BUSSAB, W. de O. **Elementos de amostragem**. Edgard Blücher, 2005. (ABE - Projeto Fisher). ISBN 9788521203674. Disponível em: <https://books.google.com.br/books?id=a_fqPwAACAAJ>. Citado na página 18.
- CUTLER, M.; STERNE, J. E-Metrics: Business Metrics For The New Economy. **NetGenesis Corp**, Cambridge,, p. 1–67, 2000. Citado na página 12.
- DINIZ, C.; LOUZADA, F. Métodos Estatísticos para Análise de Dados de Crédito. 2013. Disponível em: <http://www.mwstat.com/franciscoulouzada/Papers{_}Books{_}files/Livro{_}BICSMIF2013-TamanhoRed>. Citado nas páginas 15 e 17.
- FACELI KATTI; LORENA, A. C. G. J. C. A. C. P. d. L. F. d. **Inteligência artificial: uma abordagem de aprendizado de máquina**. [S.l.]: LTC Editora, 2011. Citado nas páginas 23 e 25.
- FINLAY, S. Using genetic algorithms to develop scoring models for alternative measures of performance. p. 1–19, 2006. Disponível em: <<http://eprints.lancs.ac.uk/27932/>>. Citado na página 27.
- FOGARTY, D. J. Using Genetic Algorithms for Credit Scoring System Maintenance Functions. **International Journal of Artificial Intelligence & Applications**, v. 3, n. 6, p. 1–8, 2012. ISSN 09762191. Disponível em: <<http://www.airccse.org/journal/ijaia/papers/3612ijaia01.pdf>>. Citado na página 27.
- LATTIN, J. D. C. e. P. E. G. J. **Análise de Dados Multivariados**. [S.l.]: CENGAGE, 2011. Citado na página 21.
- LEMESHOW, D. W. H. **Applied Logistic Regression**. [S.l.]: John Wiley and Sons, 2005. Citado na página 22.
- LIU, H.; MOTODA, H. **Feature selection for knowledge discovery and data mining**. [S.l.: s.n.], 1998. 224 p. ISSN 1098-6596. ISBN 978-1-4613-7604-0. Citado na página 19.
- NEVES, M. d. S. O Setor de Telecomunicações. **BNDES 50 Anos - Histórias Setoriais**, 2002. Disponível em: <http://www.bndes.gov.br/SiteBNDES/bndes/bndes{_}pt/Institucional/Publicacoes/Consulta{_}Expressa/Setor/Telecomunic>. Citado na página 13.
- PIRES, J. C. L.; DORES, A. B. das. Fusões e aquisições no setor de telecomunicações: características e enfoque regulatório. p. 53, 2000. Citado na página 14.
- S SILVERMAN DT, M. J. M. J. W. Selection of Controls in Case-Control Studies. **Americal Journal od Epidemiology**, v. 135, n. 5, p. 1019–1050, 1992. ISSN 1873-2585. Disponível em: <<http://www.ncbi.nlm.nih.gov/pubmed/1595688>>. Citado na página 17.

SIDDIQI, N. **Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring**. [S.l.]: SAS Institute Inc., Cary, North Carolina, USA. ALL RIGHTS RESERVED. For additional SAS resources, visit support.sas.com/bookstore., 2005. Citado na página 20.

SILVA, A. C. V. da. A Privatização no Brasil: Evolução do Mercado de Telecomunicações no Brasil no Período Pós-privatização. 2000. Citado na página 14.

SYLVESTER, J.; CHAWLA, N. V. Evolutionary Ensembles : Combining Learning Agents using Genetic Algorithms. **Engineering**, p. 46–51, 2005. Disponível em: <http://www.aaai.org/Papers/Workshops/2005/WS-05-09/WS05-09-008.pdf>. Citado na página 27.

VERGARA, J. R.; ESTÉVEZ, P. A. A review of feature selection methods based on mutual information. **Neural Computing and Applications**, v. 24, n. 1, p. 175–186, 2014. ISSN 09410643. Citado na página 20.