UNIVERSIDADE FEDERAL DE SANTA CATARINA
CENTRO TECNOLÓGICO
DEPARTAMENTO DE ENGENHARIA DE PRODUÇÃO E SISTEMAS
CURSO ENGENHARIA DE PRODUÇÃO MECÂNICA

Fernanda Guimarães Dória

UTILIZAÇÃO DE BIG DATA PARA A CONSTRUÇÃO DE FEATURES NO DESENVOLVIMENTO DE MODELOS PREDITIVOS EM ANÁLISE DE CRÉDITO

Florianópolis
2019

Fernanda Guimarães Dória

UTILIZAÇÃO DE BIG DATA PARA A CONSTRUÇÃO DE FEATURES NO DESENVOLVIMENTO DE MODELOS PREDITIVOS EM ANÁLISE DE CRÉDITO

Trabalho Conclusão do Curso de Graduação em Engenharia de Produção mecânica do Centro Tecnológico da Universidade Federal de Santa Catarina como requisito para a obtenção do título de Engenheiro mecânico habilitado em produção
Orientador: Prof. Dr. Ricardo Giglio.

Florianópolis

2019

Ficha de identificação da obra

Doria, Fernanda
UTILIZAÇÃO DE BIG DATA PARA A CONSTRUÇÃO DE FEATURES NO DESENVOLVIMENTO DE MODELOS PREDITIVOS EM ANÁLISE DE CRÉDITO / Fernanda Doria ; orientador, Ricardo Giglio, 2019.
58 p.

Trabalho de Conclusão de Curso (graduação) - Universidade Federal de Santa Catarina, Centro Tecnológico, Graduação em Engenharia de Produção Mecânica, Florianópolis, 2019.

Inclui referências.

1. Engenharia de Produção Mecânica. 2. Análise de crédito. 3. Data science. 4. Regressão logística. 5. Machine Learning. I. Giglio, Ricardo. II. Universidade Federal de Santa Catarina. Graduação em Engenharia de Produção Mecânica. III. Título.

Fernanda Guimarães Dória

UTILIZAÇÃO DE BIG DATA PARA A CONSTRUÇÃO DE FEATURES NO DESENVOLVIMENTO DE MODELOS PREDITIVOS EM ANÁLISE DE CRÉDITO

Este Trabalho Conclusão de Curso foi julgado adequado para obtenção do Título de "Engenheiro Mecânico habilitado em Produção." e aprovado em sua forma final pelo Curso de Engenharia de Produção Mecânica

Florianópolis, 29 de outubro de 2019.

Prof. Guilherme E. Vieira, Dr.
Coordenador do Curso

Banca Examinadora:

Prof Ricardo Faria giglio, Dr.
Orientador
Universidade Federal de Santa Catarina

Prof. Diego de Castro Fettermann, Dr.
Avaliador
Universidade Federal de Santa Catarina

Prof. Sérgio Fernando Mayerle, Dr.
Avaliador
Universidade Federal de Santa Catarina

RESUMO

O segmento de análise de dados tem apresentado crescimento exponencial nos mercados nacionais e internacionais nos últimos anos, possibilitando uma tomada de decisão baseada em métodos estatísticos e proporcionando uma substancial melhora nos resultados econômicos. Nesse contexto, a avaliação individual no mercado de análise de crédito vem continuamente buscando novas formas de renovação tecnológica, introduzindo diferentes aspectos de caracterização através da utilização de diversos tipos de dados pessoais. O presente trabalho busca, dessa maneira, criar novas variáveis a partir de dados de compra do varejo, de forma a categorizar comportamentos inadimplentes e complementar sistemas de avaliação já disponíveis no mercado. Os dados foram trabalhados visando a quantificação de diferentes parâmetros comportamentais através da criação de features, possibilitando sua aplicação em modelos de classificação e utilizando como variável resposta dados de uma plataforma online de modelagem preditiva. O processo interativo de análise, em que cada teste foi feito utilizando diferentes combinações de variáveis e modelos preditivos, permitiu comparar resultados através de métricas estatísticas ao final de cada ciclo, apontando quais apresentaram melhores rendimentos. Foram utilizadas ferramentas estatísticas e feature importance de diferentes modelos de classificação para a seleção de variáveis (Ridge Classifier e Balanced Random Forest), buscando comparar os modelos Random Forest, Ridge classifier e Regressão Logística quanto ao desempenho com base nas métricas ROC-AUC e MCC. O desenvolvimento dos passos apresentados resultaram em 15 combinações de grupos de variáveis e modelos estatísticos, apontando para regressão logística como modelo mais eficiente na predição de possíveis inadimplentes e a seleção de features baseada em diferentes modelos de previsão para limitação das variáveis utilizadas no algoritmo, resultando em um índice AUC-ROC de 0,647 e MCC de 0,095.

Palavras-chave: Análise de Crédito. Aprendizado de Máquina. Seleção de Features.

ABSTRACT

The data analysis segment has shown exponential growth in national and international markets in the last years, enabling decision making based on statistical methods and providing a substantial improvement in economic results. In this context, the individual evaluation in the credit analysis market has been continually seeking new ways of technological renewal, introducing different aspects of characterization through the use of different types of personal data. Thus, the present work seeks to create new variables from retail purchase data, in order to categorize default behaviors and complement evaluation systems that are already available in the market. The data were manipulated aiming the quantification of different behavioral parameters through the creation of features, enabling its application in classification models and using as an answer variable data from an online predictive modeling platform. The interactive process of analysis, in which each test was done using different combinations of variables and predictive models, allowed to compare results through statistical metrics at the end of each cycle, indicating which ones presented better yields. Statistical tools and feature importance of different classification models were used for the variable selection (Ridge Classifier and Balanced Random Forest), seeking to compare the Random Forest, Ridge Classifier and Logistic Regression models as its performance based on the ROC-AUC and MCC metrics. The development of the presented steps resulted in 15 combinations of variable groups and statistical models, pointing to logistic regression as the most efficient model for predicting possible defaults and feature selection based on different forecasting models to limit the variables used in the algorithm, resulting in an AUC-ROC index of 0,647 and MCC of 0,095.

Keywords: Credit analysis. Machine Learning. Feature Selection.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AUC - Area Under The Curve

CEP – Código de Endereçamento Postal

CPF – Cadastro de Pessoa Física

EDA - Análise Exploratória dos Dados

NLP- *Natural Language Processing*

RFE - *Recursive Feature Elimination*

ROC - *Receiver Operating Characteristic*

SCPC - Serviço Central de Proteção ao Crédito

Serasa - Centralização de Serviços dos Bancos

SPC - Serviço de Proteção ao Crédito

SUMÁRIO

1 INTRODUÇÃO

1.1 CONTEXTUALIZAÇÃO

De acordo com Liu (2019), as receitas mundiais no mercado de Big Data para softwares e serviços apontam para uma taxa de crescimento anual de 10,48%, agregando valor ao diminuir as despesas em 49,2% dos casos e criando novos caminhos para inovação (NewVantage Partners, 2017). Com a dinamicidade do mercado atual, em que compras online foram responsáveis por uma receita de R$166,2 bilhões de reais apenas em 2017 no Brasil (IPSOS, 2018), a necessidade de avaliação para a concessão de crédito individual de forma quase que instantânea demanda novos critérios e fontes de dados de forma a discernir e selecionar, de maneira otimizada, aqueles com capacidade de pagamento ou possíveis inadimplentes, buscando assim potencializar os lucros institucionais.

Visto isso, a utilização de modelos de previsão associados a dados digitais tem apresentado crescimento expressivo na avaliação de instituições para a deliberação de empréstimos, permitindo tanto o surgimento de empresas que associam Big Data ao mercado financeiro quanto o desenvolvimento tecnológico de estabelecimentos tradicionais no ramo de análise de crédito (PASSARELLI, 2016). Desde 2015, o Banco do Brasil tem utilizado dados de redes sociais como insumo para a obtenção de indicadores comportamentais sobre os clientes, adotando uma postura mais ativa na prevenção de riscos de inadimplência. Em um país que conta com 52% da população envolvida em algum tipo de dívida e um saldo de R$96,6 bilhões em inadimplência acumulada apenas em setembro de 2018 (COSTA, 2018), o desenvolvimento de métodos alternativos de avaliação de crédito aos poucos deixa de representar uma vantagem competitiva para se tornar regra em grande parte das instituições financeiras.

O presente trabalho procura, através da utilização de dados de compras realizadas tanto no varejo virtual (online, e-commerce) quanto no real (off-line, lojas físicas), criar novos parâmetros para a avaliação e concessão de crédito individual, levando em consideração diferentes aspectos sociais, regionais e comportamentais para a sua formulação através do emprego de ferramentas de aprendizado de máquina.

1.2 JUSTIFICATIVA

Para grande parte dos brasileiros, o acesso ao crédito é um requisito essencial para a mobilidade ascendente e o sucesso financeiro, sendo necessária uma boa avaliação para a aquisição de tal. Segundo Óskarsdóttir et al. (2018), a pontuação de crédito é a forma mais antiga de análise, em que credores e instituições financeiras avaliam o poder de compra do cliente visando a concessão ou não do empréstimo demandado. Segundo Chen e Xiang (2017), bancos comerciais têm historicamente confiado na experiência de agentes de crédito para a sua realização no campo da gestão de risco. Com o aumento expressivo na demanda de tais profissionais, um número crescente de abordagens baseadas em dados históricos vem sendo desenvolvido visando a otimização dos processos em questão, possibilitando a avaliação de novos candidatos ou mesmo clientes existentes.

Apesar da quantidade expressiva de tais dados, um número notável de pessoas com pouco ou nenhum histórico de crédito se encontra em situação desfavorável a esse tipo de avaliação. Segundo Hurley e Adebayo (2016), a não concessão de crédito àqueles com pequena quantidade de dados avaliados está relacionada ao alto custo de uma possível inadimplência, sendo a adoção de métodos alternativos de análise um caminho viável para facilitar o processo daqueles que se encontram em situações economicamente marginalizadas. Segundo Óskarsdóttir et al. (2018), essa abordagem pode beneficiar jovens tomadores de empréstimos, credores explorando novos mercados ou países em desenvolvimento com mercados de crédito jovens.

O marco regulatório internacional para bancos - Basileia III - que surgiu em resposta à crise americana de crédito em 2007, apresentou um conjunto abrangente de medidas de reforma para fortalecer a regulação, supervisão e gestão de risco do setor bancário, buscando novas soluções e visando maior flexibilidade e sensibilidade ao risco perante o mercado geral (ORESKI, S., ORESKI, D., ORESKI, G., 2012). Em outras palavras, os sistemas de pontuação alternativos vêm crescendo de forma institucionalizada e contínua, permitindo aos credores uma melhor tomada de decisão no gerenciamento de clientes existentes e na previsão de seu desempenho futuro (THOMAS, 2000).

Dessa forma, a adição de dados de compra na caracterização individual busca amparar e aprimorar modelos já existentes, fornecendo um panorama mais completo de avaliação e favorecendo a entrada de novos consumidores ao passo que cria barreiras para potenciais casos de inadimplência.

1.3 OBJETIVOS

1.3.1 Objetivo Geral

O objetivo geral da pesquisa consiste no desenvolvimento de um modelo de previsão capaz de identificar inadimplentes, tendo como base dados do varejo para a criação de *features* de forma melhorar resultados de modelos pré-existentes no mercado de avaliação de crédito. São utilizados modelos supervisionados de classificação através de ferramentas de aprendizado de máquina para a sua execução, comparando sua efetividade através de métricas estatísticas ao final de cada ciclo.

1.3.2 Objetivos Específicos

Alguns objetivos específicos foram propostos visando atingir o objetivo geral delimitado, sendo estes apresentados abaixo:

a) Delimitação de um escopo para a estruturação do processo de criação de *features*.

b) Criação de *features* baseadas no escopo previamente desenvolvido.

c) Seleção de *features* através de abordagens estatísticas e da aplicação de diferentes modelos de previsão.

d) Identificação de um modelo que otimize diferentes métricas de avaliação.

1.4 LIMITAÇÕES DO TRABALHO

Por apresentar dados pessoais sigilosos e features desenvolvidas como produto para uma plataforma de detecção de fraude, o presente trabalho apresenta limitações quanto a disponibilidade de informação a ser divulgada. Dessa maneira, apesar de apresentar os diferentes resultados (com base nas métricas MCC e AUC-ROC) obtidos nas diferentes combinações de métodos de classificação e seleção de variáveis, as features desenvolvidas, assim como exemplos de personas identificadas, não puderam ser explicitadas.

2 FUNDAMENTAÇÃO TEÓRICA

O presente capítulo apresentará os métodos de avaliação de crédito nos âmbitos nacionais e, brevemente, internacionais, assim como o seu desenvolvimento ao longo da história e os processos utilizados na construção do modelo apresentado.

2.1 ANÁLISE DO SISTEMA DE AVALIAÇÃO DE CRÉDITO

O primeiro modelo de análise de crédito com a concessão de uma pontuação para fins comerciais surgiu nos EUA em 1956. A pontuação FICO (ou FICO score, em inglês), desenvolvida pela empresa Fair, Isaac and Company, ainda é usada como instrumento para tomada de decisão por instituições financeiras, seguradoras e serviços terceirizados (ÓSKARSDÓTTIR et al., 2019). Além disso, o surgimento do cartão de crédito e suas tecnologias relacionadas no final da década de 1960 tornou possível a automatização da decisão de empréstimo não apenas por uma perspectiva econômica (uma vez que criou-se grande autonomia em relação a utilização de mão-de-obra para o processamento de dados), como também pela possibilidade de uma tomada de decisão imparcial e respaldada por dados e modelos legítimos baixando, assim, as taxas de inadimplência em 50% ou mais (THOMAS, 2000). Já na década de 1980, os resultados bem sucedidos provenientes do sistema de pontuação fez com que bancos passassem a utilizá-lo para empréstimos pessoais além dos empréstimos imobiliários e para pequenas empresas, como vinha acontecendo até então. Sob essa perspectiva, o interesse institucional não apenas visava a concessão ou não de crédito, como também os limites que possivelmente poderiam ser oferecidos para cada cliente de forma individual (THOMAS, 2000).

No Brasil, os métodos de avaliação de crédito se concentram em dados ou pontuações fornecidos principalmente por três instituições: Serasa (Centralização de Serviços dos Bancos), SPC (Serviço de Proteção ao Crédito) e SCPC (Serviço Central de Proteção ao Crédito). Tal serviço foi iniciado no país através da última, hoje em dia denominada Boa Vista SSPC, em resposta ao crescimento da economia e do sistema de crédito (ACIAV, 2017). Atualmente, os serviços prestados são divididos entre aqueles voltados principalmente à análises de dívidas bancárias (SERASA) e aqueles voltados à dívidas adquiridas no comércio (SPC e SCPC).

As tabelas abaixo exemplificam os dados fornecidos para cada plano pessoal oferecido pelas instituições mencionadas anteriormente:

Tabela 1 – Serviços voltados à análise de dados oferecidos pelo SPC

Plano SPC	Informações
Novo SPC Maxi	Dados cadastrais do consultado; consultas realizadas por outras empresas ao CPF/CNPJ; informações de dívidas vencidas e não pagas incluídas no banco de dados do SPC; informações sobre títulos protestados nos cartórios de todo país; informações de dívidas vencidas e não pagas incluídas no banco de dados do SPC e da Serasa; ocorrências de cheques sem fundos, roubados, sustados ou extraviados; informação sobre créditos concedidos ao cliente informados pelas empresas associadas ao SPC.
Novo SPC Mix Mais	Dados cadastrais do consultado; consultas realizadas por outras empresas ao CPF/CNPJ; informações de dívidas vencidas e não pagas incluídas no banco de dados do SPC; informações sobre títulos protestados nos cartórios de todo país; ocorrências de cheques sem fundos, roubados, sustados ou extraviados; informação sobre créditos concedidos ao cliente informados pelas empresas associadas ao SPC.
SPC Imobiliário	Dados cadastrais do consumidor; CCF – Cadastro de emitentes de cheques sem fundos; SPC; renda presumida; crédito concedido; contraordem; contumácia; dados agência bancária; ordem judicial; pendências financeiras na base Serasa; protesto; participação em empresas; ação; SPC Score 12 meses.
SPC Óbito	Dados cadastrais da pessoa física; dados cadastrais da pessoa jurídica; informação de Óbito
SPCheque Analítica	Dados cadastrais do consultado; endereço e telefone; consultas realizadas por outras empresas ao documento consultado; informações sobre cheques sem fundos, roubados, sustados ou extraviados; informação sobre os últimos endereços e telefones informados.

Fonte: SPC Brasil, acesso em 2019

Tabela 2 – Serviços voltados a análise de dados oferecidos pelo SERASA

Planos SERASA	Informações
Consulta Completa	Serasa Score; consultas à Serasa e SPC; dívidas em bancos, empresas e SPC; informações societárias; cheques; antecessores; protestos; informações cadastrais; falências e ações judiciais; CNAE e endereços
Consulta Intermediária	Serasa Score; dívidas em bancos, empresas e SPC; cheques; protestos; informações cadastrais; falências e ações judiciais; CNAE e endereços
Identificador de Dívidas	Identificador de dívidas CNPJ ou CPF (não inclui o detalhamento da dívida); endereço, telefone e demais dados CPF ou CNPJ

Fonte: SERASA Experian, acesso em 2019

Os serviços de Consulta Completa, Consulta Intermediária e Identificador de dívidas do SERASA custam, respectivamente, R$30,00, R$15,00 e R$5,00. Os dados referentes às consultas pelo SPC não são explicitados no site e não puderam ser informadas através do contato via e-mail. Para fins não empresariais, o SPC oferece consultas individuais a R$9,90 o acesso aos dados relativos ao próprio CPF e R$16,90 a consulta a CPF de terceiros. Tal serviço garante o acesso aos seguintes dados (SPC, 2019):

- Registro de inadimplência no SPC Brasil.
- Informações cadastrais, últimos endereços e telefones.
- Registro de título protestado em cartório.
- Informações do Poder Judiciário.
- Crédito concedido.
- Registro de cheque lojista.
- Consultas realizadas no documento nos últimos 180 dias.
- Alerta de documentos, mediante registro de furto junto à CDL ou à Associação Comercial.

2.2 ANÁLISE COMPORTAMENTAL CARACTERÍSTICA

Levando em consideração a grande quantidade de dados disponíveis, um estudo sobre comportamentos característicos relacionados à inadimplência foi realizado visando uma melhor compreensão e direcionamento em relação ao problema. Uma pesquisa realizada por Arya, Eckel e Wichman (2012), motivada por um maior entendimento sobre a correlação entre o credit score e atitudes sociais representativas, testou quatro fatores potencialmente ligados à inadimplência: impaciência, impulsividade, falta de confiabilidade e tolerância ao risco.

> Parece razoável esperar que credit scores mais baixos sejam associados a maiores descontos em recompensas futuras: isto é, a impaciência está associada a um desejo de transferir o consumo do futuro para o presente, e empréstimos maiores implicam em uma maior probabilidade de inadimplência. Indivíduos impulsivos são propensos a ter dificuldade em resistir à tentação de pedir emprestado para consumo atual, e mais propensos a não pagar suas dívidas. Os credit scores ruins também podem ser causados por falta de confiabilidade, pois os menos confiáveis não cumprem suas obrigações. E, finalmente, as pontuações de crédito podem ser afetadas de maneira significativa pela assunção de riscos financeiros, já que aqueles que jogam acumulam dívidas que têm dificuldade em pagar. (ARYA, ECKEL, WICHMAN, 2012).

Em relação aos resultados, apenas a tolerância ao risco não apresentou resultados que indicassem correlação com um baixo credit score. Uma análise realizada pelo SPC em agosto de 2018 apresenta os principais produtos comprados no crédito, responsáveis pela inadimplência (cujo valor médio é de R$2.934,34), apresentados abaixo (SPC, 2018):

- Roupas, calçados e acessórios: 42,0%
- Supermercado: 20,5%
- Eletrônicos (TV, DVD, aparelho de som, MP3, câmera digital, computador, notebook, tablet): 20,3%
- Celular/smartphone: 15,2%
- Eletrodomésticos (geladeira, fogão, cafeteira): 15,2%

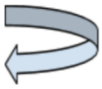
2.3 *DATA SCIENCE* NA AVALIAÇÃO DE COMPORTAMENTOS CARACTERÍSTICOS

Segundo Peng e Matsui (2016), uma pesquisa exploratória está relacionada à análise de dados visando a identificação de padrões, tendências ou relacionamento entre variáveis criando possíveis hipóteses relacionadas. Dessa forma, a identificação de padrões em comportamentos de compras característicos baseada em dados primários demanda uma análise exploratória inicial, utilizando como ferramenta simulações e previsões estatísticas qualitativas de forma a confirmar ou refutar hipóteses sobre tais comportamentos em relação à inadimplência.

Os procedimentos metodológicos relacionados a esse tipo de pesquisa são baseados em 4 atividades dinâmicas principais: Constatação e refinamento de uma hipótese inicial, análise exploratória dos dados (EDA), construção de modelos estatísticos e interpretação dos resultados, sendo eventualmente retomadas ao longo do processo. Embora existam diferentes tipos de atividades envolvidas, aspectos distintos de cada uma podem ser abordados por meio de um processo interativo denominado "ciclo de análise de dados", cujas etapas são apresentadas abaixo:

1. Estabelecimento de objetivos;
2. Coleta de dados, comparando-os com os objetivos;
3. Revisão dos objetivos, de forma a corresponder com os dados fornecidos.

Figura 1 – Procedimento metodológico da pesquisa exploratória

	Estabelecimento de objetivos	Coleta de dados	Revisão dos objetivos	
Hipótese inicial	Objetivo é factível e de interesse?	Revisão bibliográfica	Aperfeiçoamento dos objetivos	
EDA	Os dados são apropriados para o objetivo?	Análises visuais ou descritivas sobre dados obtidos	Refinamento dos objetivos ou obtenção de mais dados	
Modelos estatísticos	Modelo inicial corresponde aos objetivos iniciais?	Análise de sensibilidade sobre alterações em *features* e hiperparâmetros	Alteração de modelo de previsão	
Interpretação	Interpretação da análise concede resposta significativa e específica para o objetivo?	Interpretação da totalidade da análise relacionada a diferenças no tamanho da amostra e incertezas	Revisão da EDA e/ou modelo de previsão para prover respostas específicas e interpretáveis	

Fonte: Peng, Matsui (2016)

2.4 FEATURE ENGINEERING

Segundo Turner (1999), uma *feature* pode ser caracterizada como um agrupamento ou modularização de requisitos individuais dentro de um tema específico, enfatizando sua origem dentro do domínio do problema. A representação de tais aspectos do problema ocorre, no contexto de *Machine Learning*, através da estruturação de uma *Feature Matrix* onde cada instância a ser analisada é representada por uma linha e cada *feature* por uma coluna. Dessa forma, cada unidade observada (representada por uma linha horizontal) é caracterizada através de colunas por valores numéricos (inteiros ou contínuos), discriminadores (ou booleanos, representados por 1 ou 0) ou categóricos (representados por palavras ou números que não estejam em uma escala numérica), obtidos através da análise e tratamento dos dados recebidos. De forma geral, *feature engineering* pode ser caracterizado como o processo de transformar tais dados em recursos (ou variáveis dentro das *features*), atuando como entradas para modelos de aprendizado de máquina (SARKAR, 2018).

Para exemplificar o processo de construção de *features*, a tabela abaixo apresenta os dados obtidos a partir da coleta de *digital footprints* de usuários de uma companhia *e-commerce,* assim como as variáveis utilizadas para construir o modelo. As *digital footprints* (pegadas digitais) podem ser facilmente coletadas por lojas que prestam serviços de vendas on-line, podendo ser utilizadas para complementar métodos tradicionais de concessão de *credit score* (BERG et al., 2018):

Tabela 3 – Descrição de variáveis de um modelo de análise de crédito

feature	Descrição	Unidade
Tipo de aparelho	Tipo do aparelho utilizado na compra. Ex: desktop, tablet, celular.	Categórica
Sistema operacional	Sistema operacional utilizado. Ex: Windows, iOS, Android, Machintosh.	Categórica
E-mail host	*E-mail host.* Ex: Gmail, Hotmail, Yahoo.	Categórica

Canal	Canal o qual consumidor utilizou para direcionar ao site de compra. Ex: pago (incluindo cliques pagos e redirecionados), direto, afiliado, orgânico.	Categórica
Hora de check-out	Hora do dia em que a compra foi realizada.	Numérica (0-24h)
Configuração de não-rastreamento	Binário igual a 1 caso consumidor não permita o rastreamento do aparelho, sistema operacional ou canal de compra.	Binária
Nome no *e-mail*	Binário igual a 1 caso primeiro ou último nome do consumidor seja parte do *e-mail*	Binária
Número no *e-mail*	Binário igual a 1 caso algum número esteja presente no *e-mail*	Binária
É letra minúscula	Binário igual a 1 caso primeiro nome, último nome, nome da rua ou cidade estejam escritas em letra minúscula	Binária
Erro no *e-mail*	Binário igual a um se o endereço de e-mail contiver um erro na primeira tentativa (Observação: os clientes só podem fazer o pedido se se registrarem com um endereço de e-mail correto).	Binária

Fonte: Berg *et al.* (adaptado - 2018)

A estruturação final de uma *feature matrix,* em que cada linha é representada por uma instância individual, é apresentada abaixo:

Figura 2 – feature matrix representando exemplo anterior

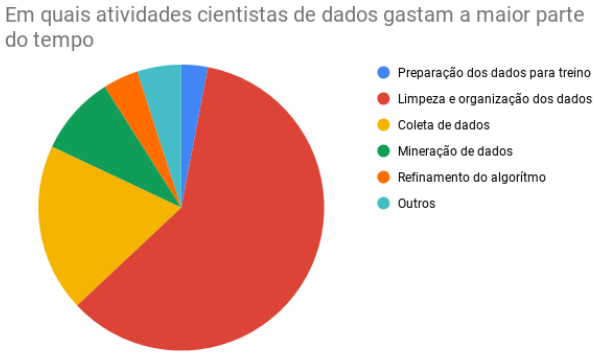
ID	Tipo de aparelho	Sistema operacional	E-mail host	Canal	Hora check-out	Configuração não-rastreamento	Nome no e-mail	Número no e-mail	É letra minúscula	Erro no e-mail
Id_0	Desktop	Windows	hotmail	Google ads	15:36	1	1	0	0	0
Id_1	Celular	iOS	gmail	facebook ads	13:07	0	0	1	0	1
Id_2	Tablet	iOS	yahoo	direto	18:00	0	1	0	1	0
Id_3	Celular	Android	gmail	facebook ads	03:00	1	0	1	1	1
Id_5	Desktop	Machintosh	gmail	direto	09:00	0	1	0	0	0

Fonte: Elaborado pelo autor (2019)

2.5 *FEATURE PROCESSING*

Para a construção de um modelo de predição consistente e eficaz, é necessário que se padronize os dados de entrada desenvolvendo heurísticas específicas para o preenchimento de variáveis que não estejam no padrão encontrado dentro das *features* elaboradas. Segundo Press (2016), a coleta e limpeza dos dados são responsáveis por aproximadamente 80% do tempo efetivo de trabalho de um cientista de dados, como evidenciado na figura abaixo:

Figura 3 – Distribuição do tempo efetivo de trabalho de um cientista de dados



Fonte: Gil Press (adaptado - 2016)

Neste processo de padronização, é inicialmente importante garantir que todas as *features* apresentem, de forma individual, objetos do mesmo tipo - dados exclusivamente numéricos (números inteiros ou decimais), *strings* (palavras), etc. Além disso, uma análise criteriosa deve ser inicialmente feita para garantir que valores discrepantes dentro de uma mesma *feature* possam ser identificados, uma vez que os mesmos podem gerar ruídos prejudiciais para o desenvolvimento do modelo. Observações que apresentam outliers numéricos devem ser examinadas para a definição de um valor válido correspondente, caso apresente claro erro de imputação, ou simples remoção da observação. Para *features* comportas por *strings,* a análise deve ser voltada para a determinação de um padrão não apenas de digitação das palavras, como também de pequenas transformações para casos de NLP (*natural language processing*). Isso corresponde a configuração das variáveis em letras apenas minúsculas e possível remoção dos radicais e dos acentos.

Após tais procedimentos de padronização, desenvolvimento de estratégias para a substituição de valores nulos ao longo da *feature matrix* deve levar em consideração o tipo de informação carregada pela *feature* avaliada. Em casos numéricos (inteiros ou contínuos),

a substituição do valor faltante por zero pode gerar falsos indicadores que, potencialmente, podem prejudicar o modelo (como a substituição de um valor nulo relacionado a idade por zero, por exemplo). Assim, a utilização de modelos matemáticos de regressão levando em consideração observações semelhantes e *features* relacionadas se faz necessária. Em casos booleanos, apenas a imputação de valores de 0 ou 1 garantem resultados satisfatórios.

Para a execução do modelo, *features* compostas por *strings* devem ser transformadas em booleanas através da transposição do valor de variáveis em novas *features.* Um exemplo de tal processo pode ser observado abaixo:

Figura 4 – Exemplo do processo descrito

ID	SEXO
Id_0	feminino
Id_1	masculino
Id_2	masculino
Id_3	feminino
Id_5	feminino

ID	feminino	masculino
Id_0	1	0
Id_1	0	1
Id_2	0	1
Id_3	1	0
Id_5	1	0

Fonte: Elaborado pelo autor (2019)

Apesar da necessidade de avaliação de cada caso, o pré-processamento das *features* pode ser feito de forma automática através da utilização de algumas bibliotecas. O pacote *sklearn.preprocessing*, por exemplo, realiza tratamentos matemáticos para a padronização de features tanto numéricas quanto em forma de strings, sendo algumas delas (SCIKIT-LEARN, 2019):

- Centralização da distribuição em 0 (*features* numéricas)
- Delimitação da escala a uma variação específica (*features* numéricas)
- Implementação dos dados em distribuição gaussiana (*features* numéricas)
- Transformação de strings em vetores (*features* categóricas)

2.6 MODELOS DE PREVISÃO

Segundo Copeland (2016), o termo *machine learning* está relacionado, de forma sucinta, ao desenvolvimento de algoritmos para coletar dados, aprender com eles e, então, determinar ou prever resultados sobre um tema especificado. Ao identificar padrões, a máquina é capaz de formar diferentes heurísticas podendo ou não conter, na sua formulação,

as variáveis resposta (configurando o modelo como supervisionado ou não-supervisionado, respectivamente). Enquanto o último apresenta melhores resultados para problemas de agrupamento ou categorização a partir de conjuntos não identificados, modelos supervisionados utilizam variáveis resposta para o desenvolvimento de um algoritmo que leva em consideração os dados apresentados na *feature matrix*. Segundo Maini (2017), o objetivo do aprendizado supervisionado é prever Y com a maior precisão possível, quando são dados novos exemplos em que X é conhecido e Y é desconhecido.

Três modelos relacionadas à resolução de problemas supervisionados serão apresentadas a seguir, sendo divididos em classificadores ou regressivos. Enquanto o primeiro remete a classificações categóricas, os modelos regressivos apresentam como *output* variáveis numéricas, provenientes de diferentes modelos matemáticos de regressão.

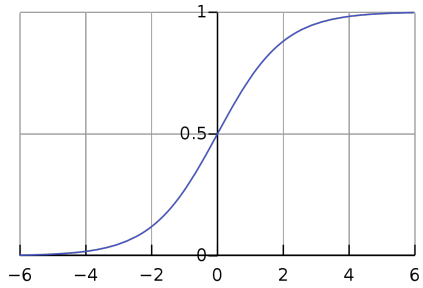
2.6.1 Regressão logística

O modelo de regressão logística consiste, no universo de *machine learning,* em uma técnica de aprendizado estatístico que visa estimar a probabilidade do acontecimento de um evento levando em consideração uma gama de diferentes variáveis, classificando-o de acordo com o problema especificado.

Assumindo Y como variável dependente de dois estados possíveis (0 e 1) e um número 'n' de variáveis independentes $(X_1, X_2, \ldots, X_n)$, podemos descrever o modelo da forma $P(Y = 1) = \frac{1}{1+e^{-f(x)}}$, sendo $f(x) = B_0 + B_1X_1 + \ldots + B_nX_n$. Os coeficientes $B_0, B_1, \ldots, B_n$são determinados através do método da máxima verossimilhança, encontrando uma combinação de coeficientes que maximize a probabilidade de a amostra ter sido observada (BARBETTA, 2011).

Considerando uma combinação específica de coeficientes e variando os valores de X, é possível obter uma curva logística (ou função sigmoide) de comportamento probabilístico, como apresentado abaixo:

Figura 5 – Curva logística de comportamento probabilístico



Fonte: Weisstein (2019)

Em problemas de classificação, a discriminação entre grupos deve ser feita através do arredondamento dos valores obtidos através da regressão: caso P(Y=1)>0,5 , Y=1. Caso contrário, se P(Y=1)<0,5, devemos considerar Y=0.

2.6.2 Ridge classifier

O modelo de classificação *Ridge* é baseada em regressões lineares, estruturando-se como uma função $\hat{y} = w[0] \times x[0] + w[1] \times x[1] + \ldots + w[n] \times x[n] + b$, sendo n o número de *features* utilizadas pelo modelo (SAPTASHWA, 2018):

A partir de um conjunto de M observações, é possível desenvolver um modelo de regressão linear visando a otimização de w e b através do método de mínimos quadrados, minimizando a função custo. Sua representação pode ser observada na equação abaixo:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{n} w_j \times b_{ij})^2$$

No modelo de regressão *Ridge,* a função custo é alterada adicionando uma penalidade equivalente ao quadrado da magnitude dos coeficientes:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{M}(y_i - \sum_{j=0}^{n} w_j \times b_{ij})^2 + \lambda \sum_{j=0}^{n} w_j^2$$

Isso significa que a função *Ridge* adiciona restrições nos coeficientes *w.* O termo de penalidade lambda (λ) realiza a regularização dos coeficientes de forma a penalizar a função

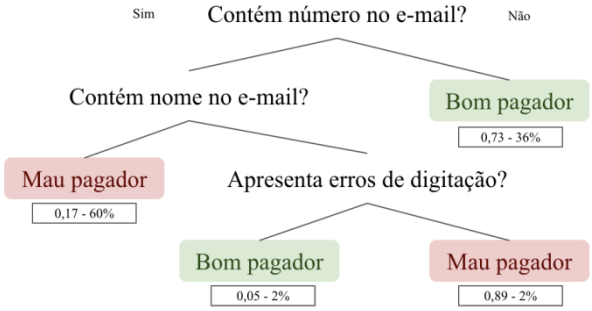
de otimização caso os mesmos assumam altos valores, reduzindo-os e baixando, assim, a complexidade e multicolinearidade do modelo. É possível observar que, quando $\lambda \rightarrow 0$, a função de custo se torna semelhante a função de regressão linear, tornando possível a alteração de tal parâmetro de acordo com a necessidade do modelo a ser implementado.

Para a implementação do modelo de regressão *Ridge* em problemas de classificação, a variável resposta é rotulada com valores +1 (casos positivos, ou 1 em modelos binários) ou -1 (casos negativos, ou 0 em classificação binária), de acordo com a classe em que pertence. Dessa forma, valores maiores que zero retornam Y=1 e valores menores que zero retornam Y=0.

2.6.3 Random Forest

O algoritmo do modelo de classificação *Random Forest* baseia-se no modelo de árvores de decisão, desenhada de cabeça para baixo com a raiz no topo. Na imagem abaixo, cada pergunta representa uma condição (ou nó) a partir da qual diferentes galhos se dividem e chegando, ao final, em uma folha (ou decisão) (GUPTA, 2017). A imagem abaixo exemplifica o processo, utilizando como modelo os dados relacionados às pegadas digitais (ou *digital footprints)* apresentados anteriormente de forma simplificada:

Figura 6 – Exemplo do modelo *Random Forest*



Fonte: elaborado pelo autor (2019)

Modelos simples de classificação (como decision trees, por exemplo) podem fazer uso de *ensemble methods* (métodos de conjunto, em português) para garantir melhores resultados, combinando diferentes modelos para a construção de uma heurística ideal. O processo denominado BAGGing (bootstrap + aggregating), combina *bootstrap* (re-amostragem iterativa de um conjunto de dados com substituição) e agregação das diferentes

previsões do modelo aplicados a dados iniciais de entrada. Uma árvore de decisão é formada em cada uma das amostras iniciais a partir de um mesmo conjunto de *features*, sendo então agregadas para formar o preditor mais eficiente. Apesar de se basear em modelos BAGGing, a aplicação do algoritmo *Random forest* seleciona *features* de forma aleatória em cada nó, gerando diversos cenários e permitindo a geração de um resultado mais preciso (LUTINS, 2017).

2.6.3.1 Balanced random forest

Em casos onde há um significativo desequilíbrio entre ocorrências na variável resposta, a probabilidade de que as amostras iniciais contenham poucas ou nenhuma observação da classe minoritária é considerável, resultando em uma árvore com baixo desempenho de previsão da classe inferior. Dessa forma, o algoritmo *Balanced Random Forest* (BRF) induz a geração de árvores em dados balanceados de baixa amostragem, a partir de uma segregação inicial dessas variáveis. Os passos relacionados ao processo são demonstrados a seguir (Chen, C. *et.al,* 2014):

1) Para cada iteração, selecione amostras *bootstrap* da classe minoritária.
2) Selecione aleatoriamente um mesmo número de amostras, com substituição, da classe majoritária.
3) Formule uma árvore de classificação de máximo tamanho de forma semelhante à apresentada anteriormente, com a seguinte modificação: Em cada nó, em vez de analisar todas as variáveis para a divisão ideal, pesquise apenas um conjunto de *m* variáveis selecionadas aleatoriamente.
4) Repita as etapas acima o número de vezes desejado. Agregue as previsões do conjunto para a previsão final.

2.7 TUNAGEM DE HIPERPARÂMETROS

Diferentemente dos parâmetros (variáveis escolhidas para realizar o ajuste dos dados através de técnicas de machine learning, como *Random forest* e *Logistic regression*), os hiperparâmetros são variáveis de configuração que controlam o próprio processo de treinamento, permanecendo constantes durante um job (Produtos Google de machine learning e IA, 2019). Segundo Prabhu (2018), os hiperparâmetros são importantes porque

controlam diretamente o comportamento do algoritmo de treinamento e têm um impacto significativo no desempenho do modelo que está sendo treinado.

2.8 MÉTRICAS DE AVALIAÇÃO

Feitos os processos anteriormente apresentados, é necessário que se avalie a qualidade do algoritmo desenvolvido baseando-se em diferentes parâmetros relacionados aos resultados apresentados. Ao testar o modelo com parte dos dados, as variáveis de uma previsão supervisionada podem facilmente ser comparadas as variáveis resposta, indicando a assertividade ou não do modelo.

2.8.1 Confusion matrix

Baseando-se na comparação entre os resultados obtidos e a variável resposta, uma matriz de confusão (*confusion matrix*) pode ser estruturada levando em consideração quatro possíveis cenários:

- Verdadeiro positivo (*True Positive* - TP): Casos apresentados como positivos (ou 1) pelo algoritmo e ratificados pela variável resposta.
- Verdadeiro negativo (*True Negative* - TN): Casos apresentados como negativos (ou 0) pelo algoritmo e ratificados pela variável resposta
- Falso positivo (*False Positive* - FP): Casos apresentados como positivos (ou 1) pelo algoritmo de forma equivocada, sendo de fato negativos
- Falso negativo (*False Negative* - FN): Casos apresentados como negativos (ou 0) pelo algoritmo de forma equivocada, sendo de fato positivos.

A *confusion matrix* é normalmente retratada através de uma tabela, apresentada abaixo:

Tabela 4 – Exemplo de *confusion matrix*

		Resposta original	
		1	0
Previsão do modelo	1	TP	FP
	0	FN	TN

Fonte: Elaborada pelo autor (2019)

A partir da avaliação dos modelos da previsão, diferentes métricas podem ser aplicadas buscando apresentar panoramas específicos de cada resultado tendo como base os indicadores apresentados na *feature matrix.* São explicitadas, na tabela abaixo, as principais métricas utilizadas para a avaliação de modelos em *machine learning:*

Tabela 5 – Exemplo de *feature matrix*

Acurácia	$\frac{TP + TN}{Total\ de\ instâncias}$
Precisão	$\frac{TP}{TP + FP}$
F1	$2 \times 1/(\frac{1}{Precisão} + \frac{1}{Recall})$
Sensibilidade (*Recall*)	$\frac{TP}{TP + FN}$
Especificidade	$\frac{TN}{TN+FP}$

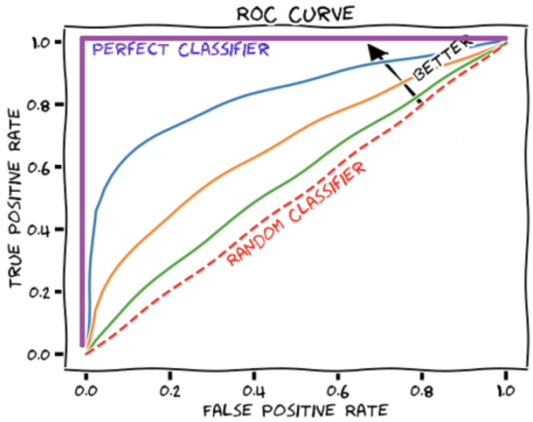
Fonte: Elaborado pelo autor (2019)

2.8.2 AUC-ROC *(Area Under the Curve - Receiver Operating Characteristics)*

A curva de probabilidade ROC (*Receiver Operating Characteristics*), aliada à medição da área sobre o gráfico formado (AUC - *Area Under the Curve*) proporciona um indicador relacionado a separabilidade das classes levando em consideração as métricas de

sensibilidade e especificidade - quanto maior o AUC, melhor o modelo em prever 0s e 1s corretamente. Uma representação da métrica pode ser observada abaixo:

Figura 7 – Representação da curva de probabilidade ROC



Fonte: Draelos (2019)

É importante que se faça uma análise sobre o significado e pertinência dos eixos horizontais e verticais. *True positive rate* (taxa de verdadeiros positivos, ou sensibilidade) é relevante no contexto em que os falsos negativos são considerados mais prejudiciais do que falsos positivos (como no caso de detecção de doenças, por exemplo). Paralelamente a isso, *False positive rate* (taxa de falsos positivos, ou especificidade) relaciona aspectos opostos àqueles apontados pela sensibilidade, complementando assim o modelo.

O índice AUC, que varia entre 0 e 1, representa a probabilidade de identificar corretamente uma observação aleatória: 1 determina chances de 100% de acerto, 0,5 aponta para um algoritmo inteiramente aleatório e 0 representa uma inversão de classes (ou seja, previsão de classes negativas para positivas e vice-versa).

2.8.3 MCC (Matthews Correlation Coefficient)

Apesar de a grande maioria das métricas utilizadas na avaliação de modelos de aprendizado de máquina funcionarem bem com variáveis quantitativamente equilibradas, a tarefa se torna desafiadora quando tratamos de dados desbalanceados, muitas vezes tendenciando a resposta ao conjunto de dados majoritário. Tais algoritmos são ineficientes nesse caso uma vez que procuram maximizar uma medida de desempenho, como precisão, negligenciando o grande desequilíbrio das variáveis.

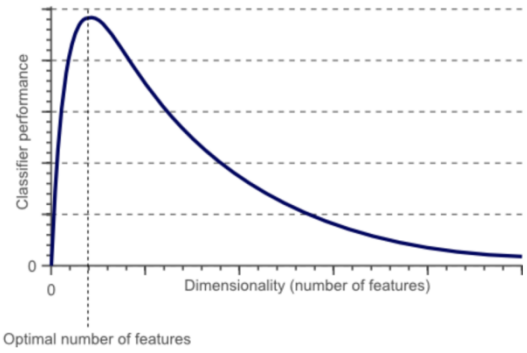
Visto isso, a métrica MCC visa solucionar tais desequilíbrios através da utilização de todos os quadrantes da *confusion matrix* (TP, TN, FP, FN), retornando valores no intervalo [-1, 1] sendo 1 uma concordância completa, -1 uma discordância completa e 0 a ausência de correlação entre as variáveis (BOUGHORBEL, 2017). A formulação da métrica pode ser observada abaixo:

$$\mathrm{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

2.9 SELEÇÃO DE *FEATURES*

Mesmo que individualmente úteis, um vasto número de *features* é capaz de prejudicar o modelo devido ao acréscimo de um grande número de dimensões. A chamada 'maldição da dimensionalidade' (*curse of dimensionality,* em inglês) se refere justamente aos fenômenos que surgem a partir da análise de dados com um elevado número de *features,* inferindo que um modelo otimizado não está necessariamente relacionado a um grande número de atributos, como apresentado na imagem abaixo:

Figura 9 – Avaliação da performance de acordo com o número de features



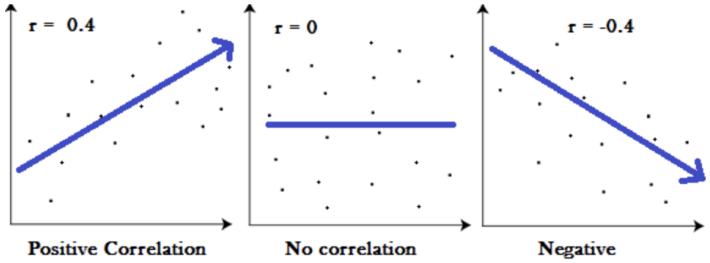
Fonte: Computer Vision for dummies (2014)

Diversas heurísticas foram desenvolvidas buscando diminuir significativamente o número de *features,* deixando no modelo apenas aquelas de fato relevantes para o problema como apresentadas a seguir.

2.9.1 Correlação entre variáveis

A correlação é uma medida estatística utilizada para medir a relação linear entre duas variáveis (BARBETTA, 2011). O ajuste dos dados pode ser representado visualmente em um gráfico de dispersão, como apresentado abaixo:

Figura 10 – Gráfico de dispersão



Fonte: Statistics How To (2019)

O coeficiente de correlação é um valor que indica a relevância da relação entre as variáveis, assumindo coeficientes entre -1 e +1, sendo interpretadas da seguinte maneira:

- -1: correlação negativa perfeita. As variáveis tendem a se mover em direções opostas (ou seja, quando uma variável aumenta, a outra variável diminui linearmente).
- 0: sem correlação.
- 1: Correlação positiva perfeita. As variáveis tendem a se mover na mesma direção (ou seja, quando uma variável aumenta, a outra variável aumenta linearmente).

A determinação do coeficiente pode ser feita através da fórmula:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \underline{x})(y_i - \underline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \underline{x})^2 \sum_{i=1}^{n}(y_i - \underline{y})^2}}$$

r_{xy}- Coeficiente de correlação linear entre as variáveis x e y

x_i- Valor de x na observação i

$\underline{x}$- Média entre os valores de x

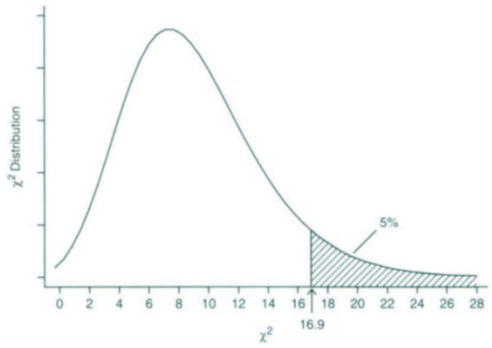
y_i- Valor de y na observação i

$\underline{y}$- Média entre os valores de y

Dessa forma, podemos observar que o coeficiente 1 é representado por uma reta a 45° do eixo das coordenadas, de forma que qualquer acréscimo em x representa uma alteração de mesma magnitude em y. No contexto de machine learning, a remoção de *features* altamente correlacionadas (90% ou mais) promove um menor esforço computacional ao rodar o modelo, tornando-o mais barato e eficiente.

Segundo Lee (2019), 'p' é uma medida estatística que auxilia na determinação de hipóteses e resultados, representada por um número entre 0 e 1 e baseada em modelos de correlação entre variáveis. O teste qui-quadrado, nesse contexto, compara *features* e variáveis resposta para verificar se estão relacionadas entre si, avaliando e confrontando suas distribuições em relação às esperadas caso não estivessem vinculadas. O resultado obtido com a fórmula de Chi-quadrado retorna o valor da probabilidade acumulada à esquerda da função, como pode ser observado abaixo:

Figura 11 – Exemplo da utilização do chi-quadrado para a obtenção do p-valor.



Fonte: Riffenburgh (2006)

Para esse tipo de teste, valores baixos de p-value estipulados a partir de um nível de significância α(normalmente abaixo de 0,05) indicam alta relevância de observações, decaindo de acordo com a sua correlação com a variável resposta.

2.9.2 RFE (*Recursive feature elimination*)

A utilização de RFE (*Recursive Feature Elimination*) consiste na eliminação consecutiva de features, chegando em um número pré-determinado que otimize ao máximo

o modelo escolhido diante de tais parâmetros. Como dito anteriormente, tal ferramenta pode ser utilizada tendo como base diferentes modelos de previsão, aplicando métodos característicos de eliminação de *features* de acordo com o modelo proposto.

Segundo Kim (2019), a análise de regressão "Ridge" consiste em um método de seleção de variáveis para modelos de regressão linear, obtendo um subconjunto de preditores que minimiza o erro de previsão para uma variável de resposta quantitativa. Tal análise impõe restrições nos parâmetros do modelo fazendo com que os coeficientes de regressão de algumas variáveis tendam a zero, sendo estas excluídas do modelo conforme os parâmetros estipulados pela utilização do RFE. A utilização do RFE em modelos de classificação *Random Forest* tem como base a avaliação de variáveis através da métrica *Gini Importance,* sendo empregada consecutivamente até chegar ao número estipulado de *features*. Para cada árvore de decisão, a relevância de cada nó é calculada assumindo dois "nós filhos" (uma vez que trata-se de um problema binário).

3 METODOLOGIA

A metodologia utilizada para o desenvolvimento da pesquisa é apresentada em três partes: na primeira, o cenário do estudo é apresentado levando em consideração o contexto e objetivos do trabalho; na segunda parte, é apresentado o enquadramento metodológico utilizado e, por fim, as etapas e procedimentos utilizados para chegar no resultado final.

3.1 CENÁRIO DE ESTUDO

A presente pesquisa tem como base dados fornecidos pela LINX (empresa especializada no ramo de tecnologia da informação), sendo estes coletados através do uso de seus produtos por mais de 50.000 empresas. Suas atividades são enquadradas em três grandes áreas: gestão e performance para varejo físico, transformação digital e meios de pagamento, possuindo 41,3% de market share do mercado varejista e 180 franquias. Suas principais atividades são apresentadas abaixo:

- Varejo físico: Sistema ERP e PDV, sistema de controle de emissão de documentos fiscais
- Transformação digital:
    - *Understand:*Captura e segmentação de clientes, além da criação de soluções a partir da automatização de *marketing*
    - *Engage:* Criação de vitrines personalizadas
    - *Reengage:* Criação de campanhas automáticas para o reengajamento de clientes.
    - OMNI OMS: Integração de canais físico e digitais
    - LINX COMMERCE: Plataforma de vendas *online*
- Meios de pagamento: Soluções de pagamento empresariais personalizadas

Além disso, dados referentes ao pagamento ou não de uma compra são obtidos através de uma plataforma online de modelagem preditiva, com foco em crédito, cobrança e fraude.

3.2 PROCEDIMENTOS METODOLÓGICOS

Os procedimentos metodológicos utilizados na pesquisa seguem o "ciclo de análise de dados" apresentado por Peng e Matsui (2016), sendo eles: constatação e refinamento de uma hipótese inicial, análise exploratória dos dados (EDA), construção de modelos estatísticos e interpretação dos resultados, apresentados a seguir.

3.2.1 Constatação e refinamento de uma hipótese inicial

Para Peng e Matsui (2016), o levantamento de um objetivo para a pesquisa deve levar em consideração 5 aspectos relevantes relacionados a hipótese inicial: ser de interesse a um segmento ou problema especificado, ainda não ter sido respondida, plausível (removendo possíveis causalidades relacionados ao tema), respondível (levando em consideração os dados existentes e questões éticas intrínsecas ao processo) e direta/específica. Dito isso, a proposta de encontrar uma possível correlação entre comportamentos característicos e inadimplência através de dados de compras no varejo confere à hipótese inicial todos os requisitos demandados, possibilitando assim o desenvolvimento da pesquisa.

3.2.2 Obtenção e análise exploratória dos dados

Segundo Medri (2011), "a análise exploratória de dados nos fornece um extenso repertório de métodos para um estudo detalhado dos dados, antes de adaptá-los. Nessa abordagem, a finalidade é obter dos dados a maior quantidade possível de informação, que indique modelos plausíveis a serem utilizados numa fase posterior, a análise confirmatória de dados ou inferência estatística". A partir dessa análise, é possível identificar quais serão os caminhos e características do banco de dados são mais pertinentes para a obtenção dos resultados desejados, assim como quais deles estão desconexos com o contexto da pesquisa ou necessitam de uma padronização.

As informações estão divididas em diferentes tabelas, cada qual relacionada a uma esfera de mercado ou atuação sendo todas indexadas pelo CPF do consumidor. Uma vez que parte das informações fora detalhada manualmente por funcionários de lojas físicas, uma avaliação criteriosa sobre possíveis erros de português e preenchimento ou outliers numéricos é necessária para que não haja a possibilidade de deturpação dos resultados. Além disso, campos com uma baixa taxa de preenchimento devem ser desconsiderados.

A seguir, são apresentados os itens utilizados para a construção do modelo, com as colunas de baixo índice de preenchimento (abaixo de 15%) já removidas:

- Tabela *"customer":* dados relativos a informações pessoais cadastrais do consumidor - CPF, nome, e-mail, telefone, dia do nascimento, gênero, país do endereço de entrega, estado do endereço de entrega, cidade do endereço de entrega, rua do endereço de entrega, CEP do endereço de entrega.

- Tabela *"payments":* dados relativos à loja, momento da compra e forma de pagamento (perspectiva fiscal da transação) - CPF, ano da compra, mês da compra, hora da compra, CNPJ da loja, cidade da loja, estado da loja, CEP da loja, número e descrição de cada divisão do CNAE (Classificação Nacional de Atividades Econômicas), valor do pagamento, método de pagamento, tipo de pagamento (condições de parcelamento) e marca do cartão de crédito (se utilizado).

- Tabela *"offline sales":* dados relativos à loja, produto e valores (perspectiva organizacional da venda) - CPF, ano da compra, mês da compra, hora da compra, CNPJ da loja, cidade da loja, estado da loja, CEP da loja, número e descrição de cada divisão do CNAE, código EAN do produto, SKU do produto, código do produto no sistema da instituição, categoria do produto no sistema da instituição, quantidade de produtos vendidos, preço comercial do produto, valor de desconto e valor total da compra

- Tabela *"online users":* dados relativos ao consumidor, nome da loja e detalhes sobre dispositivo utilizado e valores relacionados à compra - CPF, e-mail, gênero, CEP do cliente, credenciais da loja, *browser* utilizado para efetuar a compra, sistema operacional utilizado para efetuar a compra, número de produtos comprados.

Além disso, foi utilizado como variável resposta (ou Y) dados provenientes da plataforma online citada anteriormente, em que 0 representa inadimplência e 1 o devido pagamento das compras realizadas. Vale ressaltar que tais variáveis se apresentam de forma desproporcional: apenas aproximadamente 10% dos CPFs cadastrados apresentam 1 como variável resposta.

Feita a familiarização acerca das tabelas apresentadas, foi determinada a segmentação de alguns tópicos para uma melhor análise do problema: divisão entre dados e *features* relacionadas ao cliente (características relacionadas a dados individuais cadastrais de cada CPF), compra (características gerais relacionadas a produtos e grupos de produtos, assim como padrões de compra), loja (características gerais relacionadas a loja onde a compra

foi realizada) e transação (características gerais relacionadas ao método de pagamento), permitindo assim diferentes enfoques entre as categorias. É importante salientar que se trata de uma divisão abstrata, em que grande parte das *features* se enquadra em mais de uma categoria.

3.2.3 Modelagem formal

Segundo Oreski, S., Oreski, D. e Oreski, G. (2012), a precisão das previsões de um cliente bom ou ruim em termos de risco de crédito pode ser otimizada através de uma boa seleção de dados de entrada e da combinação dos resultados de diferentes métodos de classificação. Dessa forma, o método de construção do modelo preditivo foi inicialmente focado amplamente no desenvolvimento de diferentes *features* visando caracterizar, de forma individual, comportamentos relacionados à inadimplência no Brasil. A partir da segmentação feita em uma primeira análise exploratória dos dados, foi determinado quais aspectos dentro de cada subgrupo deveriam ser observados através de cada *feature,* baseando-se em alguns comportamentos identificados na literatura. Além disso, os dados de compra foram reduzidos de forma a apresentar apenas observações de pessoas cujo CPF constava na base de dados da plataforma online, retornando informações referentes a um total de 20.000 indivíduos.

A partir da tabela *customer*, diferentes dados relacionados ao cadastro individual por CPF permitiram a construção de um perfil individual majoritariamente enquadrado no grupo "Cliente". A escrita do *e-mail*, por exemplo, permite a análise de aspectos relacionados à escolaridade, profissionalismo e necessidade de identificação pessoal, enquanto a avaliação do CEP do endereço de entrega possibilita uma investigação sobre características regionais relacionadas ao ambiente de vida da pessoa em questão.

Diferentemente da primeira tabela apresentada acima, as tabelas *payments, offline sales* e *online users* apresentam um histórico temporal de compras encabeçadas pelo CPF do cliente, em que diversas informações referentes à loja e produtos comprados são agregadas de acordo com o seu acontecimento. Em relação ao grupo "Compras", dados relacionados ao custo de itens semelhantes permitem a classificação de lojas e produtos a partir de uma comparação de preços, podendo classificá-los como *premium* ou populares. Além disso, a classificação de um produto como promocional e a quantidade de compras realizadas utilizando cupons de desconto permitem a criação numérica de um perfil quantificando possíveis características impulsivas.

A análise sobre compras realizadas em diferentes lojas ao longo do tempo permite a criação de *features* enquadradas nas categorias "Loja" e "Compra", possibilitando a formulação de um cenário em que se investiga hábitos ou mudanças comportamentais e sociais. Uma compra realizada fora da cidade em período de férias possivelmente indica um maior acesso pessoal ao lazer, enquanto gastos online no período da madrugada potencialmente estão ligados a atitudes impulsivas, por exemplo. Compras concentradas no início do mês potencialmente caracterizam um menor poder individual de compra, diferentemente da distinção de indivíduos que as realiza em lojas com poucos concorrentes.

Por fim, a construção de variáveis relacionadas à temática “Transação” está diretamente ligadas à forma de pagamento, parcelamentos e métodos utilizados para a realização das compras, possibilitando a identificação e enquadramento individual em um comportamento padrão característico possivelmente inadimplente. Além disso, valores numéricos foram extensamente trabalhados a fim de extrair diferentes parâmetros quantitativos (valores máximos, médios, mínimos, moda, desvio padrão).

Por questões de sigilo, o conjunto de *features* desenvolvidas não pôde ser apresentado no presente trabalho.

Após a construção de um número significativo de *features*, foi necessário padronizá-las de forma a termos apenas variáveis numéricas na *feature matrix* para a implementação do modelo. *Features* categóricas foram transpostas em booleanas, e valores nulos foram substituídos por 0. Feitos tais processos de *feature engineering* e *feature processing,* diversos testes foram realizados levando em consideração variações nos modelos preditivos, seus respectivos hiperparâmetros e seleções de *features.*

Segundo Óskarsdóttir *et al* (2018), o modelo de regressão logística para modelos tradicionais de pontuação de crédito tem demonstrado resultados significativamente positivos, sendo técnicas de classificação mais recentes passíveis de oferecer apenas ganhos de desempenho marginais. Dito isso, foram feitos testes utilizando, além do modelo tradicional, os modelos de classificação *Ridge* e *Random Forest,* com diferentes combinações entre alguns de seus hiperparâmetros (variações do lambda no primeiro e no número de árvores no segundo, além de diferentes pesos entre as variáveis resposta em ambos os casos) para a validação de tal hipótese inicial. Para validação das respostas e aplicação das métricas escolhidas, a *feature matrix* foi dividida no modelo 70-30, de modo a possibilitar o desenvolvimento do modelo nos 70% dos dados iniciais e aplicá-lo nos 30% restantes.

Por fim, diversos métodos de seleção de variáveis foram utilizados a fim de evitar os problemas relacionados à "maldição da dimensionalidade", estipulando um valor de 150 para o número máximo de *features* a serem utilizadas.

3.2.4 Interpretação

Ao analisar os resultados ao final de cada teste, diferentes métricas de avaliação foram utilizadas visando uma melhor percepção acerca dos diferentes parâmetros relacionados à resposta. Isso têm importância significativa sob a perspectiva de que as variáveis resposta se apresentam em forma desproporcional. Dessa forma, além da métrica ROC-AUC (utilizada pela plataforma de análise de análise preditiva e amplamente difundida na literatura), o coeficiente de correlação Matthews (MCC) foi utilizado a fim de analisar os efeitos do desbalanceamento de variáveis na interpretação dos modelos de previsão.

4 RESULTADOS

A fim de constatar a hipótese de que a adição de *features* provenientes de dados do varejo auxilia na construção de um modelo de previsão para a análise de crédito, diferentes modelos e seleções de *features* foram utilizadas e combinadas de forma distintas. A seguir, são apresentados os resultados de cada combinação, de forma a serem comparadas no final da presente seção. Em relação aos hiperparâmetros, uma vez apontada a distribuição desbalanceada entre as respostas, foram determinados diferentes pesos entre as variáveis 1 e 0 de forma a não prejudicar a previsão de ocorrências não populares, sendo o primeiro 100 vezes maior que o segundo. Além disso, foi estipulado um número de 1000 árvores para a produção do modelo *Random Forest,* e um $\lambda = 0{,}01$no modelo de classificação *Ridge*.

4.1 CENÁRIO ATUAL

Os resultados obtidos a partir dos testes realizados sobre os dados fornecidos pela plataforma de detecção de fraude são apresentados abaixo. Não foi feito nenhum tipo de tratamento ou seleção inicial sobre as 101 *features* recebidas, de forma a estabelecer um parâmetro inicial de comparação.

Tabela 6 – Resultados dos testes realizados sobre as variáveis da plataforma

Modelo	ROC-AUC	MCC
Regressão Logística	0,584	0,018
Ridge Classifier	0,575	0,023
Random Forest	0,537	0,009

Fonte: Elaborado pelo autor (2019)

Paralelamente a isso, um teste utilizando as 3033 *features* desenvolvidas na pesquisa também foi feito, de forma a comparar os resultados iniciais obtidos.

Tabela 7 – Resultado do teste com 3303 features

Modelo	ROC-AUC	MCC

Regressão logística	0,553	0,006
Ridge Classifier	0,559	0,029
Random Forest	0,512	0,002

Fonte: Elaborado pelo autor (2019)

Por fim, a fim de avaliar o desempenho da combinação de *features* juntamente com o impacto gerado por uma seleção posterior, um novo teste foi feito mantendo as mesmas configurações iniciais, agora com um total de 3404 *features*. Os resultados são apresentados à seguir:

Tabela 8 – Resultado do teste com 3404 features

Modelo	ROC-AUC	MCC
Regressão Logística	0,559	-0,009
Ridge Classifier	0,564	0,033
Random Forest	0,562	-0,004

Fonte: Elaborado pelo autor (2019)

É possível perceber que, apesar de apresentar indicadores de ROC-AUC maiores do que uma seleção aleatória de dados (0,5), essa pequena discrepância está associada ao desbalanceamento das variáveis aleatórias, confirmando-se ao observar a métrica MCC. Parte do mau desempenho do modelo está associado ao grande número de variáveis utilizadas para a sua construção, ocasionando a denominada "maldição da dimensionalidade".

4.2 FEATURE SELECTION

Para que seja possível conciliar informações provenientes da plataforma e as *features* produzidas de forma otimizada, é necessário que haja uma remoção significativa do número de variáveis, utilizando métodos já apresentados anteriormente. Para isso, dois caminhos foram inicialmente adotados levando em consideração tanto ferramentas intrínsecas dos modelos adotados (selecionando as *features* mais relevantes dos modelos *Random Forest* e

Ridge Classifier) quanto métodos estatísticos de correlação (p-value e correlação entre *features*). Para que fosse possível comparar resultados aparentemente similares de ROC-AUC, uma nova métrica não utilizada originalmente pela plataforma contratante do serviço foi adicionada, levando em consideração aspectos negligenciados pela primeira. Os resultados obtidos são apresentados abaixo:

4.2.1 Modelos estatísticos

Para que fosse possível reduzir significativamente o número de variáveis foram inicialmente observadas quais *features* estavam correlacionadas entre si, possibilitando assim um menor esforço computacional no processamento e uma diminuição das dimensões. Posteriormente, o p-value sobre cada *feature* foi calculado de forma a possibilitar a criação de um ranking de acordo com seus respectivos valores, escolhendo os 150 menores resultados para a introdução das respectivas *features* no modelo. Os valores obtidos no processamento de tais dados são apresentados a seguir:

Tabela 9 – Resultados da seleção por modelos estatísticos

Modelo	ROC-AUC	MCC
Regressão Logística	0,636	0,015
Ridge Classifier	0,653	0,021
Random Forest	0,569	0,020

Fonte: Elaborado pelo autor (2019)

4.2.2 *Feature importance* de modelos de previsão

Outro caminho utilizado para a seleção de um número reduzido de *features* foi abordado ao empregar métodos de seleção provenientes de modelos de previsão, visando chegar em um total final de 150 *features.* Para isso, a ferramenta RFE foi utilizada tendo como base os modelos Balanced *Random Forest* (seleção das 500 *features* mais relevantes)

e *Ridge Classifier* (seleção das 150 *features* finais). A tabela abaixo apresenta os resultados obtidos em cada teste:

Tabela 10 – Resultados da seleção por feature importance

Modelo	ROC-AUC	MCC
Regressão Logística	0,647	0,095
Ridge Classifier	0,652	0,080
Random Forest	0,555	0,042

Fonte: Elaborado pelo autor (2019)

Apesar de apresentarem resultados de ROC-AUC similares, uma análise sobre a métrica MCC aponta grande disparidade nos métodos de seleção de variáveis observados, indicando aqueles provenientes de modelos de previsão mais indicados para o caso especificado. Além disso, como já previamente apontado pela literatura, o modelo de regressão logística apresentou resultados expressivos ao lidar com os casos apresentados, respondendo de forma positiva perante os problemas de desbalanceamento existentes no modelo. Na tabela abaixo, os métodos de seleção de variáveis (assim como as métricas aplicadas) são apresentados, de forma a facilitar a comparação:

Tabela 11 – Resultados dos diferentes métodos de seleção de variáveis

Seleção de *features*	Número de *features* utilizadas	Modelo	ROC-AUC	MCC
Features plataforma (sem seleção)	101	Regressão Logística	0,584	0,018
Features plataforma (sem seleção)	101	*Ridge Classifier*	0,575	0,023
Features plataforma (sem seleção)	101	*Random Forest*	0,537	0,009
Features desenvolvidas no projeto (sem seleção)	3033	Regressão Logística	0,553	0,006

Features desenvolvidas no projeto (sem seleção)	3033	*Ridge Classifier*	0,559	0,029
Features desenvolvidas no projeto (sem seleção)	3033	*Random Forest*	0,512	0,002
Combinação de *features* (sem seleção)	3404	Regressão Logística	0,559	-0,009
Combinação de *features* (sem seleção)	3404	*Ridge Classifier*	0,564	0,033
Combinação de *features* (sem seleção)	3404	*Random Forest*	0,562	-0,004
Seleção por métodos estatísticos	150	Regressão Logística	0,636	0,015
Seleção por métodos estatísticos	150	*Ridge Classifier*	0,653	0,021
Seleção por métodos estatísticos	150	*Random Forest*	0,569	0,020
Seleção por *feature importance* de modelos de previsão	150	Regressão Logística	0,647	0,095
Seleção por *feature importance* de modelos de previsão	150	*Ridge Classifier*	0,652	0,080
Seleção por *feature importance* de modelos de previsão	150	*Random Forest*	0,555	0,042

Fonte: Elaborado pelo autor (2019)

5 CONCLUSÃO

A pesquisa apresentada teve como objetivo principal a criação de um modelo de previsão capaz de identificar pessoas inadimplentes através da criação de diversas *features* provenientes de dados do varejo, utilizando-as como insumo em diferentes modelos de classificação e comparando-os no final de cada ciclo apresentado.

A primeira etapa do projeto foi representada por uma vasta pesquisa bibliográfica, em que duas perspectivas principais foram levadas em consideração: aspectos socioculturais relacionados a comportamentos inadimplentes e diferentes abordagens de *Machine Learning* sobre temas relacionados.

Posteriormente, uma análise meticulosa sobre os dados apresentados foi feita na segunda etapa, observando quais características sobressaíam em relação ao padrão e quais apresentavam possível relação com a variável resposta. Dessa forma, foi possível delimitar os passos seguintes através da divisão e coordenação dos esforços na delimitação de características específicas relacionadas a 4 grandes áreas principais: cliente, compra, loja e transação.

A partir da compreensão holística do assunto e de uma análise criteriosa sobre os dados iniciais, a terceira etapa teve como abordagem principal a criação de diversas *features* capazes de transcrever, de forma tanto quantitativa quanto qualitativa, aspectos significativos observados na etapa anterior. A caracterização individual buscou, dessa forma, identificar padrões inadimplentes brasileiros observados na literatura, levando em consideração perspectivas sociais, regionais e comportamentais.

Por fim, a quarta etapa teve como foco o processo interativo de combinação entre diferentes *features* e modelos de previsão, comparando-os ao final de cada estágio através da métrica AUC-ROC e MCC, permitindo uma análise geral sobre os resultados individuais de verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos. Ao levar em consideração a desproporcionalidade entre as variáveis respostas, foi possível observar indicadores MCC díspares mesmo entre valores próximos de AUC-ROC, conduzindo a avaliação à tal métrica como critério de desempate. Dessa forma, foi constatado que a seleção de *features* baseadas em modelos de previsão (redução de 3404 *features* para 500 através do modelo *Balanced Random Forest* e posteriores 150 através do modelo *Ridge Classifier*) apresentou os melhores resultados, sendo o modelo de regressão linear o melhor entre os avaliados.

Apesar de conter uma quantidade expressiva de dados, a pesquisa foi limitada por uma quantidade considerável de informações faltantes, o que repercutiu tanto na etapa da construção de *features* quanto no bom desempenho dos diferentes modelos. No primeiro caso, a falta de um histórico completo de compras gerou significativas interferências no sentido de estimar um comportamento econômico financeiramente ascendente ou descendente, uma vez que a regressão linear ao longo do período analisado sofre alterações

significativas caso apresente dados faltantes. Da mesma maneira, na construção do modelo, a ausência de dados referentes à *features* numéricas e sua respectiva substituição por 0 gera a falsa compreensão de que não houveram atividades no período em detrimento da falta de informação do caso original.

Os modelos construídos na pesquisa podem servir como base para o desenvolvimento de métodos alternativos de avaliação de crédito, possibilitando uma melhor identificação de clientes inadimplentes e facilitando o acesso ao crédito de pessoas com pouco ou nenhum histórico relacionado.

REFERÊNCIAS

ARYA, Shweta; ECKEL, Catherine; WICHMAN, Colin. Anatomy of the credit score. **Journal of Economic Behavior & Organization**, v. 95, p. 175-185, 2013.

BARBETTA, Pedro. Estatística aplicada às ciências sociais. 7. ed. Florianópolis: Editora UFSC, 2011

BERG, Tobias et al. **On the rise of fintechs–credit scoring using digital footprints**. National Bureau of Economic Research, 2018.

Boughorbel S, Jarray F, El-Anbari M. **Optimal classifier for imbalanced data using Matthews Correlation Coefficient metric.** Tianjin University, China, 2017.

CFI - Corporate Finance Institute. **Correlation: A statistical measure of the relationship between two variables.** Disponível em: <https://corporatefinanceinstitute.com/resources/knowledge/finance/correlation/>. Acesso em: 27 set. 2019.

CHEN, Chao et al. Using random forest to learn imbalanced data. **University of California, Berkeley**, v. 110, n. 1-12, p. 24, 2014.

CHEN, Hongmei; XIANG, Yaoxin. The study of credit scoring model based on group lasso. **Procedia computer science**, v. 122, p. 677-684, 2017.

COMPUTER VISION. **The Curse of Dimensionality in classification**. abr. 2014 Disponível em: <https://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/>. Acesso em: 27 set. 2019.

COPELAND, Michael. **What's the Difference Between Artificial Intelligence, Machine Learning, and Deep Learning?** *NVIDIA Blog*, 29 julho 2016. Disponível em: <https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/>. Acesso em: 31 agosto 2019.

COSTA, Gilberto. **Inadimplência atinge 62 milhões de brasileiros e afeta 3% do crédito.** 12 nov. 2018. Disponível em: <http://agenciabrasil.ebc.com.br/economia/noticia/2018-11/inadimplencia-atinge-62-milhoes-de-brasileiros-e-afeta-3-do-credito>. Acesso em: 27 set. 2019.

DRAELOS, Rachel. **Measuring Performance: AUC (AUROC)**. 23 fev. 2019. Disponível em: <https://glassboxmedicine.com/2019/02/23/measuring-performance-auc-auroc>. Acesso em: 27 set. 2019.

GUPTA, Prashant. **Decision Trees in Machine Learning**. 17 mai. 2017. Disponível em: <https://towardsdatascience.com/decision-trees-in-machine-learning-641b9c4e8052>. Acesso em: 27 set. 2019.

HURLEY, Mikella; ADEBAYO, Julius. Credit scoring in the era of big data. **Yale JL & Tech.**, v. 18, p. 148, 2016.

IPSOS (Brasil). **4ª edição da pesquisa PayPal/Ipsos: o perfil do consumidor online.** 12 set. 2018. Disponível em: <https://www.paypal.com/stories/br/4-edico-da-pesquisa-paypal-ipsos-o-perfil-do-consumidor-online>. Acesso em: 12 set. 2019.

KIM, Kyoosik. **Ridge Regression for Better Usage.** 2 jan. 2019 Disponível em: <https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db>. Acesso em: 27 set. 2019.
LEE, Admond. **P-values Explained.** 13 jul. 2019. Disponível em: <https://towardsdatascience.com/p-values-explained-by-data-scientist-f40a746cfc8>. Acesso em: 27 set. 2019.

LIU, Shanhong. **Forecast of Big Data market size, based on revenue, from 2011 to 2027.** 9 ago. 2019. Disponível em: <https://www.statista.com/statistics/254266/global-big-data-market-forecast/>. Acesso em: 27 set. 2019.

MAINI, Vishal. **Machine Learning for Humans, Part 2.1: Supervised Learning**. 19 ago. 2017. Disponível em: <https://medium.com/machine-learning-for-humans/supervised-learning-740383a2feab>. Acesso em: 27 set. 2019.

MEDRI, Waldir. **Análise exploratória de dados**. Mar. 2011 Disponível em: <www.uel.br/pos/estatisticaeducacao/textos_didaticos/especializacao_estatistica.pdf> . Acesso em 29 de maio de 2019.

ORESKI, Stjepan; ORESKI, Dijana; ORESKI, Goran. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. **Expert systems with applications**, v. 39, n. 16, p. 12605-12617, 2012.

ÓSKARSDÓTTIR, María et al. The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analytics. **Applied Soft Computing**, v. 74, p. 26-39, 2019.

PENG, Roger D.; MATSUI, Elizabeth. The Art of Data Science. **A Guide for Anyone Who Works with Data. Skybrude Consulting, LLC**, 2015.

PRABHU. **Understanding Hyperparameters and its Optimisation techniques**. 3 jul. 2018 Disponível em: <https://towardsdatascience.com/understanding-hyperparameters-and-its-optimisation-techniques-f0debba07568>. Acesso em: 27 set. 2019.

PRESS, Gil. Cleaning Big Data: Most Time-Consuming, Least Enjoyable Data Science Task, Survey Says. **Forbes**, 2016. Disponível em: <www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/>. Acesso em: 26 de agosto de 2019.

Produtos Google de Machine Learning e IA. **Visão geral do ajuste de hiperparâmetro**. Disponível em:< https://cloud.google.com/ml-engine/docs/hyperparameter-tuning-overview>. Acesso em: 27 set. 2019.

RONAGHAN, Stacy. **The Mathematics of Decision Trees, Random Forest and Feature Importance in Scikit-learn and Spark.** 11 mai. 2018 Disponível em: <https://medium.com/@srnghn/the-mathematics-of-decision-trees-random-forest-and-feature-importance-in-scikit-learn-and-spark-f2861df67e3>. Acesso em: 27 set. 2019.

SAPTASHWA. **Ridge and Lasso Regression: A Complete Guide with Python Scikit-Learn.** 26 set, 2018 Disponível em: <https://towardsdatascience.com/ridge-and-lasso-regression-a-complete-guide-with-python-scikit-learn-e20e34bcbf0b>. Acesso em: 27 set. 2019.

SARKAR, D. **Understanding Feature Engineering**. *Towards Data Science*, 04 jan. 2018. Disponível em: <https://towardsdatascience.com/understanding-feature-engineering-part-1-continuous-numeric-data-da4e47099a7b>. Acesso em: 31 agosto 2019.

SCIKIT-LEARN. Versão 0.21.2. Código Aberto. 2019.

SCPC BOA VISTA**. 60 anos de história como fornecedor de informações de crédito.** Disponível em: <http://aciav.org.br/3545-2/>. Acesso em: 29 de maio de 2019.

SEGAL, M. R. **Machine learning benchmarks and random forest regression**. University of California, 2004.

SERASA EXPERIAN. **Consulta Serasa**. 2019. Disponível em: <https://empresas.serasaexperian.com.br/consulta-serasa>. Acesso em: 29 de maio de 2019.

SILVA, José Pereira da. **Gestão e Análise de Risco de Crédito**. 2 ed. São Paulo: Atlas, 1998.

SPC BRASIL**. Consulta análise de crédito SPC**, 2019. Disponível em: <https://www.spcbrasil.org.br/produtos/categoria/1-analise-de-credito>. Acesso em: 29 de maio de 2019.

STATISTICS HOW TO. **Correlation Coefficient: Simple Definition, Formula, Easy Steps.** Disponível em: <https://www.statisticshowto.datasciencecentral.com/probability-and-statistics/correlation-coefficient-formula/#Pearson>. Acesso em: 27 set. 2019.

SUBASI, Caglar. **LOGISTIC REGRESSION CLASSIFIER**. 4 mar. 2019. Disponível em: <https://towardsdatascience.com/logistic-regression-classifier-8583e0c3cf9>. Acesso em: 27 set. 2019.

THOMAS, Lyn C. A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers. **International journal of forecasting**, v. 16, n. 2, p. 149-172, 2000.

TURNER, C. Reid et al. A conceptual basis for feature engineering. **Journal of Systems and Software**, v. 49, n. 1, p. 3-15, 1999.

WEISSTEIN, Eric W.. **Chi-Squared Distribution.** Disponível em: <http://mathworld.wolfram.com/Chi-SquaredDistribution.html>. Acesso em: 27 set. 2019.