UNIVERSIDADE FEDERAL DE SÃO CARLOS

Uso de aprendizado de máquina supervisionado para mensurar provisão no mercado de crédito estruturado: uma comparação entre modelos

Matheus de Brito Soares Porto
Orientador: Prof. Dr. Alexandre Luis Magalhães Levada

São Carlos
2023

RESUMO

A expansão do crédito tem sido fundamental para o crescimento econômico global, permitindo que pessoas e empresas tenham acesso a mais recursos financeiros para investimentos e consumo dentro de um sistema econômico cada vez mais complexo e competitivo. Dentre as diversas modalidades de crédito, evidencia-se neste trabalho o crédito estruturado, em que dívidas de várias fontes são unificadas em um só produto financeiro. Dado o risco que cada um desses créditos possui, mensurar uma provisão justa para essas estruturas é um grande desafio para os agentes do mercado. Seguindo este contexto, o objetivo do presente estudo é aplicar e comparar diferentes técnicas de aprendizado supervisionado para prever eventos de calote em carteiras de crédito estruturado, de forma a auxiliar instituições financeiras a realizarem provisões de forma mais precisa e segura. Após as análises, foi possível concluir que o algoritmo XGBoost se sobressaiu nos testes de precisão e sensibilidade, entretanto todos os outros obtiveram resultados satisfatórios.

Palavras-chave: Crédito estruturado. Provisionamento de crédito. Aprendizado de máquina. Ciência de dados.

ABSTRACT

The credit expansion has been fundamental to global economic growth, allowing people and businesses to access more financial resources for investment and consumption within an increasingly complex and competitive economic system. Among the various types of credit, this work highlights structured credit, in which debts from various sources are unified into a single financial product. Given the risk associated with each of these credits, accurately measuring a fair provision for these structures is a major challenge for market agents. In this context, the objective of this study is to apply and compare different supervised learning techniques to predict default events in structured credit portfolios, in order to assist financial institutions in making more accurate and secure provisions. After the analyses, it was possible to conclude that the XGBoost algorithm outperformed the others in terms of accuracy and sensitivity, although all of the others obtained satisfactory results.

Keywords: Structured credit. Credit provisioning. Machine learning. Data science.

LISTA DE FIGURAS

SUMÁRIO

1 INTRODUÇÃO

Este capítulo tem o objetivo de apresentar a contextualização e motivação deste trabalho, bem como seus objetivos gerais e específicos. Além disso, o capítulo exibe a organização geral do texto.

1.1 Contextualização

O fim dos acordos de Bretton Woods em 1971 foi um marco importante para o sistema financeiro global. Com o fim do padrão-ouro, que ligava o valor do dólar americano ao ouro e permitia que outras moedas fossem ligadas ao dólar, os países podiam ajustar livremente suas taxas de câmbio, o que ocasionou um aumento generalizado de empréstimos intercontinentais e iniciou um período de expansão do crédito na economia mundial (KYUTEG, 2022).

O gerenciamento de risco é um dos princípios básicos da economia, e desde a crise do *subprime* em 2008, o risco de crédito é um dos maiores alvos de estudo da ciência econômica. Segundo JP Morgan em Riskmetrics (1996) risco de crédito é o risco de perda financeira devido ao não cumprimento das obrigações de uma parte em uma operação. Nesse sentido, uma das formas de mitigar riscos e evitar catástrofes financeiras é mensurar corretamente a provisão para devedores duvidosos (PDD).

A PDD é importante porque permite que uma empresa ou fundo de investimento reconheça e contabilize perdas potenciais de crédito antes que elas sejam realmente incorridas. Isso fornece uma visão mais precisa da situação financeira da entidade e ajuda a garantir que ela tenha recursos suficientes para cobrir essas perdas potenciais.

No contexto do crédito estruturado, onde, de maneira simplificada, diversos recebíveis são colocados em um só veículo de investimento, como Fundos de Investimento em Direitos Creditórios (FIDC), e disponibilizados no mercado, o PDD é muito importante para evitar perdas para os investidores. Estimar a perda dos diversos créditos dentro de um FIDC não é uma tarefa fácil, e muito embora a Comissão de Valores Mobiliários (CVM) obrigue, através da Instrução CVM 489, os agentes do mercado a estabelecerem metodologias claras de provisionamento, estas metodologias geralmente só levam em conta os dias de atraso dos créditos.

Muito embora o *machine learning* ou aprendizado de máquina, não seja um tema novo, empresas de crédito ainda estão relutantes em aceitar estes métodos em suas políticas de crédito (EZEIZ, 2019). Frente a esse contexto surge o tema de pesquisa que motivou o desenvolvimento deste trabalho: “Uso de aprendizado de máquina supervisionado para mensurar provisão no mercado de crédito estruturado: uma comparação entre modelos”. Neste sentido, na seção 1.2, é apresentado o objetivo buscado.

1.2 Objetivo

O objetivo deste trabalho é implementar um *pipeline* de aprendizado de máquina, ou seja, uma sequência de passos interconectados entre a obtenção dos dados brutos até a treinamento de um modelo, avaliar e comparar a eficácia de diferentes algoritmos supervisionados na previsão de eventos de *default* no contexto de carteiras de crédito estruturadas.

1.3 Objetivos específicos

De forma específica, os objetivos são:

- Expor os principais algoritmos de aprendizado de máquina supervisionados, bem como suas técnicas estatísticas e casos de uso indicados;
- Obter dados históricos de carteiras de crédito estruturadas do mercado brasileiro;
- Implementar um *pipeline* de aprendizado de máquina com esses algoritmos utilizando bibliotecas para Data Science da linguagem Python;
- Consultar e analisar os resultados obtidos pelos modelos e discutir sobre as conclusões obtidas ao final deste trabalho.

1.4 Organização do trabalho

No Capítulo 1, é apresentada a contextualização do trabalho, evidenciando a importância da expansão do crédito para a economia e a importância de se discutir novos métodos de controle sobre o provisionamento destes. Também são apresentados os objetivos gerais e específicos deste trabalho.

No Capítulo 2, é realizada a fundamentação teórica dos dois principais conceitos abordados no trabalho. A primeira parte é destinada à discussão sobre crédito e sua importância para a economia global, depois é discutido mais especificamente o crédito estruturado, evidenciando a relevância dessa modalidade de crédito e a importância de se ter controle dos seus riscos associados. A segunda parte da fundamentação teórica discorre sobre o aprendizado de máquina, com ênfase nos algoritmos de aprendizado supervisionado.

No Capítulo 3, é feita uma descrição sobre a metodologia utilizada para este trabalho, desde o detalhamento sobre como foram levantados os dados para a análise até quais foram as ferramentas e tecnologias utilizadas.

No Capítulo 4, é discorrido sobre os passos que compuseram a parte prática do trabalho: o tratamento dos dados, criação das features para treinamento e a criação do pipeline de aprendizado de máquina.

No Capítulo 5, são apresentados e comparados os resultados obtidos após o treinamento dos modelos.

Por fim, o Capítulo 6 apresenta as conclusões deste trabalho, bem como sugestões para trabalhos futuros.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem o objetivo de detalhar a teoria econômica sobre a qual o projeto está inserido, e expor a relevância do crédito e da prática de provisionamento para o sistema econômico. Também são abordados conceitos importantes de aprendizado de máquina supervisionado, além da apresentação dos seus principais algoritmos, pois a compreensão desses auxiliará no entendimento do projeto descrito posteriormente.

2.1 Crédito

O mercado de crédito é um sistema onde indivíduos e organizações podem emprestar ou tomar emprestado dinheiro. Ele tem uma importância fundamental na economia mundial, fornecendo aos negócios e aos indivíduos a oportunidade de obter recursos necessários para financiar projetos, expandir suas atividades e fazer gestão dos seus fluxos de caixa.

O crédito refere-se a um acordo em que um credor fornece dinheiro ou propriedade a um tomador de empréstimo, e o tomador de empréstimo concorda em reembolsar o valor principal juntamente com juros. O mercado de crédito oferece uma ampla gama de produtos e serviços, incluindo empréstimos, títulos e derivativos de crédito, para atender às necessidades financeiras diversas de seus participantes (TIROLE, 2006).

Na economia global, o crédito é essencial para o crescimento e desenvolvimento econômico: permite que as empresas invistam em novos projetos, contratem novos funcionários e comprem novos equipamentos, enquanto também permite que os indivíduos financiem compras importantes, como uma casa ou carro. Além disso, o crédito também é um fator chave para permitir que os intermediários financeiros forneçam financiamento para projetos e investimentos que podem ser muito arriscados ou custosos para os investidores individuais (VAN GESTEL, BAESENS, 2009)

No Brasil, o crédito também teve um papel importante no desenvolvimento econômico do país. Segundo Borça Júnior e Guimarães (2015) o ciclo expansionista de crédito livre à pessoa física, entre 2004 e 2013, foi responsável, na média, por 45% do crescimento do consumo das famílias e, consequentemente, em função de sua grande participação no Produto Interno Bruto (PIB), por um terço do crescimento médio da economia.

Apesar de sua importância para a economia nacional e internacional, o mercado de crédito não é isento de riscos. Um dos maiores desafios enfrentados pelos participantes do mercado é o risco de inadimplência, também conhecido como calote. Nesse sentido, a gestão de risco de crédito é crucial para garantir a estabilidade e integridade do mercado de crédito (BLUHM, 2010). Para gerenciar este risco, as instituições financeiras usam vários métodos, incluindo análise de pontuação de crédito, análise de garantia e diversificação de portfólio.

2.1.1 Risco e gestão do risco de crédito

O risco, que se configura como uma parte fundamental nos temas abordados neste trabalho, é um conceito que, no contexto econômico, pode ser dividido em quatro principais categorias, segundo JP Morgan em Riskmetrics (1996):

- Risco de Crédito: é o risco de perda financeira devido ao não cumprimento das obrigações de uma parte em uma operação;
- Risco Operacional: é o risco de perda devido a falhas humanas ou processuais na rotina da instituição;
- Risco de Liquidez: é o risco de não se conseguir vender ou comprar o ativo com relativa rapidez no preço estipulado.
- Risco de Mercado: é o risco de perda do valor do ativo devido a variação de parâmetros de mercado como taxa de juros, câmbio e inflação.

O conhecimento e estudo do conceito de risco perdura pela história da análise de probabilidades, mas somente com o surgimento das finanças modernas o estudo mais sistemático da gestão do risco ganhou força, a ponto de ser considerado “uma das mais importantes inovações do século XX” segundo Steinherr (2000).

Em uma definição mais recente, Securato (2007) diz que o risco gira em torno da ideia da incerteza dos eventos futuros e assimetria das probabilidades de ocorrência de cada evento gerador de perda.

Analisando mais a fundo o risco de crédito, que é o tipo de risco que esse trabalho se propõe a analisar através do aprendizado de máquina, Securato (2002) defende que por mais que as instituições sejam competentes quanto às suas avaliações com relação a conceder ou não crédito para um determinado cliente, o resultado de uma operação de crédito só será conhecido em seu vencimento, quando o acordo for cumprido ou não pela contraparte.

Dada a importância do gerenciamento de risco para a economia e a baixa aderência dos agentes do mercado em admitir novas metodologias para esse fim, conforme discutido por Ezeis (2019), observa-se a relevância de propor a utilização de tecnologias como o aprendizado de máquina nesse contexto. Nesse sentido, Aziz e Dowling (2018) analisam a evolução da utilização de inteligência artificial e aprendizado de máquina e como estes estão transformando a gestão do risco no mercado financeiro, trazendo benefícios tanto para o risco de crédito quanto para outros tipos de risco como de mercado e operacional.

2.1.2 Crédito estruturado

O crédito estruturado é uma classe de ativos financeiros que possuem como lastro uma série de recebíveis, ou seja, são títulos de dívida que unificam em um só produto diversos fluxos de pagamentos de origens diferentes. Esse tipo de crédito é construído através de um processo chamado de securitização. Esse processo, por sua vez, envolve o agrupamento de dívidas semelhantes, como hipotecas ou empréstimos para carros, e a emissão de títulos lastreados pelos pagamentos feitos sobre esses ativos, e se tornou uma ferramenta importante na economia global, pois permite que as instituições financeiras levantem capital e distribuam risco, aumentando a disponibilidade de crédito no mercado.

Segundo Fabozzi e Mann (2012), a securitização tem sido usada na indústria financeira desde a década de 1970, mas tornou-se mais ampla na década de 1990, quando bancos e outras instituições financeiras começaram a usá-la para emitir títulos lastreados por hipotecas e empréstimos pessoais. O aumento da securitização foi facilitado por avanços na tecnologia de computação e melhorias em modelagem financeira, o que tornou mais fácil analisar e precificar os títulos.

Uma das principais aplicações da securitização é permitir que diversas áreas da economia tenham mais oportunidades de financiamento, graças à distribuição do risco através da diversificação. O processo de securitização permite que as instituições acessem os mercados de capital e levantem fundos a partir de uma ampla gama de investidores, incluindo fundos de pensão, seguradoras e outros investidores institucionais (KENDALL, FISHMAN, 2000). Isso, por sua vez, torna mais crédito disponível para consumidores e empresas, o que ajuda a sustentar o crescimento econômico.

Apesar de suas muitas vantagens, o processo de securitização e o crédito estruturado também foram alvos de críticas por contribuir para a crise financeira de 2008 (DEKU et al.,

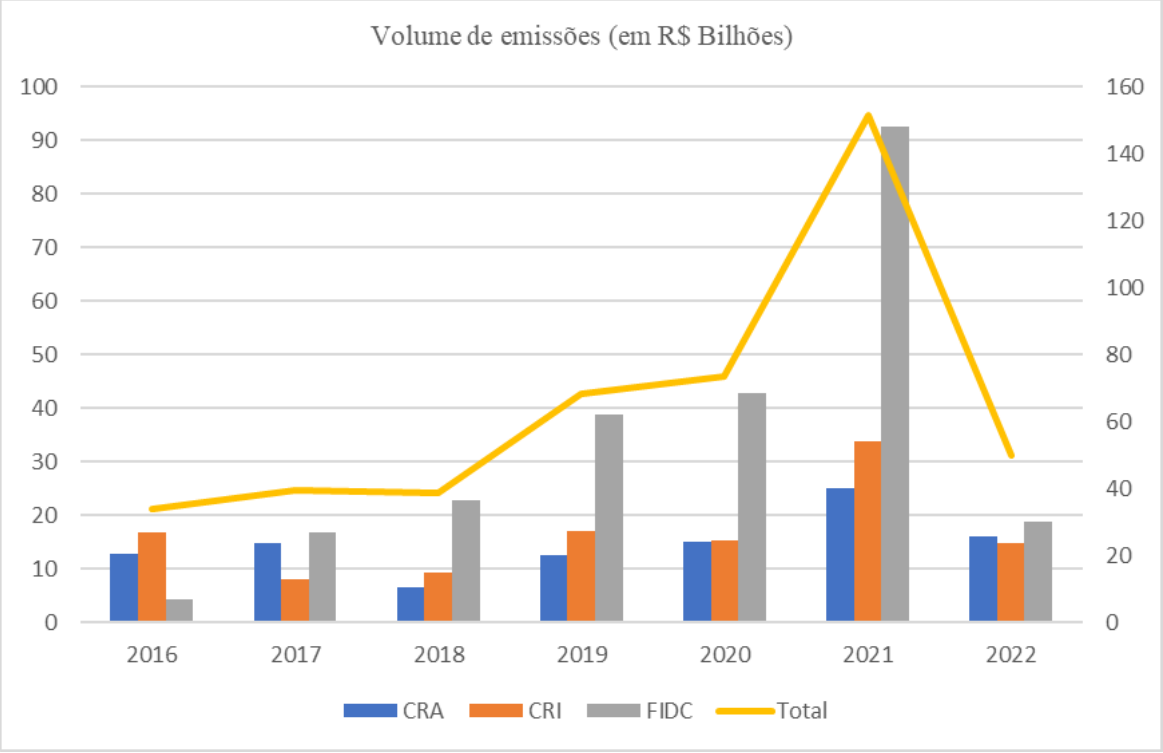
2019). Muitos dos títulos emitidos através da securitização eram baseados em hipotecas de credores que não ofereciam garantias o suficiente para honrar com suas dívidas, e isso omitia o verdadeiro risco atrelado aos produtos securitizados, que na verdade apresentavam um alto risco de inadimplência. Quando o mercado imobiliário começou a declinar, muitos desses títulos sofreram uma perda abrupta de valor, o que culminou numa grande crise financeira.

Nesse sentido, observa-se a importância de ter gerência sobre os riscos embutidos dentro de um produto de crédito estruturado, e de propor novas formas de estimar a perda esperada em veículos de securitização.

2.1.3 Principais veículos de securitização

Com a publicação da Instrução Normativa CVM nº 476 em 2009, houve uma grande simplificação e redução dos custos do processo de emissão de ativos securitizados no mercado brasileiro, e o resultado disso foi um rápido crescimento do volume de captações de CRAs, CRIs e FIDCs. A partir da publicação da CVM, o total de emissões desses produtos passou de R$ 79 bilhões até 2009 para R$ 630 bilhões até de 2022, com um acelerado crescimento nos últimos anos, como mostra o gráfico da figura abaixo:

Figura 1 – Volume de emissões de FIDCs, CRIs e CRAs (2016 – 2022) em bilhões de reais.

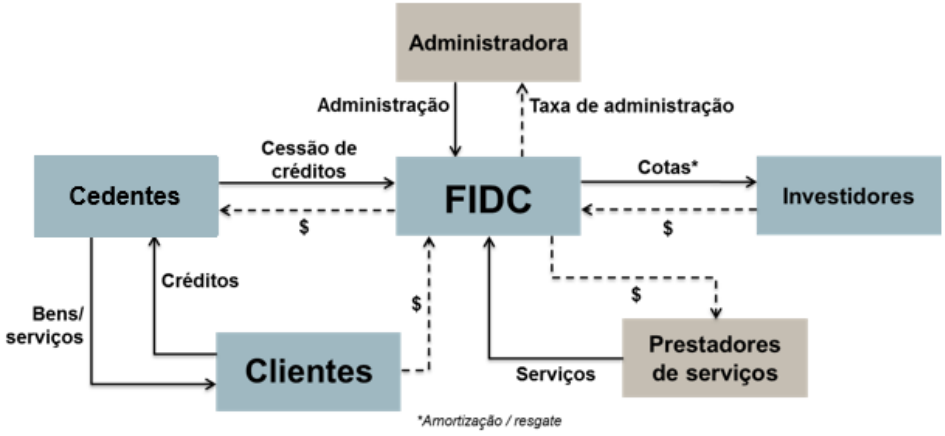


Fonte: Anbima (2023)

Dada a importância dessa classe de ativos para o mercado e o seu crescimento pujante nos últimos anos, as descrições e características dos principais veículos de securitização foram dispostas abaixo.

- FIDC: Os Fundos de Investimento em Direitos Creditórios são veículos de investimento coletivo, destinados à aplicação em direitos e títulos representativos de crédito, também denominados direitos creditórios. O objetivo do FIDC é a valorização das suas cotas e geração de ganhos para os investidores, através da aquisição desses direitos creditórios. Cada FIDC possui uma personalidade jurídica distinta, constituídos sob forma de condomínio aberto, e estão sob a responsabilidade fiduciária de um administrador, que é o responsável legal pelo fundo e pela prestação de contas aos cotistas e órgãos reguladores. Além da figura do administrador, outros dois agentes essenciais são o gestor, responsável pela alocação da carteira conforme política de investimento, e o custodiante, responsável pela custódia, controladoria e liquidação dos direitos creditórios. As principais vantagens dos FIDCs são a inexistência de impostos no nível do fundo, ou seja, não há incidência de IOF, IR ou PIS/COFINS sobre o rendimento da carteira. Segue uma figura de um fluxograma simplificado que ilustra o funcionamento de um FIDC, em que as setas pontilhadas representam fluxos de pagamento, e as seta contínuas representam bens ou serviços.

Figura 2 – Diagrama simplificado do funcionamento de um FIDC



Fonte: Associação Nacional dos Participantes em Fundos de Investimento em Direitos Creditórios Multicedentes e Multissacados (s.d.)

- CRA e CRI: os Certificados de Recebíveis Imobiliários (CRI) são ativos lastreados em direitos creditórios imobiliários. Por outro lado, os Certificados de Recebíveis do Agronegócio (CRA) são considerados equivalentes dos CRIs para recebíveis relacionados ao agronegócio. Securitizadoras são os únicos agentes que podem emitir CRIs e CRAs, instituições essas que possuem como única finalidade a aquisição de direitos creditórios e emissão dos valores mobiliários correspondentes. Conforme a Lei 9.514/97, os direitos creditórios adquiridos pela securitizadora não integram seu patrimônio e não podem ser alcançados por credores da securitizadora. Nesse caso, cada operação está vinculada a um patrimônio separado sob o qual é constituído regime fiduciário em favor dos investidores dos CRIs e CRAs, garantindo a estes acessos legais irrestritos sobre os créditos que lastreiam a respectiva emissão (BOLOGNINI, 2016). As principais vantagens dos CRAs e CRIs são a inexistência de impostos sobre as operações no nível da securitizadora, assim como nos FIDCs. Adicionalmente, há maior flexibilidade de estruturas quando comparados aos FIDCs e possibilidade de custos de captação mais baixos, em decorrência da isenção de imposto de renda nos rendimentos dos CRAs e CRIs adquiridos por investidores pessoas físicas.

2.1.4 Provisão para devedores duvidosos

Como discutido anteriormente, o PDD é um instrumento importante para que as instituições financeiras reconheçam e contabilizem perdas potenciais antes que elas sejam realmente incorridas. Em uma definição de Viceconti e Neves (2013) o PDD é um montante financeiro contabilizado como uma expectativa de perdas que uma pessoa jurídica possui em virtude da possibilidade de nem todos os devedores, havendo recebimentos a prazo, honrarem com seus compromissos de pagamento. Por se tratar de uma expectativa de perda, esse valor pode ser obtido através de diferentes metodologias.

O Banco Central do Brasil (BACEN), através da resolução nº 2682 criada em 1999, instrui que as instituições financeiras classifiquem suas operações de crédito e estabeleçam regras para a constituição do provisionamento para as liquidações duvidosas. Muito embora esse tenha sido um importante passo para a proteção do sistema financeiro, o BACEN não

impõe uma metodologia específica de PDD a ser utilizada, deixando a critério da instituição o modelo mais adequado.

No contexto das operações de crédito estruturado, os modelos adotados para provisionamento dos créditos adquiridos pelo veículo de securitização geralmente utilizam apenas o atraso atual dos contratos, conforme ilustra a figura a seguir:

Tabela 1 – Regra de provisionamento de um FIDC

Faixas de atraso	% Provisão
Risco nível A: atraso entre 3 e 15 dias:	0,50%
Risco nível B: atraso entre 15 e 30 dias:	1,00%
Risco nível C: atraso entre 31 e 60 dias:	3,00%
Risco nível D: atraso entre 61 e 90 dias:	10,00%
Risco nível E: atraso entre 91 e 120 dias:	30,00%
Risco nível F: atraso entre 121 e 150 dias:	50,00%
Risco nível G: atraso entre 151 e 180 dias	70,00%
Risco nível H: atraso superior a 180 dias	100,00%

Fonte: elaboração própria

2.2 Aprendizado de máquina

A aprendizagem de máquina é uma subárea da inteligência artificial que permite que os computadores aprendam a partir de dados e melhorem seu desempenho em uma tarefa específica ao longo do tempo. Segundo Allende-Cid (2019), ela pode ser descrita como uma área de pesquisa computacional que visa estudar um conjunto de métodos que identificam padrões em um conjunto de dados, e os utilizam para prever eventos futuros.

Em uma definição mais antiga, Konar (1999) relaciona o aprendizado de máquina ao processo humano de adquirir conhecimento, pois quando executamos tarefas semelhantes, adquirimos a capacidade de melhorar seu desempenho. Essa habilidade, quando aplicada a sistemas computacionais, é chamada de aprendizado de máquina.

De acordo com Russel e Norvig (2013), existem três tipos principais de aprendizagem de máquina: aprendizagem supervisionada, aprendizagem não supervisionada e aprendizagem por reforço. Na aprendizagem supervisionada, o computador recebe um conjunto de dados rotulados em que cada ponto de dados está associado a uma variável de destino. O objetivo é aprender um mapeamento das variáveis de entrada para a variável de saída. Na aprendizagem não supervisionada, o computador recebe um conjunto de dados não rotulados, e o objetivo é encontrar padrões ou estrutura nos dados. Na aprendizagem por reforço, o computador aprende interagindo com um ambiente e recebendo recompensas ou penalidades por suas decisões.

Um exemplo para ilustrar o processo de aprendizado de máquina, em linha com o tema de pesquisa deste trabalho, seria um conjunto de dados referentes a empréstimos dentro de uma estrutura de securitização. Cada linha na tabela 2 abaixo corresponde à informação de um crédito, com suas características ou atributos em determinado momento, que podem ser: volume financeiro, número de parcelas, quantidade de parcelas pagas, região do credor, prazo. Um desses atributos será considerado o atributo de saída ou alvo, cujos valores devem ser estimados utilizando os valores dos outros atributos, que são chamados de atributos de entrada (*features*). O objetivo de um algoritmo de aprendizado de máquina é aprender, a partir de dados, denominado conjunto de treinamento, gerar um modelo para ter hipóteses capazes de relacionar valores de atributos de entrada de um exemplo do conjunto de treinamento ao valor de seu atributo de saída (CARVALHO et al., 2011)

Tabela 2 – Exemplo de créditos

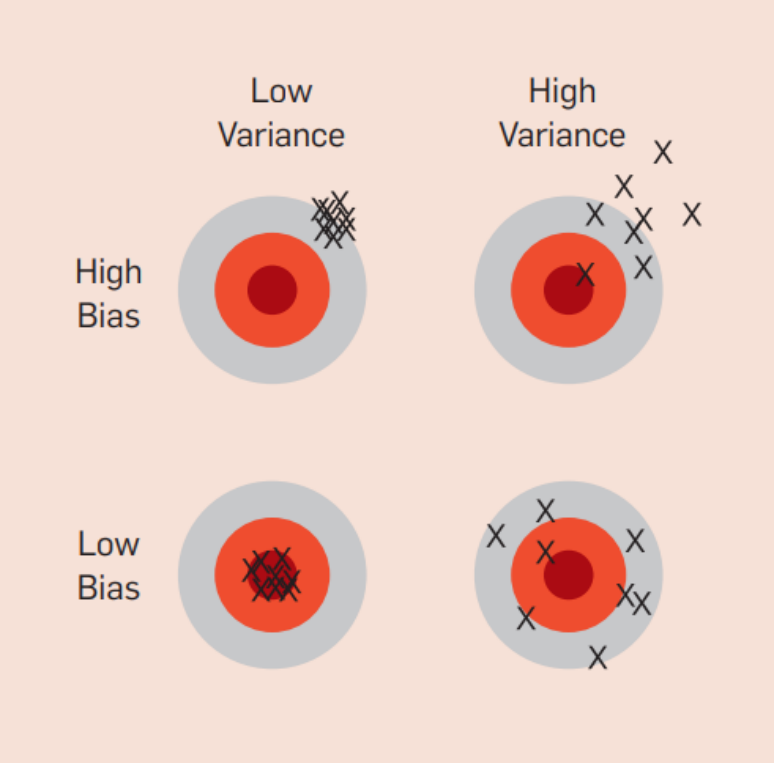
ID do Crédito	**Qtd. de Parcelas**	**Volume (R$)**	**Região**	**Calote**
1	12	1000,00	Norte	Sim
2	24	3000,000	Nordeste	Não
3	12	1800,00	Centro-oeste	Não

Fonte: elaboração própria

Além da etapa de pré-processamento, que consiste em preparar dados brutos para uso pelos algoritmos de aprendizado de máquina por meio da limpeza, transformação e criação de novas *features*, a fim de melhorar a qualidade dos dados para treino, outra importante etapa a ser aplicada é o *data splitting*. Quando os resultados gerados pelo modelo estão super

ajustados ao conjunto de dados, mostra que o modelo se especializou no conjunto de dados, mas não necessariamente é capaz de prever os atributos de saída de novos dados (*overfitting*). Em contrapartida, quando o modelo gerado pode não conseguir capturar padrões dos dados, gerando uma taxa de acurácia baixa (*underfitting*). (CARVALHO et al., 2011). O *data splitting* é o processo de dividir um conjunto de dados em dois ou mais subconjuntos separados para treinar e testar o modelo, e uma boa maneira de fazer isso é dividir os subconjuntos aleatoriamente (IZBICKI e SANTOS, 2020). O conjunto de treinamento é usado para treinar o modelo de aprendizado de máquina, enquanto o conjunto de testes é usado para avaliar o desempenho do modelo em novos dados. O conjunto de treinamento é geralmente maior do que o conjunto de testes, e a divisão comum é de 70% para treinamento e 30% para testes, mas essas proporções podem variar dependendo do tamanho e natureza do conjunto de dados.

Figura 3 – Ilustração que mostra a relação entre viés e variância



Fonte: Domingos (2012)

Outra definição importante quando se trata de aprendizado de máquina e *overfitting* são os conceitos de viés e variância. Viés é a tendência do modelo de aprendizado a aprender consistentemente a mesma coisa errada. Variância é a tendência a aprender coisas aleatórias,

independentemente do sinal real (DOMINGOS, 2012). A Figura 2 ilustra isso pela analogia clássica de dardos em um alvo.

Um modelo com baixo viés e alta variância pode se ajustar muito bem aos dados de treinamento, mas pode não ser capaz de performar bem com novos dados. Por outro lado, um modelo com alto viés e baixa variância pode não se ajustar bem aos dados de treinamento, mas pode ser mais robusto e generalizar melhor para novos dados. O objetivo é encontrar um equilíbrio entre viés e variância que permita que o modelo se ajuste bem aos dados de treinamento e generalize bem para informações novas (DOMINGOS, 2012).

Por fim, o aprendizado de máquina é uma vasta área de conhecimento e possui diversos tipos de algoritmos, que devem ser escolhidos conforme o tipo de problema para ter uma melhor eficiência nos resultados. Um dos critérios de escolha, seria se o problema se trata de um paradigma de aprendizado preditivo ou descritivo. Conforme ilustrado na figura 3, a hierarquia de aprendizado, dividida em aprendizado supervisionado, que seriam as tarefas preditivas e os não supervisionados que seriam as descritivas. (CARVALHO et al., 2011).

Figura 4 – Hierarquia do aprendizado de máquina

Fonte: Carvalho et. al. (2012)

2.2.1 Análise de desempenho dos algoritmos

Em um problema como a previsão de calote em carteiras de crédito, o modelo pode cometer dois tipos de erros: ele pode atribuir incorretamente um indivíduo que não paga como adimplente, ou pode atribuir incorretamente um indivíduo que paga como

inadimplente. É frequentemente de interesse do pesquisador determinar quais desses dois tipos de erros estão sendo cometidos. Uma matriz de confusão, mostrada para os dados de inadimplência, é uma maneira conveniente de exibir essas informações (JAMES et al., 2013). Segue uma tabela que ilustra essa matriz:

Tabela 3 – Hierarquia do aprendizado de máquina

Valor Predito	**Valor verdadeiro**	
	$Y = 0$	$Y = 1$
$Y = 0$	VN (verdadeiro negativo)	FN (falso negativo)
$Y = 1$	FP (falso positivo)	VP (verdadeiro positivo)

Fonte: Izbicki e Santos (2020)

Em que,

VN: Verdadeiro negativo, valor observado de não inadimplência foi classificado corretamente como não inadimplência;

VP: Verdadeiro positivo, valor observado de inadimplência foi classificado corretamente como inadimplência;

FN: Falso negativo, valor observado de inadimplência foi classificado incorretamente como não inadimplência;

FP: Falso positivo, valor observado de não inadimplência foi classificado incorretamente como inadimplência.

Com base na matriz de confusão, pode-se calcular algumas métricas, como por exemplo:

- Sensibilidade: dos créditos inadimplentes, quantos foram corretamente identificados.

$$Sensibilidade = \frac{VP}{VP + FN} \quad (1)$$

- Especificidade: dos créditos não inadimplentes, quantos foram corretamente classificados.

$$Especificidade = \frac{VN}{VN + FP} \quad (2)$$

- Precisão: dos créditos classificados como inadimplentes, quantos foram classificados corretamente.

$$Precisão = \frac{VP}{VP + FP} \quad (3)$$

- Estatística F1: média harmônica entre sensibilidade e precisão.

$$F1 = \frac{2 \times Precisão \times Sensibilidade}{Precisão + Sensibilidade} \quad (4)$$

No caso do problema tratado por este trabalho, a previsão de eventos de calote em carteiras de crédito, a matriz de confusão se mostra uma ferramenta eficiente para avaliar o desempenho do modelo. Por exemplo, em um cenário hipotético, após o treinamento do modelo, testa-se em um conjunto de dados de 100 créditos que já foram rotulados como calote ou não calote. A matriz de confusão para este modelo mostra que ele classificou corretamente 80 créditos como não calote e 10 créditos como calote, mas classificou erroneamente 5 créditos que sofreram calote como não calote e 5 créditos não sofreram calote como calote.

Sem usar uma matriz de confusão, pode-se concluir que o modelo tem uma precisão de 90%. No entanto, isso não conta toda a história. Ao olhar para a matriz de confusão, percebe-se que o modelo tem uma taxa alta de falsos negativos, ou seja, classifica incorretamente os créditos que sofreram calote. Isso aponta uma ineficiência do modelo em detectar eventos de calote em uma carteira de crédito, que é um justamente o objetivo almejado pelo algoritmo.

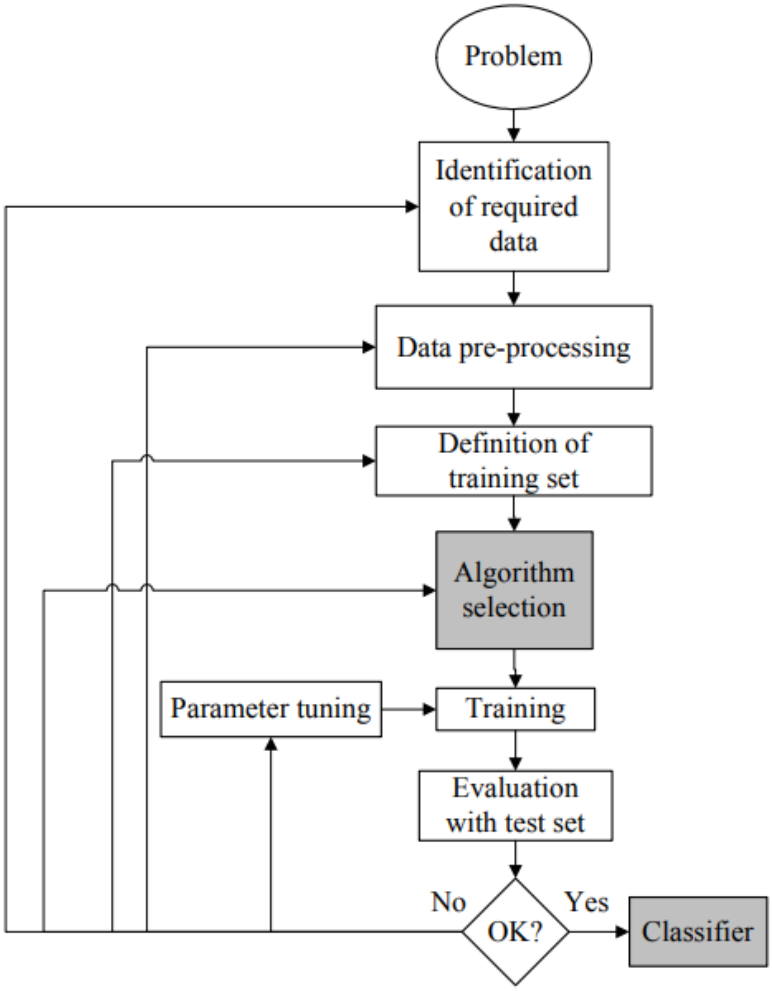
Ao usar uma matriz de confusão, é possível obter mais insights sobre o desempenho do seu modelo e identificar áreas para melhorias. No exemplo supracitado, pode-se tentar melhorar a sensibilidade do modelo incorporando mais *features* que possam distinguir entre créditos que sofreram calote ou não. Sem a matriz de confusão, este aspecto importante do desempenho do modelo poderia ser perdido, e acabar com um sistema de previsão de calote menos eficaz.

2.2.2 Aprendizado supervisionado

A aprendizagem de máquina supervisionada é o processo de aprender um conjunto de regras a partir de exemplos de um conjunto de treinamento, em outras palavras, criar um

classificador que possa ser usado para generalizar a partir de novos exemplos (KOTSIANTIS, 2007). Na aprendizagem supervisionada, o algoritmo recebe um conjunto de dados que consiste em recursos de entrada, ou chamados de variáveis independentes, e suas etiquetas correspondentes, ou chamadas de variáveis dependentes. O objetivo do algoritmo é aprender uma função que mapeia os recursos de entrada para as etiquetas de saída corretas, de modo que ele possa fazer previsões precisas em novos dados não vistos (SINGH, THAKUR e SHARMA, 2016). Segue uma figura que ilustra a aplicação do aprendizado de máquina supervisionado em um problema, desde sua identificação e aquisição dos dados necessários e pré-processamento, até a seleção e teste dos algoritmos. É pertinente notar a característica iterativa do processo, em que etapas podem ser repetidas dependendo dos resultados obtidos com os modelos.

Figura 5 – Diagrama do processo de um projeto de aprendizado de máquina



Fonte: Kotsiantis (2007)

O termo "supervisionado" refere-se ao fato de que os dados de treinamento são rotulados com os valores de saída corretos, o oposto do que acontece com a aprendizagem não supervisionada, onde os dados não são rotulados e o algoritmo é encarregado de descobrir padrões ou estrutura ocultos nos dados (JAIN et al. 1999).

A aprendizagem supervisionada pode ser usada para uma ampla gama de tarefas, incluindo classificação, regressão e previsão. A classificação é a tarefa de atribuir uma entrada a uma das várias categorias possíveis, enquanto a regressão envolve a previsão de um valor de saída numérico (CARVALHO et al., 2011).

2.2.3 Algoritmos de aprendizado supervisionado

Conforme dito anteriormente, existem diversos algoritmos de aprendizagem supervisionada, para cada tipo de problema um dos algoritmos vai se encaixar melhor e trará melhores resultados. Este subcapítulo tem o objetivo de trazer à tona a teoria por trás de alguns desses algoritmos, para que na seção 4 deste trabalho, os conceitos sejam aplicados à prática de previsão de calote em carteiras de crédito para mensuração da PDD.

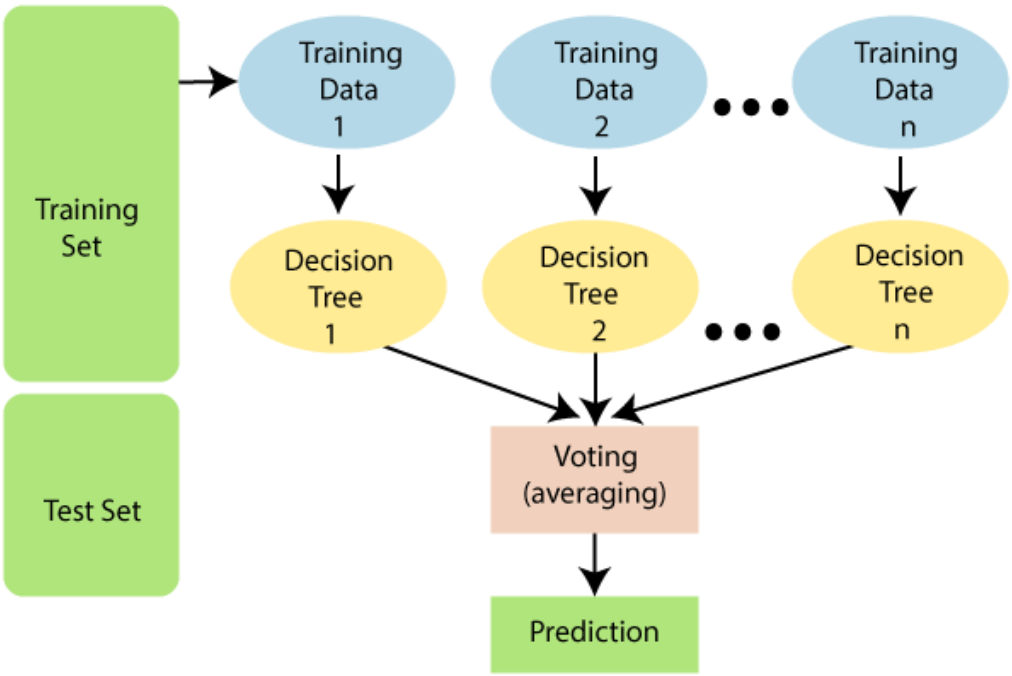
2.2.3.1 Random Forest

Árvores de decisão, ou *decision trees*, têm um aspecto que as impedem de ser a ferramenta ideal para o aprendizado preditivo, e esse aspecto é a imprecisão. O *random forest* surge como um método de aprendizado que aperfeiçoa o método de *decision trees*, pois combina múltiplas árvores para criar um modelo mais preciso e robusto. A ideia básica por trás da *random forest*, ou floresta aleatória, é criar um conjunto de árvores de decisão que são treinadas em diferentes subconjuntos dos dados e, em seguida, agregar as previsões dessas árvores para fazer uma previsão final (HASTIE, TIBSHIRANI e FRIEDMAN, 2001).

Segundo Leo Breiman (2001), criador do algoritmo, o *random forest* funciona tanto para tarefas de classificação quanto de regressão, e funciona com base no uso de diversas árvores de decisão e no conceito de *bagging* para melhorar a precisão e reduzir o *overfitting*. *Bagging* significa *bootstrap aggregating* e envolve a amostragem dos dados para criar vários subconjuntos de dados a partir do conjunto de treinamento inicial. Ao treinar cada árvore em um subconjunto diferente dos dados e votando nas decisões de cada árvore, a floresta

aleatória é capaz de melhorar a precisão do modelo. Segue uma figura que ilustra de maneira simplificada o processo.

Figura 6 – Diagrama do processo de um projeto de aprendizado de máquina



Fonte: Javatpoint (s.d.)

Além do *bagging*, outra característica relevante para a boa performance da floresta aleatória é a seleção aleatória do subconjunto de *features* em cada divisão das árvores, técnica introduzida por Ho (1995). No seu trabalho, Ho partiu de árvores de decisão e seguiu os princípios da modelagem estocástica, propondo um método para construir classificadores baseados em árvore cuja capacidade pode ser expandida arbitrariamente para aumentar a precisão tanto para treinamento quanto para dados não vistos.

2.2.3.2 Logistic Regression

A regressão logística é um algoritmo de aprendizado de máquina para problemas de classificação, e é comumente usado para prever resultados binários, com base na função logística:

$$p(x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{e^{\beta_0+\beta_1 x}}} \quad (5)$$

Assim, para qualquer valor de x na equação, o modelo de regressão logística garante que a resposta de saída seja um valor contido no intervalo de 0 a 1 (JAMES et al., 2013).

A ideia principal por trás da regressão logística é modelar a relação entre as características de entrada e a classe de saída utilizando uma função logística, que mapeia as características de entrada para um valor de probabilidade entre 0 e 1. O algoritmo então usa esse valor de probabilidade para tomar uma decisão binária sobre qual classe a entrada pertence, geralmente utilizando um *threshold* de 0,5.

Para treinar o modelo de regressão logística, utiliza-se uma abordagem de estimativa de máxima verossimilhança ou, *maximum likelihood estimation* (MLE), onde se procura encontrar os parâmetros que maximizam a função de verossimilhança:

$$l(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} \left(1 - p(x_{i'})\right) \quad (6)$$

Os coeficientes β_0 e β_1 são desconhecidos e devem ser estimados com base nos dados de treinamento. Em um cenário em que se utiliza o modelo para prever eventos de inadimplência em uma carteira, como a proposição deste trabalho, a intuição básica por trás do uso da máxima verossimilhança para ajustar um modelo de regressão logística seria: buscar estimativas para β0 e β1 tal que a probabilidade prevista de inadimplência para cada indivíduo corresponde o mais próximo possível ao indivíduo que de fato inadimpliu. Em outras palavras, tenta-se encontrar β0 e β1 tais que essas estimativas no modelo produzam, na função logística, um número próximo de um para todos os inadimplentes e um número próximo de zero para todos os indivíduos não inadimplentes. (JAMES et al., 2013).

2.2.3.3 XGBoost

XGBoost é um algoritmo de aprendizado de máquina baseado no *gradient boosting* utilizado tanto para classificação quanto para regressão, e o seu modelo é baseado em árvores de decisão agregadas através dos métodos *bagging* e *boosting*.

O método *bagging*, tal como explicado anteriormente, gera subconjuntos de árvores de decisão e as utiliza de maneira agrupada para prever um determinado resultado. As várias versões das árvores de decisão são formadas a partir de amostras *bootstrap* do conjunto de dados originais (JAMES et al. 2009).

Entretanto, mesmo utilizando diferentes amostras para construir diferentes árvores, os subconjuntos tendem a ser parecidos. Para contornar isso, somente algumas *features* do modelo são utilizadas na construção de cada árvore, tal qual uma *random forest*, pelo fato da decisão final ser tomada através da agregação de várias árvores de decisão (CHEN e GUESTRIN, 2016).

Já o método *boosting* surge no algoritmo a partir de uma primeira árvore construída com alto viés e baixa variância, ou seja, com alto grau de *overfitting*. Deste ponto em diante as novas árvores são construídas a partir de um aprimoramento da árvore anterior, à medida em que busca reduzir este grau de sobreajuste a cada nova rodada de ajuste (CHEN e GUESTRIN, 2016).

É através da combinação dessas técnicas que o XGBoost consegue operar com bilhões de exemplos de dados usando muito menos recursos do que outros modelos existentes, e que é usado amplamente por cientistas de dados para alcançar resultados de ponta em muitos desafios de aprendizado de máquina. (CHEN e GUESTRIN, 2016).

3 METODOLOGIA

Este capítulo detalha o caminho percorrido para a realização deste trabalho, apresentando as ferramentas e os procedimentos utilizados para obter os resultados.

3.1 Ferramentas utilizadas

A linguagem Python foi escolhida para a implementação do projeto pois além de ser uma linguagem com sintaxe simples e de rápido aprendizado, ela possui uma vasta quantidade de bibliotecas e frameworks que a tornam muito eficiente para trabalhar com dados. A grande vantagem e motivo da escolha do Python, no que diz respeito ao desenvolvimento deste trabalho, é a possibilidade de trabalhar com a mesma linguagem em todas as etapas do processo, como a visualização dos dados brutos, o tratamento dos dados e criação das *features*, aplicação dos algoritmos de *machine learning* e avaliação das métricas de desempenho dos modelos. As seções a seguir descrevem as bibliotecas utilizadas no projeto.

3.1.1 Pandas

Pandas é uma biblioteca poderosa e flexível para manipulação e análise de dados. Ele fornece estruturas de dados como Séries e DataFrames que permitem trabalhar com dados de forma conveniente e eficiente. A biblioteca também fornece uma ampla variedade de ferramentas para lidar e transformar dados, incluindo funções para filtrar, agrupar e agregar dados, além de lidar com valores ausentes e trabalhar com datas e horas. Além disso, ela tem uma API intuitiva e fácil de usar e está bem documentado, o que o torna uma escolha popular para cientistas de dados e analistas. (PANDAS, 2023)

3.1.2 Scikit-learn

Scikit-learn, também conhecido como sklearn, é uma biblioteca poderosa e fácil de usar de aprendizado de máquina para Python. Ele fornece uma ampla variedade de ferramentas para tarefas como classificação, regressão, agrupamento, redução de dimensionalidade e *feature engineering*. Além disso, o sklearn fornece muitas ferramentas

úteis para avaliação de modelos, incluindo validação cruzada e métricas de desempenho. (SCIKIT-LEARN, 2023)

3.1.3 XGBoost

XGBoost é uma biblioteca de código aberto para aprendizado de máquina, especificamente projetada para aplicação do modelo do extreme gradient boost. Conforme explicado anteriormente, o *extreme gradiente boosting* é uma técnica poderosa e flexível de aprendizado de máquina que pode ser usada tanto para problemas de regressão quanto para classificação. O código original foi desenvolvido em C++, porém possui interface com Python e pode ser importada via gerenciador de pacotes, contando com características como regularização, processamento paralelo, poda de árvores e tratamento de valores faltantes. (XGBOOST, 2022)

3.1.4 Imblearn

Imbalanced-learn, ou imblearn, é uma biblioteca de código aberto para lidar com conjuntos de dados desbalanceados em aprendizado de máquina. Conjuntos de dados desbalanceados são conjuntos de dados nos quais a distribuição de classes é altamente desigual, com uma classe sendo muito mais prevalente do que a outra, sendo esse um desafio particular para tarefas de predição onde deseja-se aprender sobre a classe com menor proporção, como é o caso dos calotes em carteiras de crédito. (IMBLEARN, 2022)

3.1.5 Plotly

Plotly é uma biblioteca para criar visualizações e gráficos interativos. Ele fornece uma ampla variedade de tipos de gráficos, desde gráficos básicos de linha e de dispersão até gráficos 3D e estatísticos. A utilização do Plotly é bastante útil no processo de aprendizado de máquina, principalmente para explorar os dados e para avaliar resultados. (PLOTLY, 2023)

3.2 Obtenção dos dados

Para o desenvolvimento do projeto foram utilizados arquivos de estoque e liquidações de Fundos de Investimento em Direitos Creditórios (FIDC). Esses dados foram obtidos através das administradoras dos fundos, que cumprem o papel de custodiar os títulos

financeiros e defender os interesses dos cotistas. Toda análise foi realizada com os dados de fundos os quais a empresa que eu trabalho possui cotas.

Os arquivos supracitados possuem informações sobre todos os recebíveis que estão contidos no FIDC. O estoque reflete todos os ativos pendentes de pagamento, ou seja, todas as parcelas de pagamento dos recebíveis que o fundo ainda não recolheu. Já o arquivo de liquidações descreve todos os fluxos já recebidos pelo fundo. Ambos os arquivos são retirados de um sistema web da administradora do fundo e possuem informações de cada parcela dos créditos adquiridos pelo FIDC. Segue uma tabela com descrição de cada campo deste arquivo.

Tabela 4 – Descrição dos campos dos arquivos de estoque e liquidações

Campo	**Descrição**
CedenteCnpjCpf	Documento do agente que concedeu o crédito.
CedenteNome	Nome do agente que concedeu o crédito.
NotaPDD	Uma nota de A a D que classifica o risco do crédito. Utilizado pelo administrador do FIDC para definir a provisão de cada crédito.
SacadoCnpjCpf	Documento da pessoa ou empresa que tomou o empréstimo.
SacadoNome	Nome da pessoa ou empresa que tomou o empréstimo.
DataAquisicao	Data da aquisição do crédito pelo FIDC.
DataVencimento	Data de vencimento da parcela.
NumeroTitulo	Identificador do crédito. Único para cada empréstimo
ValorAquisicao	Valor de aquisição da parcela.
ValorNominalAtual	Valor futuro da parcela, ou seja, valor que o tomador vai ter que pagar na data do vencimento
ValorPresente	Valor presente da parcela, ou seja, valor atual com o juro calculado na data de referência do arquivo.
DataPosicao	Data de referência, ou seja, a data da “foto” da carteira do FIDC.
DataLancamento	Data em que a parcela foi paga pelo sacado. Essa coluna só está presente no arquivo de liquidações.

Fonte: elaborado pelo autor

3.3 Tratamento dos dados e criação de *features*

Nesta segunda etapa os dados estão disponíveis para uso, entretanto, podem estar em diferentes formatos e padrões, precisando então realizar conversões para cada característica observada. Sendo assim, esta etapa se resume em retirar dados indevidos, tratar variáveis e garantir que elas estejam no formato correto para utilização nos modelos de aprendizado de máquina, ou seja, deixar o conjunto de dados 'limpo' para o treinamento. Os processos necessários para tratar os dados são:

- Converter identificadores de contrato para texto;
- Converter dados que são data para tipo *date*;
- Retirar linhas com valores nulos;
- Converter dados numéricos para o tipo *int* ou *float*.

Além disso, conforme exposto na seção 3.2, o conjunto de dados descreve as condições de cada parcela dos contratos, separados em dois arquivos, estoque e liquidações, que contém as parcelas a pagar e as parcelas pagas, respectivamente. Outra etapa que precisa ser aplicada aos dados, agora tratados, é unir essas bases, de forma que todas as parcelas do contrato sejam encontradas em apenas uma tabela. Esse passo se faz necessário pois, dado que a característica que este trabalho se propõe a prever é o calote de um contrato, a tabela utilizada pelo modelo precisa estar agrupada por contrato, apresentando características de cada contrato ao longo do tempo, diferente da base de dados original que descreve as parcelas.

Com a tabela agrupando os dados de cada contrato ao longo do tempo, de forma similar a tabela 2, deve-se então obter as informações derivadas das parcelas para utilizá-las como *features* de treinamento do modelo. O principal atributo que precisa ser criado é o atributo de saída, que será utilizado pelo modelo como alvo da predição, que é se o crédito sofreu um calote ou não. Outros atributos auxiliares para o treinamento também devem ser criados nessas etapas, como os *payment defaults*, que indicam se o crédito sofreu atraso de pagamento em uma das três primeiras parcelas, que podem ser um bom indicativo para maus pagadores.

3.4 Aplicação dos modelos de aprendizado supervisionado

Para aplicação dos algoritmos de aprendizado supervisionado, devem ser utilizadas as bibliotecas scikit-learn e XGboost, a primeira para os modelos *random forest* e *logistic regression*, e a segunda para o *extreme gradient boost*. Para treinar os modelos, será feito o *data splitting* com proporção de 70% dos dados para o treinamento e 30% para os testes, utilizando uma função de treinamento e teste do sklean. Para a escolha das *features* e dos parâmetros utilizados para cada modelo, serão feitos testes de precisão e sensibilidade com os resultados obtidos, e fazendo ajustes para obter os melhores resultados em cada algoritmo.

4 PRÁTICA

Neste capítulo são apresentados todos os passos práticos do desenvolvimento do projeto. Seguindo os procedimentos metodológicos descritos da seção 3, obtiveram-se os dados brutos do sistema web da administradora do FIDC analisado, e fez-se o tratamento das colunas com as funções *astype*, *to_numeric* e *to_datetime* da biblioteca Pandas. Não havia nenhum dado em branco nos arquivos analisados.

Em seguida foram agrupadas as informações das parcelas para que cada linha na tabela representasse a informação de um único contrato. Para isso, foram criadas *features* a partir das informações de cada parcela dos contratos no final de cada mês desde o início do fundo até uma data de corte (31/01/2023), que foi a data das últimas informações disponíveis da carteira do FIDC. A figura a seguir ilustra a tabela resultante.

Figura 7 – Informações dos contratos agrupadas por contrato e data de referência

	ccb	database	fpd	spd	tpd	dias_atraso_atual	nominal_aberto_perc	meses_decorridos	default3meses
0	00007720-000	2022-04-29	0	0	0	3.0	0.500000	12	0
1	00007721-000	2022-04-29	0	0	0	0.0	0.000000	12	1
2	00007724-000	2022-04-29	0	0	0	3.0	0.500000	12	0
3	00007725-000	2022-04-29	0	0	0	-211.0	0.791667	12	1
4	00007726-000	2022-04-29	0	0	0	3.0	0.500000	12	0
...	...	...	...	...	...	...	...	...	...
10072	00021737-000	2022-09-30	0	0	0	28.0	1.000000	0	0
10073	00021743-000	2022-09-30	0	0	0	32.0	1.000000	0	0
10074	00021745-000	2022-09-30	0	0	0	28.0	1.000000	0	0
10075	00021842-000	2022-09-30	0	0	0	31.0	1.000000	0	0
10076	00021844-000	2022-09-30	0	0	0	31.0	1.000000	0	0

Fonte: elaboração própria

Em que,

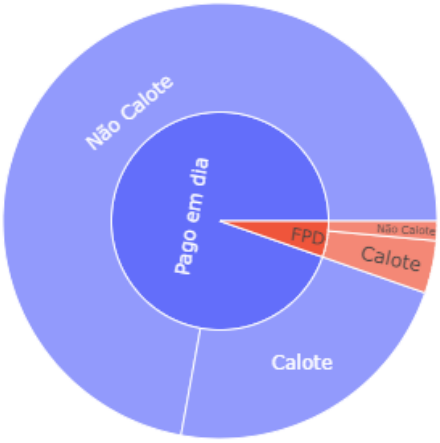
- ccb: Código identificador de cada contrato, único para cada empréstimo;
- database: Data de referência da informação, ou seja, a data que o contrato tinha aquelas características;
- fpd: *First payment default*, indica o atraso de pelo menos 30 dias no pagamento da primeira parcela do empréstimo. A intuição dessa *feature*, assim com a spd e tpd é de que se o devedor atrasar as primeiras parcelas do contrato, ele não deve ter fôlego

financeiro para arcar com empréstimo, tornando-o mais arriscado e com mais chance de calote.;

- spd: *Second payment default*, indica o atraso de pelo menos 30 dias no pagamento da segunda parcela do empréstimo;
- tpd: *Third payment default*, indica o atraso de pelo menos 30 dias no pagamento da terceira parcela do empréstimo;
- dias_atraso_atual: Número de dias de atraso da parcela mais antiga ainda não paga. Esse valor deve ser negativo para o caso de contratos com parcelas atrasadas e positivo para contratos adimplentes. A intuição dessa *feature* é que quanto mais atrasados os contratos, maior a chance de calote;
- nominal_aberto_perc: Relação percentual entre o valor total a ser pago pelo devedor e o valor que já foi pago. Exemplo: Um contrato de 10 parcelas de R$100, em que o devedor já pagou 4 parcelas, o nominal_aberto_perc será de 0,6 ou 60%. A intuição dessa *feature* é que o risco de calote é maior para contratos em que o possuem um maior volume financeiro que ainda precisa ser pago;
- meses_decorridos: Número de meses entre a originação do crédito e a data de referência. A intuição dessa *feature* é de que contratos mais antigos que ainda não foram totalmente liquidados são mais arriscados.
- default3meses: Atributo de saída, ou alvo do modelo. Indica se o crédito sofreu calote ou não, sendo “1” para calote e “0” para não calote. Neste trabalho, utilizamos a definição de calote como sendo: crédito não teve nenhum evento de pagamento em um espaço de tempo de 3 meses.

Como o atributo de saída, default3meses, precisa de três meses de informação futura para ser definido, os dados das últimas três datas com informações disponíveis do FIDC foram descartados do treinamento. Além disso, para avaliar outras *features* criadas para enriquecer o modelo, como o *first payment default*, foram realizados gráficos do tipo explosão solar para visualizar o impacto dessas características nos eventos de calote, conforme ilustra a figura 8. Percebe-se, pelo gráfico, que dos créditos que sofreram atraso na primeira parcela (fpd), a proporção de créditos que sofreram calote é muito maior do que entre os créditos que pagaram a primeira parcela em dia, o que indica que a intuição de que o *first payment default* é um bom indicador de créditos arriscados é verdadeira.

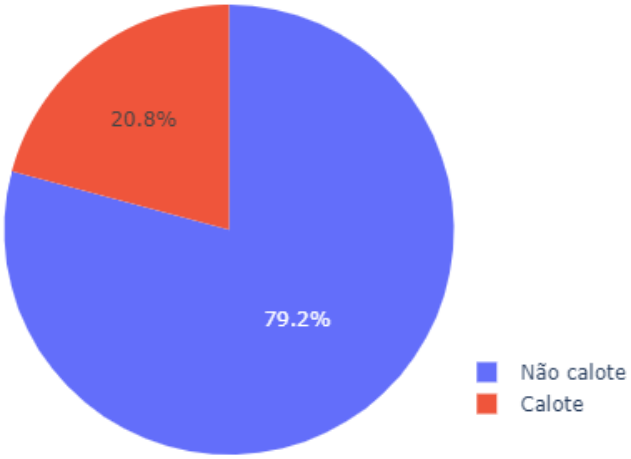
Figura 8 – Relação entre o FPD e os calotes



Fonte: elaboração própria

Um problema comum que se enfrenta em problemas de classificação como a previsão de eventos de calote em crédito é ter um conjunto de dados desbalanceado, onde a classe que se tenta prever, o calote, acontece com muito menos frequência que a outra classe, em que o crédito não sofre calote. Segue uma figura que ilustra a distribuição das classes com a base de créditos utilizada neste trabalho.

Figura 9 – Distribuição de créditos que sofreram ou não calote



Fonte: elaboração própria

Para lidar com esse desbalanceamento foi utilizado a técnica SMOTE (*Synthetic Minority Over-sampling Technique*) da biblioteca imblearn, cuja finalidade é balancear a classe com menor proporção ao criar instâncias sintéticas desta, a fim de melhorar a capacidade de previsão do modelo (MENG, ZHOU e LIU, 2020).

Outra etapa aplicada aos dados de treinamento foi a uniformização dos dados através do StantandScaler, da biblioteca sklean. Esta etapa tem como objetivo amplificar a qualidade e reduzir a ambiguidade por meio da padronização nos dados, para evitar que características com grandes variações influenciem uma métrica. (RAJU et al., 2020)

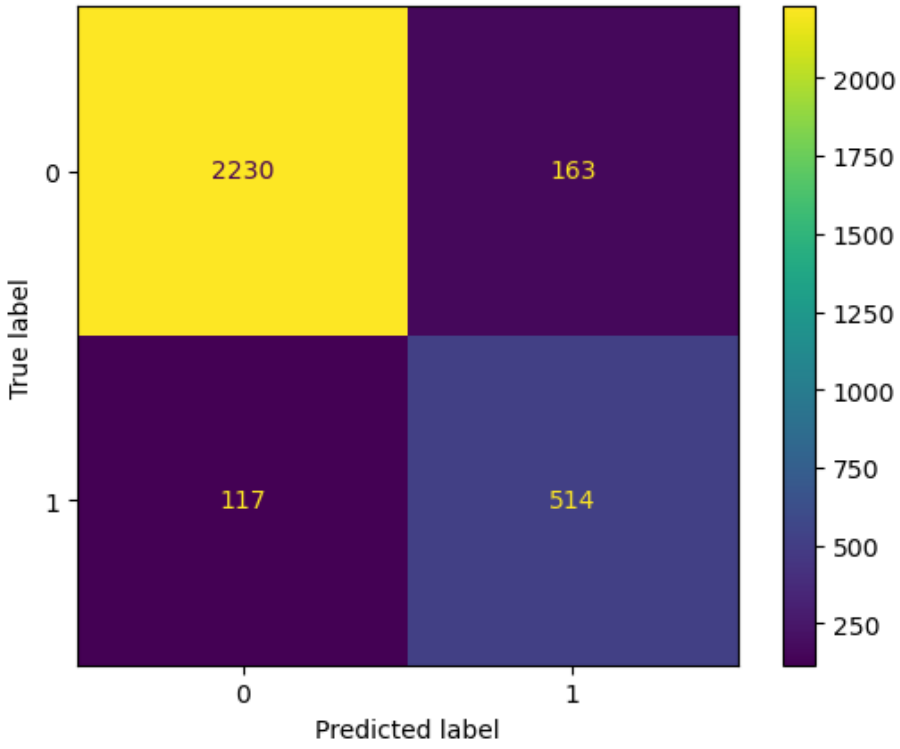
Para treinar os modelos, foi feito o *data splitting* com proporção de 70% dos dados para o treinamento e 30% para os testes, utilizando a função train_test_split do sklearn. As variáveis ccb e database foram descartadas para o treinamento, pois, apesar de ter significado prático, não trazem valor para a solução do problema. Os três modelos foram treinados e testados através das funções *fit* e *predict*, e as matrizes de confusão foram exibidas com o auxílio da função *ConfusionMatrixDisplay*, para assim serem calculadas as métricas de precisão e sensibilidade para dos modelos, conforme as fórmulas exibidas na seção 2.2.1.

5 RESULTADOS

Para os valores de saída de precisão e sensibilidade dos algoritmos, foram consideradas apenas duas casas decimais, por exemplo: 80,12%.

Com o modelo de regressão logística, obteve-se a seguinte matriz de confusão, e a partir dela foi calculada uma precisão de 95,01% e uma sensibilidade de 93,18%.

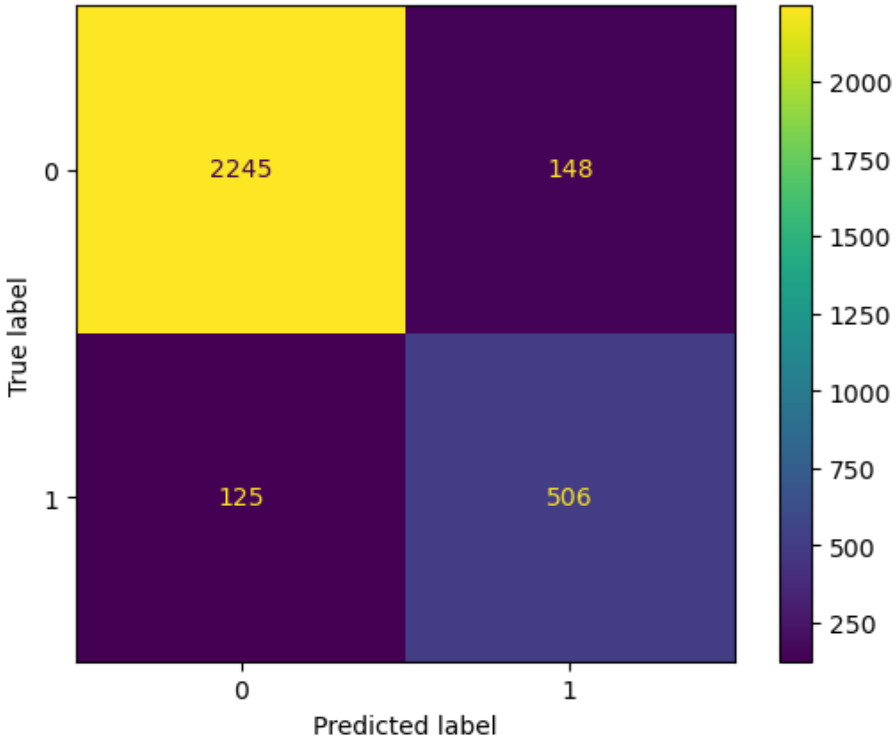
Figura 10 – Matriz de confusão obtida com a regressão logística



Fonte: elaboração própria

Já com o algoritmo random forest, obteve-se a seguinte matriz de confusão, e com isso foi calculada uma precisão 94,72% de e uma sensibilidade de 93,81%.

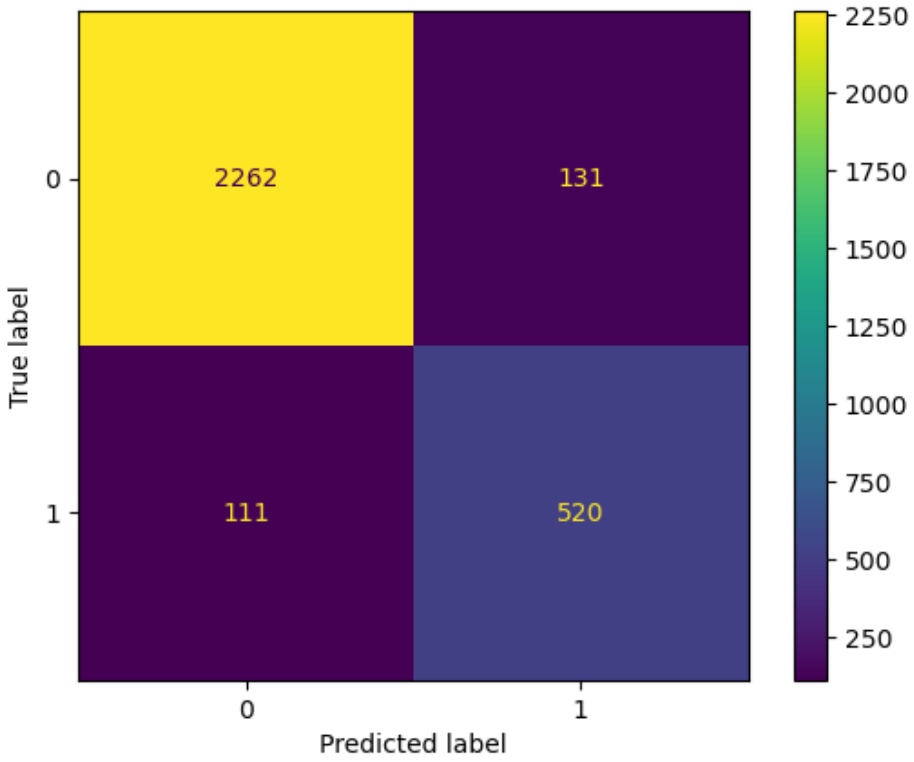
Figura 11 – Matriz de confusão obtida com a random forest



Fonte: elaboração própria

Por fim, com o XGBoost obteve-se a seguinte matriz de confusão, e com ela foi calculada uma precisão 95,32% de e uma sensibilidade de 94,52%.

Figura 12 – Matriz de confusão obtida com o xgboost



Fonte: elaboração própria

De forma geral, todos os algoritmos de aprendizagem supervisionada tiveram bons resultados, com mais de 90% de precisão e acurácia, entretanto os melhores resultados obtidos foram com o XGBoost, com o destaque para sua melhor sensibilidade, que é bastante relevante para o contexto do problema por evitar que créditos que sofreram calote sejam classificados incorretamente.

6 CONCLUSÃO

O crédito é uma ferramenta fundamental para o crescimento econômico, permitindo que pessoas e empresas tenham acesso a recursos financeiros para investimentos e consumo. No entanto, a concessão e o gerenciamento do risco de crédito precisam ser feitos de forma responsável e eficiente, garantindo que estes riscos não afetem a estabilidade do sistema econômico. Nesse sentido, a utilização de novas técnicas, como o aprendizado de máquina, pode ser extremamente benéfica, pois ao ter mais capacidade de previsão de créditos sofrerem calote, as instituições financeiras podem realizar provisões de forma mais precisa e segura.

Este trabalho teve como intuito implementar um pipeline de aprendizado de máquina, avaliar e comparar a eficácia de diferentes algoritmos supervisionados na previsão de eventos de calote no contexto de carteiras de crédito estruturadas, concluindo assim seu objetivo geral. Os objetivos específicos também foram alcançados, dado que foram expostos alguns dos principais algoritmos de aprendizado de máquina supervisionados, foram obtidos dados históricos de carteiras de crédito estruturado do mercado brasileiro e esses algoritmos foram utilizados em um pipeline e comparados os resultados na previsão de calote.

Finalmente, este trabalho contribuiu para o conhecimento de conceitos importantes relacionados ao crédito, com ênfase no crédito estruturado, e ao aprendizado de máquina, com destaque para os algoritmos supervisionados. Como trabalhos futuros sugere-se um possível aprofundamento no estudo de predição de calote ao incorporar mais informações do crédito para treinar os modelos, como localização geográfica, faturamento ou renda, e nível de endividamento. Além disso, a pesquisa pode ser expandida para outros FIDCs do mercado, como também pode utilizar outros algoritmos de aprendizado supervisionado para avaliar se os resultados são similares.

REFERÊNCIAS

ALLENDE-CID, H. Machine Learning: Catalisador da Ciência. **Machine Learning: Desafios para um Brasil competitivo**, Porto Alegre, n.39, ed.1, p.16-18, 2019.

AZIZ, S & DOWLING, M. **AI and Machine Learning for Risk Management**. SSRN Electronic Journal, 2018.

BANCO CENTRAL DO BRASIL. **Resolução nº 2.682**. Disponível em: <https://www.bcb.gov.br/pre/normativos/res/1999/pdf/res_2682_v2_L.pdf>. Acesso em: 05 fev. 2023.

BLUHM, C.; OVERBECK, L.; WAGNER, C. **Introduction to Credit Risk Modeling**. 2 ed. New York: Chapman and Hall/CRC, 2010.

BOLOGNINI, L. G. A. **Determinantes da Remuneração do Spread de Certificados de Recebíveis Imobiliários no Mercado Brasileiro**. Dissertação de mestrado Fundação Getúlio Vargas, 2016.

BORÇA JUNIOR, G. R.; GUIMARÃES, D. Impacto do ciclo expansionista de crédito à pessoa física no desempenho da economia brasileira 2004-2013. **Revista do BNDES**, São Paulo, v. 43, p. 119-159, jun. 2015.

CARVALHO, A et. al. **Inteligência Artificial**: Uma abordagem de Aprendizado de Máquina. Rio de Janeiro: Ltc, 2011.

CHEN, T.; GUESTRIN, C. XGBoost: a scalable tree boosting system. In: PROCEEDINGS OF THE 22nd ACM SIGKDD INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA, 22, 2016, San Francisco. **Anais** [...]. New York, Association for Computing Machinery, 2016.

CHEN, Z.; WU, Z.; YE, W.; WU, S. An Artificial Neural Network-Based Intelligent Prediction Model for Financial Credit Default Behaviors. **Journal of Circuits, Systems and Computers**, jan. 2023.

CVM. **Instrução CVM 489**. Disponível em: <https://conteudo.cvm.gov.br/legislacao/instrucoes/inst489.html>. Acesso em: 13 fev. 2023.

DEKU, S. Y.; KARA, A.; ZHOU, Y. Securitization, bank behaviour and financial stability: A systematic review of the recent empirical literature. **International Review of Financial Analysis**, United States, v. 61, p. 245-254, jan. 2019.

DOMINGOS, Pedro. A few useful things to know about machine learning. **Communications of the ACM**, v. 55, n. 10, p. 78-87, 2012.

EREIZ, Z. Predicting Default Loans Using Machine Learning (OptiML). **Telecommunications Forum (TELFOR)**, Belgrade, Serbia, v. 27, p. 1-4, 2019.

FABOZZI, F. J.; MANN, S. V. **The handbook of fixed income securities**. New York: McGraw-Hill Education, 2012.

GESTEL, T, V; BAESENS, B. **Credit Risk Management**: Basic Concepts: Financial Risk, Components, Rating Analysis, Models, Economic and Regulatory Capital. New York: Oxford University Press, 2008.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning**. New York: Springer, 2001.

IMBLEARN: Toolbox for imbalanced dataset in machine learning. [S. l.], 2022. Disponível em: https://imbalanced-learn.org/stable/. Acesso em: 15 fev. 2023.

HO, T.K. Random decision forests. **Proceedings of 3rd International Conference on Document Analysis and Recognition**, Montreal, v.1, p. 278-282, 1995.

HULL, J. **Risk Management and Financial Institutions**. New Jersey: Wiley, 2006.

IZBICKI, R.; SANTOS, T. M. dos. **Aprendizado de máquina**: uma abordagem estatística. 1ª ed. 2020.

JAIN, A. K.; MURTY, M. N.; FLYNN, P. Data clustering: A review. **ACM Computing Surveys**, New York, v.31, n.3, p. 264–323, 1999.

JAMES, G.; WITTEN, D.; HASTIE T.; TIBSHIRANI, R. **An Introduction to Statistical Learning**: with Applications in R. New York: Springer, 2013.

JAVATPOINT. Random Forest Algorithm. [s.d.]. Disponível em <https://www.javatpoint.com/machine-learning-random-forest-algorithm>. Acesso em: 15 fev. 2023.

KENDALL, L. T.; FISHMAN, M. J. **A primer on securitization**. Massachusetts: MIT press, 1996.

KONAR, A. **Artificial Intelligence and Soft Computing**: Behavioral and Cognitive Modeling of the Human Brain. [S. l.]: Crc Press, 1999.

KOTSIANTIS, S. B. **Supervised Machine Learning**: A Review of Classification Techniques. [S. l.]: IOS Press, 2007.

LIM, K. Monetary transformation: revisiting the end of the Bretton Woods order. **Globalizations**, Londres, Inglaterra, v. 19, n. 7, p. 989-1012, jan. 2022.

MENG, C.; ZHOU, L; LIU, B. A case study in credit fraud detection with SMOTE and XGboost. **Journal of Physics: Conference Series**, United Kingdom, v. 1601, n. 5, ago. 2020.

MORGAN, J. P. **RiskMetrics**: Technical document. 4 ed. New York Morgan Guaranty Trust Company of New York. 1996.

PANDAS: Python Data Analysis Library. [S. l.], 2023. Disponível em: https://pandas.pydata.org/index.html. Acesso em: 15 fev. 2023.

PLOTLY: Interactive graphing library for Python. [S. l.], 2023. Disponível em: https:/plotly.com/. Acesso em: 15 fev. 2023.

RAJU, V. N. G.; LAKSHMI, K. P.; JAIN, V. M.; KALIDINDI, A.; PADMA, V. Study the Influence of Normalization/Transformation process on the Accuracy of Supervised Classification. **Third International Conference on Smart Systems and Inventive Technology (ICSSIT)**, Tirunelveli, India, 2020, p. 729-735, 2020.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial.** 3. ed. Rio de Janeiro: Elsevier, 2013.

SCIKIT-LEARN: Machine Learning in Python. [S. l.], 2023. Disponível em: https://scikitlearn.org/. Acesso em: 15 fev. 2023.

SECURATO, J. R. **Crédito**: análise e avaliação do risco - pessoas físicas e jurídicas São Paulo: Saint Paul Editora, 2002.

SECURATO, J. R. **Decisões Financeiras em Condições de Risco**. 2 ed. São Paulo: Saint Paul Editora, 2007.

SHARMA, A. K.; LI, L. H.; AHMAD, R. Default Risk Prediction Using Random Forest and XGBoosting Classifier. **2021 International Conference on Security and Information Technologies with AI, Internet Computing and Big-data Applications**, 2022.

SINGH, A.; THAKUR, N.; SHARMA, A. **A review of supervised machine learning algorithms**. India: INDIACom, 2016.

STEINHERR, A. **Derivatives**: The Wild Beasty of Finance. 1st ed. Chichester: John Wiley & Sons, 2000.

TIROLE, J. **The theory of corporate finance**. Princeton: Princeton University Press, 2006.

VICECONTI, P.; DAS NEVES, S. **Contabilidade Avançada e Análise de Demonstrações Financeiras**. São Paulo: Saraiva, 2013.

XGBOOST: XGBoost Library for Python. [S. l.], 2022. Disponível em: https://xgboost.readthedocs.io/en/stable/index.html. Acesso em: 15 fev. 2023.