UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uso de *machine learning* na avaliação de risco para aumento do limite de crédito a partir do histórico de compras acima do limite original

Hian Clisman de Medeiros Costa

Trabalho de Conclusão de Curso
MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Uso de *machine learning* na avaliação de risco para aumento do limite de crédito a partir do histórico de compras acima do limite original

Hian Clisman de Medeiros Costa

USP - São Carlos

2023

Hian Clisman de Medeiros Costa

Uso de *machine learning* na avaliação de risco para aumento do limite de crédito a partir do histórico de compras acima do limite original

Trabalho de conclusão de curso apresentado ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Ricardo Rodrigues Ciferri

USP - São Carlos

2023

Esta página deve conter a ficha catalográfica e deve ser impressa no verso da folha de rosto.
Para elaborar, acesse o endereço:
https://www.icmc.usp.br/institucional/estrutura-administrativa/biblioteca/servicos/ficha
ou procure um bibliotecário na Seção de Atendimento ao Usuário da Biblioteca do ICMC

RESUMO

COSTA, H. C. **Uso de *machine learning* na avaliação de risco para aumento do limite de crédito a partir do histórico de compras acima do limite original**. 2023. 64 f. Trabalho de conclusão de curso (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

A necessidade de modelos automatizados para prever a capacidade de pagamento dos clientes e a importância de características diversas para uma avaliação sólida levou a transformação das análises de risco de crédito, indo além do tradicional score de crédito e incorporando abordagens inovadoras como o score de comportamento, especialmente relevante no setor varejista. O objetivo central dessa pesquisa é desenvolver um modelo de aprendizado de máquina capaz de aumentar os limites de crédito conforme as necessidades individuais dos clientes. Para atingir esse propósito, objetivos específicos são delineados, incluindo a análise de modelos de score de crédito, a seleção criteriosa de dados de transações excedentes ao limite e a identificação de atributos pertinentes para caracterizar essas transações. O resultado final desejado é um modelo que avalie o risco de elevar os limites de crédito a fim de alinhar as necessidades atuais do consumidor. Assim, foi realizada uma avaliação reunindo métodos de geração de score de crédito de diferentes pesquisas e elencados os melhores resultados tendo algoritmos ensemble como *Random Forest* e *XGBoost* como um dos promissores nessa área. Um fator determinante para alcançar o objetivo dessa pesquisa foi o processo de manipulação e seleção de *features*, impactando no tipo de resposta que mais se aproxime ao que foi proposto. Como resultado, os dois modelos obtiveram treinos com avaliações e métricas satisfatórias, tendo destaque para o *Random Forest*.

Palavras-chave: *Machine Learning*, crédito, varejo, algoritmos *ensemble*, *feature engineering*, *xgboost*, *Random forest*.

ABSTRACT

COSTA, H. C. **Utilization of machine learning in risk assessment for credit limit increase based on the history of purchases exceeding original limit**. 2023. 64 f. Monograph (MBA in Artificial Intelligence and Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

The necessity for automated models to predict customers' repayment capacity and the significance of diverse characteristics for robust assessment has led to the transformation of credit risk analysis, extending beyond the conventional credit score and integrating innovative approaches like behavioral scoring, particularly relevant in the retail sector. The primary objective of this research is to develop a machine learning model capable of increasing credit limits according to individual customer needs. To achieve this aim, specific objectives are delineated, including the analysis of credit score models, the judicious selection of transactions data exceeding limits, and the identification of relevant attributes to characterize these transactions. The ultimate desired outcome is a model that assesses the risk of raising credit limits to align with current consumer requirements. Thus, an evaluation was conducted, pooling credit score generation methods from various studies, and the best outcomes were identified, with ensemble algorithms such as Random Forest and Gradient Boosting emerging as promising in this field. A pivotal factor in achieving the research goal was the manipulation and feature selection process, impacting the type of response that closely aligns with the proposed objective. Consequently, both models underwent training with satisfactory assessments and metrics, with Random Forest standing out.

Keywords: Machine Learning, credit, retail, ensemble algorithms, feature engineering, XGBoost, Random Forest.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Neste capítulo será apresentada a evolução das análises de risco de crédito no mercado. De que forma esse score vem se expandindo para além do mercado financeiro surgindo novos scores complementares, tão importantes quanto score de crédito, como o score de comportamento muito presente no Varejo. Assim, será descrita a motivação para a geração de um modelo de aprendizado de máquina unindo features de crédito e comportamento para evolução do processo de concessão de crédito da maior varejista da Amazônia Ocidental: Bemol. Com isso, o objetivo geral será traçado a fim de responder o problema de pesquisa definido seguindo os objetivos específicos. E, por fim, será estruturado o corpo deste trabalho de conclusão de curso do MBA de IA e BIG DATA..

1.1 Contexto

Um dos desafios da relação de credores com seus clientes é a capacidade de determinar de forma eficiente e assertiva a necessidade de crédito versus a capacidade de pagamento individual. Tradicionalmente essa análise da capacidade individual de recebimento dentro do prazo estipulado, risco de crédito, era atribuído apenas ao score de crédito, valor apontado para cada cliente medindo sua capacidade de pagamento.(RODRIGUES, 2021).

Em um novo formato de avaliação de crédito, com a evolução de técnicas de análise e previsão de dados, os credores estão quebrando essa tradicionalidade e adicionando alternativas ao score de crédito para se manterem competitivos (FREAS, 2018). A exemplo disso temos a utilização do score de comportamento (*Behavior Score*) utilizado bastante por credores não bancários, como o varejo, os quais avaliam, além da capacidade de pagamento através de dados transacionais, a relação entre cliente e credor.

Tendo em vista a crucial importância da avaliação de risco e se manter atual no mercado de concessão de crédito, faz-se necessário a utilização de modelos automatizados que consigam prever, de forma assertiva e constante, a capacidade de pagamento de cada cliente. E, para uma avaliação robusta e com acurácia, é muito importante o levantamento de *features* (entradas do modelo) que consigam englobar diversos atributos relacionados aos clientes.(TRIVEDI, 2020)
.

Manter alinhado o limite de crédito concedido às necessidades atuais de cada cliente é uma das necessidades do varejo. Dia a dia solicitações de aumento de crédito chegam através de diversos canais digitais e atender essa demanda não é simples pois é necessário unificar essas solicitações, identificar essa necessidade e conseguir avaliar de forma personalizada para cada cliente.

1.2 Motivação

A Bemol, maior varejista da região norte, disponibiliza produtos para todo o Brasil com maior força de atuação na região da Amazônia Ocidental (Amazonas, Acre, Rondônia e Roraima). A empresa tem como primeiro princípio a integridade, possuindo um código de ética que busca fortalecer a prestação de serviços e a boa conduta empresarial perante a sociedade. Assim, o principal motivador para esse projeto de pesquisa é reforçar a boa relação e a fidelidade dos clientes bemol, recebendo e atendendo suas necessidades de crédito, mantendo esse diferencial importante em relação ao mercado.

Apesar de cada cliente ter um crediário com limite máximo fixo para compras, existe a possibilidade de eles conseguirem ultrapassar esse limite inicial. Quando essa situação acontece é necessário aguardar duas formas de aprovação, dependendo do quão acima for o valor da compra, podendo passar por aprovação da gerência local na loja onde se está realizando a compra ou, em caso de valores maiores, ir para uma análise de um time de crédito dentro do escritório central.

A falta de um modelo inteligente automatizado para avaliar o risco de crédito para ajustar o limite inicial às necessidades atuais do cliente é algo presente no cenário da empresa. Apesar de existir um modelo robusto de regras que funcione bem de forma conservadora, ele acaba por não ser tão assertivo na avaliação individual de cada cliente e possuir um processo de manutenção ainda mais dificultoso de se promover qualquer alteração rápida.

1.3 Objetivos

O objetivo geral desse projeto de pesquisa é atender necessidades de aumento de limite de crédito através de um modelo de aprendizado de máquina. De que forma seria possível alinhar o limite de crédito concedido às necessidades atuais de cada cliente? Para alcançar esse objetivo e responder essa pergunta será necessário atingir os seguintes objetivos específicos:

1.3.1 Objetivos Específicos

- Analisar modelos que são comumente utilizados para geração de scores de crédito;
- Selecionar dados de transações de compras realizadas acima do limite;
- Identificar atributos que avaliem e caracterizem compras realizadas acima do limite.

Ao fim, o modelo deverá informar o risco de ser aumentado um nível acima do limite atual de clientes que fizeram solicitações de aumento de crédito.

1.4 Organização

Além deste capítulo introdutório esse projeto de pesquisa seguirá com uma fundamentação teórica, onde será descrito os conceitos fundamentais para entendimento e prosseguimento com este trabalho. Em adicional, terá uma seção de trabalhos relacionados, onde será feita uma análise de modelos que vêm sendo mais utilizados em outras pesquisas com a mesma finalidade.

Após mapeamento do que vêm sendo aplicado será iniciado um capítulo de metodologia, onde será abordado o modelo utilizado assim como atributos selecionados e etapas para atingir cada objetivo específico. E assim, finalizando com resultados e considerações finais, onde será avaliado a efetividade do modelo e os resultados obtidos para atingimento do objetivo geral.

2 FUNDAMETAÇÃO TEÓRICA

Para aprofundar o entendimento desta pesquisa será descrito neste capítulo como funciona o processo de concessão de crédito e como ele está ligado ao mercado varejista. Com base na literatura será definido o conceito de *Credit Score* e *Behavior Score* e a evolução da avaliação de riscos utilizando atributos complementares para enriquecimento de modelos de crédito.

2.1 Crédito e Risco de Crédito

Crédito refere-se à concessão de um bem ou valor presente com a expectativa de pagamento futuro. O credor, fornecedor desse crédito, utiliza de sua carteira de crédito, valor total que ele está disposto a financiar, como uma ferramenta para potencializar as receitas fornecendo mais poder de compra aos clientes. Em contraponto, o risco de crédito mede a incerteza de que essa concessão de crédito possa resultar em perdas para o credor, o risco associado aos aumentos de crédito.

Adicionalmente existe o chamado risco de perda, indicador que avalia a saúde de uma carteira de crédito. Nesse caso é avaliado o quão provável o limite de crédito disponibilizado pelo credor tem de não ser recuperado, o qual já deve ser previsto para diversificação da carteira com taxas de juros diferenciadas para amenização de perdas.

Existem várias etapas a serem seguidas até ser efetivamente concedido determinado crédito. Segundo Rodrigues (2021), comumente esse processo é iniciado com a coleta de informações que podem variar de acordo com o tipo de crédito a ser oferecido e a natureza do credor. A partir dessas informações é avaliado o risco de crédito, o qual é responsável por definir taxas de juros para equiparar possíveis perdas, e capacidade de pagamento, a qual pode ser utilizada para definir limites de crédito.

Definido e liberado o crédito para sua base de clientes vem então a etapa operacional, a qual embarca todo seguimento transacional de liberação do crédito do credor ao cliente final. Nesta etapa também existem riscos a serem considerados, como casos de fraude que podem impactar diretamente em perdas para carteira.

Por fim, todo esse processo deve ser monitorado através de dados para avaliação constante da saúde do produto. O principal indicador de saúde de uma carteira de crédito é a inadimplência. A medida adotada pelo Serasa Experian e seguida por financeiras é de até 60 dias para quitação dos compromissos financeiros, após isso os valores financiados passam ser

considerados inadimplentes, mas esse tempo pode variar de acordo com cada credor. (SERASA, 2022).

Com mais dados e informações reais de recebimentos e perdas entra-se em um ciclo de reavaliação de toda essa cadeia de concessão de crédito.

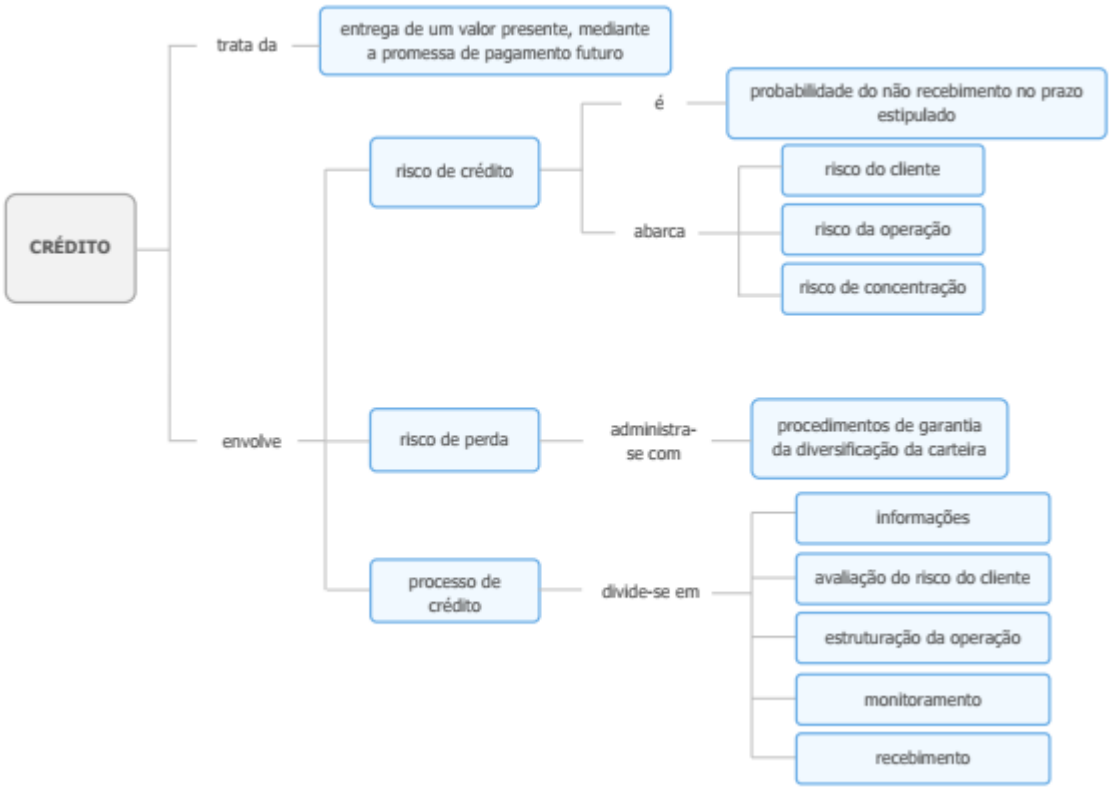

Figura 1. Diagrama da cadeia de concessão de crédito (Confederação Nacional das Instituições Financeiras, 2021)

2.2 Concessão de crédito no varejo

A concessão de crédito dentro do varejo é motivada pela necessidade de estimular vendas (ROSS; WESTERFIELD;JORDAN, 2022). As transações de crédito comumente eram feitas a partir dos programas em parceria entre instituições financeiras, as quais assumiam vários formatos: presença de financeiras dentro das lojas assumindo todo o processo de crédito; sociedades entre os varejistas para concessão de crédito; e até mesmo instituições financeiras criadas especificamente para atender ao varejo (FREIRE, 2009).

Segundo Crespi (2014), muitas varejistas caminharam na captação de investimentos para um crediário próprio, uma das mais importantes ferramentas de financiamento no varejo brasileiro, onde clientes mesmo não bancarizados passam a ter a capacidade de realizar compras a prazo. Nesse sistema a varejista se comporta como uma rede financeira e passa a arcar com

todo o risco de concessão de crédito. Assim, englobando a operacionalidade da disponibilização, avaliação e monitoramento de sua carteira de crédito.

2.3 Scores de crédito

Ao incorporar o sistema de crediário próprio a varejista passa a ter que reunir informações para avaliação do risco de seus clientes. Para fugir dos processos tradicionais de grandes financeiras na aquisição dessas informações, processo esse bem burocrático e que dificultaria a captação de clientes, a liberação inicial de crédito costuma se dar por um conjunto de informações cadastrais básicas fornecidas e complementadas pelo chamado Crédito Bureau.

O crédito bureau é um relatório completo sobre consumidores, com dados cadastrais, de comportamento e de relacionamento com o mercado. Atualmente no Brasil existem quatro principais birôs de crédito: Serasa, SPC Brasil, Boa Vista Serviços e Quod. Contudo, de acordo com Schneider (2021), apesar de fornecerem dados ricos para se pressupor um risco de crédito, a forma como o cliente irá se comportar com a varejista pode ser bem diferente do seu comportamento perante o mercado.

Com a descentralização da informação, se torna mais importante a inteligência e o uso da tecnologia para complementar essa concessão de crédito. Com isso as varejistas precisam trabalhar as informações do próprio crediário e desenvolver um score próprio especialmente para seu nicho de mercado (SCHNEIDER, 2021).

Desta forma, conforme o incremento de informações financeiras ao cadastro do cliente, através uso de algoritmos, ou modelos estatísticos, capazes de guiar decisões de crédito, a varejista passa a criar seu próprio score de crédito (*Credit Score*). Uma ferramenta de classificação, em escala numérica, do conjunto de informações históricas consideradas para a análise de crédito, definindo então uma previsão de risco mais direcionada.

Em adicional a esses scores cadastrais o mercado desenvolveu um novo formato de score, o score comportamental (*Behavior Score*). O que passa a ser avaliado é o comportamento do cliente ao longo do tempo avaliando todo dado complementar que reflita a relação entre o cliente e o credor: periodicidade de pagamento, tempo de cadastro, montante gasto em produtos, frequência de compras etc.

Por serem scores baseados na experiência da instituição, os modelos de *credit score* e *behavior score* apresentam resultados consistentes, facilidade de uso e maior eficiência do processo, características que refletem bem o perfil do consumidor no varejo (CRESPI, 2014).

2.4. Feture Engineering

Feature Engineering é um processo importante no desenvolvimento de modelos de aprendizado de máquina, pois envolve a geração de variáveis necessárias para análise ou previsão. De acordo com Forecast Global (2023) é o processo de selecionar, transformar, extrair, combinar e manipular dados brutos para gerar as variáveis desejadas para análise ou modelagem preditiva.

O objetivo de todo modelo de aprendizado de máquina é prever o valor de uma variável alvo (*target*) usando um conjunto de variáveis preditoras. Assim, com o processo de *Feature Engineering*, a construção de novas variáveis pensadas e manipuladas visando o *target* a ser alcançado se torna importante para a melhoria do desempenho do modelo de aprendizado de máquina (PATEL, 2021).

2.4.1 Seleção de Atributos (*Features)*

Com a evolução das técnicas estatísticas e dos avanços nos meios de processamento e aprendizado de máquina, os métodos numéricos baseados em critérios objetivos passaram a conquistar gradualmente espaço na análise de crédito (CRESPI, 2014). Com o uso de dados e sistemas estatísticos os credores conseguem diminuir custos e ampliar a oferta de crédito ao consumidor e, quanto maior o número de dados disponíveis, maior a capacidade de um modelo ajustado capaz de prever perfis de crédito distintos.

Para iniciar o processo de aprendizado de máquina, utiliza-se um conjunto de dados ao qual se quer obter conclusões, previsões ou reconhecer padrões. De acordo com Correia (2019) um atributo é representado por uma coluna pertencente a esse conjunto de dados onde é esperado extrair um novo conhecimento.

	Atributos				
Exemplos	X_1	X_2	...	X_M	Classe (Y)
E_1	x_{11}	x_{12}	...	x_{1M}	y_1
E_2	x_{21}	x_{22}	...	x_{2M}	y_2
E_3	x_{31}	x_{32}	...	x_{3M}	y_3
⋮	⋮	⋮	⋱	⋮	⋮
E_N	x_{N1}	x_{N2}	...	x_{NM}	y_N

Figura 2. Formato de um conjunto de dados comporto por atributos (LEE, 2005)

De Acordo com Lee (2005), a seleção de atributos tem como objetivo ordená-los de acordo com sua importância, reduzir a dimensionalidade do espaço de busca e remover dados ruidosos. Além disso, quantidade muito grandes de atributos pode exigir muito processamento tornando o processo caro, uma a seleção mais cuidadosa, ou até mesmo unificando um ou mais atributos, pode permitir a escolha de um subconjunto menor, mas ainda representativo, do conjunto original.

3 Trabalhos relacionados

O ponto de partida para a resolução do problema dessa pesquisa será dado a partir da análise de trabalhos com temas relacionados. Neste capítulo será discorrido sobre como foram encontradas e selecionadas pesquisas semelhantes ao tema proposto e, a partir de suas conclusões, levantar qual seria o melhor caminho para atingir o objetivo final.

3.1 Estratégia e expressão de busca

Inicialmente foram utilizados dois repositórios digitais para filtrar trabalhos relacionados, IEEE e Google Scholar. A fim de fazer se identificar publicações acadêmicas que se enquadram a proposta dessa pesquisa, assim como seus objetivos específicos, foi definada a seguinte *string* de busca: "*credit*"& "*score*" & ("*features*" OR "*feature selection*") & ("models" OR "*machine learning*") & "*grant*".

Assim, foi filtrado apenas um artigo relacionado pela IEEE. Quanto ao Google Scholar, mesmo com o filtro, foram retornadas várias páginas de artigos acadêmicos. Foi selecionado apenas artigos da primeira página e, pela resumo e objetivo geral, realizado mais um filtro que melhor se encaixaria ao objetivo desta pesquisa.

3.2 Trabalhos analisados

3.2.1. Statistical and machine learning models in credit scoring: A systematic literature survey

No artigo *Statistical and machine learning models in credit scoring: A systematic literature survey* de Dastile et al. (2020), é empregado uma abordagem sistemática de modelos estatísticos e de aprendizado de máquina para geração de um score de crédito, propondo ao final um guia de aplicação de *machine learning* nesse ramo para direções emergentes (Figura 3).

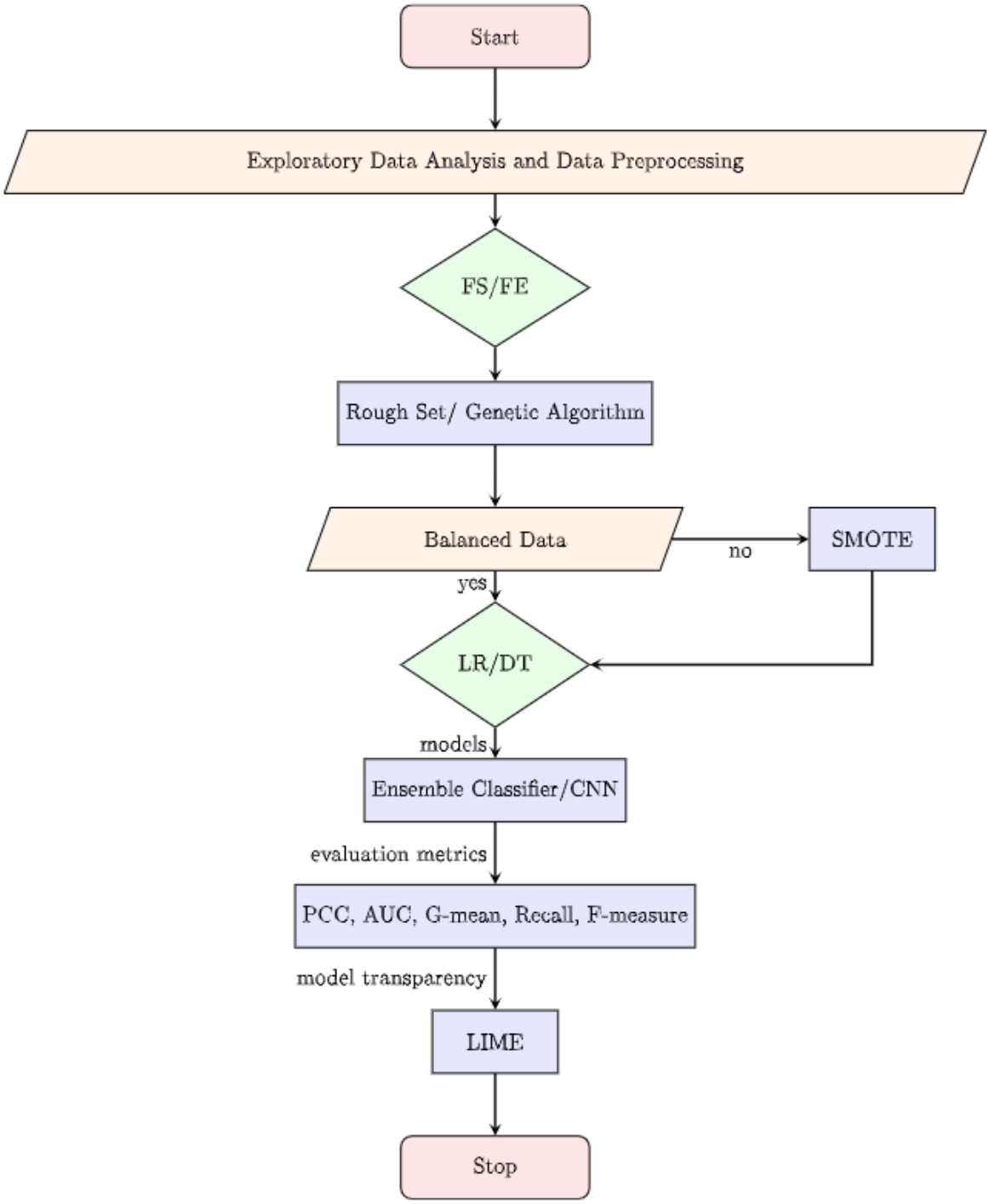

Figura 3. Guia para geração de score de crédito comumente utilizado em projetos de pesquisa. Dastile et al. (2020)

O estudo apresenta uma revisão das técnicas de seleção de *features* e engenharia de *features* mais comuns utilizadas na avaliação de crédito. O texto discute estudos que mostram que a remoção de *features* redundantes pode melhorar o desempenho do modelo na avaliação do score de crédito. Também são apresentadas as diferentes categorias de métodos de seleção

de features e engenharia de features. A técnica de Conjunto Aproximado foi a mais utilizada, seguida por Stepwise, Algoritmo Genético e PCA (Figura 4).

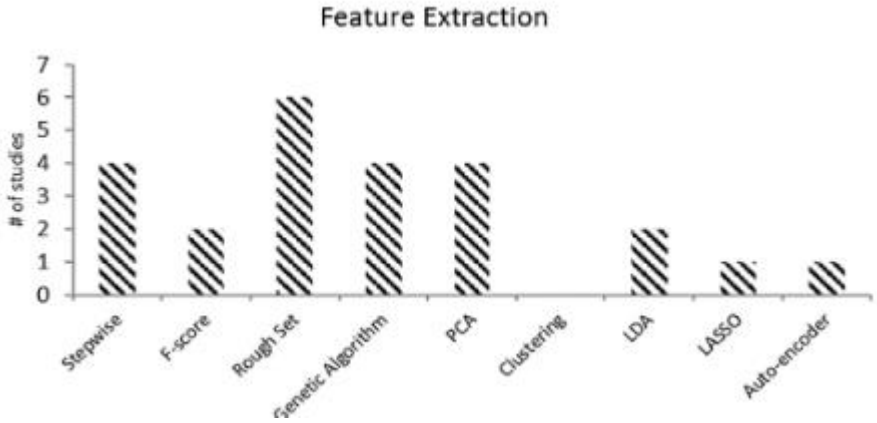

Figura 4. Técnicas de seleção e engenharia de features. Dastile et al. (2020)

Para avaliação das métricas dos modelos aplicados para o score de crédito o estudo identifica que, em maior parte, os estudos da literatura utilizam técnicas de *Probability of Correct Classification* (PCC) e *Area Under the Curve* (AUC) (Figura 5).

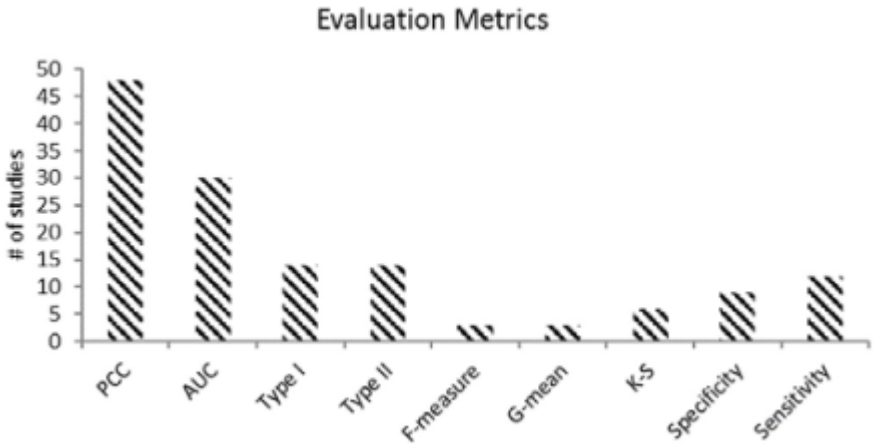

Figura 5. Técnicas de Avaliação de Métricas de Modelos. Dastile et al. (2020)

Ao final, o estudo mostrou que utilizar um conjunto de classificadores como *boosting* é mais eficaz do que usar apenas um e, embora modelos de aprendizado profundo sejam pouco usados na geração do score de crédito, como redes neurais artificiais (ANN), eles apresentaram resultados promissores (Figura 6).

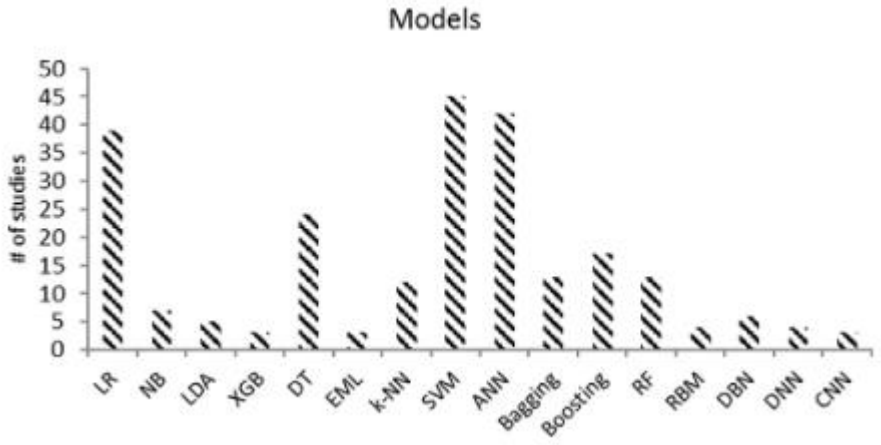

Figura 6. Modelos mais utilizados em *Credit Scoring*. Dastile et al. (2020)

3.2.2. A study on credit scoring modeling with different feature selection and machine learning approaches

Nesse artigo é avaliado diferentes abordagens para a geração de um score de crédito. Em seu estudo, Trivedi (2020) se concentra em construir um modelo de previsão utilizando um conjunto de dados públicos alemão.

Diferentemente do estudo citado no tópico anterior, Trivedi (2020) utiliza um outro framework para a geração do score de crédito. Neste, conforme mostrado na figura 6, é apresentado três métodos de seleção de *features*(*Chi-Square*, *Informatinon Gain* e *Gain Ratio*) a fim de reduzir as 20 features iniciais contidas no data set disponível para 15 ao final.

A respeito dos modelos de aprendizado de máquina para identificar o melhor para geração do score de crédito é descrito e utilizado quatro para testar a eficácia: *Bayesian, Decision Tree, Support Vector Machine* (RBF) e *Random Forest*. E, para avaliação cinco métricas foram utilizadas: acurácia de desempenho, *F-value*, taxa de falso positivo, taxa de falso negativo e tempo de treinamento.

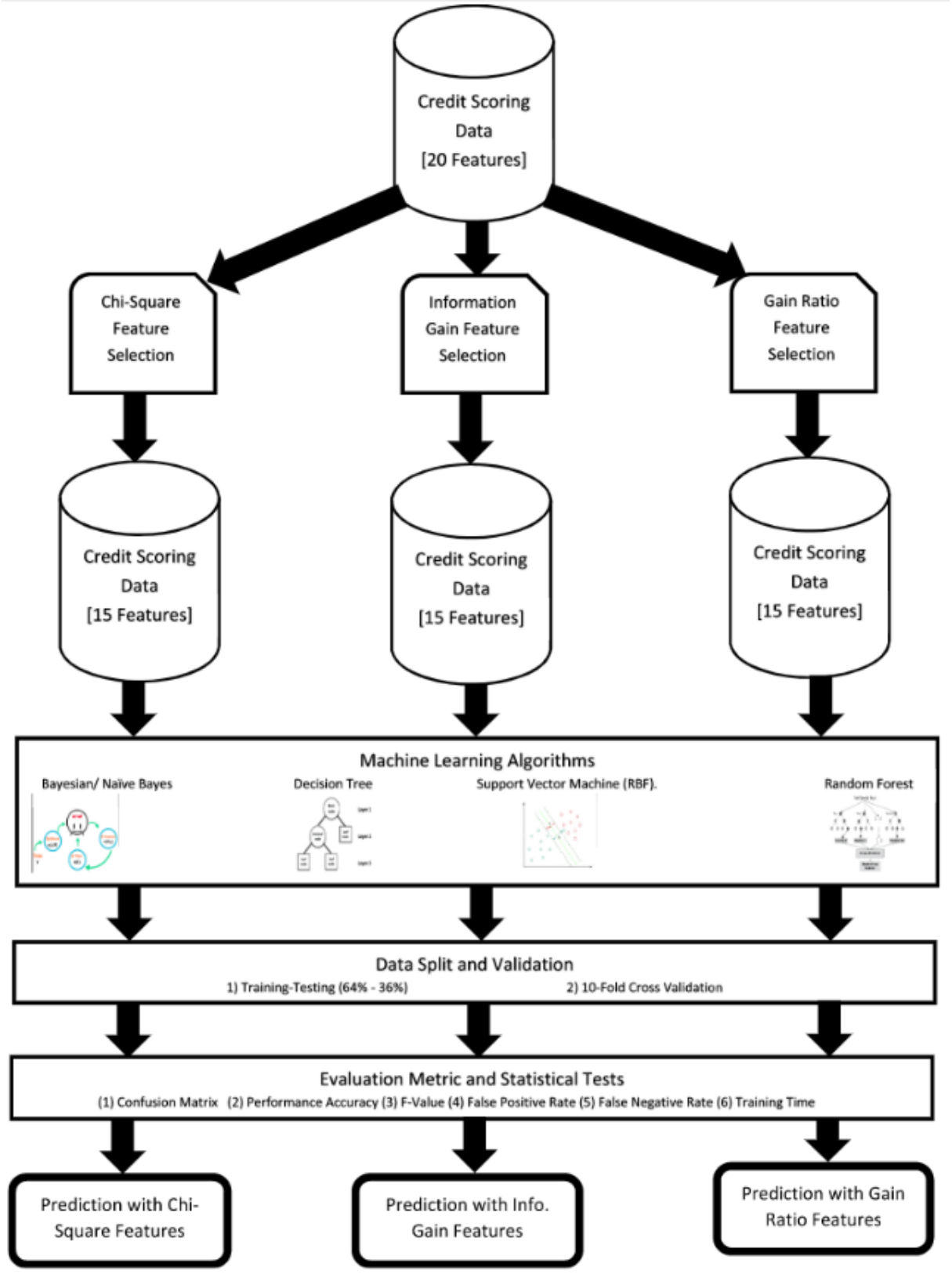

Figura 7. Guia utilizado por Trivedi (2020) para geração do score de crédito

Apesar do tempo de treinamento alto, o estudo apontou uma combinação do modelo de aprendizado de máquina Random Forest e o método de seleção de recursos Chi-Square como uma boa escolha para geração de um score de crédito robusto e preciso.

3.2.3. A benchmark of machine learning approaches for credit score prediction

Em seu artigo, Moscato et al (2020) propõe um estudo de alguns dos modelos de pontuação de risco de crédito mais usados para prever se um empréstimo será pago em uma plataforma P2P. Um conjunto de dados de uma plataforma de empréstimo social real é usado para realizar a análise experimental, considerando diferentes métricas de avaliação e comparando os resultados obtidos nos três melhores métodos conforme Figura 8.

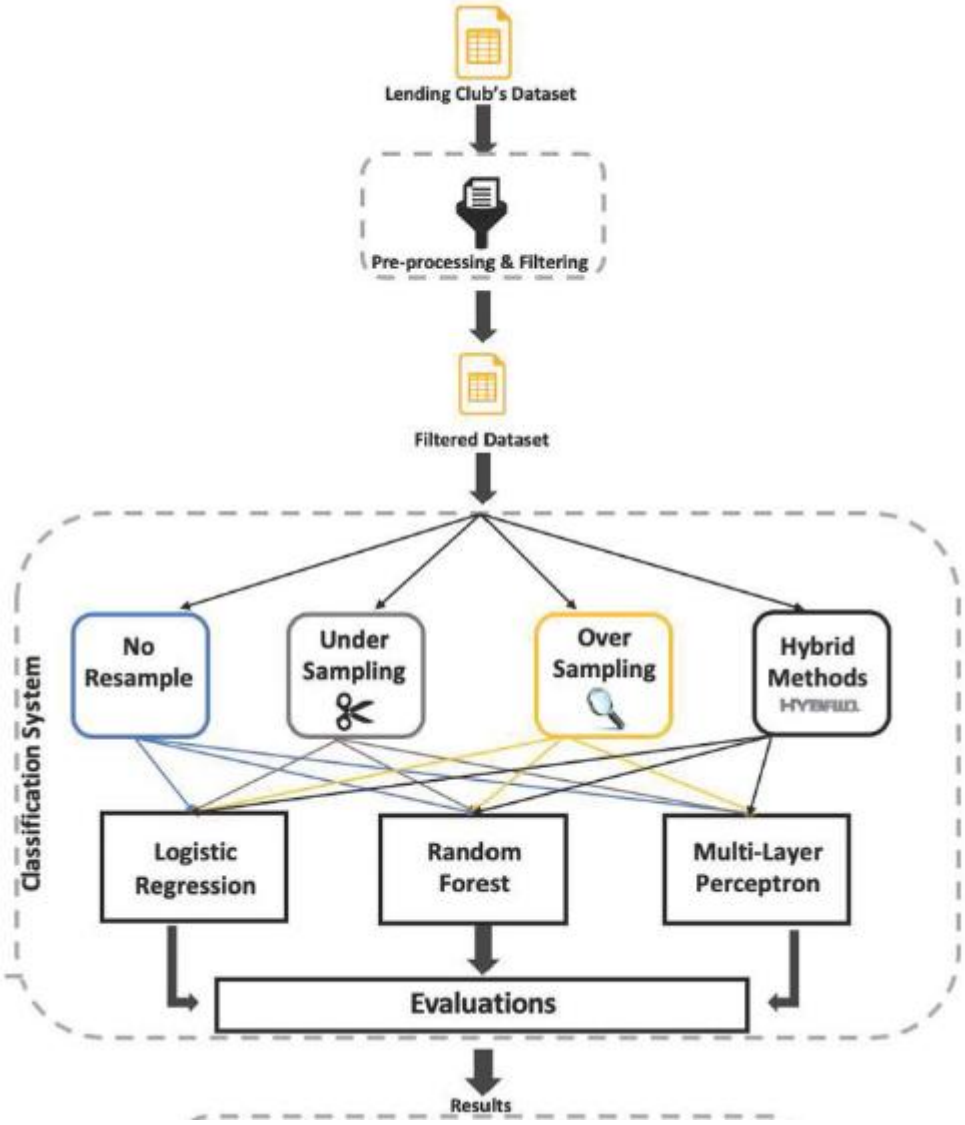

Figura 8. Guia utilizado por Moscato et al. (2020) para geração do score de crédito

O estudo conclui que o modelo de Random Forest (RF-RUS) é o melhor método para prever o status de bom ou mau pagador. Em adicional ele aponta que, além da métrica de AUC (considerada como uma das métricas mais utilizadas em modelos de scoragem de crédito) é importante se atentar para a métrica de falsos positivos pois resulta em impactos de custo muito maiores para o mercado financeiro.

Classifier	AUC	TPR	TNR	FP-Rate	G-Mean	ACC
RF - RUS	**0.717**	**0.630**	**0.680**	**0.320**	**0.6560**	**0.640**
LR - ROS	0.710	0.659	0.642	0.360	0.6503	0.650
LR - SmoteToken	0.710	0.660	0.640	0.360	0.6500	0.656
Logistic regression	0.685	0.983	0.069	0.960	0.2600	0.770
Random forest	0.720	0.983	0.084	0.920	0.2870	0.773
MLP	0.704	0.990	0.040	0.945	0.2060	0.771

Figura 9. Métricas de avaliação dos classificadores utilizados por Moscato et al (2020)

3.2.4. A comparative study on base classifiers in ensemble methods for credit scoring

Abellán e Castellano (2016) se concentram na seleção do melhor classificador usado em ensembles em conjuntos de dados de crédito. Os resultados mostram que um classificador base simples alcança um melhor equilíbrio entre alguns aspectos importantes. O estudo considera a acurácia direta e a área sob a curva ROC como medidas comparativas.

Os modelos utilizados foram: *AdaBoost*, *Bagging*, *Random Subspace*, *DECORATE* e *Rotation Forest*.

Base	AdaBoost	Bagging	Random Subspace	DECORATE	Rotation Forest	Average
LogR	2.8	3.3	3.5	**2.4**	3.3	3.06
MLP	3.9	**2.7**	2.4	4.4	3.6	3.40
SVM	3.2	3.3	4.8	2.8	3.8	3.58
C4.5	2.6	3.0	2.5	2.8	**2.0**	2.58
CDT	**2.5**	**2.7**	**1.8**	2.6	2.3	**2.38**

Figura 10. Acurácia de cada classificador. Abellán e Castellano (2016)

Base	AdaBoost	Bagging	Random Subspace	DECORATE	Rotation Forest	Average
LogR	4.5	**2.3**	**2.0**	**1.5**	2.3	2.52
MLP	3.8	2.7	**2.0**	3.2	3.3	3.00
SVM	**2.0**	5.0	5.0	5.0	5.0	4.40
C4.5	2.5	2.6	3.25	3.2	2.7	2.85
CDT	2.2	**2.3**	2.5	2.2	**1.7**	**2.19**

Figura 11. AUC para cada classificador. Abellán e Castellano (2016)

3.2.5. The effect of feature selection on financial distress prediction

Sobre o impacto da seleção de features, no artigo The effect of feature selection on financial distress prediction, Liang, Tsai e Wu (2014) apontam que esse é um passo importante e crucial no desenvolvimento de modelos de previsão de risco de crédito.

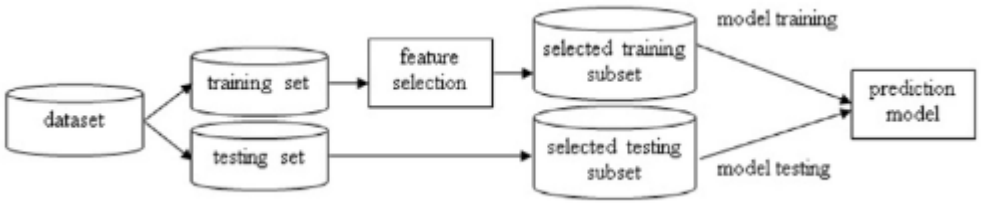

Figura 12. Modelo utilizado por Liang, Tsai e Wu (2014) para avaliação do impacto de seleção de features em modelos de score de crédito

O artigo apresenta um estudo abrangente sobre o efeito de realizar seleção de features em modelos de previsão usando diferentes técnicas de seleção e classificação. Os resultados experimentais mostram que não há uma combinação ideal de técnica de seleção e classificação e que a realização de seleção de features nem sempre melhora o desempenho mas que, em média, a utilização de algoritmo genético e regressão logística pode melhorar a previsão em dados de crédito

3.3. Consideração finais

Na tabela 1 é apresentado os artigos selecionados nessa pesquisa, trazendo um resumo das principais características de cada publicação.

ARTIGOS	TITULO	PUBLICACAO	BASE DE DADOS	SELEÇÃO DE FEATURES	MELHORES RESULTADOS
A01	Statistical and machine learning models in credit scoring: A systematic literature survey	25/03/2020	German and Australian credit data	Rough Set, Stepwise, Genetic Algorithm e PCA	Convolutional Neural Networks, Random Forests, Bagging e Boostin
A02	A study on credit scoring modeling with different feature selection and machine learning approaches	28/09/2020	German credit data	Chi-square	Random forest e Decision tree (C5.0)
A03	A benchmark of machine learning approaches for credit score prediction	09/09/2020	Lending Club data-set	Matrix correlation	Gradient Boost e Random Forest
A04	A comparative study on base classifiers in ensemble methods for credit scoring	14/12/2016	6 data-sets	-	AdaBoost e Rotation Forest
A05	The effect of feature selection on financial distress prediction	17/10/2014	4 data-sets	Genetic algorithms	-

Tabela 1: Resumo estruturado de trabalhos relacionados. Autor

Com base nos estudos levantados será descrito no capítulo seguinte o guia a ser adotado para a geração do score de crédito para avaliar a possibilidade de aumento de limite dos clientes conforme problema proposto neste projeto de pesquisa.

4 GERAÇÃO DE SCORE DE CRÉDITO PARA AUMETO DE LIMITE

Baseado nas pesquisas citadas anteriormente e nos conhecimentos adquiridos ao decorrer do curso de Inteligência Artificial e Big Data (*ICMC/USP*), este capítulo discorrerá sobre a metodologia a ser seguida para atingir aos objetivos propostos neste trabalho. Além disso será detalhada a proposta de desenvolvimento iniciando com um guia geral e percorrendo cada etapa do processo até atingir o score de crédito e avaliações dos modelos.

4.1. Metodologia

A metodologia proposta nesta pesquisa será uma unificação das metodologias apresentadas no capítulo anterior, selecionando algumas técnicas que obtiveram bons resultados em cada uma delas.

A estrutura para realização da geração da base de dados e dos experimentos será sustentada pela plataforma *Databricks*, a qual integra-se ao armazenamento em nuvem dos dados fornecidos pela Bemol, empresa à qual está sendo destinada essa pesquisa. A plataforma é utilizada para construir e implantar fluxos de trabalho de engenharia de dados, modelos de aprendizado de máquina, painéis de análise, entre outras funcionalidades.

Ao final a estrutura será formulada por safras mensais onde cada safra conterá um linha por cliente com informações focadas em transações realizadas acima do limite e complementadas com outras informações como dados cadastrais e até mesmo comportamento em compras dentro do limite original.

Com os dados já estruturados serão aplicadas técnicas comumente usando em análises exploratórias como *boxplot* para identificar a distribuição dos dados a partir da média, máximos e mínimos e quartis, assim como outliers os quais serão removidos da base de dados.

Através da matriz de correlação, assim como utilizado no artigo A03 (Tabela 1), serão eliminadas *features* com alto índice de correlação. A faixa de corte definida será níveis acima de 80% priorizando manter *features* focadas em transações monetárias ocorridas acima do limite original do cliente.

Para gerar o score de crédito final serão aplicados dois modelos presentes nos artigos A01, A02, A03 e A04 (Tabela 1): modelos de *boosting* (XGBoost) e *Random Forest*. Para os dois modelos será utilizada a mesma base de dados de treino e teste geradas após as etapas anteriores.

Para cada modelo será realizado 8 ajustes experimentais nos parâmetros: *n_estimators*, *max_depth* e *learning_rate* para o XGBoost e 4 ajustes para o Random Forest nos parâmetros: *n_estimators* e *max_depth* . Em resumo:

- *n_estimators* - define o número de árvores de decisão que serão construídas, serão testados os valores de 300 e 150;
- *max_depth* - controla a profundidade máxima das árvores, serão testados os valores de 10 e 15;
- *learning_rate* - ajusta a taxa de aprendizado do modelo, serão testados os valores de 0.1 e 0.01.

Ao final será definido o melhor modelo para o problema levantado, avaliados considerando três principais métricas: AUC, KS e matriz de confusão. Esta última focando nos verdadeiros positivos pois o objetivo final é identificar quais clientes são bons a ponto de terem seus limites aumentados em um nível.

4.2. Proposta de Desenvolvimento

O desenvolvimento desta pesquisa, em suma, será dividido em quatro etapas formadas por: identificação das bases de dados, estruturação e tratamento de dados, seleção de features e análise final dos modelos.

Segue abaixo o guia diagramado e, em seguida, detalhes de cada uma das etapas de desenvolvimento adotadas.

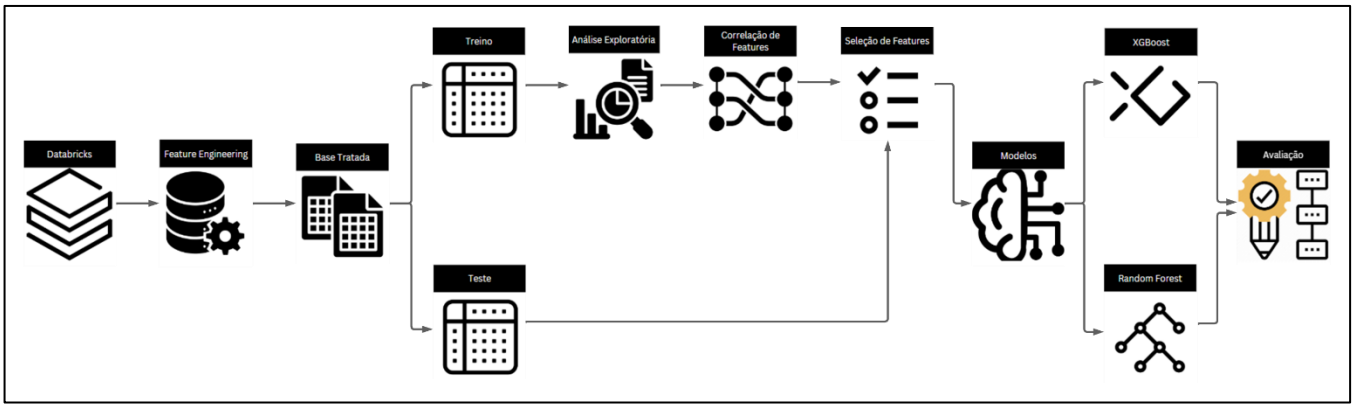

Figura 13. Guia gerado pelo autor para geração do score de crédito. Autor

4.2.1. DataSets

O primeiro passo para gerar as *features* que seriam utilizadas no modelo para se alcançar o score de crédito final foi o mapeamento de diversos tipos de dados que conseguissem prover informações necessárias para este fim.

Para conseguir essas informações foi consumido dados de um *DataLake* através da plataforma *Databricks*, plataforma de dados já disponibilizada na Bemol para todos os colaboradores. Ao todo foram levantados 15 base de dados distintas para compor a base de dados final disponibilizando informações a respeito de cada clientes como:

- Dados de crediário: informações do crediário do cliente;
- Dados de relacionamento: informações de tempo de relacionamento com a empresa e participações nos programas de bônus;
- Dados de pagamentos: informações sobre valor, quantidade e pagamentos de parcelas em dia ou em atraso;
- Dados de contratos: informações de contratos parcelados pelo crediário e contratos de compras à vista
- Dados de mercado: informações sobre negativações no SPC/SERASA e dados de compras feitas com cartão de crédito

4.2.2. Feature Engineering

A partir das bases de dados já selecionadas inicia-se o processo de engenharia de *features*, onde serão manipulados os dados para gerar informações mais ricas e precisas. Além da definição das *features* finais a serem utilizadas um dos passos importantes para gerar a base final a qual será aplicada ao modelo é como ela estaria estruturada.

Ao todo foram geradas 82 *features*, dentre elas quantitativas e categóricas, às quais não poderão ser descritas neste projeto por questão de confidencialidade e LGPD. Contudo, segue como estão divididas dentro da base de dados:

Dados de Crediário	24 features
Dados de Contratos	13 features
Dados de Pagamentos	21 features
Dados de Relacionamento	15 features
Dados de Mercado	9 features

Tabela 2: Quantidade de *features* presentes na base de dados. Autor

Para geração da base será criado safras mensais onde cada safra terá uma linha por cliente possuindo *features* agrupadas de dados em uma janela de 12 meses. Ou seja, para safra 1 que começará no período de 2021-10 terá suas *features* compostas por dados de 2020-10 até 2021-09, completando 12 meses de dados retroativos. Assim sucessivamente até a última safra 2022-11 que será composta por dados de 2021-11 a 2022-10.

O target de cada safra será dado por duas categorias: adimplente (1) e inadimplente (0). A inadimplência é calculada verificando se após 6 meses da safra analisada o cliente possui algum pagamento pendente com atraso a mais de 60 dias. Ou seja, para a primeira safra de 2021-10 a inadimplência será analisada 6 meses após, em 2022-03, caso nesse período o cliente tenha pelo menos uma parcela com atraso a mais de 60 dias ele é considerado inadimplente. Esse processo é realizado até a última safra de 2022-11 onde a inadimplência será observada 6 meses após, em 2023-05.

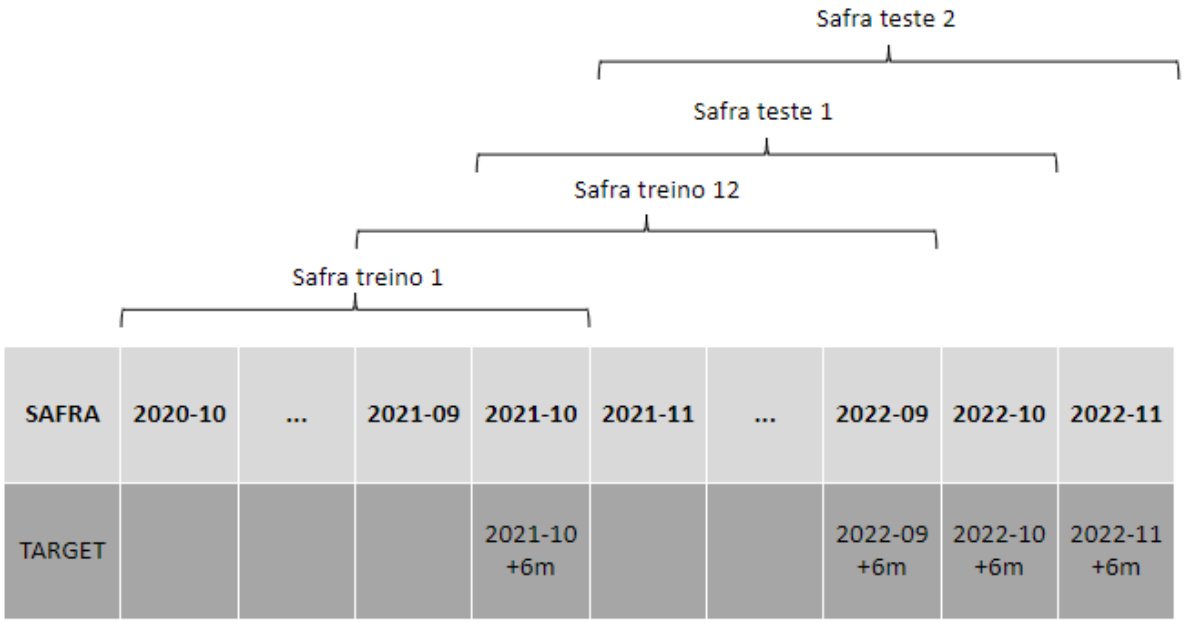

Figura 14. Estruturação da base de dados. Autor

Após os dados estruturados existirão ao todo 14 safras onde será unificada as 12 primeiras safras para treino do modelo (safra 2021-10 a 2022-09), gerando uma base de dados de *9.066.193 linhas x 82 colunas*. As safras 2022-10 e 2022-11 serão utilizadas para avaliação, formadas respectivamente por: *891.513 linhas x 82 colunas* e *902.871 linhas x 82 colunas*.

4.2.3. Seleção de *Features*

Com os dados já estruturados e as *features* selecionas o próximo passo será uma análise exploratória a fim de reduzir a dimensionalidade para levar ao modelo e eliminar possíveis outliers.

Nesta etapa será utilizado a visualização das *features* através de *boxplot* onde será possível identificar valores mínimos, máximos e suas distribuições através dos quartis. Apenas duas *features* foram identificadas com dados que fugiam muito do padrão relacionado às suas características e as linhas foram totalmente eliminadas.

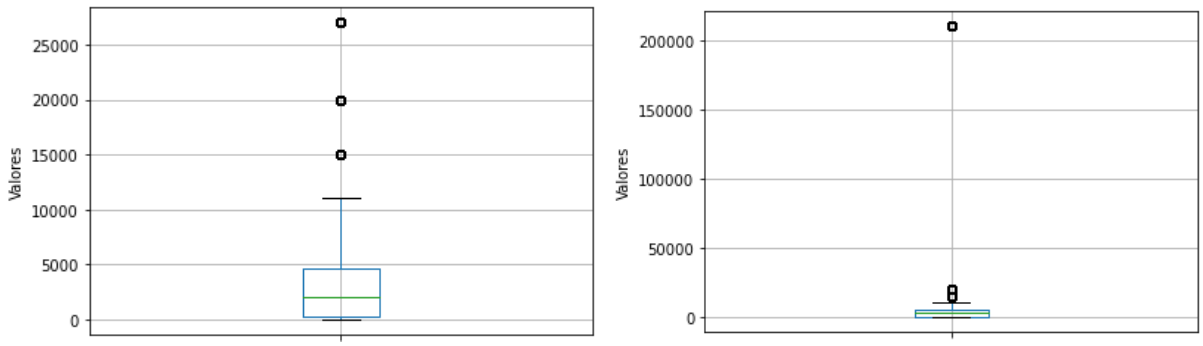

Figura 15. *Features* com outliers presentes em sua composição. Autor

Após este tratamento inicial é aplicada a correlação entre as features. Conforme descrito na metodologia, valores de correlação acima de 80% apenas uma dessas duas features deverá ser mantida e utilizada para compor o modelo. Como resultado 10 features foram eliminadas resultando em uma base de treino de *9.066.180 linhas x 72 colunas* (2021-10 a 2022-09) e duas bases de teste de *891.512 linhas x 72 colunas* (2021-10) e *902.870 linhas x 72 colunas* (2021-11).

Dados de Crediário	24 features
Dados de Contratos	11 features
Dados de Pagamentos	16 features
Dados de Relacionamento	13 features
Dados de Mercado	8 features

Tabela 3: Quantidade de *features* presentes na base de dados após redução por correlação. Autor

5 RESULTADOS E DISCUSSÕES

Neste capítulo serão apresentados os resultados da aplicação dos modelos de *XGBoost* e Random Forest sob a base de dados modelada conforme descrito no capítulo anterior. Após a geração do primeiro experimento e analisando os indicadores foi identificado alguns pontos de melhoria na base de dados inicial a fim de se obter um resultado que pudesse responder melhor a problemática proposta nessa pesquisa. Esse novo experimento também será abordado neste tópico bem como a seleção do modelo que melhor se comportou.

5.1. Experimento 1

5.1.1. *XGBoost*

A tabela a seguir mostra os resultados variando os parâmetros no modelo de *XGBoost*. No geral pode ser observado ótimas métricas de KS e AUC em todos os treinos (safras 2021-10 a 2022-09). Contudo, ao olhar as métricas das duas bases de teste (safra 2022-10 e 2022-11, respectivamente) tem uma queda acentuada em relação ao treino indicando um *overfitting*, ou seja, não generalizando muito bem dados não vistos. Esses casos podem ser observados mais acentuados nos treinos 5 e 6.

	EXPERIMENTOS				TREINO			TESTE1			TESTE2		
	LEARNING_RATE	MAX_DEPTH	N_ESTIMATORS	COLSAMPLE_BYTREE	KS	ACC	AUC	KS	ACC	AUC	KS	ACC	AUC
1	0,01	10	300	0,5	68,00%	84,00%	92,39%	70,50%	87,00%	92,95%	70,40%	86,00%	92,54%
2	0,01	10	150	0,5	66,10%	83,00%	91,52%	69,70%	86,00%	92,55%	69,90%	85,00%	92,22%
3	0,01	15	300	0,5	85,50%	93,00%	98,19%	73,60%	88,00%	94,04%	71,20%	86,00%	92,82%
4	0,01	15	150	0,5	81,20%	90,00%	96,95%	72,60%	87,00%	93,60%	70,80%	86,00%	92,58%
5	0,1	15	150	0,5	96,70%	98,00%	99,87%	75,60%	88,00%	94,55%	70,60%	85,00%	92,82%
6	0,1	15	300	0,5	99,00%	99,00%	99,99%	75,90%	88,00%	94,61%	70,40%	85,00%	92,78%
7	0,1	10	300	0,5	80,70%	90,00%	97,11%	72,50%	87,00%	93,72%	70,60%	85,00%	92,79%
8	0,1	10	150	0,5	76,00%	88,00%	95,61%	72,10%	87,00%	93,61%	70,90%	85,00%	92,89%

Tabela 4: Métricas do modelo de *XGboost* para o experimento 1. Autor

Apesar de obter boas métricas no modelo de XGBoost no geral, avaliando as *features importances*, o top 3 *features* representam em torno de 70% na geração do score do modelo. Sendo que, dessas 3 *features* mais importantes, nenhuma delas representam *features* construídas baseadas em *features* de contratos acima do limite.

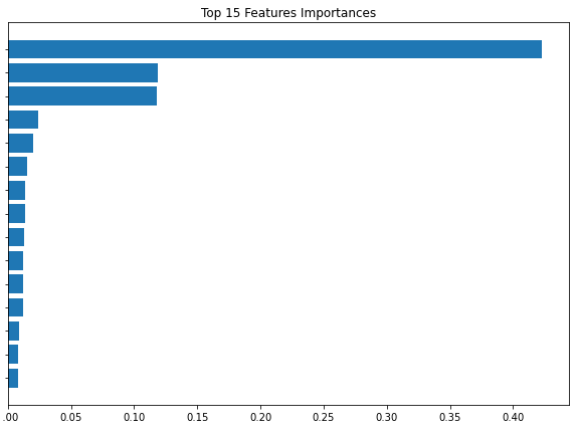

Figura 16. Gráfico de *Features Importances* do modelo *XGBoost* do experimento 1. Autor

	FEATURE1	FEATURE2	FEATURE3	FEATURE4	FEATURE5
TREINO 1	CONTRATOS_26	PAGAMENTOS_58	CREDIARIO_20	CONTRATOS_ACIMA_LIMITE_35	PAGAMENTOS_45
TREINO 2	CONTRATOS_26	PAGAMENTOS_58	CREDIARIO_20	CONTRATOS_ACIMA_LIMITE_35	PAGAMENTOS_45
TREINO 3	CONTRATOS_26	CREDIARIO_20	PAGAMENTOS_58	CONTRATOS_28	CONTRATOS_ACIMA_LIMITE_35
TREINO 4	CONTRATOS_26	CREDIARIO_20	PAGAMENTOS_58	CONTRATOS_28	CONTRATOS_ACIMA_LIMITE_35
TREINO 5	CONTRATOS_26	CREDIARIO_20	PAGAMENTOS_58	CONTRATOS_29	PAGAMENTOS_49
TREINO 6	CONTRATOS_26	CREDIARIO_20	PAGAMENTOS_58	CONTRATOS_29	PAGAMENTOS_49
TREINO 7	CONTRATOS_26	CREDIARIO_20	PAGAMENTOS_58	CONTRATOS_29	PAGAMENTOS_49
TREINO 8	CONTRATOS_26	CREDIARIO_20	PAGAMENTOS_58	PAGAMENTOS_49	CONTRATOS_29

Tabela 5. Cinco principais *Features Importances* do modelo *XGBoost* do experimento 1. Autor

Como o modelo está se baseando para geração do score de crédito em informações de contratos dentro do limite do cliente (*feature* de contratos 26), informações do crediário do cliente (*feature* de crediário 20) e informações de pagamentos de parcelas (*feature* de pagamentos 58), aparenta-se que na verdade ele esteja respondendo uma questão diferente da proposta nessa pesquisa. O score parece estar avaliando o quão bom ou mal pagador esse cliente será dentro do próprio limite atual dele.

Essa percepção levou a um ajuste em como seria abordado esse experimento visando um score ao qual respondesse melhor o objetivo geral deste trabalho: aumentar o nível de limite no crediário do cliente.

5.1.2. Random Forest

Com as variações dos parâmetros no modelo *Random Forest* obteve-se resultados similares aos do *XGBoost* com uma pequena queda no KS, ou seja, o modelo possui menor capacidade para separar as classes positivas e negativas, contudo ainda com uma AUC alta.

	EXPERIMENTOS			TREINO			TESTE1			TESTE2		
	N_ESTIMATORS	MAX_DEPTH	RANDOM_STATE	KS	ACC	AUC	KS	ACC	AUC	KS	ACC	AUC
1	350	10	0	61,18%	80,65%	88,96%	68,82%	85,78%	91,82%	68,11%	85,35%	91,38%
2	350	15	0	70,73%	85,39%	93,21%	70,24%	86,63%	92,67%	69,78%	85,81%	92,14%
3	300	10	0	61,18%	80,65%	88,96%	68,83%	85,79%	91,83%	68,11%	85,35%	91,39%
4	300	15	0	70,72%	85,39%	93,22%	70,72%	86,44%	92,73%	69,78%	85,81%	92,14%

Tabela 6: Métricas do modelo de *Random Forest* para o experimento 1. Autor

Nesse modelo as *features* ficaram melhores em questão de balanceamento. Uma das *features* de contratos acima do limite apareceram no top 5 das *features importances* (treino 1 e 3) o que é desejável para ter maior influência no resultado do score de crédito.

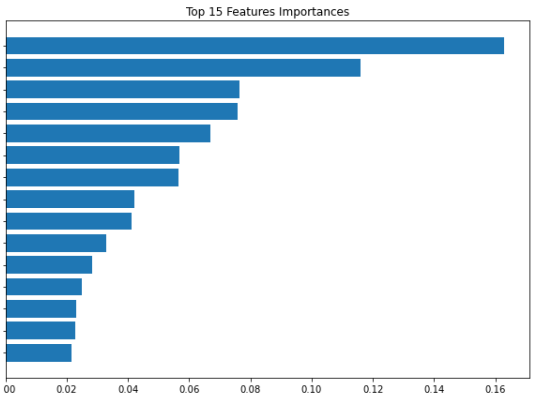

Figura 17. Gráfico de *Features Importances* do modelo *Random Forest* do experimento 1. Autor

	FEATURE1	FEATURE2	FEATURE3	FEATURE4	FEATURE5
TREINO 1	PAGAMENTOS_PARCELAS_58	CREDIARIO_20	PAGAMENTOS_45	PAGAMENTOS_49	CONTRATOS_ACIMA_LIMITE_35
TREINO 2	PAGAMENTOS_PARCELAS_58	CREDIARIO_20	PAGAMENTOS_53	PAGAMENTOS_45	CONTRATOS_26
TREINO 3	PAGAMENTOS_PARCELAS_58	CREDIARIO_20	PAGAMENTOS_45	CONTRATOS_26	CONTRATOS_ACIMA_LIMITE_35
TREINO 4	PAGAMENTOS_PARCELAS_58	CREDIARIO_20	PAGAMENTOS_53	PAGAMENTOS_45	CONTRATOS_26

Tabela 7. Cinco principais *Features Importances* do modelo *Random Forest* do experimento 1. Autor

Para a variação do *max_depth* no caso do *Random Forest* não se tem grandes alterações das métricas do modelo, treinos 1 e 3 ficaram bem similares entre si assim como os treinos 2 e 4. Abaixo algumas métricas a mais, matriz de confusão e distribuição dos scores, dos treinos em potencial (treino 3) para avaliação final:

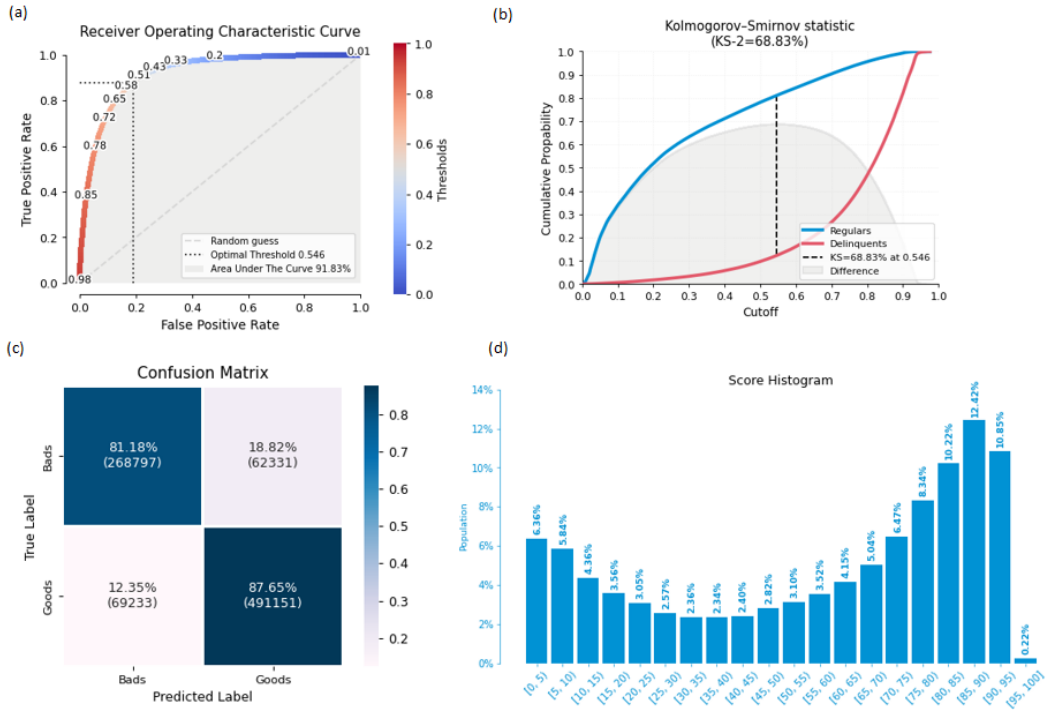

Figura 18: Métricas do Experimento 1 *Random Forest*, treino 3, teste 1. (a) AUC (b) KS (c) Matriz de confusão (d) distribuição dos scores. Autor

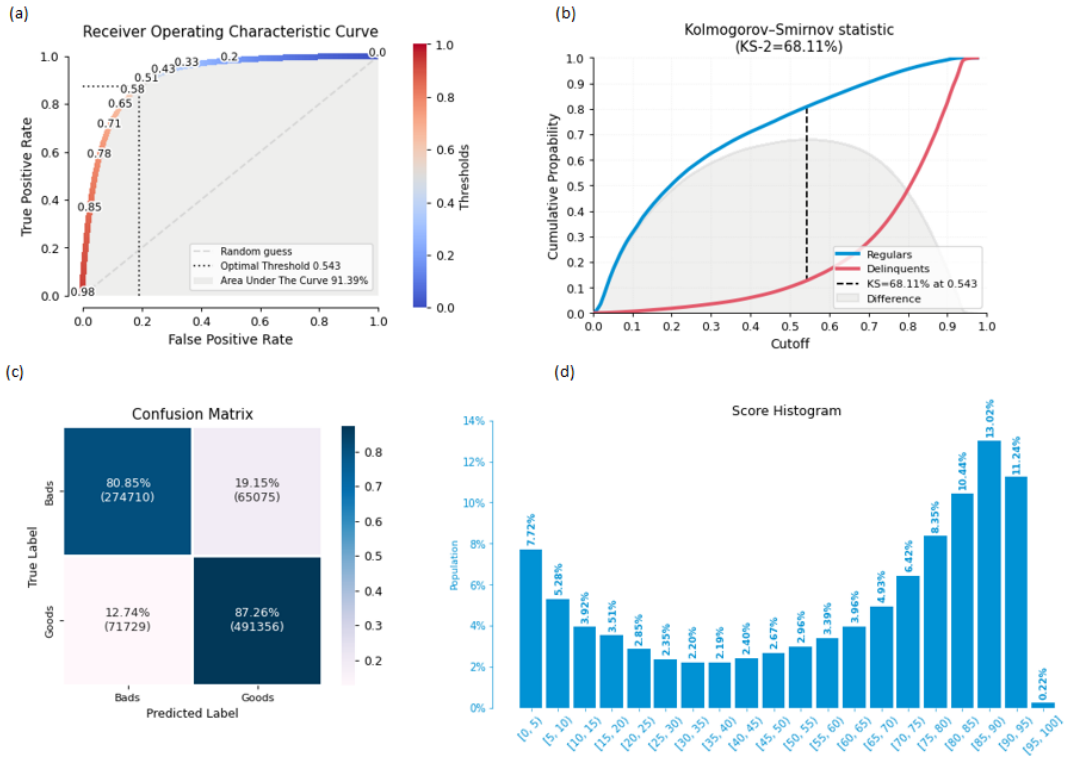

Figura 19: Métricas do Experimento 1 *Random Forest*, treino 3, teste 2. (a) AUC (b) KS (c) Matriz de confusão (d) distribuição dos scores. Autor

5.2. Experimento 2

Para esse experimento foi realizada uma pequena modificação na proposta inicial da metodologia da base de treino. Como mencionado no resultado do *XGBoost* no experimento 1, o objetivo agora é realizar alterações a fim de se obter uma maior influência de informações de compras acima do limite sob o score de crédito final. Para isso, o modelo foi treinado agora apenas para instâncias onde há compras acima do limite que representem um valor maior ou igual a 70% do limite original do cliente. Assim, a base de treino passou dos seus *1.543.404 linhas x 72 colunas* (2021-10 a 2022-09) para *563.291 linhas x 72 colunas* (2021-10 a 2022-09).

5.2.1. *XGBoost*

Após rodar o modelo com a nova base para todas as mesmas variações de parâmetros do experimento 1 pode ser observado uma melhoria nas métricas do treino porém um decaimento ainda maior das métricas de teste, indicando o mesmo problema do experimento 1 onde o modelo se mostra não inferir tão bem dados ainda não vistos.

	EXPERIMENTOS				TREINO			TESTE1			TESTE2		
	LEARNING_RATE	MAX_DEPTH	N_ESTIMATORS	COLSAMPLE_BYTREE	KS	ACC	AUC	KS	ACC	AUC	KS	ACC	AUC
1	0,01	10	300	0,5	75,08%	87,00%	94,56%	69,90%	84,00%	92,49%	66,10%	82,00%	90,85%
2	0,01	10	150	0,5	72,88%	85,00%	93,66%	71,60%	71,00%	93,14%	65,10%	82,00%	90,37%
3	0,01	15	300	0,5	92,29%	95,00%	99,24%	89,30%	90,00%	98,69%	67,00%	82,00%	91,26%
4	0,01	15	150	0,5	88,37%	93,00%	98,35%	85,60%	89,00%	97,68%	66,20%	82,00%	90,93%
5	0,1	15	150	0,5	99,54%	100,00%	100,00%	98,40%	93,00%	99,97%	67,40%	82,00%	91,61%
6	0,1	15	300	0,5	99,95%	100,00%	100,00%	99,70%	94,00%	100,00%	67,90%	82,00%	91,76%
7	0,1	10	300	0,5	91,16%	95,00%	99,20%	86,60%	91,00%	98,29%	67,90%	82,00%	91,69%
8	0,1	10	150	0,5	85,45%	92,00%	98,02%	81,80%	89,00%	97,03%	67,80%	83,00%	91,54%

Tabela 8: Métricas do modelo de *XGboost* para o experimento 2. Autor

A porcentagem de impacto das *features* de contratos acima do limite adquiriu um pouco mais de relevância nesse novo experimento. O top 3 *features* ainda correspondem a cerca de 50% no impacto do score de crédito final, porém mais bem distribuídos.

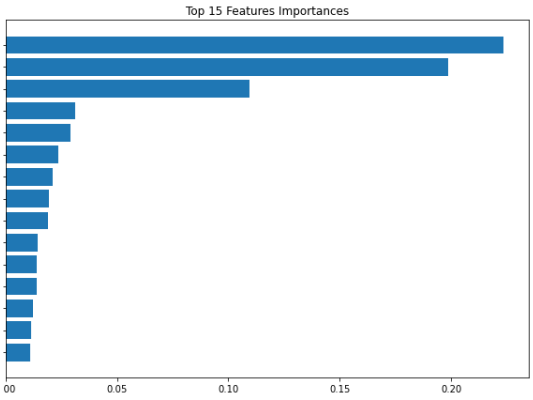

Figura 20. Gráfico de *Features Importances* do modelo *XGBoost* do experimento 2. Autor

	FEATURE1	FEATURE2	FEATURE3	FEATURE4	FEATURE5
TREINO 1	CONTRATOS_26	PAGAMENTOS_58	CREDIARIO_20	CONTRATOS_ACIMA_LIMITE_35	PAGAMENTOS_50
TREINO 2	CONTRATOS_26	PAGAMENTOS_58	CREDIARIO_20	CONTRATOS_ACIMA_LIMITE_35	PAGAMENTOS_50
TREINO 3	CONTRATOS_26	PAGAMENTOS_58	CREDIARIO_20	PAGAMENTOS_53	CREDIARIO_03
TREINO 4	CONTRATOS_26	PAGAMENTOS_58	CREDIARIO_20	PAGAMENTOS_53	PAGAMENTOS_52
TREINO 5	CREDIARIO_20	CONTRATOS_26	PAGAMENTOS_58	CREDIARIO_03	PAGAMENTOS_49
TREINO 6	CREDIARIO_20	CONTRATOS_26	PAGAMENTOS_58	CREDIARIO_03	PAGAMENTOS_49
TREINO 7	CREDIARIO_20	PAGAMENTOS_58	CREDIARIO_03	PAGAMENTOS_49	PAGAMENTOS_52
TREINO 8	CREDIARIO_20	PAGAMENTOS_58	CREDIARIO_03	PAGAMENTOS_49	PAGAMENTOS_53

Tabela 9. Cinco principais *Features Importances* do modelo *XGBoost* do experimento 2. Autor

Os treinos 1 e 2 se destacaram por trazer *features* de contratos acima do limite com um maior impacto no score de crédito, com o treino 1 como melhores métricas de KS e AUC.

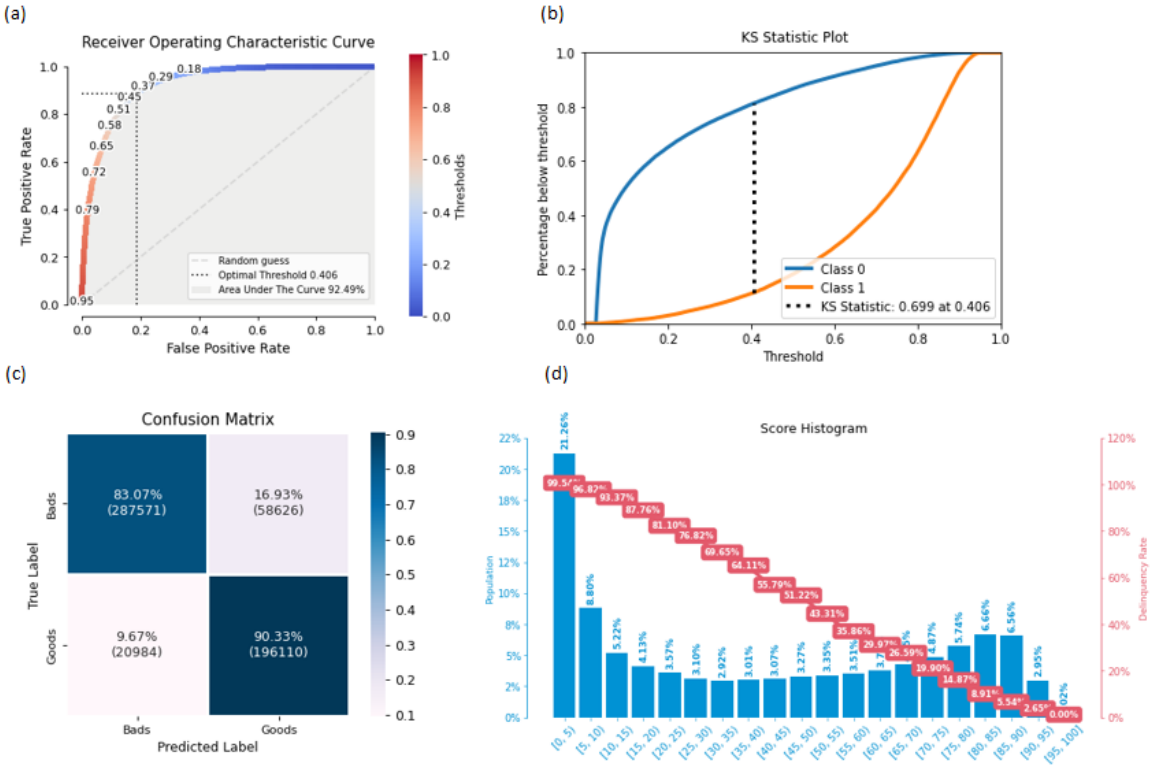

Figura 21: Métricas do Experimento 2 *XGBoost*, treino 1, teste 1. (a) AUC (b) KS (c) Matriz de confusão (d) distribuição dos scores. Autor

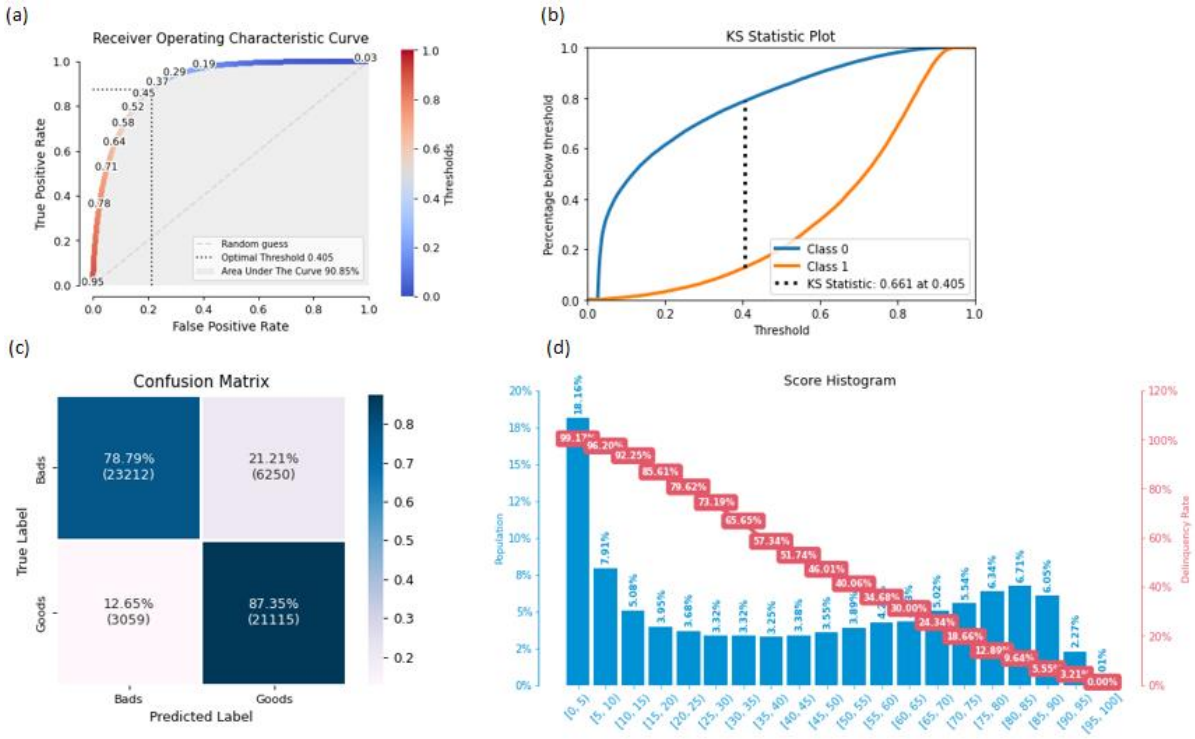

Figura 22: Métricas do Experimento 2 *XGBoost*, treino 1, teste 2. (a) AUC (b) KS (c) Matriz de confusão (d) distribuição dos scores. Autor

5.2.2. Random Forest

Para esse segundo experimento utilizando o modelo de Random Forest as métricas decaíram um pouco, apesar de continuarem boas com um AUC estável tanto no teste 1 quanto no teste 2.

	EXPERIMENTOS			TREINO			TESTE1			TESTE2		
	N_ESTIMATORS	MAX_DEPTH	RANDOM_STATE	KS	ACC	AUC	KS	ACC	AUC	KS	ACC	AUC
1	350	10	0	66,78%	83,22%	91,40%	58,33%	68,06%	84,91%	62,96%	64,20%	86,70%
2	350	15	0	79,63%	89,51%	96,19%	60,23%	69,02%	86,11%	63,91%	64,81%	87,63%
3	300	10	0	66,77%	83,21%	91,42%	58,28%	68,00%	84,95%	62,94%	64,13%	86,75%
4	300	15	0	79,51%	89,46%	96,17%	68,47%	83,02%	92,08%	64,56%	80,16%	90,20%

Tabela 10: Métricas do modelo de *Random Forest* para o experimento 2. Autor

Em todos os treinos apareceu uma *feature* de compras acima do limite no top 5 *features* e todas elas bem distribuídas, assim como a distribuição do experimento 1.

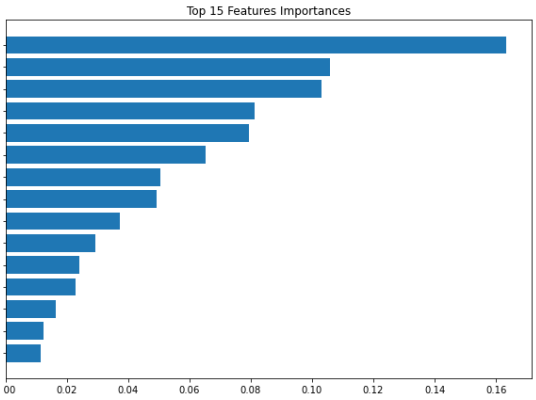

Figura 23. Gráfico de *Features Importances* do modelo *Random Forest* do experimento 2. Autor

	FEATURE1	FEATURE2	FEATURE3	FEATURE4	FEATURE5
TREINO 1	PAGAMENTOS_58	CREDIARIO_20	PAGAMENTOS_53	PAGAMENTOS_49	CONTRATOS_ACIMA_LIMITE_35
TREINO 2	PAGAMENTOS_58	PAGAMENTOS_53	CREDIARIO_20	PAGAMENTOS_49	CONTRATOS_ACIMA_LIMITE_35
TREINO 3	PAGAMENTOS_58	CREDIARIO_20	PAGAMENTOS_53	CONTRATOS_ACIMA_LIMITE_35	PAGAMENTOS_49
TREINO 4	PAGAMENTOS_58	PAGAMENTOS_53	CREDIARIO_20	PAGAMENTOS_49	CONTRATOS_ACIMA_LIMITE_35

Tabela 11. Cinco principais *Features Importances* do modelo *Random Forest* do experimento. Autor

Nesse experimento, em todos os treinos a *feature* que está sendo focada (contratos acima do limite) aparece na posição 4 e 5, as quais possuem pesos bem similares. Assim, a seleção entre os modelos será dada pelas métricas de KS e AUC, tendo assim o treino 4 para o experimento 2 do *Random Forest*.

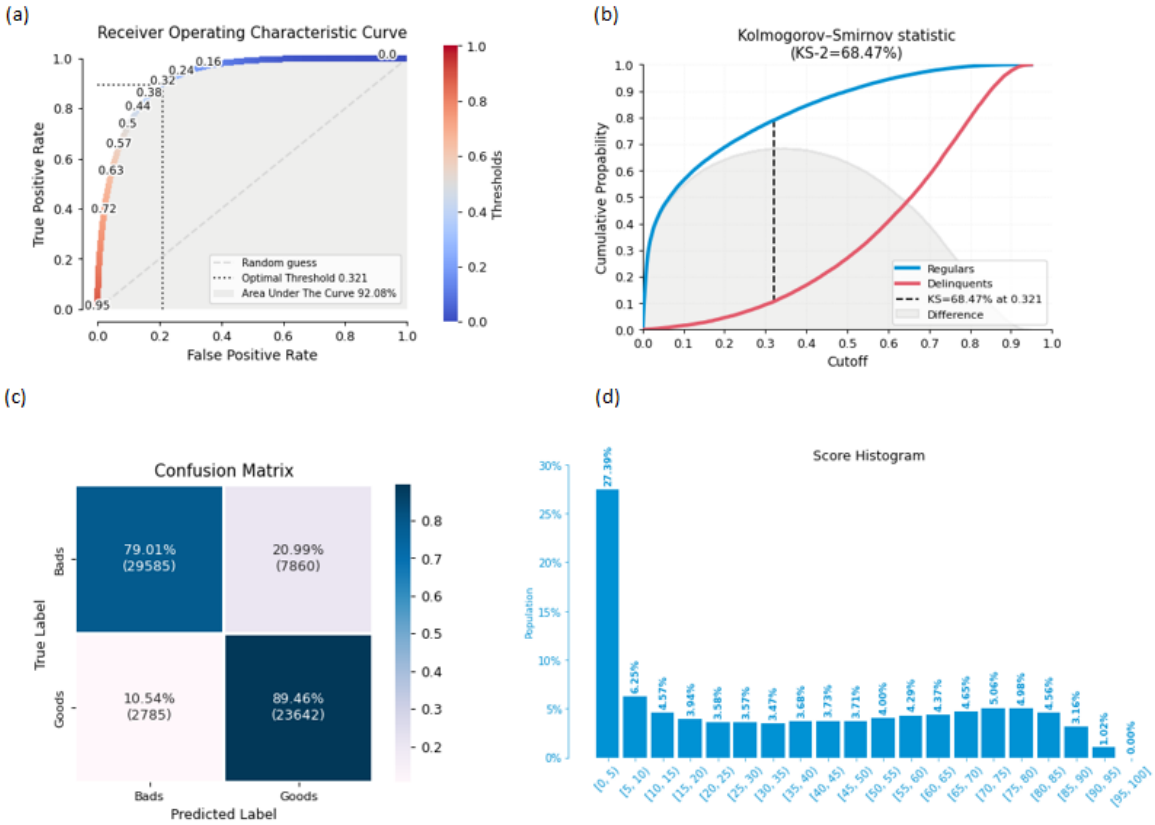

Figura 24: Métricas do Experimento 2 *Random Forest*, treino 4, teste 1. (a) AUC (b) KS (c) Matriz de confusão (d) distribuição dos scores. Autor

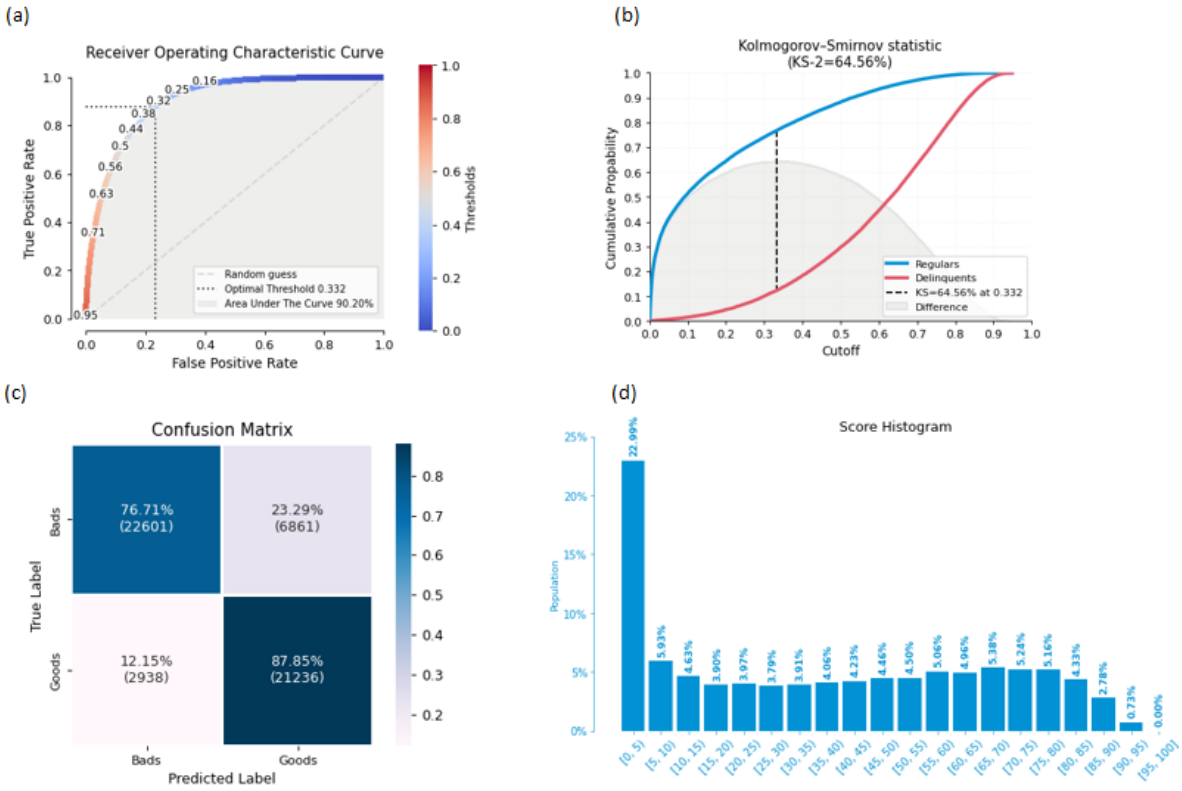

Figura 25: Métricas do Experimento 2 *Random Forest*, treino 4, teste 2. (a) AUC (b) KS (c) Matriz de confusão (d) distribuição dos scores. Autor

5.2. Escolha do modelo

Como descrito na metodologia, baseado nas pesquisas relacionadas, a escolha do modelo se dá pelas métricas de KS, AUC e Verdadeiros Positivos entre os modelos que se mostraram mais promissores.

EXPERIMENTO	MODELO	TREINO	KS	AUC	VP_1	VP_2
EXPERIMENTO 1	RANDOM FOREST	3	61,18	88,96	87,65	87,26
EXPERIMENTO 2	XGBOOST	1	75,08	94,56	90,33	87,35
EXPERIMENTO 2	RANDOM FOREST	4	79,51	96,17	89,46	87,85

Tabela 12: Resumo de métricas dos melhores modelos

No geral os modelos possuem ótimas métricas e distribuições dos scores por faixa. Contudo o mais promissor se mostrou o Random Forest utilizando-se da base alterada no experimento 2 para se atingir o objetivo geral da pesquisa: alinhar os limites dos clientes às suas necessidades atuais.

6 CONCLUSÃO

Existem diversas formas no mercado de se construir e avaliar o perfil de crédito de clientes, e vem se tornando comum o uso de modelos de *machine learning* para essa avaliação em diversos contexto. O desenvolvimento dessa pesquisa possibilitou o levantamento de técnicas e fluxos bem-conceituados de aprendizado de máquina para geração de scores de crédito. Permitindo assim realizar uma comparação e avaliar modelos diferentes para a geração de um score de crédito específico para aumento de limite de crediário de clientes da varejista BEMOL SA.

Nos trabalhos relacionados nota-se bons resultados na construção desses modelos utilizando algoritmos de ensemble como o *Random Forest* e de *Gradient Boosting*. As duas técnicas foram utilizadas para a geração do score final e de fato se mostraram promissoras ao separar as classes de bons e de maus pagadores.

Outro tema apontado nos trabalhos relacionados a respeito da melhoria dos resultados de um modelo de score de crédito é a etapa de *feature engineering*. A construção e seleção das *features* é essencial para fazer com o que modelo seja capaz de gerar um score específico para o objetivo esperado. Ao realizar o primeiro experimento proposto na metodologia esse ponto ficou bastante evidente, como diversas *features* de dimensões diferentes a respeito dos clientes são utilizadas no modelo para a geração do score, as *features* que mais impactam no modelo podem levar ele a responder muito bem sobre um objetivo diferente ao qual se está querendo apontar.

Assim, modelos de aprendizado de máquina, como o Random Forest que se mostrou ao final o mais promissor nessa pesquisa, conseguem gerar ótimos resultados na análise de crédito tendo sempre alinhado para a sua construção o conhecimento técnico do fluxo de dados ideal para a geração de *features* e construção das bases de treino e teste com as expectativas e conhecimento da área de negócio.

Como etapas futuras que poderão acrescentar mais valor a essa pesquisa pode ser apontado:

1. Aumentar o número de atributos referentes a transações acima do limite original do cliente
2. Testar algoritmos de redes neurais para comparação de resultados

REFERÊNCIAS

ABELLÁN, J.; CASTELLANO, J. G. A comparative study on base classifiers in ensemble methods for credit scoring. Expert systems with applications, v. 73, p. 1–10, 2017.

CAROLINE, A.; POTRICH, G. Análise da Concessão de Crédito em uma Empresa Varejista de Materiais de Construção. *In*: SIMPÓSIO DE EXCELÊNCIA EM GESTÃO E TECNOLOGIA, 9., 2012, Rio de Janeiro.

CORREIA, A. Seleção de Atributos para Data Science e Machine Learning. Disponível em: <https://medium.com/@airtonneto/sele%C3%A7%C3%A3o-de-atributos-para-data-science-e-machine-learning-2842c63fc59f>. Acesso em: 13 mai. 2023.

CRESPI JÚNIOR, Hugo. Gerenciamento do ponto de corte para a concessão de crédito no varejo brasileiro. 2014. 98 f. Dissertação (Mestrado em Ciências Contábeis) - Universidade Presbiteriana Mackenzie, São Paulo, 2014.

DASTILE, X.; CELIK, T.; POTSANE, M. Statistical and machine learning models in credit scoring: A systematic literature survey. Applied soft computing, v. 91, n. 106263, p. 106263, 2020.

LEE, Huei Diana. Seleção de atributos importantes para a extração de conhecimento de bases de dados. 2005. Tese (Doutorado em Ciências de Computação e Matemática Computacional) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2005. doi:10.11606/T.55.2005.tde-22022006-172219.

LIANG, D.; TSAI, C.-F.; WU, H.-T. The effect of feature selection on financial distress prediction. Knowledge-based systems, v. 73, p. 289–297, 2015.

FREAS, T. Evolution of consumer credit: Credit risk decisioning. Disponível em: <https://www.equifax.com/business/blog/-/insight/article/evolution-of-consumer-credit-a-glimpse-into-the-future-of-credit-risk-decisioning/>. Acesso em: 2 fev. 2023.

FREIRE, Boanerges Ramos. A era do varejo financeiro. Revista Empreendedor [online]. Entrevista dada a Paulo Clóvis Schmitz. 9 dez. 2009. Disponível em: <http://empreendedor.com.br/pt-br/artigos/a-era-do-varejo-financeiro>. Acesso em: 5 fev. 2023.

FORECAST GLOBAL. Feature engineering. Disponível em: <https://corporatefinanceinstitute.com/resources/data-science/feature-engineering/>. Acesso em: 13 mai. 2023.

MOSCATO, V.; PICARIELLO, A.; SPERLÍ, G. A benchmark of machine learning approaches for credit score prediction. Expert systems with applications, v. 165, n. 113986, p. 113986, 2021.

PATEL, Harshil. What is Feature Engineering — Importance, Tools and Techniques for Machine Learning. Disponível em: <https://towardsdatascience.com/what-is-feature-engineering-importance-tools-and-techniques-for-machine-learning-2080b0269f10>. Acesso em: 13 mai. 2023.

ROSS, Stephen A.; WESTERFIELD, Randolph W.; JAFFE, Jeffrey F. Administração financeira. São Paulo: Editora Atlas, 1995

SCHNEIDER, J. Por que o varejo ainda erra na concessão de crédito aos consumidores? Disponível em: <https://www.ecommercebrasil.com.br/artigos/por-que-o-varejo-ainda-erra-na-concessao-de-credito-aos-consumidores>. Acesso em: 5 fev. 2023.

THEODORO, L. Setor do varejo concentra o maior percentual de dívidas pagas em maio por empresas inadimplentes, mostra. Disponível em: <https://www.serasaexperian.com.br/sala-de-imprensa/analise-de-dados/setor-do-varejo-concentra-o-maior-percentual-de-dividas-pagas-em-maio-por-empresas-inadimplentes-mostra-serasa-experian-2/>. Acesso em: 2 fev. 2023.

TRIVEDI, S. K. A study on credit scoring modeling with different feature selection and machine learning approaches. Technology in society, v. 63, n. 101413, p. 101413, 2020.