UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS DA COMPUTAÇÃO

HÉLIO GONÇALVES DE SOUZA JÚNIOR

Comparação de métodos de inferência dos rejeitados em modelos de *Credit Scoring*

Recife

2022

HÉLIO GONÇALVES DE SOUZA JÚNIOR

Comparação de métodos de inferência dos rejeitados em modelos de *Credit Scoring*

Trabalho apresentado ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Mestre em Ciência da Computação.

Área de Concentração: Inteligência computacional

Orientador (a): Prof. Dr. Germano Crispim Vasconcelos

Coorientador (a): Prof. Dr. Rodrigo Carneiro Leão Vieira da Cunha

Recife

2022

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

S729c Souza Júnior, Hélio Gonçalves de
Comparação de métodos de inferência dos rejeitados em modelos de *Credit Scoring* / Hélio Gonçalves de Souza Júnior. – 2022.
81 f.: il., fig, tab.

Orientador: Germano Crispim Vasconcelos.
Dissertação (Mestrado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2022.
Inclui referências.

1. Inteligência computacional. 2. Aprendizagem de máquina. I. Vasconcelos, Germano Crispim (orientador). II. Título.

006.31 CDD (23. ed.) UFPE - CCEN 2023-57

Hélio Gonçalves de Souza Júnior

“"Comparação de métodos de inferência dos rejeitados em modelos de Credit Scoring”

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Pernambuco, como requisito parcial para a obtenção do título de Mestre em Ciência da Computação. Área de Concentração: Inteligência Computacional.

Aprovado em: 11/03/2022.

BANCA EXAMINADORA

_____

Profa. Dra. Renata Maria Cardoso Rodrigues de Souza
Centro de Informática/UFPE

_____

Prof. Dr. Rosalvo Ferreira de Oliveira Neto
Departamento de Engenharia da Computação/UNIVASF

_____

Prof. Dr. Germano Crispim Vasconcelos
Centro de Informática/UFPE
(**Orientador**)

Dedico este trabalho a Deus, minha família e meus amigos que foram porto seguro e essenciais para o meu sucesso.

AGRADECIMENTOS

Agradeço a Deus, pela saúde dada para encarar diariamente os desafios. Por me guiar nesta importante fase da minha vida, e me conceder forças para conciliar os estudos e minha vida profissional.

Aos meus pais, que sempre mostraram a importância do estudo na vida de uma pessoa.

Aos familiares e amigos pelo apoio, compreensão e acolhimento

Ao meu orientador Germano Vasconcelos e coorientador Rodrigo Cunha, pela confiança, esforço, apoio no dia dia, companheirismo e amizade durante a elaboração do trabalho.

Ao Centro de Informática, por conceder a oportunidade de desenvolvimento e aperfeiçoamento nos meus estudos.

E a todos que de alguma forma contribuíram e me apioaram nessa jornada.

RESUMO

Os modelos de $Credit\ Scoring$ têm desempenhado por muitos anos um papel importante na sociedade, contribuindo para a saúde financeira e a oferta de crédito no mercado, com benefícios para credores e tomadores de empréstimos em geral. No entanto, na prática, esses modelos são normalmente construídos numa amostra da população de créditos aprovados e não consideram os clientes que foram rejeitados, causando um viés amostral. A Inferência dos Rejeitados é uma abordagem para estimar como os requerentes de crédito rejeitados teriam se comportado se tivessem sido aprovados, incorporando essas informações na reconstrução do modelo de Credit Scoring. Esta dissertação investiga e compara os métodos considerados estado da arte para inferência dos rejeitados, com dados reais em problemas de larga escala de análise de crédito: Reclassificação, $Augmentation$, $Cluster$ e Parcelamento. Além disso, propõe uma nova abordagem para inferência dos rejeitados com um algoritmo de $Deep\ Learning$ usado em outras aplicações, o $Deep\ Learning$, o $Deep\ Embedded\ Clustering$ (DEC), para extração de características dos dados originais. Os métodos são avaliados por diversas métricas de performance, tais como: área sobre a curva ROC, Teste Kolmogorov-Smirnov, F1 $score$, Acurácia, Diferença entre taxas de inadimplência. Também são empregados o teste não paramétrico de Kruskal-Wallis e o teste post-hoc de Nemenyi para análise da relevância estatística dos resultados. Os métodos são analisados em três conjuntos de dados oriundos de empresas do varejo e mercado financeiro, com diversos perfis de taxa de reprovação. É investigado o quanto a inclusão de parcela dos rejeitados pode impactar em ganhos de performance nos métodos avaliados. Os experimentos realizados evidenciaram que existe diferença significativa entre os métodos estudados e que o método DEC teve desempenho superior que os demais métodos para a maioria das métricas avaliadas.

Palavras-chaves: inferência dos rejeitados; aprendizagem de máquina; $deep\ learning$; $credit\ scoring$; risco de crédito.

ABSTRACT

Credit Scoring models have played an important role in society for many years, contributing to financial health and the supply of credit in the market, with benefits for creditors and borrowers in general. However, in practice, these models are usually built on a sample of the population of approved credits and do not consider the customers that were rejected, causing a sample bias. Reject Inference is an approach to estimate how rejected credit applicants would have behaved had they been approved, incorporating this information into the reconstruction of the Credit Scoring model. This dissertation investigates and compares state-of-the-art methods for inference of rejects, with real data in large-scale problems of credit analysis: Reclassification, Augmentation, Cluster and Parcelation. In addition, it proposes a new approach for reject inference with a Deep Learning algorithm used in other applications, Deep Learning, Deep Embedded Clustering (DEC), to extract features from the original data. The methods are evaluated by several performance metrics, such as: AUC, KS, F1, Accuracy, DTI. The non-parametric Kruskal-Wallis test and the post-hoc Nemenyi test are also used to analyze the statistical relevance of the results. The methods are analyzed in three sets of data from retail and financial market companies, with different failure rate profiles. It is investigated how much the inclusion of rejects can impact on performance gains in the evaluated methods. The experiments carried out showed that there is a significant difference between the methods studied and that the DEC method performed better than the other methods for most of the evaluated metrics.

Keywords: reject inference; machine learning; credit scoring; credit risk.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

ACC	Acurácia
ALVO	Definição da variável resposta
CRISP-DM	*Cross Industry Standard for Data Mining*
DEC	*Deep Embedded Clustering*
DTI	Diferença entre taxas de inadimplência
F1	*F-score*
Fintechs	financial technology
KS	Estatística do teste de Kolmogorov-Sminorf
MAR	*Missing at random*
MCAR	*Missing completely at random*
MNAR	*Missing not at random*
PC	Ponto de Corte
PO	Ponto de Operação
ROC	*Receiver Operating Characteristics*

SUMÁRIO

1 INTRODUÇÃO

Este capítulo tem como objetivo apresentar uma visão geral sobre o trabalho desenvolvido nesta pesquisa, destacando a motivação para o seu desenvolvimento. Além disso serão apresentados os objetivos e como a pesquisa está estruturada.

1.1 APRESENTAÇÃO E MOTIVAÇÃO

O conceito de concessão de crédito data de 2000 A.C., isto é, da época dos Babilônios. Existem evidências de que os agricultores já recorriam a empréstimos a fim de regular seu fluxo de caixa devido às safras de colheita. Entretanto, foi apenas na década de 1920 que a revolução da indústria de crédito ocorreu impulsionada pela demanda por veículos. (Lewis E., 1992)

O crédito tem um papel fundamental na economia de um país, sendo um dos principais negócios no cotidiano das instituições financeiras. A definição de crédito depende do contexto em que se esteja tratando. Num sentido restrito e específico, o crédito se refere à entrega de um bem ou de um valor presente, mediante compromisso de pagamento no futuro. No comércio, uma venda à crédito possibilita ao cliente adquirir um bem (produto) mediante uma promessa de pagamento, que pode ser na modalidade rotativo ou na modalidade parcelado. No caso de banco, o crédito consiste em emprestar dinheiro, isto é, colocar à disposição do cliente um determinado valor monetário, sob promessa de reembolso futuro.

O crédito cumpre um papel importante tanto econômico quanto social, possibilitando às empresas alavancarem seu nível de atividade, facilitando a execução de projetos para os quais não disponham de recursos financeiros, ajudando, então, a população na obtenção de moradia, bens e até alimentos. Mas, por outro lado, dependendo do uso, o crédito pode tornar as pessoas físicas e jurídicas altamente endividadas. (Pires I.R., Teixeira T.R., Souza J.B., 2008)

Com o crescimento da indústria de crédito, as instituições que o concedem necessitam de um procedimento para decidir se emprestarão ou não o capital a um proponente, devido ao risco do não pagamento. Essa decisão é fundamental para o resultado financeiro da instituição, já que o lucro dos credores está diretamente associado à proporção de candidatos aprovados e ao percentual de clientes que pagam as dívidas contratadas.

Por muitos anos, o procedimento para analisar o risco na concessão de crédito era baseado

apenas na experiência adquirida dos analistas da área, no conhecimento técnico e nas informações disponibilizadas (internas e externas) que possibilitam a identificação de fatores de risco que possam comprometer a capacidade de pagamento do crédito.

Nesta linha, (Silva J., 2016) Silva utiliza cinco variáveis, chamadas de C's de crédito, como importantes aspectos relacionados ao risco de crédito. São elas: o caráter, que refere-se à idoneidade do cliente no mercado de crédito; a capacidade, que se trata da habilidade ou competência em converter negócios em renda; as condições, que são os impactos de fatores externos na geração de fluxos de caixas; o capital, que refere-se à situação econômico-financeira; e, por fim, o colateral, que é a capacidade do cliente em oferecer garantias complementares.

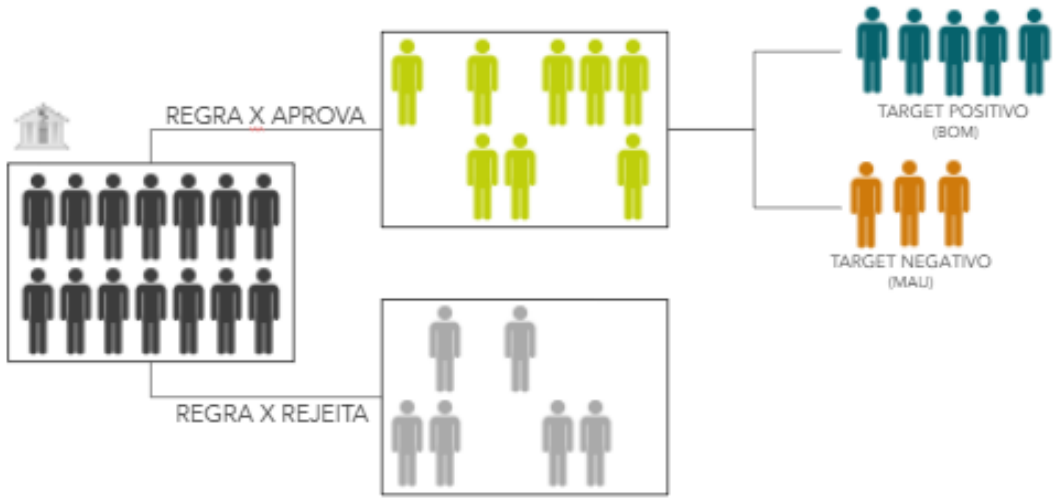
Em virtude disso, a aprovação de um pedido de crédito, nessa forma de análise, é bastante subjetiva, pois é, basicamente, no julgamento de um ou mais analistas, estando sujeito a pré-julgamentos que aumentam as chances de decisões erradas para a concessão de crédito. Com o aumento na velocidade da concessão do crédito, e as empresas buscando uma melhor rentabilidade, percebeu-se que era necessário, cada vez mais, o uso de modelos estatísticos para aumentar o volume dos créditos concedidos sem perder a agilidade e a qualidade, controlando assim o risco de crédito.

Apesar da concessão de crédito ser praticada há 4.000 anos, o conceito de *Credit Scoring*, como o conhecemos, existe desde a década de 40. Os modelos de *Credit Scoring*, , modelos de pontuação para crédito, têm sido cada vez mais utilizados para avaliar com melhor rapidez e exatidão o risco de conceder crédito para novos clientes. Por definição, o objetivo dos modelos de *Credit Scoring* é identificar o perfil dos bons e maus pagadores, qualquer que seja o conceito de bom e mau. (Durand D., 1941) desenvolveu os primeiros trabalhos para problemas de crédito, em que a técnica de análise discriminante de Fisher (Fisher R.A., 1936) foi empregada para discriminar bons e maus empréstimos.

Após a Segunda Guerra Mundial, os Estados Unidos experimentaram uma época de crescimento, e o uso de métodos automáticos para a concessão de crédito passou a ser uma necessidade real. Fundada em 1956, a FICO introduziu soluções analíticas, tais como escores de crédito, que tornaram o crédito mais amplamente disponível.

Thomas et al. (Thomas L., Edelman D., Crook J, 2002) afirmam que a motivação por trás do estudo de *Credit Scoring* é o pragmatismo e o empirismo. Os modelos não têm como objetivo explicar o risco de inadimplência, mas sim predizê-lo. Ainda de acordo com os autores, a essência dos modelos de *Credit Scoring* para consumidores deve se basear na utilização de dados históricos e toda variável relevante na predição do risco deve ser incluída no modelo.

Figura 1 – Esquema da distribuição dos dados para um modelo de *Credit Scoring*.



Fonte: Autor

Esses modelos são desenvolvidos, geralmente, com base em clientes previamente aceitos, porque a instituição sabe se o cliente pagou ou não o crédito. O problema é que essa amostra de dados é tendenciosa, pois exclui os proponentes rejeitados sistematicamente. Dessa forma, as amostras usuais, formadas apenas pelos clientes aceitos, não são totalmente representativas da população de interesse e, possivelmente, existe um viés amostral intrínseco. Dado isto, a amostra pode não ser uma representação da população de interesse, pois as características dos clientes que não foram aceitos não estão presentes na base de dados de modelagem (Alves M.C., 2008).

Esse viés pode ser mais ou menos influente no modelo de acordo com a proporção de rejeitados em relação ao total de proponentes. Quanto maior essa proporção, mais importante é o uso de alguma estratégia para a correção deste viés. A Figura 1 apresenta um esquema da distribuição dos dados para um modelo de *Credit Scoring*. (Diniz C.A.R., Louzada F., 2013)

É nesse contexto que surge a necessidade de revisitar e aprimorar os modelos baseados em inferência dos rejeitados. As técnicas de inferência trazem uma grande importância para minimizar o viés dos modelos utilizados tradicionalmente, baseados apenas em clientes aprovados anteriormente. A inferência dos rejeitados é a associação de uma resposta para o indivíduo não observado de forma que seja possível utilizar suas informações em um novo modelo.

Permitindo o uso de toda a população para o processo de modelagem, possibilitando assim o uso das características da população rejeitada. E também reduzir o viés "discriminatório" com a inclusão dos rejeitados na construção do modelo. Na figura 2 abaixo é apresentado o esquema da distribuição do processo de inferência dos rejeitados.

Figura 2 – Esquema da distribuição dos dados para um processo de inferência dos rejeitados.

Fonte: Autor

1.2 OBJETIVO

Esta pesquisa tem como objetivo principal investigar o impacto da incorporação de solicitações rejeitadas de crédito na amostra dos casos usados nos modelos de credit scoring. Dentro desse contexto, este trabalho aborda métodos de inferência de rejeitados considerados como estado da arte, com o intuito de compará-los como $benckmarks$. Como consequência, uma nova abordagem baseada em um método empregado em outras aplicações é também proposta para enfrentamento do problema.

Os objetivos secundários deste trabalho foram divididos nas seguintes fases:

- Investigar o estado da arte das técnicas de Inferência dos Rejeitados e mapear um conjunto de algoritmos de aprendizagem de máquina mais utilizados no problema;
- Identificar e potencialmente trazer novas ideias para ganhos de desempenho;
- Realizar experimentos das técnicas de inferência dos rejeitados e dos algoritmos mapeados no estado da arte em problemas de crédito reais;
- Realizar uma análise comparativa dos experimentos realizados para avaliação de desempenho e apontamento dos melhores resultados.

2 REVISÃO DA LITERATURA

Este capítulo tem como objetivo apresentar um levantamento geral sobre as pesquisas e trabalhos desenvolvidos que possuem um propósito similar ao desta dissertação. Foram consideradas pesquisas similares a este trabalho, onde abrangiam assuntos relacionados à inferência dos rejeitados. Diante disso, a seguir, serão apresentados trabalhos relacionados ao tema. Vale evidenciar que não é objetivo deste capítulo apresentar uma revisão sistemática formal, e sim a introdução e apresentação de trabalhos científicos que tiveram como objetivo solucionar problemas de inferência dos rejeitados.

2.1 TRABALHOS RELACIONADOS

Os modelos de *Credit scoring* são ferramentas essenciais para as instituições financeiras tomarem melhor decisão, distinguindo os proponentes por risco de inadimplência. No entanto, como a maioria dos modelos foram construídos com base apenas em proponentes de crédito aceitos, há uma chance de problemas de viés de seleção de amostra. Estudos anteriores tendiam a tratar esses vieses como problemas de dados ausentes, como mecanismos ausentes, sendo divididos em ausentes aleatoriamente (MAR-*Missing at Random*) e faltando não ao acaso (MNAR-*Missing not at random*) (Rubin D., Little R., 2020).

Na Tabela 1 abaixo, apresentamos uma pesquisa, visão geral, sobre inferência de rejeição nos modelos de *credit scoring*. No decorrer deste capítulo, os artigos apresentados na tabela 1 serão revisados.

Tabela 1 – Visão geral da pesquisa sobre inferência dos rejeitados

Ano	Autor	Método de Inferência dos rejeitados
1978	Hsia	Reclassication e Augmentation
1993	Joanes	Reclassification
2000	Feelders	EM
2004	Crook and Banasik	Augmentation e Extrapolation
2005	Hsieh et al.	Cluster (K-means)
2006	Sohn and Shin	Reclassication
2007	Banasik and Crook	Augmentation e Heckman's model
2010	Banasik and Crook	Augmentation
2010	Maldonado e Paredes	Extrapolation
2012	Siddiqi	Cutoff augmentation e Parceling
2103	Anderson e Hardin	Augmentation, EM
2016	Nguyen	Augmentation, Extrapolation
2017	Li et al.	Extrapolation
2108	Tian et al.	K-means, S^3VM e FQSSVM
2019	Kozodo et al.	Parceling, Cutoff augmentation e Self-Learning
2020	(2020) Ehrhardt et al.	Parceling, Augmentation e S^3VM

Fonte: Autor

De modo geral os modelos de *credit scoring*, não surpreendentemente, são na maioria dos casos rejeitados MAR, onde a aceitação é baseada nos valores das características, bem como alguns pontos de corte arbitrários (Feelders A., 2000).

Existem duas abordagens amplas para estimar a probabilidade de inadimplência: o modelo de estimativa de função, como a regressão logística, e a abordagem de estimativa de densidade, por exemplo, a análise discriminante linear. Esta última é mais suscetível a fornecer estimativas de parâmetros tendenciosos quando os pedidos rejeitados são ignorados (Feelders A., 2000).

De acordo com (Feelders A., 2000), a inferência dos rejeitados representa vários desafios. Em primeiro lugar, aprender com amostras não aleatórias é um problema de crucial importância. Segundo, a construção de um classificador apenas com base em proponentes aprovados pode levar a estimativas tendenciosas. E por fim, modelos mistos são bem interessantes para a prática de inferência dos rejeitados, contudo, estudos mais aprofundados devem ser realizados para determinar tal afirmação.

Dempster et al. (Dempster A.P., Laird N.M., 1977) apresentam o tratamento dos rejeitados como dados incompletos. Uma abordagem simples para inferência dos rejeitados é a reclassificação. (Joanes D.N., 1993) sugeriu que a reclassificação fosse utilizada a depender da taxa de rejeitados presentes e também combinada com outra técnica para amplificar suas forças

mútuas. (Joanes D.N., 1993) descobriu que nem o desempenho nem a forma simples de implementação foram significativamente melhorados pela inclusão da amostra rejeitada.

Hsia (Hsia D. C., 1978) apresenta um método definido por *augmentation*. O método ajusta os pesos dos proponentes aprovados do modelo bom/mau por uma estimativa da probabilidade de aceitação, ou seja, a probabilidade de ser incluído na população conhecida.

Diversos autores usaram o método (*augmentation*) e nenhum dos estudos empíricos mostraram melhorias significativas quando comparados com outras técnicas. Entretanto, o método tem performance superior ao processo tradicional de modelagem. (Crook J., Banasik J., 2004) (Verstraeten G., 2005) (Banasik J., Crook J., Thomas L., 2003) (Bücker M.,Van Kampen M., 2013). Entretanto, Kim e Sohn (Kim Y., Sohn S.Y., 2007) mostram, empiricamente, que esta suposição está errada.

Hand e Henley (Hand D.J. ,Henley, W.E., 1993) explanaram que a inferência dos rejeitados só pode ser efetiva se os métodos se basearem na extrapolação do modelo de aceitos para os rejeitados *parceling* ou textitaugmentation. Caso contrário, informações complementares como o status real dos rejeitos devem estar disponíveis. Embora a aquisição de informações complementares, às vezes, possa ser cara (Banasik J., Crook J., 2010) (Banasik J., Crook J., Thomas L., 2003) (Kim Y., Sohn S.Y., 2007).

Crook e Banasik (Banasik J., Crook J., 2005) apresentaram o método de ponderação e exploraram métodos de *parceling* e *augmentation*, realizando um estudo mais aprofundado e revelando que os métodos só beneficiaram o modelo de *credit scoring* (Banasik J., Crook J., 2007). Bücker, modificaram o método de inferência e os resultados modificados superam o modelo desenvolvido de forma tradicional (Bücker M.,Van Kampen M., 2013).

Ash (Ash D., Meesters S., 2002) apresentam o método de parcelamento, discutem alguns tipos de extrapolação para inferir o desempenho dos rejeitados, e apontam como conclusão que, quanto maior a proporção de rejeitados, maior é a necessidade de obter alguma estratégia para reduzir o vício amostral.

Chen (Chen W., Liu Y., Xiang G., Liu Y., 2012) destaca que os resultados mostram que o método de *K-means* para determinar a classificação dos proponentes rejeitados, realmente funciona; é útil porém tem alto custo computacional.

Lim e Hsieh (Lim M.K., Sohn S.Y., 2007) (Hsieh N.C., 2005) relatam que o agrupamento baseado em *clustering* contribui para uma melhor precisão, pois o método, de certa forma, consegue discriminar quem, na população de rejeitados, tem perfil de bom e mau pagador.

Uma comparação de diferentes métodos de inferência dos rejeitados é apresentada por

Nguyen (Nguyen H. T., 2016). Os resultados do artigo não mostram maior desempenho do modelo usando a rejeição métodos de inferência. O autor mostra que a reponderação e o parcelamento fornecem resultados mais precisos do que o $fuzzy\ augmentation$ e o método de Heckman. Em segundo lugar, os proponentes rejeitados podem afetar a precisão da previsão do modelo de $credit\ scoring$: o modelo completo produz estimativas de parâmetros estatisticamente diferentes, podendo também ser uma maneira eficaz de reduzir o $overfitting$ na seleção de modelos. E, por fim, demonstra que a inferência dos rejeitados acaba por produzir uma representação melhorada da população, uma vez que reduz drasticamente o erro padrão.

Tian et al. (Tian Y., Yong Z., Luo J., 2018) propõem um novo método, o *kernel-free fuzzy quadratic surface support vector machines (FQSSVM)*. Os autores descrevem que o método apresenta quatro vantagens quando são comparadas com métodos de SVM e suas extensões. Primeiramente, o método apresenta um processo de detecção de $outliers$ para eliminar qualquer efeito ruim. Segundo, o método pode capturar melhor as nuances da inferência de rejeição com mais precisão. Em terceiro lugar, o novo método pode evitar a demorada tarefa de busca no modelo SVM tradicional. Por fim, a estrutura convexa do modelo $FQSSVM$ garante sua alta eficiência na implementação.

Kozodoi et al. (Kozodoi N., Katsas P., 2019)relatam que os resultados empíricos indicam que a estrutura de modelagem proposta, $Self - Learning$, supera os métodos convencionais de inferência dos rejeitados em termos de três medidas de desempenho (AUC, BS e RP).

Ehrhardt e Biernacki (Ehrhardt A., Biernacki C., 2020) destacam que, em todos os métodos estudados, apenas os métodos de $augmentation$ e parcelamento melhoraram os resultados dos modelos de $credit\ scoring$. No geral, os resultados destacam que a inferência dos rejeitados pode reduzir o viés amostral, especialmente no caso de uma carteira com alta taxa de rejeição.

3 FUNDAMENTAÇÃO TÉORICA

O presente capítulo apresenta a fundamentação teórica do trabalho desenvolvido. É descrita uma visão geral dos métodos de inferência dos rejeitados baseados em modelo que foram propostos na literatura. Não apresentaremos métodos que exijam a coleta de informações suplementares sobre os indivíduos rejeitados, como por exemplo tag de bom/mau de Bureau.

3.1 INFERÊNCIA DOS REJEITADOS

Uma premissa fundamental na modelagem estatística é que a amostra selecionada para o modelo represente a população total de interesse. Porém, nos problemas de *Credit Scoring*, geralmente, essa premissa é violada, pois são utilizados apenas os proponentes aceitos, cujos comportamentos foram observados. Os rejeitados, por sua vez, não são observados e, usualmente, são descartados do processo de modelagem. Uma amostra que represente bem a população de clientes de uma determinada empresa deve considerar aqueles que tiveram a solicitação de crédito negada a fim de reduzir o viés amostral.

Segundo Siddiqi(Siddiqi N., 2006), a inferência dos rejeitados é um processo pelo qual o desempenho dos rejeitados, anteriormente, é analisado para estimar seu comportamento, ou seja, para atribuir uma classe de desempenho. Assim como existem proponentes maus na população de aprovados, alguns bons serão reprovados. A inferência dos rejeitados é a associação de uma resposta para o indivíduo não observado de forma que seja possível utilizar suas informações em um novo modelo. Os principais métodos podem ser vistos em Ash (Ash D., Meesters S., 2002), Crook e Banasik (Banasik J., Crook J., 2005)(Banasik J., Crook J., 2004)(Banasik J., Crook J., 2007), Feelders (Feelders A., 2003), Hand (Hand D., 2001) e Zhiyong (Li Z., Tian Y., Li K. Zhou F., Yang W., 2017) .

Logo, o uso da inferência dos rejeitados em problemas de $credit\ scoring$ pode proporcionar muitos benefícios. O primeiro é a relevância de ignorar os rejeitados num modelo de $credit\ scoring$, produzindo assim um viés por não trabalhar com a população total de proponentes. A inferência pode neutralizar os efeitos distorcivos da escolha seletiva, e até mesmo regras de política – também em ambientes onde a concessão é dada manualmente por gerentes de filiais das empresas. De uma perspectiva de tomada de decisão, a inferência dos rejeitados permite previsões mais precisas e realistas para todos os proponentes.

Segundo Ash e Meesters (Ash D., Meesters S., 2002), as técnicas de inferência dos rejeitados possuem a característica de serem mais ineficazes à medida que a proporção de rejeitados aumenta e, quanto maior ela for, maior é a necessidade de alguma estratégia para reduzir o viés amostral.

É importante destacar que trabalhar com inferência dos rejeitados envolve o trabalho com a previsão do desconhecido, carregando assim um certo grau de incerteza. O nível de incerteza pode ser reduzido pelo uso de melhores técnicas e pelo uso criterioso. Vale mencionar que deve-se entender que o processo de inferência pode levar a uma melhora na tomada de decisão, mas não é 100% precisa.

Para estruturar a discussão a seguir, é necessário, antes de aplicar qualquer técnica de inferência dos rejeitados, entender o mecanismo de seleção que determina se o proponente é aprovado ou reprovado. Graham et al.(Shevock A., Cumsille P., Graham J., 2012)descrevem que os rejeitados podem ser causados por combinações de três motivos: processos aleatórios, processos mensuráveis e processos não mensuráveis. Little e Rubin (Rubin D., Little R., 2020) os classificam em três categorias: *Missing completely at random* (MCAR), *Missing at random* (MAR) e *Missing not at random* (MNAR).

Segundo Rubin (Rubin D., Little R., 2020), seja $Y_i \in \{0, 1\}$ a variável resposta para os proponentes aprovados i e para os clientes reprovados i na concessão será *missing*. Além disso, definimos uma variável auxiliar a, com $a = 1$ se o proponente for aprovado e $a = 0$ se o proponente for reprovado. Observe que y é observado se $a = 1$ e ausente se $a = 0$.

Definida por:

- 1, se o i-ésimo cliente (y) aprovado $(a = 1)$ for bom pagador
- 0, se o i-ésimo cliente (y) aprovado $(a = 1)$ for mau pagador
- *Missing(na)*, se o i-ésimo cliente (y) for reprovado $(a = 0)$

Um vetor de variáveis $x = (x1, ..., x_k)$ é completamente observado para cada observação. Num cenário de concessão de crédito é atribuído uma classificação para cada proponente $g(x) = f(x_k)$. Onde é definido um corte h, tal que, quando $g(x) \geq h$ o proponente é aprovado (τ_a), caso contrário o proponente é reprovado (τ_r). Abaixo a tabela 2, o conjunto de dados pode ser definido em forma de matriz τ.

Tabela 2 – Detalhamento matricial do conjunto de dados

	x_1	x_2	...	x_k	$g(x)$	$Público$	a	Y
1					1		0	na
2					.		0	na
.					.	τ_r	0	na
.					.		0	na
.					$h-1$		0	na
.					h		1	0
.					.		1	0
.					.	τ_a	1	1
.					.		1	1
i					i		1	1

Fonte: Chen G.G., Astebro T. (2012)

3.1.1 Missing At Random (MAR)

Na classe *missing at random*(MAR) os aceitos dependem de x, mas condicional em x e não depende de y, ou seja,

$$P(a = 1|x, y) = P(a = 1|x) \tag{3.1}$$

É importante ressaltar que, apesar de sua nomenclatura, essa hipótese não se assemelha a um comportamento randômico, mas significa que há uma relação entre as variáveis e a probabilidade de ser incompleta na amostra. Essa situação ocorre com frequência na prática, uma vez que muitas empresas utilizam um modelo e/ou regra de seleção. Nesse caso, y é observado apenas se alguma função g das variáveis que ocorrem em x exceder um valor limiar definido, digamos $g(x) \geq h$ (Li Z., Tian Y., Li K. Zhou F., Yang W., 2017). Na prática de concessão de crédito, não surpreendentemente, a maioria dos casos rejeitados são MAR.

3.1.2 Missing Not At Random (MNAR)

Na classe *missing not at random*(MNAR) os aceitos dependem de y, mesmo condicionado em x, ou seja,

$$P(a = 1|x, y) \neq P(a = 1|x) \tag{3.2}$$

Isso é discutido em Heckman(Rubin D., Little R., 2020)e, normalmente, ocorre quando a aceitação é parcialmente baseada em características que não estão registradas em x. Por exemplo, um modelo de *credit scoring* formal é usado, mas, às vezes, é rejeitado por uma outra regra com base em características que não estão registradas em x, por alguma decisão do credor, que tem o direito de alterar a decisão de acordo com suas impressões com base na experiência pessoal ou estratégia de negócios. Um outro exemplo é quando o proponente solicita um empréstimo e é aceito através de um determinado modelo, contudo, o proponente não efetua o empréstimo.

3.1.3 Missing Completely At Random (MCAR)

A classe *missing completely at random*(MCAR) é a probabilidade de y ser observado e não dependa do valor de y, nem do valor de x. Em outras palavras, a falta da variável é completamente assistemática, ou seja,

$$P(a = 1|x, y) = P(a = 1) \tag{3.3}$$

Por exemplo, quando a falta da informação se deve à perda de cadastro do proponente, em virtude de ele ter mudado de residência por motivos totalmente alheios. Hsia (Hsia D.C., 1978) explica que essa forma tem sido usada por instituições de crédito, embora existam fatores econômicos óbvios que restringem seu uso. A maioria das instituições de crédito tem outras regras de política de aceitação.

3.1.4 Técnicas de Inferência dos Rejeitados

Conforme discutido na seção anterior, uma maneira rigorosa de usar toda a amostra observada τ no processo de estimativa implica em algumas modelagens desafiadoras e etapas de suposição. Um método que usa toda a amostra τ é tradicionalmente chamado de inferência dos rejeitados, uma vez que não usa apenas proponentes aprovados (τ_a), mas também rejeitados (τ_r). Com base nisso, algumas estratégias de inferência dos rejeitados são propostas para realizar uma melhoria de desempenho sobre o processo de modelagem.

Neste trabalho vamos introduzir um algoritmo de $deep\ learning$, o $deep\ embedded\ clustering$ (DEC) (Xie J., Girshick R., 2016), no processo de inferência dos rejeitados. Uma vez que muitos do métodos de inferência são baseados em metologias de $clustering$.

O trabalho concentra em uma metodologia que aprende simultaneamente representações de recursos e atribuições de $cluster$ usando $deep\ learning$. Ao contrário dos algoritmos de $clustering$ tradicionais que se concentram em funções de distância e algoritmos de agrupamento. O DEC primeiro utiliza o espaço de dados fornecido em um espaço de recursos de dimensão inferior. Em seguida, otimiza um objetivo de agrupamento neste espaço de dimensão inferior.

Os autores, com base em seus experimentos, observam adicionalmente que o DEC é significativamente menos sensível à escolha de hiperparâmetros em comparação com métodos de última geração. No artigo, os autores trabalharam com dados de imagem e texto, e seus resultados são bem expressivos. No problema de imagem o DEC tem performance superior em 58% quando comparado com o $K-means$, já para o problema de texto o DEC tem performance superior em 40%.

3.1.4.1 Reclassificação

Esta abordagem é mais simples para inserir os proponentes rejeitados na construção do modelo e considerar toda população dos rejeitados como maus pagadores. Ou seja, para todo $a = 0$ o vetor da variável resposta $Y_i = 0$.

Esse método tem como principal objetivo reduzir o viés amostral centrado na ideia de que, na população dos rejeitados, espera-se que a maioria dos clientes se tornem maus pagadores. Contudo, os problemas com essa abordagem são óbvios: reforça os preconceitos de más decisões do passado. Uma vez que algum grupo de proponente em potencial recebeu uma classificação ruim, não importa o quão erroneamente.

Segundo Siddiqi (Siddiqi N., 2006)), a única situação em que isso seria aceitável é quando a taxa de aprovação é muito alta, beirando os 100%, e também há um alto grau de confiança no processo de concessão. Neste caso, a suposição de que todos os rejeitados são ruins pode ser feita com alguma confiança, e vale destacar que esse evento não é tão comum.

3.1.4.2 Augmentation

Também conhecido como $hard\ cutoff$, o método consiste a priori em desenvolver um modelo inicial usando apenas os proponentes aprovados. Abaixo é apresentado o passo a passo do método.

1. Desenvolva um modelo usando os proponentes aprovados;
2. Pontue os pronentes rejeitados com o modelo dos aprovados;
3. Defina um valor de corte de taxa de $maus$ no qual quem estiver abaixo desse corte será classificado com mau, casso contrário bom;
4. Aplique esse valor de corte nos proponentes rejeitados e rótule seguindo a regra;
5. Junte os proponentes aprovados e rejeitados e desenvolva o modelo final;

Vale destacar que um bom valor de corte se da quando é possível maximizar a distinção entre os $bons$ e $maus$ clientes.

3.1.4.3 Cluster (K-means)

Essa técnica usa agrupamento para identificar os bons e maus em uma amostra rejeitada e não depende de nenhum modelo previamente desenvolvido.

1. Crie dois conjuntos de $cluter$ $(K - means)$ na população reprovada, a partir das características do proponente;

2. Classifique esse dois conjuntos em população mau e boa;

3. Combine os aprovados e reprovados para criar um conjunto de dados inferido e remodelar.

De modo geral, os algoritmos de agrupamento podem ser divididos em duas classes amplas: abordagens centróides e abordagens hierárquicas. As abordagens do centróide estimam os centróides ou o ponto central em cada $cluster$ e atribuem pontos ao $cluster$ do centróide mais próximo.

Abordagens hierárquicas assumem que cada ponto é um cluster por si só, que, repetidamente, mescla $clusters$ próximos, usando alguma medida de quão próximos dois $clusters$ estão (por exemplo, distância entre seus centróides), ou quão bom um $cluster$ no grupo resultante seria. Uma classificação não supervisionada amplamente conhecida por ser um algoritmo que se baseia em agrupar os dados em regiões locais é o algoritmo $K - means$.

$K - means$ é um dos métodos de aprendizagem não supervisionada mais connehcidos, o algoritmo foi descoberto por vários pesquisadores em diferentes disciplinas, mais notavelmente Lloyd (Lloyd, 1982).

O algoritmo é um método interativo que consiste em particionar um determinado conjunto de dados em n conjuntos em que $K > 1$ $cluster$, de modo que os dados em um $cluster$ são semelhantes entre si e são diferente daqueles em outros $clusters$ (MacQueen J., 1967).

Seja $N = (x_1, .., x_n)$ um o conjunto de n objetos a serem agrupados por um critério de similaridade, onde $x_i \in \mathbf{R}^d$, onde $i = 1, ..., n$ e $d \geq 1$ é o número de dimensões. Além disso, seja $k \geq 2$ um inteiro e $K = (1, \ldots, k)$. Para $k - partições = G(1), ..., G(k)$ de N, seja μ_j a centróide do $cluster$ $G(j)$, para $j \in K$, e $M = (\mu_1, ..., \mu_k)$ e $W = (w_{11}, ..., w_{ij})$.

Portanto, o problema de agrupamento pode ser formulado como um problema de otimiza-

ção (MacQueen J., 1984), que é descrito

$$P : minimize\ z(W, M) = \sum_{i=1}^{n} \sum_{j=1}^{k} w_{ij}(x_i, \mu_j)$$

$$sujeito\ a\ \sum_{j=1}^{k} w_{ij} = 1 \tag{3.4}$$

i=1,...,nej=1,...,k.

onde $w_{ij} = 1$ implica que o objeto x_i pertence ao agrupamento $G(j)$ e $d(x_i, \mu_j)$ denota a distância euclidiana entre x_i e μ_j para $i = 1, \ldots, n$ e $j = 1, \ldots, k$.

A distância euclidiana dos dois pontos é a norma euclidiana do vetor apontando de um dos pontos para o outro. O cálculo da norma euclidiana de um vetor $\mathbf{x} \in R^n$ pode ser visto na equação

$$\|\mathbf{x}\| = \sqrt{x_1^2 + ... + x_n^2} \tag{3.5}$$

A distância euclidiana de uma amostra $\mathbf{x} \in R^n$ e um centróide $\mu \in R^n$ é $\|\mathbf{x} - \mu\|$. Além da distância euclidiana, outras métricas podem ser usadas como o Manhattan ou a distância de Mahalanobis (Rui X., Wunsch D., 2005).

A atualização dos centróides dos $clusters$ é realizada após cada elemento do conjunto de dados foi associado a um dos $clusters$.A localização inicial dos centróides do $cluster$ podem ser colocados aleatoriamente, ou amostras aleatórias podem ser selecionadas como centróides do $cluster$ inicial para garantir que eles estão localizados no espaço de onde os dados são extraídos.

De modo geral o algoritmo $K - means$ consiste nas etapas, conforme mostrado no pseudocódigo no Algoritmo 3.1.4.3.

[H] Pseudocódigo $K - means$

Inicialização:

N:=$x_1, ..., x_n$;

$M := \mu_1, ..., \mu_k$;

Classificação:

$For\ x_i \geq N and \mu_{k \geq M}$

$Calcule distância euclidiana de cada x_i$ para os k centroides;

Atribua o objeto x_i ao centróide mais próximo μ_k;

Calculo Centróide:

Calcule centróide μ_k;

Convergência:

If $M := \mu_1, ..., \mu_k$ permanece inalterado em duas consecutivas iterações

then:

Pare o algoritmo;

else:

Vá para a Classificação

End

Desde os estudos conduzidos por Lloyd (Lloyd, 1982), MacQueen (MacQueen J., 1967) e Jancey (Jancey R., 1966), muitas investigações têm como objetivo encontrar uma partição k de N que resolve o problema P, definido pela Eq. 3.4.

A obtenção de uma solução ótima é geralmente um problema complexo. Consequentemente, uma variedade de algoritmos foram propostas para obter a solução mais próxima possível do ótimo de P, sendo o mais importante daqueles projetados como algoritmos do tipo K-means (MacQueen J., 1967).

3.1.4.4 Parcelamento

De acordo com Parnitzke (Parnitzke T., 2005), o método consiste a priori em desenvolver um modelo inicial usando apenas os proponentes aprovados. Ash e Meesters (Ash D., Meesters S., 2002) caracetrizam o método como um processo de reclassificação por risco. Este método é semelhante ao $simple\ augmentation$, mas, em vez de classificar todas as rejeições em uma determinada pontuação como bons ou maus, ele as atribui proporcionalmente à taxa mais esperada nessa pontuação (Siddiqi N., 2006).

Tendo o modelo inicial disponível, verifica-se o escore de cada indivíduo da população dos aprovados. O próximo passo é categorizar os indivíduos por faixas de escore, geralmente, distribuídos de modo uniforme. Em cada faixa de escore, observa-se a taxa de inadimplência, e com os escores dos rejeitados distribuídos nas mesmas faixas de escore construídas. Depois, associa-se às respostas de bom/mau aleatoriamente, segundo as taxas de inadimplência observadas nas faixas dos proponentes aprovados (Diniz C.A.R., Louzada F., 1977). A partir dessa classificação, é construído um modelo com os clientes aprovados e rejeitados, tendo as respostas inferidas para os clientes rejeitados. Na tabela 3, temos um exemplo do método.

Tabela 3 – Distribuição por faixa de escore

Faixa de Escore	Aprovados			Reprovados		
(Quintil)	Total	Bons	Maus	Total	Bons	Maus
Q1	29746	20462	9284 (31.2%)	2200	1513	687
Q2	29745	22531	7214 (24.5%)	2080	1576	504
Q3	29746	24040	5706 (19.2%)	2000	1616	384
Q4	29745	25163	4582 (15.4%)	1785	1510	275
Q5	29746	26599	3147 (10.6%)	1550	1386	164

Fonte: Autor

Na faixa Q_1, temos 31.2% de maus, então os reprovados, total do intervalo, receberão o mesmo percentual ateatoriamente. Então os 687 (2200 x 31.2%) proponentes rejeitados serão classificados como clientes maus e o complemento do total será bons clientes.

Já na faixa Q_2, temos 24.5% de maus, então os reprovados, total do intervalo, receberão o mesmo percentual ateatoriamente. Então os 504 (2080 x 24.5%) proponentes rejeitados serão classificados como clientes maus e o complemento do total será bons clientes. O processo é repetido para todas as faixas de escore.

3.1.4.5 Deep Embedded Clustering (DEC)

O *Deep Embedded Clustering* (DEC) é um método que, simultaneamente, aprende representações e atribuições de $cluster$, usando redes neurais profundas. O DEC aprende um mapeamento do espaço de dados para uma dimensão inferior ao espaço de recursos, no qual otimiza iterativamente um objetivo de agrupamento.

O método é proposto inicialmente por Xie(Xie J., Girshick R., 2016) e modificado por Guo (Guo X., Gao L., 2017) e Li (Li F., Qiao H., 2017) em vários aspectos. Para facilitar a descrição, é usado DEC para representar a família de algoritmos que realiza $clustering$ nos recursos incorporados de um autoencoder.

O objetivo do DEC é minimizar a função de perda dada por:

$$L = \alpha L_r + \beta L_c, \tag{3.6}$$

onde L_r é a perda de reconstrução de um autoencoder, L_c é a perda de agrupamento e (α, β) são coeficientes para equilibrar essas duas funções de perda. Dado um conjunto de amostras de treinamento $X = (x_i \in R^d)$ para $i = 1, ..., n$, onde D e n são a dimensão e o número de amostras. A perda de reconstrução é definida pelo Erro Quadratico Médio (MSE):

$$L_r = \frac{1}{n}\sum_{i=1}^{n} \|x_i - gU(fW(x_i))\|_2^2, \tag{3.7}$$

onde fW e gU são codificador e decodificador, respectivamente. A perda de agrupamento, no entanto, pode ser variada entre os membros da família (Yang B., Hong M., 2017).

$$L_c = \frac{1}{n}\sum_{i=1}^{n} \|fW(x_i - Ms_i\|_2^2, \tag{3.8}$$

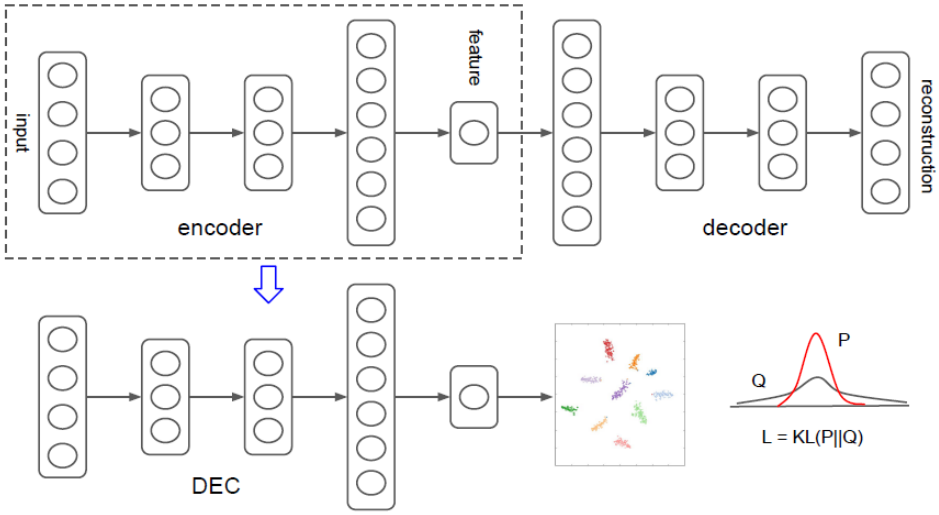
onde $s_i \in \{0,1\}^K$, $s_i = 1$ o indicador de atribuição do $cluster$ para x_i, e $\mathbf{M} = [\mu_1, ..., \mu_K \in \mathbf{R}^{dxK}$ contém K centróides de $cluster$ no espaço de recurso com dimensão d. A perda do $clustering$ também pode ser a divergência KullbackLeibler(KL) ((Xie J., Girshick R., 2016), (Li F., Qiao H., 2017), (Dizaji K., Huang H., 2017))

$$L_c = KL(P\|Q) = \sum_i \sum jp_{ij} log \frac{p_{ij}}{q_{ij}} \tag{3.9}$$

onde q_{ij} é a similaridade entre z_i e o centro do $cluster$ μ_j da distribuição $t - Student's$ (Maaten L., Hinton G., 2008)

$$q_{ij} = \frac{(1 + \|z_i - \mu_j\|^2)^{-1}}{\sum_j (1 + \|z_i - \mu_j\|^2)^{-1}} \tag{3.10}$$

Figura 3 – Esquema da Estrutura da Rede



Fonte: Xie J., Girshick R. (2016)

e p_{ij} é definido por

$$q_{ij} = \frac{q_{ij}^2 / \sum_i q_{ij}}{\sum_j (q_{ij}^2 / \sum_i q_{ij})} \tag{3.11}$$

O DEC consiste em dois estágios: pré-treinamento e $finetuning$. O pré-treinamento aprende recursos válidos que são utilizados para inicializar os centros do $cluster$. No $finetuning$, o agrupamento e aprendizado são realizados em conjunto. Desta forma, o recurso aprendido será específico da tarefa, ou seja, adequado para $clustering$.

Após um treinamento em camadas, concatenamos todos os codificadores – camadas seguidas por todas as camadas do decodificador, ao contrário da ordem de treinamento em camadas – para formar um $autoencoder$, e, em seguida, ajustar para minimizar a perda. O resultado é um $autoencoder$ de multicamadas com uma camada de codificação no meio. Em seguida, descartamos as camadas do decodificador e usamos as camadas do codificador como nosso mapeamento inicial, como mostrado na Figura 3.

3.2 APREDIZAGEM DE MÁQUINA

3.2.1 Regressão Logística

A regressão logística nos últimos anos tem tido um papel muito importante na área de modelagem estatística. Quando a variável resposta é binária, a regressão logística é indicada, ou seja, no contexto em que o objetivo é discriminar bons e maus clientes o modelo logístico se encaixa de forma adequada.

A regressão logística pode ser escrita como um caso particular dos Modelos Lineares Generalizados, ver (Paula, G.A., 2004), com função de ligação logit e variável resposta Y_i seguindo uma distribuição bernoulli e com probabilidade de sucesso π_i.

Considere a variável resposta Y_i para o cliente i, definida por:

- 1, se o i-ésimo cliente for BOM
- 0, se o i-ésimo cliente for MAU

O modelo de regressão logístico pode ser escrito como,

$$ln(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p, \tag{3.12}$$

onde $\pi(x)$ é definido como

$$\pi(x) = \frac{exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}{1 + exp(\beta_0 + \beta_1 x_1 + ... + \beta_p x_p)}, \tag{3.13}$$

em que $x_i = (1, x_{i1}, x_{i2}, ..., x_{ip}$ o vetor de variáveis explicativas do cliente e $\beta = (\beta_0, \beta_1, ..., \beta_p$ o vetor dos parâmetros do modelo.

Como a variável resposta Y_i tem distribuição Bernoulli com probabilidade de sucesso π_i, então:

- $E(Y_i = 1|x_1, ..., x_p) = P(Y_i = 1|x_1, ..., x_p) = \pi_i$, que é a probabilidade de que o cliente seja bom dado as variáveis explicativas
- $E(Y_i = 0|x_1, ..., x_p) = P(Y_i = 0|x_1, ..., x_p) = 1 - \pi_i$, que é a probabilidade de que o cliente seja mau dado as variáveis explicativas

Para representar esta probabilidade em forma de um escore S, multiplicamos esse número por 100, sendo que, quanto menor o escore, maior a probabilidade do proponente se tornar inadimplente.

3.2.1.1 Estimação dos Coeficientes da Regressão

Os parâmetros da regressão logística são, geralmente, estimados por máxima verossimilhança. A função de verossimilhança pode ser escrita da seguinte maneira:

$$L(\beta) = \prod_{i=1}^{n} [\pi(x_i)]^{y_i} [1 - \pi(x_i)]^{1-y_i} \tag{3.14}$$

Como Y_i tem distribuição Bernoulli(π_i), sendo $\pi_i = \pi(x_i)$ e usando a função de ligação $logit$ dada por:

$$ln(\frac{\pi(x)}{1-\pi(x)}) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} \tag{3.15}$$

por conveniência de cálculos, é trabalhado com a $log - verossimilhança$, ou seja, o logaritmo da função de verossimilhança, que é dada por:

$$l(\beta) = ln[L(\beta)] = \sum_{i=1}^{n} y_i (x_i^T \beta) - \sum_{i=1}^{n} ln[1 + exp(x_i^T \beta)] \tag{3.16}$$

Para a estimação dos parâmetros pode-se utilizar o método de máxima verossimilhança, encontrando o valor de β que maximiza $l(\beta)$. Nelder e Wedderburn (Nelder, J.A., Wedderburn, R., 1972) desenvolveram um algoritmo que se baseia no método de Newton-Raphson. O processo interativo de Newton-Raphson para obtenção das estimativas é definido pela expansão da função escore $U(\beta)$ em torno do valor inicial $\beta^{(0)}$, onde $U(\beta)$ é

$$U(\beta) = X^T (y - \pi) \tag{3.17}$$

tal que,

$$U(\beta) \cong U(\beta^{(0)}) + U'(\beta^{(0)})(\beta - \beta^{(0)}) \tag{3.18}$$

Isto resulta em um processo interativo de mínimos quadrados reponderados

$$\beta^{(m+1)} = (\mathbf{X}^{(T)} \mathbf{V}^{(m)} \mathbf{X})^{-1} \mathbf{X}^{(T)} \mathbf{V}^{(m)} \mathbf{z}^{(m)} \tag{3.19}$$

em que $\mathbf{V} = diag[\pi_1(1-\pi_1), ..., \pi_n(1-\pi_n)]$, $\mathbf{z} = (z_1, ..., z_n)^T$ é a variável resposta modificada, $z_i = \frac{\eta_i + (y_i - \pi_i)}{\pi_i(1-\pi_i)}$, $m = 0, 1, ..., n$.

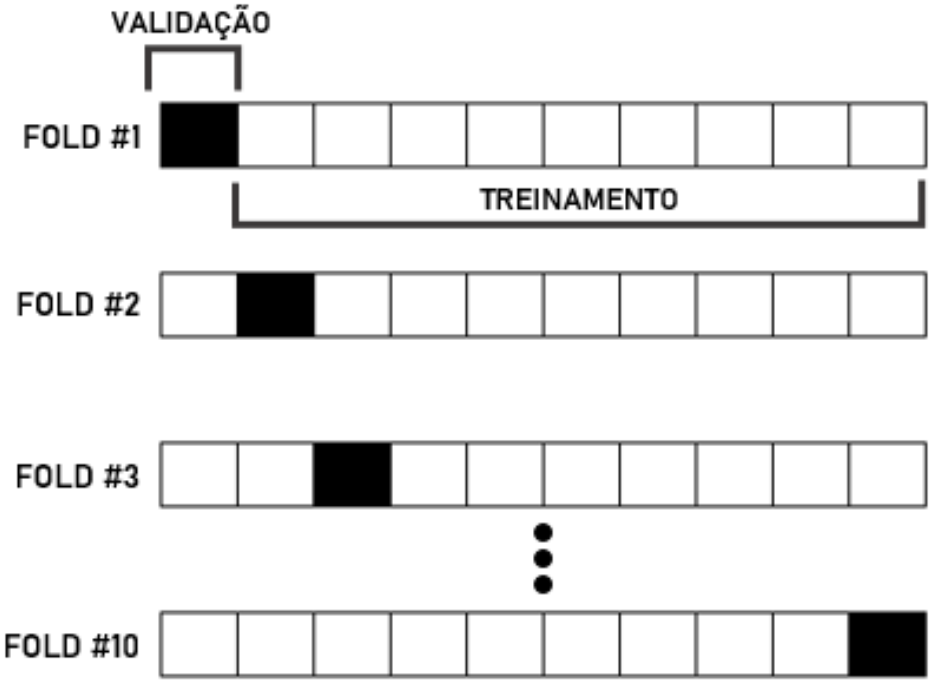
3.3 VALIDAÇÃO CRUZADA

O $k - fold\ Cross - Validation$ é um importante método que utiliza uma parte dos dados para desenvolvimento do modelo e outro subconjunto para avaliá-lo. Os dados disponíveis são divididos em k subconjuntos de tamanhos, aproximadamente, iguais e disjuntos. Cada um dos k subconjuntos é utilizado como validação do modelo, e os demais $k - 1$ subconjuntos

são designadas para a construção do modelo. Este procedimento é repetido até que cada um dos k subconjuntos tenha servido como conjunto de validação. A média das k medições de desempenho nos k conjuntos de validação é o desempenho de validação cruzada (Kohavi R., 1995).

A Figura 4 ilustra esse processo para $k = 10$, ou seja, validação cruzada de 10 $folds$. No primeiro $fold$, o primeiro subconjunto serve como conjunto de validação e os nove subconjuntos restantes servem como conjunto de treinamento. No segundo $fold$, o segundo subconjunto é o conjunto de validação e os subconjuntos restantes são o conjunto de treinamento e assim por diante.

Figura 4 – Validação cruzada de 10 $folds$.



Fonte: Kohavi R. (1995)

3.4 MÉTRICAS DE AVALIAÇÃO

Uma vez construído um modelo, a próxima etapa é avaliar o seu desempenho, ou seja, o quanto os escores produzidos pelo modelo conseguem distinguir os bons e os maus clientes, uma vez que, o que se deseja é identificar, previamente, esses grupos e tratá-los de forma distinta. Nesta seção, são descritas algumas medidas que são utilizadas para avaliar a performance do modelo de classificação binária, tais como: matriz de confusão, a estatística de Kolmogorov-Smirnov (KS), área da curva ROC, $precision$, $recall$ e a acurácia

3.4.1 Matriz de Confusão

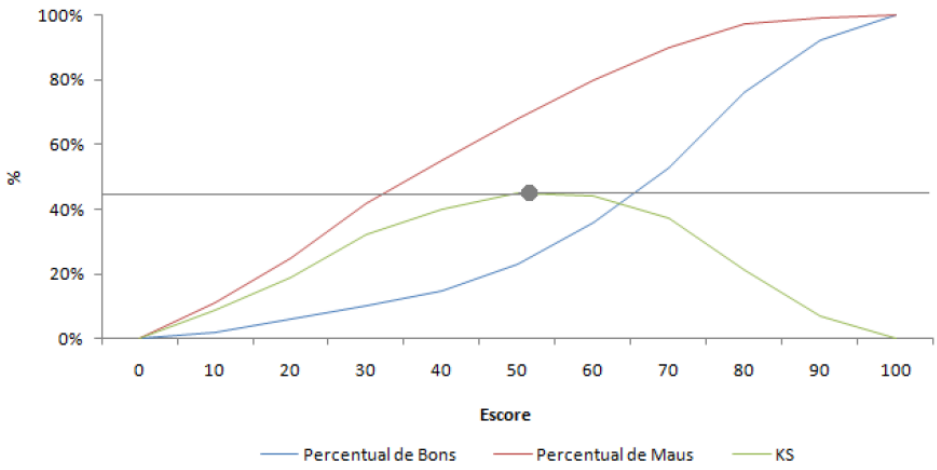
A capacidade preditiva de um modelo está relacionada com suas medidas de desempenho que podem ser calculadas a partir da matriz de confusão. A matriz de confusão é uma maneira de observar se o modelo está prevendo adequadamente os bons e maus clientes, a partir de um modelo de classificação.

Pode-se determinar para cada indivíduo i do conjunto de dados, um escore S_i, onde o escore é a chance do cliente ser bom pagador, ou seja, quanto maior for o escore, maior a probabilidade do indivíduo vim a pagar seu débito.

Suponha que um indivíduo seja classificado como bom, se $S_i > P_c$ e, caso contrário, classificado como mau, onde P_c é chamado de Ponto de Corte (PC) ou Ponto de Operação (PO).). Para obter a decisão sobre a estratégia de concessão do crédito, é preciso determinar um PC aceitável para a carteira, ou seja, um valor de escore para a tomada de decisão.

Assim, valores abaixo desse ponto de corte indicam que a operação foi negada, e valores acima desse ponto de corte, que a operação foi aprovada. Existem várias formas de determinar o melhor PC, dependendo muito do negócio e domínio de aplicação. Neste trabalho, foi definido como PC o ponto que maximiza a Estatística do teste de Kolmogorov-Sminorf (KS). A Figura 5 apresenta uma ideia de como é a análise a partir desse princípio.

Figura 5 – Analise do Ponto corte atráves do KS



Fonte: Autor

A curva azul representa a distribuição acumulada dos bons clientes e, a partir dos pontos de corte, é possível verificar o percentual de bons clientes rejeitados e aprovados. Já a curva vermelha representa a distribuição acumulada dos maus clientes e é possível também tirar as

mesmas conclusões. A curva verde representa os valores da estatística de KS e o ponto em destaque é o valor que a maximiza.

Fixando-se um ponto de corte podemos construir a matriz de confusão (Duda R., Hart P., Stork D., 2001) dada pela tabela 4.

Tabela 4 – Matriz de Confusão

	Real		
Previsto	Bom	Mau	Total
Bom	a	b	$a+b$
Mau	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Fonte: Duda R., Hart P., Stork D. (2001)

onde temos que,

- a: Representa o número de bons clientes, classifcados corretamente como bons, ou seja, Verdadeiro Positivo(VP);
- b: Representa o número de maus clientes, classifcados incorretamente como bons, ou seja, Falso-Positivo(FP);
- c: Representa o número de bons clientes, classifcados incorretamente como maus, ou seja, Falso-Negativo(FN);
- d: Representa o número de maus clientes, classifcados corretamente como maus, ou seja, Verdadeiro Negativo(VN).

Por meio da matriz de confusão é factível determinar as taxas de acertos do modelo, das quais podemos citar algumas métricas:

- $Sensibilidade$ ou $Recall$: É a proporção de Verdadeiros Positivos classificados entre todos da sua característica. Ou seja, avalia a capacidade do modelo de detectar com sucesso resultados classificados como positivos;

$$Recall = \frac{VP}{VP + FN} \tag{3.20}$$

- $Especificidade$: É a proporção de Verdadeiros Negativos classificados entre todos da sua característica. Ou seja, avalia a capacidade do modelo de detectar resultados negativos;

$$Especificidade = \frac{VN}{VN + FP} \tag{3.21}$$

- $Precision$: A quantidade de Verdadeiros Positivos sobre a soma de todos os valores positivos. Ou seja, avalia a capcidade da predição do modelo de detectar resultados positivos;

$$Precision = \frac{VP}{VP + FP} \tag{3.22}$$

- *F-score* (F1): É uma média harmônica calculada com base no $Precision$ e $Recall$;

$$F_1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.23}$$

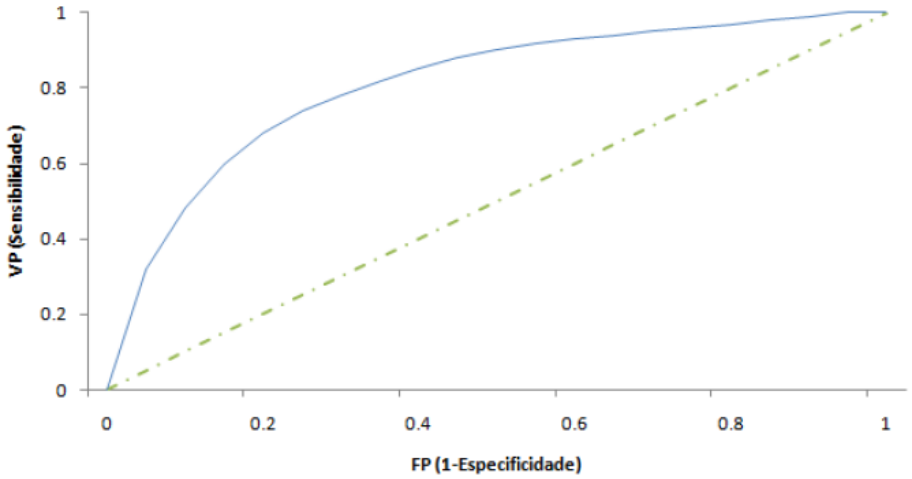
- Acurácia (ACC): É a proporção de acerto do modelo. Isto é, a capacidade total de acerto do modelo.

$$Precision = \frac{VP + FN}{VP + FP + VN + FN} \tag{3.24}$$

3.4.2 Curva ROC

A curva *Receiver Operating Characteristics* (ROC), na área de risco de crédito, é uma métrica bastante utilizada para a avaliação de modelos. A curva é baseada nos conceitos da sensibilidade e especificidade do modelo e varia os pontos de corte ao longo dos escores. Assim, a curva ROC é obtida, tendo no eixo x os valores de especificidade e no eixo y os valores de sensibilidade (Metz C.E., 1986). A Figura 6 ilustra a curva ROC.

Figura 6 – Curva ROC



Fonte: Autor

A área abaixo da curva ROC está associada ao poder discriminante de um teste de diagnostico. (Hanley J.A., 1988) relata que a área abaixo da curva ROC pode ser determinada através de:

- Métodos de resolução númerica, tipo regra de trapézio;
- Métodos estatísticos: relação com a estatística de Wilcoxon Mann-Witney e estimativa de máxima verosimilhança.

3.4.3 Estatística Kolmogorov Smirnov (KS)

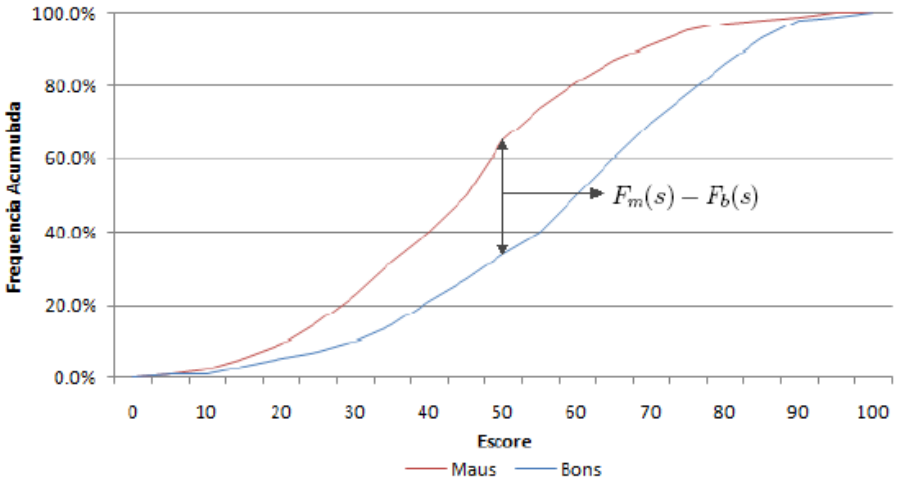
A estatística de *Kolmogorov-Smirnov* KS é descrita pela teoria estatística não-paramétrica para testar se as funções de distribuição de uma variável são iguais em dois grupos (Connover W.J., 1999). Em modelos de *credit scoring*, a estatística é utilizada para para comparar a função de distribuição empírica dos escores, dos grupos de bons e maus clientes.

Logo o KS mede a máxima separação entre a frequência acumulada dos bons clientes, definido por $F_b(s)$ e a frequência acumulado dos maus, definida por $F_m(s)$. Em modelos bem ajustados, os bons clientes são atribuídos, em sua maioria, altas probabilidades, enquanto os maus clientes, a maior concentração estão nas baixas probabilidades. A estatística de KS é definida por:

$$KS = max_s(|F_m(s) - F_m(s)|), 0 \leq KS \leq 100. \tag{3.25}$$

Na Figura 7 é apresentado um exemplo de cálculo da estatística KS. A maior separação entre as distribuições acumuladas de bons e maus representa o valor do KS.

Figura 7 – Estatística de Kolmogorov-Smirnov(KS)



Fonte: Connover W.J. (1999)

3.4.4 Diferença entre Taxas de Inadimplência (DTI)

Tomazela (Tomazela S.M.O., 2007) cita que uma característica importante em um modelo de *credit scoring* é a capacidade de ordenação da taxa de inadimplência que o modelo proporciona de acordo com a variação do escore. Com isso, pode-se criar indicadores baseados na capacidade de ordenação de um modelo.

Aplicando a taxa de inadimplência em cada faixa de intervalo, previamente definido, há uma diferença entre as taxas no primeiro intervalo em relação ao último intervalo, obtendo assim uma capacidade de separação entre os bons e maus do modelo proposto.

Para calcular a Diferença entre taxas de inadimplência (DTI), primeiro, define o tipo de intervalo das faixas de escores, se vai ser decil, quintil e demais. Após a definição e construído

os intervalos, calcula-se a frequência total de observações e as frequências totais de bons e maus para cada intervalo.

Tabela 5 – Calculo da taxa de inadimplência por intervalo de escore

Faixa de Escore (Quintil)	Total	Bons	Maus	Taxa de Inadimplência (tm)
Intervalo_1	t_1	b_1	m_1	$tm_1 = \frac{m_1}{t_1} * 100$
Intervalo_2	t_2	b_2	m_2	$tm_2 = \frac{m_2}{t_2} * 100$
Intervalo_3	t_3	b_3	m_3	$tm_3 = \frac{m_3}{t_3} * 100$
Intervalo_4	t_4	b_4	m_4	$tm_4 = \frac{m_4}{t_4} * 100$
Intervalo_5	t_5	b_5	m_5	$tm_5 = \frac{m_5}{t_5} * 100$

Fonte: Tomazela S.M.O. (2007)

em que:

- t_k é o total de clientes no $k - ésimo$ intervalo;
- b_k é o total de $bons$ clientes no $k - ésimo$ intervalo;
- m_k é o total de $maus$ clientes no $k - ésimo$ intervalo;
- tm_k é a taxa de maus em relação ao total para o $k - ésimo$ intervalo.

Para a Tabela 5 o $k = 1, 2, 3, 4, 5$. A diferença entre as taxas de inadimplência (DTI) é dada por

$$DTI(Min(Intervalo_k); Max(Intervalo_k)) = Min(tm_k) - Max(tm_k) \tag{3.26}$$

No exemplo da Tabela 5 o DTI seria dado por $DTI(1; 5) = tm_1 - tm_5$, ou seja, a taxa de inadimplência do 1° quintil subritaído da taxa do 5° (último) quintil. E quanto maior o valor observado, melhor a separação do modelo.

3.5 ESTATÍSTICA NÃO-PARAMÉTRICA

Em todos os problemas de inferência estatística considerados, assume-se que a distribuição da variável aleatória que está sendo amostrada seja conhecida, ao menos, para alguns parâmetros. Na prática, a forma funcional da distribuição é raramente ou nunca conhecida.

De modo geral, um procedimento não paramétrico é um procedimento estatístico que tem certas propriedades desejáveis, que se sustentam sob suposições em relação às populações das quais os dados são obtidos.

Segundo Hollander (Hollander M., Wolfe D.A., 2014), o desenvolvimento rápido e contínuo de métodos de estatística não paramétrica, ao longo das décadas, se deve a uma série de vantagens técnicas, tais como:

1. Os métodos não paramétricos requerem poucas suposições sobre as populações das quais os dados são obtidos. Em particular, os procedimentos não paramétricos renunciam à suposição tradicional de que as populações são normais;

2. Os métodos permitem que o usuário obtenha $P-valor$ exatos para testes, probabilidades de cobertura exatas para intervalos de confiança, taxas de erro de experimentos para procedimentos de comparação múltipla e probabilidades de cobertura exatas para bandas de confiança sem depender de suposições de que as populações sejam normais;

3. Os métodos não paramétricos não são relativamente sensíveis a observações do tipo $outliers$;

4. Procedimentos não paramétricos geralmente são fáceis de entender;

5. Os métodos não paramétricos são aplicados em muitas situações em que os procedimentos da teoria normal não podem ser utilizados. Muitos métodos não paramétricos requerem apenas os $ranks$ das observações, ao invés da magnitude real das observações;

6. Com os avanços da computação os métodos tornaram-se muito simples e velozes na sua construção.

3.5.1 Teste de Kruskal-Wallis

O teste de Kruskal-Wallis, proposto por Kruskal e Wallis (Kruskal W.H., Wallis W.A., 1952), é um teste não paramétrico baseado em $rank$, que pode ser usado para determinar se existem diferenças estatisticamente significativas entre dois ou mais grupos de uma variável independente e uma variável dependente contínua ou ordinal. É considerada a alternativa não paramétrica à ANOVA e uma extensão do teste U de Mann-Whitney para permitir a comparação de mais de dois grupos independentes.

Os dados consistem em $N \sum_{j=1}^{k} n_j$ observações, com n_j observações do $j - ésimo$ tratamento, $j = 1, ..., k$. Na tabela 6 é detalhado um esboço dos dados.

Tabela 6 – Detalhamento do delineamento amostral

Tratamento			
1	2	...	k
X_{11}	X_{12}	...	X_{1k}
X_{21}	X_{22}	...	X_{2k}
.	.		.
.	.		.
.	.		.
$X_{n_1 1}$	$X_{n_2 2}$	...	$X_{n_k k}$

Fonte: Kruskal W.H., Wallis W.A. (1952)

Seja N variáveis aleatórias $\{X_{1j}, X_{2j}, ..., X_{n_j j}\}$, $j = 1, 2, ...k$, mutualmente independentes. Para cada $j \in \{1, ..., k\}$, as n_j variáveis aleatórias $\{X_{1j}, X_{2j}, ..., X_{n_j j}\}$ são uma amostra aleatória de uma distribuição contínua com função de distribuição F_j.

A função de distribuição $F_1, ..., F_k$ estão conectadas através da relação

$$F_j(t) = F(t - \tau_j) \tag{3.27}$$

para $j = 1, ..., k$, onde F é uma função de distribuição para uma distribuição contínua com mediana desconhecida θ e τ_j o efeito de tratamento desconhecido para a $j - ésima$ população.

$$X_{ij} = \theta + \tau_j + e_{ij} \tag{3.28}$$

para $i = 1, ..., n_j$ e $j = 1, ..., k$. Onde θ é a mediana geral, τ_j é o efeito do tratamento j, e os N formam uma amostra aleatória de uma distribuição contínua com mediana zero (0).

A hipótese a ser testada é a de não existe diferença entre os efeitos dos tratamentos $\tau_1, ..., \tau_k$, contra a hipótese alternativa de que pelo menos dois dos efeitos do tratamento não são iguais. Ou seja,

$$\begin{cases} H_0 : \tau_i = \cdots = \tau_j \\ H_1 : \tau_i \neq \cdots \neq \tau_j \end{cases} \tag{3.29}$$

onde $i, j \in [1, ..., k]$, com $(i \neq j)$

Para calcular a estatística de Kruskal-Wallis (H), primeiro combinamos todas as N observações de k amostras e ordene-as do menor para o maior. r_{ij} Denota a classificação de X_{ij} nesta classificação conjunta e defina

$$R_j = \sum_{i=1}^{n_j} r_{ij} \tag{3.30}$$

onde pode escrever R_j como $R_j = \frac{R_j}{n_j}$. A estatística de Kruskal–Wallis (H) pode ser definida por

$$H = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{Rj^2}{n_j} \right) - 3(N+1) \tag{3.31}$$

no nível α de significância, rejeita H_0 se

$$H \geq h_\alpha \tag{3.32}$$

caso contrário, não rejeita. Onde a constante h_α é escolhida para tornar a probabilidade de erro tipo I igual a α. Se a hipótese nula não for verdadeira, então H deverá tomar valores grandes e portanto, rejeitamos H_0 em a favor de H_1.

3.5.2 Teste de Nemenyi

O teste de Nemenyi (Nemenyi P., 1963) é um teste $post - hoc$, ou seja, é feita uma comparação múltipla com o erro padrão ajustado para amostras iguais, que é usado após a aplicação de teste não paramétricos com três ou mais fatores. O teste consiste em fazer comparações em pares, grupos dois a dois, com o intuito de verificar qual dos fatores que diferem entre si.

Os dados consistem em $N \sum_{j=1}^{k} n_j$ observações, com n_j observações do $j - ésimo$ tratamento, $j = 1, ..., k$. Seja N variáveis aleatórias $\{X_{1j}, X_{2j}, ..., X_{n_j j}\}$, $j = 1, 2, ...k$, mutualmente independentes.

A hipótese a ser testada é a de não existe diferença entre os efeitos dos tratamentos $\tau_1, ..., \tau_k$, contra a hipótese alternativa de que grupos dois a dois. Ou seja,

$$\begin{cases} H_0 : \tau_i = \tau_j \\ H_1 : \tau_i \neq \tau_j \end{cases} \tag{3.33}$$

onde $i, j \in [1, ..., k]$, com $(i \neq j)$.

Os postos das N observações ordenadas de acordo com o tamanho $X(1), ..., X(N)$, são tomados assinalando 1 para $X(1)$, 2 para X(2) e assim em diante. Onde R_{ij} é dado pela equação 3.30. O teste H_0 rejeita quando:

$$Hk - 1 \geq h_{k-1}^{\alpha} \tag{3.34}$$

em que:

$$H_{k-1} = \left(\frac{12}{N(N+1)} \sum_{j=1}^{k} \frac{Rj^2}{n_j} \right) - 3(N+1) \tag{3.35}$$

para N suficientemente grande h_{k-1} segue uma distribuição χ^2_{k-1}, e para N pequeno existe tabelas em (Hollander M., Wolfe D.A., 2014).

4 EXPERIMENTAÇÃO E ANALISE DOS RESULTADOS

4.1 METODOLOGIA DE EXPERIMENTAÇÃO

Como metodologia para ter um plano completo para a realização de um projeto de mineração de dados, foi utilizado o *Cross Industry Standard for Data Mining* (CRISP-DM). O CRISP-DM é um modelo de mineração de dados não proprietário, neutro, documentado e disponível livremente. É um modelo específico de processo que descreve abordagens comumente usadas em mineração de dados para resolver problemas desse domínio (Shearer C., 2000). O mesmo é dividido em seis fases:

1. **Entendimento do negócio**: Talvez a fase mais importante de qualquer projeto de mineração de dados, a fase inicial de entendimento do negócio, concentra-se em entender os objetivos do projeto a partir de uma perspectiva de negócios, convertendo esse conhecimento numa definição de problema de mineração de dados e, em seguida, desenvolvendo um plano preliminar projetado para atingir os objetivos.

2. **Entendimento dos dados**: A fase começa com uma coleta de dados inicial. O analista, então, passa a aumentar a familiaridade com os dados, identificar problemas de qualidade de dados, descobrir insights iniciais sobre os dados ou detectar subconjuntos interessantes para formar hipóteses sobre informações ocultas. A fase de compreensão dos dados envolve quatro etapas, incluindo a coleta dos dados iniciais, a descrição dos dados, a exploração dos dados e a verificação da qualidade dos dados.

3. **Preparação dos dados**: A fase abrange todas as atividades para construir o conjunto de dados final. As tarefas incluem seleção da base, bem como transformação e limpeza de dados para modelagem. As cinco etapas na preparação de dados são: seleção de dados, limpeza de dados, construção de dados, integração de dados e formatação de dados.

4. **Modelagem**: Nesta fase, diversas técnicas de modelagem são selecionadas/aplicadas e seus parâmetros são calibrados. Algumas técnicas têm requisitos específicos sobre a forma de dados. Portanto, podemos voltar à fase de preparação de dados para adequação dos dados. As etapas de modelagem incluem a seleção da técnica de modelagem, a geração do projeto de teste, a criação de modelos e a avaliação de modelos.

5. **Avaliação**: Nesta fase, é importante avaliar detalhadamente o modelo e revisar a construção do modelo para ter certeza de que ele atinge adequadamente os objetivos de negócios. No final desta fase do projeto, deve-se decidir exatamente como usar os resultados. Os principais passos aqui são a avaliação dos resultados e a revisão do processo.

6. **Implantação**: Por fim, após a construção e avaliação do modelo, este deve ser disponibilizado para poder ser aplicado. Dependendo dos requisitos, a fase pode ser tão simples quanto gerar um relatório ou tão complexa quanto a implementação mais robusta do modelo em toda a empresa.

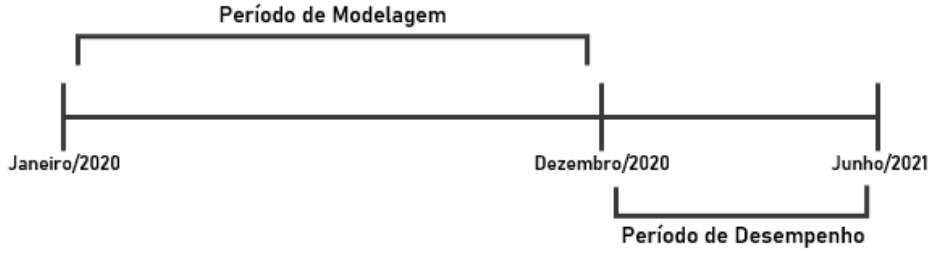
4.2 CARACTERIZAÇÃO E DELINEAMENTO AMOSTRAL

O objetivo com os experimentos é duplo. Primeiro, comparamos o desempenho de nossos modelos propostos com uma gama de técnicas que estão presentes no estado da arte em inferência dos rejeitados para *credit scoring*, incluindo duas técnicas de agrupamento (*K-Means* e *Deep embedded clustering*) em cenários realistas, preservando as taxas de aprovação originais em três conjuntos de dados reais. Segundo, para entender melhor o comportamento dos modelos de inferência dos rejeitados, foi testado o desempenho do modelo em diferentes cenários variando a taxa de aprovados e rejeitados. Durante o processo de modelagem, foi usado o modelo de aprendizado de máquina supervisionado: regressão logística.

Os modelos de *credit scoring* são desenvolvidos a partir de bases históricas do comportamento do cliente. Definimos como período de modelagem o período em que vamos observar os clientes na data em que ele solicitou o crédito. Já o período de desempenho é o período para determinar o alvo do cliente. A Definição da variável resposta (ALVO), geralmente, é fixada em dias ou meses, de acordo com a necessidade para a solução do problema.

A Figura 8 mostra um exemplo de delineamento do período dos dados.

Figura 8 – Período dos dados para um modelo de *Credit Scoring*



Fonte: Autor

Para definição do delineamento amostral, é importante que o planejamento e definição do problema sejam observadas. As bases de dados utilizadas nos experimentos para o desenvolvimento do modelo de *credit scoring* é formada por clientes que foram às empresas solicitar algum tipo de crédito.

4.2.1 Conjunto de Dados

Três conjuntos de dados de crédito de diversos setores foram usados para demonstrar de forma abrangente os resultados dos métodos propostos. Na tabela 7 é detalhado as informações dos conjuntos de dados.

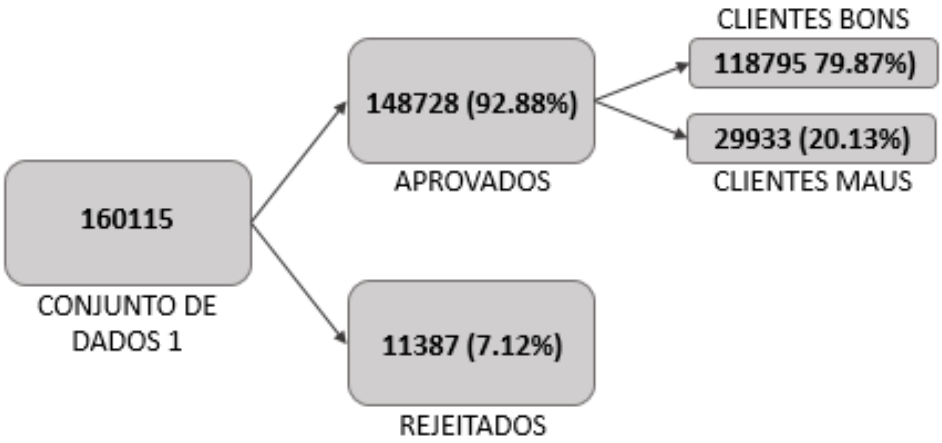
Tabela 7 – Informações/Definições dos experimentos

Conjunto de dados	I	II	III
Periodo dos dados	Jul/2018-Dez/2018	Jun/20-Nov/20	Ago/2019-Jan/2020
Tamanho da base	160115	219822	176904
Taxa de Aprovados	92.88%	78.23%	63.11%
Taxa de rejeitados	7.12%	21.77%	36.89%

Fonte: Autor

O conjunto de dados I é derivado de um varejista do setor de supermercado, onde é oferecido produtos alimentícios, produtos domésticos e eletroeletrônicos. Lá, a empresa oferece cartão de crédito, conhecido como $private\ label$, para seus clientes comprarem os produtos destacados. Os dados estudados estão no período de Julho/2018 até Dezembro/2018. Neste período, foram observados cerca de 160 mil proponentes interessados no cartão, dos quais cerca de 92% foram aprovados, ou seja, tiveram acesso ao cartão de crédito. Na figura 9 mostra como é a distribuição dos dados.

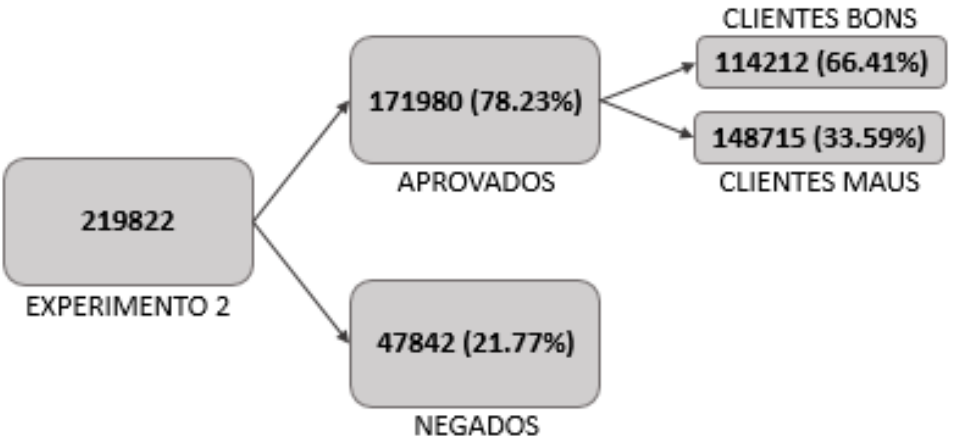
Figura 9 – Esquema da distribuição dos dados dos Conjunto de dados 1



Fonte: Autor

No Conjunto de dados II, os dados são de uma instituição financeira que tem uma operação de cartão de crédito. O cartão é uma forma de empréstimo com prazo de pagamento mensal que a instituição disponibiliza. A financeira oferta o cartão e o cliente recebe um limite de crédito para fazer compras de bens e serviços. Os dados estudados estão no período de Junho/2020 até Novembro/2020. Neste período, foram observados cerca de 219 mil proponentes interessados no cartão, onde cerca de 78% foram aprovados, ou seja, tiveram acesso ao cartão de crédito. Na figura **??** mostra como é a distribuição dos dados.

Figura 10 – Esquema da distribuição dos dados dos Conjunto de dados 2

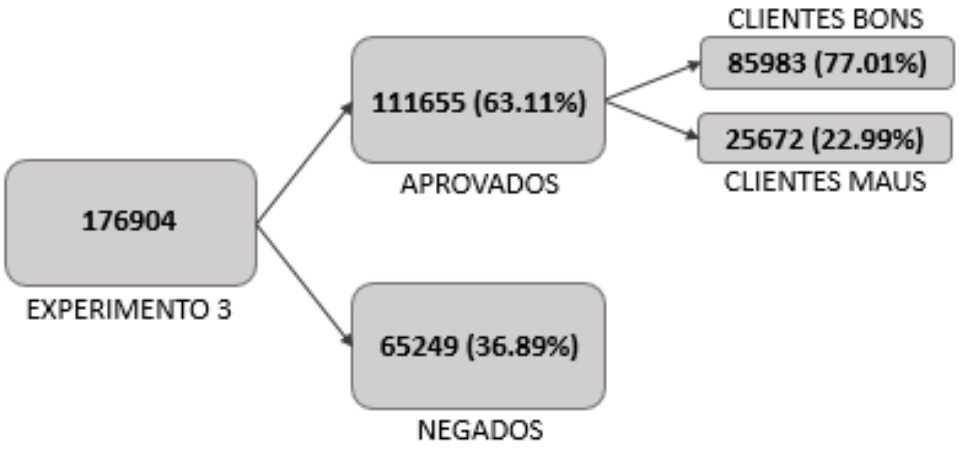


Fonte: Autor

Já no Conjunto de dados III, os dados são oriundos de um banco digital, o banco oferece cartão 100% digital para seus clientes. Geralmente, o diferencial dessas financial technology (Fintechs) é a facilidade de comunicação com os clientes e acesso aos bancos e os serviços gratuitos ofertados. Os dados estudados estão no período de Agosto/2019 até Janeiro/2020.

Neste período, foram observados cerca de 76 mil proponentes ao cartão, do qual cerca de 63% foram aprovados, ou seja, tiveram acesso ao cartão digital. Na figura **??** mostra como é a distribuição dos dados.

Figura 11 – Esquema da distribuição dos dados dos Conjunto de dados III



Fonte: Autor

Os diversos conjuntos de dados são constituídos das informações no momento de solicitação do crédito, ou seja, o cliente vai até a empresa que destina o cartão e solicita o mesmo. Neste momento, as empresas adquirem uma série de variáveis do proponente para decidir se deve conceder o cartão ou não. Essas variáveis observam uma série de características do proponente. Na tabela 8, é apresentada as diversas informações disponíveis dos proponentes no momento da concessão do crédito.

Tabela 8 – Definições e informações das características dos dados

Características	Definição
Demográficas	Características populacionais e sociais
Geolocalizadas	Localização e tipos de estabelecimentos próximos
Informações financeiro	Informações sobre comportamento financeiro
Capital Humano	Visão de educação por perfis da população, socioeconômicas
Ocupacional	Perfil de ocupação, com informações socioeconômicas
Poder de Compra	Comportamento financeiro e informações relacionadas à renda

Fonte: Autor

De modo geral, as empresas estão observando os proponentes do cartão em seis características. Em características demográficas, informações de idade, gênero e classe social são estudadas.

Para a característica geolocalizada, as empresas observam como é o entorno de onde o proponente reside, por exemplo, se sua região é de muita área verde (parques) e muitos pontos comerciais (bares, lojas, galerias, $shopping$).

Para a informação financeira e poder de compra, as empresas observam se o proponente tem um cartão em outra instituição do tipo $prime$, qual sua renda, se a renda do proponente é maior do que a região que ele reside, se o proponente está há muito tempo empregado.

Na característica capital humano, as instituições tendem a observar como está o modo educacional/populacional da região que o proponente reside. Por exemplo: se a região em que ele reside é composta por pessoas de alta instrução educacional ou não ou se nas residências da sua região o número de residentes é elevado ou não.

Na característica ocupacional, é observado como está o proponente de modo socioeconômico, por exemplo, qual é o rendimento médio da região em que o proponente reside, se sua região de residência é composta por pessoal empregadas (taxa de população ativa no mercado).

4.3 RESULTADOS EXPERIMENTAIS

Nesta seção, os métodos de inferência dos rejeitados foram aplicados em três conjuntos de dados diferentes, para cada conjunto o processo de validação cruzada foi considerado e com a utilização de 10 $folds$.

Foram avaliadas diversas métricas de performance. Tais como: teste de KS, a área da curva ROC (AUC), a acurácia, o $precision$, o $recall$, o F_1 e o DTI que foram realizados para testar a capacidade dos métodos propostos para resolver o problema de inferência dos rejeitados. Os métodos avaliados são:

- Tradicional: Método comumente utilizado, onde é utilizado no processo de modelagem apenas os proponentes aprovados;
- Reclassificação: (Seção 3.1.4.1)
- $Augmentation$ (Seção 3.1.4.2)
- $Cluster$ (Seção 3.1.4.3)
- Parcelamento (Seção 3.1.4.4)

- DEC (Seção 3.1.4.5)

4.3.1 Experimento I

Nesta seção, serão apresentadas as análises dos resultados alcançados pelos métodos nas diversas métricas avaliadas para o Experimento I, o cartão de crédito para um varejista do setor de supermercado. Nas tabelas 9 e 10, a seguir, poderemos observar os valores alcançados pelos métodos para a métrica de KS e ROC para cada $fold$ e suas estatísticas: Mínimo (Min), Média, Máximo (Max) e Desvio-padrão (DP).

Tabela 9 – Resultados do KS para os 10 $folds$ do Experimento I (%)

$Fold$	Tradicional	Reclassificação	$Augmentation$	Cluster	Parcelamento	DEC
1	19.48	17.49	16.94	18.32	19.79	**21.13**
2	**19.55**	16.62	16.84	18.41	19.52	19.01
3	**20.52**	17.50	19.38	18.34	18.28	19.61
4	19.51	17.45	18.16	18.47	**20.88**	18.28
5	19.63	17.11	19.00	17.00	19.88	**21.88**
6	19.75	16.29	19.01	18.20	**20.19**	19.96
7	18.43	17.75	18.02	16.22	**19.99**	18.02
8	19.22	16.53	18.52	16.63	17.99	**20.04**
9	20.46	18.03	18.62	18.71	**20.71**	18.99
10	**19.61**	17.67	18.10	19.10	19.14	19.41
Min	18.43	16.29	16.84	16.22	17.99	18.02
Média	19.52	17.24	18.26	17.94	19.64	19.63
Max	20.52	18.03	19.38	19.10	20.88	21.88
DP	0.59	0.58	0.85	0.96	0.94	1.19

Fonte: Autor

Tabela 10 – Resultados do ROC para os 10 $folds$ do Experimento I (%)

$Fold$	Tradicional	Reclassificação	$Augmentation$	Cluster	Parcelamento	DEC
1	62.73	61.90	61.76	61.85	63.37	**65.67**
2	**62.66**	61.46	61.64	62.60	62.30	60.66
3	64.18	61.74	62.68	61.24	63.25	**67.85**
4	**63.41**	61.94	62.17	61.84	**63.41**	60.53
5	63.07	61.82	63.06	61.93	62.43	**66.70**
6	62.46	60.93	62.34	61.76	**63.84**	63.15
7	62.92	61.47	61.88	61.57	**62.99**	60.76
8	62.93	60.93	62.01	61.32	63.06	**65.23**
9	**63.78**	61.67	62.48	61.79	63.22	60.98
10	63.07	61.90	61.89	61.11	63.40	**64.30**
Min	62.46	60.93	61.64	61.11	62.30	60.53
Média	63.12	61.58	62.19	61.70	63.13	63.58
Max	64.18	61.94	63.06	62.60	63.84	67.85
DP	0.53	0.38	0.45	0.43	0.46	2.75

Fonte: Autor

Na tabela 9 e 10, podemos observar a variação dos resultados para todos os métodos propostos para cada $fold$. Nota-se que os métodos de parcelamento, DEC e tradicional tiveram desempenhos superiores para a maioria dos $folds$.

Entretanto para a métrica KS (Tabela 9), dos 10 $folds$ o método de Parcelamento teve resultados superiores para 4 $folds$, o DEC 3 $folds$ e o Tradicional 3 $folds$. Já para a métrica ROC (Tabela 10), o DEC teve melhor resultados em 5 $folds$, o Parcelamento em 3 $folds$ e por fim o Tradicional foi superior em 3 $folds$.

Sendo assim, foi realizado o teste de Kruskal-Wallis (4.1), para verificar a existência de diferenças estatisticamente significativas entre os métodos estudados. Todavia o teste de Kruskal-Wallis será avaliada para as métricas de KS e ROC uma vez que são as métricas mais observadas nos cenários acadêmico (ROC) e do mundo de negócio (KS)

$$\begin{cases} H_0 : \text{Todos os métodos são iguais para a métrica avaliada} \\ H_1 : \text{Os métodos são diferentes para a métrica avaliada} \end{cases} \tag{4.1}$$

O p-valor do teste de Kruskal-Wallis foi inferior à 0,0001 (menor que um nível de significância de 0,05), nas métricas KS e ROC, nos levando a rejeitar a hipótese nula (H_0) de que existe diferença entre os métodos de inferência dos rejeitados. Após a identificação de que os métodos são diferentes, é importante avaliar se existe alguma diferença entre os métodos dois

a dois. Com isso, foi aplicado o Teste comparações múltiplas de Nemenyi, buscando identificar as diferenças nos métodos individualmente.

Tabela 11 – Resultados do Teste de Nemenyi (p-valor) do Experimento I para a métrica KS

Método	Tradicional	Reclassificação	$Augmentation$	$Cluster$	Parcelamento
Reclassificação	0.00013	-	-	-	-
$Augmentation$	0.07874	0.48734	-	-	-
Cluster	0.02742	0.72978	0.99908	-	-
Parcelamento	1	0.00018	0.09584	0.03459	-
DEC	0.99901	0.00066	0.19201	0.08006	0.99967

Fonte: Autor

Na tabela 11 Avaliando os p-valores das comparações duas a duas do teste de Nemenyi, podemos ver que o teste só revelou haver diferença significativa com um nível de 5% (α =0.05) entre os métodos (Caselas verdes-métrica KS):

- Tradicional e Reclassificação (p-valor=0.00013)
- Tradicional e $Cluster$ (p-valor=0.02742)
- Reclassificação e Parcelamento (p-valor=0.00018)
- Reclassificação e DEC (p-valor=0.00066)
- $Cluster$ e Parcelamento (p-valor=0.03459)

Avaliando num nível de significância de 10% (α =0.1) os métodos que revelaram diferença significativa foram (Caselas laranjas-métrica KS):

- Tradicional e $Augmentation$ (p-valor=0.07874)
- $Augmentation$ e Parcelamento (p-valor=0.09584)
- $Cluster$ e DEC (p-valor=0.08006)

O teste não apresentou evidências de diferença significativa entre os métodos para as demais combinações dois a dois (Caselas brancas-métrica KS).

Tabela 12 – Resultados do Teste de Nemenyi (p-valor) do Experimento I para a métrica ROC

Método	Tradicional	Reclassificação	$Augmentation$	$Cluster$	Parcelamento
Reclassificação	0.00444	-	-	-	-
$Augmentation$	0.32547	0.62372	-	-	-
Cluster	0.01032	0.9999	0.77112	-	-
Parcelamento	0.9999	0.00268	0.25486	0.00645	-
DEC	0.8785	0.12877	0.94104	0.2163	0.81616

Fonte: Autor

Na tabela 12 Avaliando os p-valores das comparações duas a duas do teste de Nemenyi, podemos ver que o teste só revelou haver diferença significativa com um nível de 5% (α =0.05) entre os métodos (Caselas verdes-métrica ROC):

- Tradicional e Reclassificação (p-valor=0.00444)
- Tradicional e $Cluster$ (p-valor=0.01032)
- Reclassificação e Parcelamento (p-valor=0.00268)
- $Cluster$ e Parcelamento (p-valor=0.00645)

Para as demais o teste não apresentou evidências de diferença significativa entre os métodos para as demais combinações dois a dois (Caselas brancas-métrica ROC).

Na Tabela 13, são apresentadas as pontuações médias dos métodos e os desvios padrão, com os resultados em negrito indicando o melhor desempenho de cada métrica de avaliação.

Conforme mostrado na Tabela 13, o método de DEC teve um desempenho superior em quase todas as métricas de avaliação, 75% dos métricas foram favoráveis para o método citado. Com exceção do indicador de KS, onde o método de parcelamento foi superior, a média dos dois métodos são bem semelhantes. O método DEC tem uma acurácia (ACC) e F_1 bem superior aos demais métodos, uma diferença próxima a 20 pontos percentuais para ambas as métricas.

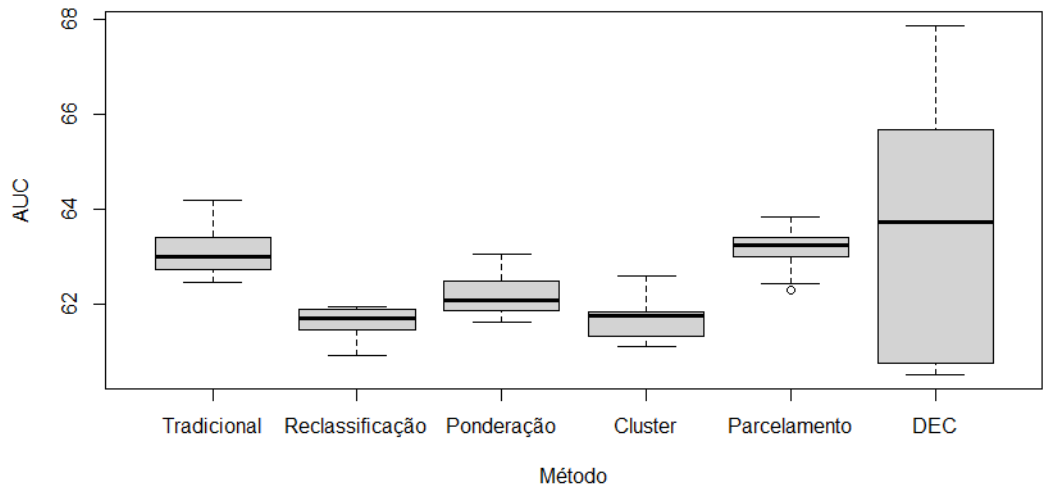
Tabela 13 – Resultados dos métodos do Experimento I (%)

Método	Métrica			
	AUC	KS	ACC	F_1
Tradicional	63.12 ± 0.53	19.52 ± 0.59	58.47 ± 0.30	68.92 ± 0.29
Reclassificação	61.58 ± 0.38	17.24 ± 0.58	57.59 ± 0.25	66.49 ± 0.27
Augmentation	62.19 ± 0.45	18.26 ± 0.85	57.70 ± 0.42	68.24 ± 0.41
Cluster	61.70 ± 0.43	17.94 ± 0.96	57.81 ± 0.43	68.42 ± 0.39
Parcelamento	63.13 ± 0.46	**19.64 ± 0.94**	58.63 ± 0,41	69.12 ± 0.36
DEC	**63.58 ± 2.75**	19.63 ± 1.19	**78.58 ± 0.18**	**87.75 ± 0.11**

Fonte: Autor

Na Figura 12, o *boxplot* apresenta a dispersão das métricas estudadas. Curiosamente, o método em destaque (DEC), para todas as métricas em variação e desvio padrão, é bem maior do que os demais métodos. Nota-se que a variação (desvio padrão) da AUC dos métodos está em torno de 0.4 e 0.5, exceto o DEC com desvio padrão de 2.75. Essa variação pode ser devido a característica desse experimento, a taxa de clientes aprovados é bem superior (92.8%) quando comparado com os demais experimentos. Nas Tabelas 19 e 25 o desvio padrão não é elevado como neste experimento.

Figura 12 – Variação dos métodos do experimento 1 (AUC)



Fonte: Autor

Tabela 14 – ExpI: Inadimplência por intervalo de quintil

Faixa Escore (Quintil)	Tradicional	Reclassificação	*Augmentation*	Cluster	Parcelamento	DEC
Q1	31.21%	31.02%	30.22%	29.47%	31.26%	31.80%
Q2	24.25%	24.14%	23.87%	24.00%	24.24%	24.37%
Q3	19.18%	19.30%	19.44%	20.09%	19.17%	19.14%
Q4	15.40%	15.59%	15.88%	15.84%	15.41%	15.28%
Q5	10.58%	10.58%	11.14%	11.23%	10.55%	10.03%
DTI(1;5)	20.63%	20.44%	19.08%	18.24%	20.72%	21.77%
	Fonte: Autor					

A Tabela 14 mostra a medida da diferença entre as taxas de inadimplência nos intervalos extremos definidos pelos quintis (DTI). Nela, são apresentados os resultados dos métodos propostos.

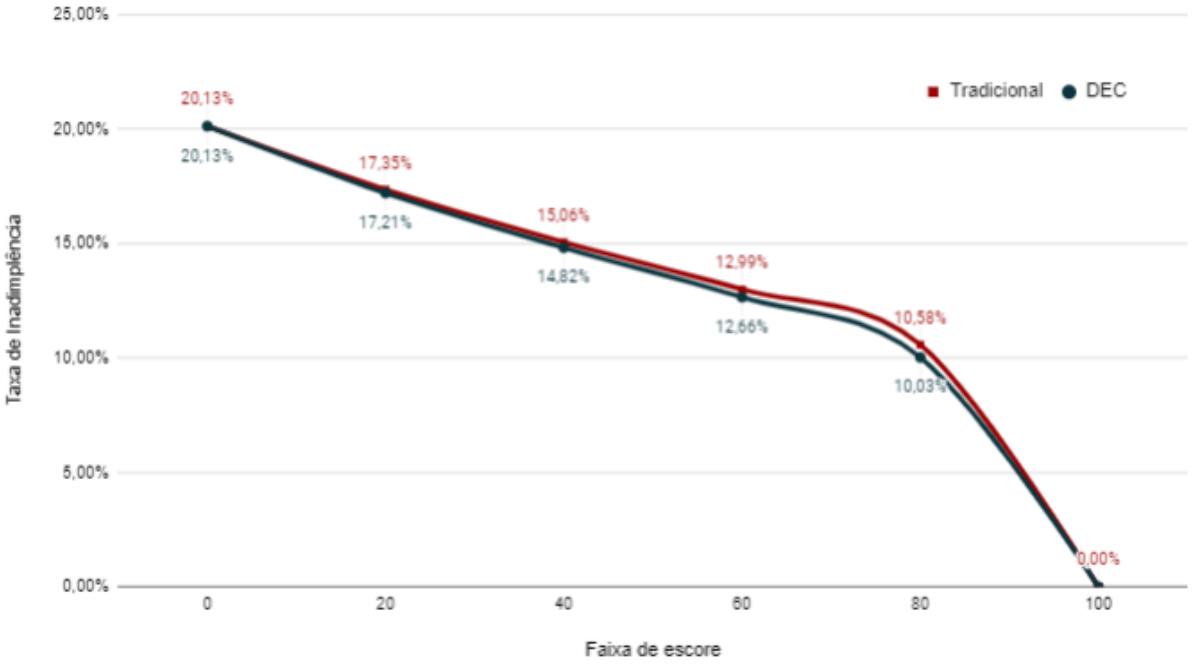
No método DEC, a sua diferença foi de 21,77%. Já para os demais métodos, seus DTI foram inferiores a 21%. Mostrando, assim, que o método DEC tem menor sensibilidade no indicador, proporcionando um melhor suporte na estratégia de uso do modelo.

Na Figura 13, observa-se um gráfico onde é feito um processo de simulação de aprovação para estimar qual será a projeção de inadimplência após a definição desse ponto de corte. No mesmo gráfico, os métodos avaliados foram o tradicional e o DEC, uma vez que o método DEC teve melhor performance em praticamente todos os indicadores previamente apresentados.

Supondo um ponto de corte no escore 40, ou seja, aprovando todos os proponentes que tiverem escore superior a 40, teríamos uma taxa de inadimplência para o método tradicional de 15,06% e, para o DEC, de 14,82%. Isso significa que o DEC teria uma redução de inadimplência de 2%, mantendo a mesma aprovação comparado com o tradicional.

Com o aumento do ponto de corte, podemos observar que o DEC aumenta o impacto na redução da inadimplência: 3% para o ponto de corte de 60, e 5% para o de 80.

Figura 13 – Simulação aprovação e inadimplência por ponto de corte - Escore percentil



Fonte: Autor

4.3.2 Experimento II

Nesta seção, serão apresentadas as análises dos resultados alcançados pelos métodos nas diversas métricas avaliadas para o Experimento II, uma operação que oferta cartão de crédito para seus proponentes. Nas tabelas 15 e 16, a seguir, poderemos observar os valores alcançados pelos métodos para a métrica de KS e ROC para cada $fold$ e suas estatísticas: Mínimo (Min), Média, Máximo (Max) e Desvio-padrão (DP).

Tabela 15 – Resultados do KS para os 10 $folds$ do Experimento II (%)

$Fold$	Tradicional	Reclassificação	$Augmentation$	Cluster	Parcelamento	DEC
1	16.22	13.09	16.71	14.50	15.10	**19.82**
2	16.44	14.50	15.79	15.88	16.69	**19.89**
3	16.67	13.12	14.44	14.10	15.80	**19.48**
4	15.11	12.50	16.05	13.89	16.28	**20.53**
5	17.03	13.13	16.85	13.33	16.58	**20.04**
6	15.84	15.13	15.32	13.98	17.20	**19.39**
7	16.70	12.04	15.18	13.55	16.61	**20.35**
8	14.48	13.26	15.93	14.31	16.77	**19.77**
9	15.14	13.35	14.22	13.83	15.35	**20.06**
10	16.09	13.25	17.34	13.50	17.33	**19.64**
Min	14.48	12.04	14.22	13.33	15.10	19.39
Média	15.97	13.34	15.78	14.09	16.37	19.90
Max	17.03	15.13	17.34	15.88	17.33	20.53
DP	0.83	0.89	1.02	0.73	0.74	0.36

Fonte: Autor

Tabela 16 – Resultados da ROC para os 10 $folds$ do Experimento II (%)

$Fold$	Tradicional	Reclassificação	$Augmentation$	Cluster	Parcelamento	DEC
1	60.70	58.64	61.03	59.57	60.78	**63.08**
2	60.85	59.57	60.42	60.48	61.52	**63.13**
3	61.00	58.66	59.53	59.31	61.09	**62.85**
4	59.97	58.25	60.59	59.17	61.03	**63.55**
5	61.24	58.67	61.12	58.80	61.39	**63.23**
6	60.46	59.99	60.11	59.23	61.71	**62.80**
7	61.02	57.95	60.02	58.95	61.66	**63.43**
8	59.55	58.75	60.51	59.44	61.20	**63.05**
9	59.99	58.81	59.38	59.13	60.64	**63.24**
10	60.62	58.74	61.44	58.91	61.65	**62.96**
Min	59.55	57.95	59.38	58.80	60.64	62.80
Média	60.54	58.80	60.42	59.30	61.27	63.13
Max	61.24	59.99	61.44	60.48	61.71	63.55
DP	0.54	0.59	0.67	0.48	0.38	0.24

Fonte: Autor

Em ambas tabelas 15 e 16, podemos observar a variação dos resultados para todos os métodos propostos para cada $fold$. Nota-se que o método DEC obteve desempenho superior para todos $folds$, nas métricas de KS e ROC. Sendo assim, foi realizado o teste de Kurskal-Wallis (4.2)para verificar a existência de diferenças estatisticamente significativas entre os métodos estudados.

$$\begin{cases} H_0 : \text{Todos os métodos são iguais para a métrica avaliada} \\ H_1 : \text{Os métodos são diferentes para a métrica avaliada} \end{cases} \tag{4.2}$$

O p-valor do teste de Kruskal-Wallis foi inferior à 0,0001 (menor que um nível de significância de 0,05), nas métricas KS e ROC, nos levando a rejeitar a hipótese nula (H_0) de que não existe diferença entre os métodos de inferência dos rejeitados. Entretanto, após a identificação de que os métodos não são semelhantes, é importante avaliar se existe alguma diferença entre os métodos dois a dois. Com isso, foi aplicado o teste de comparações múltiplas de Nemenyi, buscando identificar as diferenças nos métodos individualmente.

Tabela 17 – Resultados do Teste de Nemenyi (p-valor) do Experimento II para a métrica KS

Método	Tradicional	Reclassificação	$Augmentation$	$Cluster$	Parcelamento
Reclassificação	0.0131	-	-	-	-
$Augmentation$	0.99996	0.02435	-	-	-
$Cluster$	0.16497	0.94435	0.2456	-	-
Parcelamento	0.99463	0.00184	0.97832	0.04174	-
DEC	0.06094	0	0.03525	0	0.21912

Fonte: Autor

Na tabela 17, avaliando os p-valores das comparações duas a duas do teste de Nemenyi, podemos ver que o teste só revelou haver diferença significativa com um nível de 5% (α =0.05) entre os métodos (Caselas verdes-métrica KS):

- Tradicional e Reclassificação (p-valor=0.0131)
- Reclassificação e $Augmentation$ (p-valor=0.02435)
- Reclassificação e Parcelamento (p-valor=0.00184)
- Reclassificação e DEC (p-valor=0.00000)
- $Augmentation$ e DEC (p-valor=0.03525)
- $Cluster$ e Parcelamento (p-valor=0.04174)
- $Cluster$ e DEC (p-valor=0.00000)

Avaliando num nível de significância de 10% (α =0.1) os métodos que revelaram diferença significativa foram (Caselas laranjas-métrica KS):

- Tradicional e DEC (p-valor=0.06094)

O teste não apresentou evidências de diferença significativa entre os métodos para as demais combinações dois a dois (Caselas brancas-métrica KS).

Tabela 18 – Resultados do Teste de Nemenyi (p-valor) do Experimento II para a métrica ROC

Método	Tradicional	Reclassificação	$Augmentation$	$Cluster$	Parcelamento
Reclassificação	0.03459	-	-	-	-
$Augmentation$	0.99998	0.0558	-	-	-
$Cluster$	0.28724	0.95206	0.38159	-	-
Parcelamento	0.66557	0.00011	0.55546	0.00434	-
DEC	0.02435	0	0.01426	0	0.60249

Fonte: Autor

Na tabela 18, avaliando os p-valores das comparações duas a duas do teste de Nemenyi, podemos ver que o teste só revelou haver diferença significativa com um nível de 5% (α =0.05) entre os métodos (Caselas verdes-métrica ROC):

- Tradicional e Reclassificação (p-valor=0.03459)
- Tradicional e DEC (p-valor=0.02435)
- Reclassificação e Parcelamento (p-valor=0.00011)
- Reclassificação e DEC (p-valor=0.00000)
- $Augmentation$ e DEC (p-valor=0.01426)
- $Cluster$ e Parcelamento (p-valor=0.00434)
- $Cluster$ e DEC (p-valor=0.00000)

Avaliando num nível de significância de 10% (α =0.1) os métodos que revelaram diferença significativa foram (Caselas laranjas-métrica ROC):

- Reclassificação e $Augmentation$ (p-valor=0.0558)

O teste não apresentou evidências de diferença significativa entre os métodos para as demais combinações dois a dois (Caselas brancas-métrica ROC).

Na tabela 19, é visto que os métodos que têm performance próxima ou superior ao processo tradicional de modelagem de inferência dos rejeitados foram o parcelamento e o DEC. Especificamente, em todas as métricas, o método de DEC supera significativamente os demais métodos de inferência dos rejeitados. Foi observado um incremento de 24.6%, 14.9% e 18.9% no ganho das méticas KS, ACC e F_1, respectivamente, quando comparado com o método tradicional.

Tabela 19 – Resultados dos métodos do Experimento II (%)

Método	Métrica			
	AUC	KS	ACC	F_1
Tradicional	60.54 ± 0.54	15.97 ± 0.83	58.89 ± 0.47	66.43 ± 0.45
Reclassificação	55.80 ± 0.59	13.34 ± 0.89	58.84 ± 0.36	67.42 ± 0.33
$Augmentation$	60.42 ± 0.67	15.78 ± 1.02	59.48 ± 0.49	67.47 ± 0.46
$Cluster$	59.30 ± 0.48	14.09 ± 0.73	57.25 ± 0.42	64.33 ± 0.50
Parcelamento	61.27 ± 0.38	16.37 ± 0.74	60.00 ± 0.26	66.41 ± 0.33
DEC	**63.13 ± 0.24**	**19.90 ± 0.36**	**67.64 ± 0.22**	**78.99 ± 0.14**

Fonte: Autor

A Tabela 20 mostra de que modo a taxa de inadimplência apresenta boa ordenação para todos os métodos ajustados, uma vez que as classes de escore mais baixas apresentaram a maior taxa de inadimplência e essa taxa foi decrescente a medida que aumentava o escore, demostrando uma boa discriminação entre adimplentes e inadimplentes. Sendo que os métodos de parcelamento e DEC se destacaram no indicador DTI, ambos tiveram diferença superior a 24%, uma diferença de 4% a 5% quando observado os demais métodos.

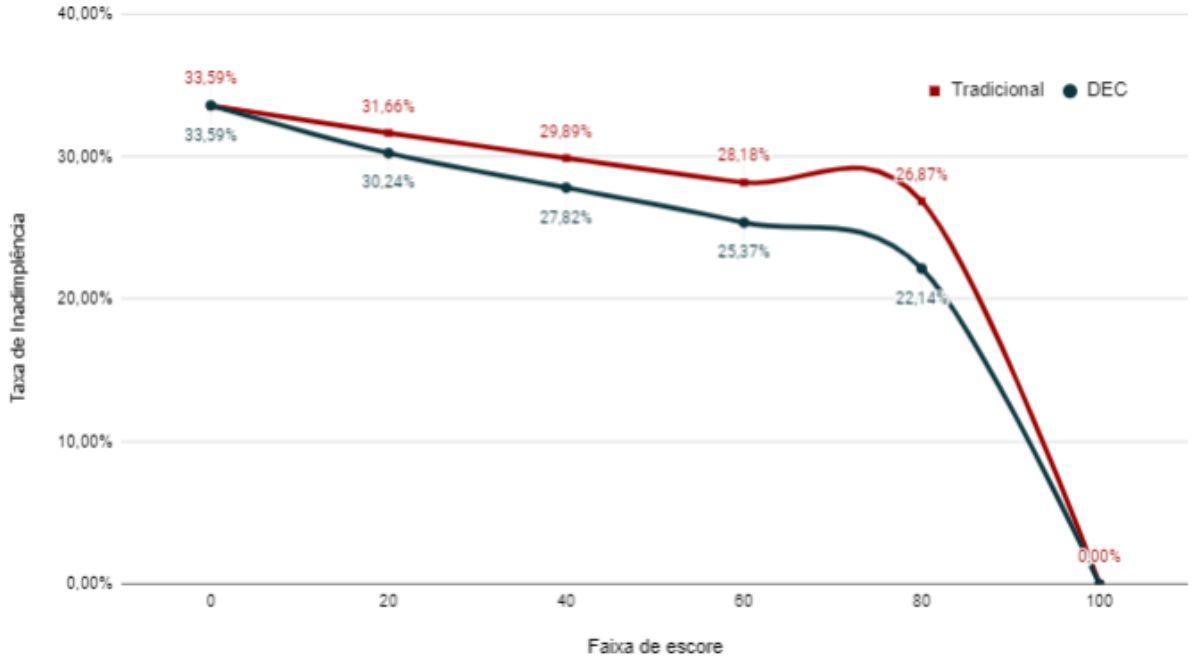
Tabela 20 – ExpII: Inadimplência por intervalo de quintil

Faixa Escore (Quintil)	Tradicional	Reclassificação	$Augmentation$	Cluster	Parcelamento	DEC
Q1	41.32%	43.73%	43.32%	45.33%	46.97%	46.97%
Q2	36.95%	37.34%	36.99%	36.78%	37.45%	37.53%
Q3	33.32%	33.66%	33.17%	32.69%	32.81%	32.71%
Q4	29.49%	29.33%	30.08%	28.49%	28.56%	28.62%
Q5	26.87%	23.91%	24.21%	24.67%	22.16%	22.14%
DTI(1;5)	14.45%	19.82%	19.11%	20.65%	24.81%	24.85%
	Fonte: Autor					

Na Figura 14, os métodos avaliados foram o tradicional e o DEC, uma vez que o método DEC teve melhor performance comparados com os demais métodos.

Supondo um ponto de corte no escore 40, teríamos uma taxa de inadimplência para o método tradicional de 29,89% e para o DEC de 27,82%; isto significa que o DEC teria uma redução de inadimplência de 7%, mantendo a mesma aprovação comparado com o tradicional. Com o aumento do ponto de corte, podemos observar que o DEC aumenta o impacto na redução da inadimplência: 10% para o ponto de corte de 60, e 18% para o de 80.

Figura 14 – Simulação aprovação e inadimplência por ponto de corte - Escore percentil.



Fonte: Autor

4.3.3 Experimento III

Nesta seção, serão apresentadas as análises dos resultados alcançados pelos métodos nas diversas métricas avaliadas para o Experimento III, uma operação que oferta cartão digital para seus proponentes. Nas tabelas 21 e 22, a seguir, poderemos observar os valores alcançados pelos métodos para a métrica de KS e ROC para cada $fold$ e suas estatísticas: Mínimo (Min), Média, Máximo (Max) e Desvio-padrão (DP).

Tabela 21 – Resultados do KS para os 10 $folds$ do Experimento III (%)

$Fold$	Tradicional	Reclassificação	$Augmentation$	Cluster	Parcelamento	DEC
1	12.73	7.96	11.38	8.42	12.52	**15.55**
2	12.67	8.27	12.29	11.80	13.08	**17.85**
3	10.98	7.90	11.19	9.94	12.56	**16.29**
4	13.39	7.36	12.37	10.78	12.99	**15.84**
5	14.21	8.56	14.07	9.51	12.81	**15.69**
6	12.05	7.73	10.22	8.95	11.90	**16.51**
7	11.23	7.63	11.23	9.99	14.51	**15.94**
8	12.98	8.39	12.46	11.19	12.26	**16.01**
9	12.32	8.60	12.60	11.67	11.87	**15.65**
10	13.09	8.50	12.38	9.04	11.80	**16.16**
Min	10.98	7.36	10.22	8.42	11.80	15.55
Média	12.57	8.09	12.02	10.13	12.63	16.15
Max	14.21	8.60	14.07	11.80	14.51	17.85
DP	0.97	0.43	1.05	1.19	0.81	0.67

Fonte: Autor

Tabela 22 – Resultados da ROC para os 10 $folds$ do Experimento III (%)

$Fold$	Tradicional	Reclassificação	$Augmentation$	Cluster	Parcelamento	DEC
1	58.63	55.26	57.51	55.56	58.26	**60.26**
2	58.63	55.46	58.11	57.79	58.63	**61.78**
3	57.51	55.22	57.39	56.56	58.29	**60.75**
4	58.61	54.86	58.17	57.12	58.57	**60.46**
5	59.34	55.65	59.28	56.28	58.46	**60.36**
6	58.24	55.10	56.75	55.91	57.86	**60.90**
7	57.77	55.04	57.41	56.59	59.57	**60.52**
8	58.71	55.54	58.22	57.39	58.09	**60.57**
9	57.99	55.68	58.31	57.70	57.83	**60.33**
10	58.51	55.61	58.17	55.96	57.79	**60.67**
Min	57.51	54.86	56.75	55.56	57.79	60.26
Média	58.39	55.34	57.93	56.69	58.34	60.66
Max	59.34	55.68	59.28	57.79	59.57	61.78
DP	0.53	0.29	0.70	0.78	0.53	0.44

Fonte: Autor

Em ambas tabelas 21 e 22, podemos observar a variação dos resultados de todos os métodos propostos para cada $fold$. Nota-se que o método DEC obteve desempenho superior para todos $folds$, nas métricas de KS e ROC. Sendo assim, foi realizado o teste de Kurskal-Wallis (4.3), para verificar a existência de diferenças estatisticamente significativas entre os métodos estudados.

$$\begin{cases} H_0 : \text{Todos os métodos são iguais para a métrica avaliada} \\ H_1 : \text{Os métodos são diferentes para a métrica avaliada} \end{cases} \tag{4.3}$$

O p-valor do teste de Kruskal-Wallis foi inferior à 0,0001 (menor que um nível de significância de 0,05), nas métricas KS e ROC, nos leva a rejeitar a hipótese nula (H_0) de que existe diferença entre os métodos de inferência dos rejeitados. Após a identificação de que os métodos são diferentes, é importante avaliar se existe alguma diferença entre os métodos dois a dois. Com isso, foi aplicado o Teste comparações múltiplas de Nemenyi, buscando identificar as diferenças nos métodos individualmente.

Tabela 23 – Resultados do Teste de Nemenyi (p-valor) do Experimento III para a métrica KS

Método	Tradicional	Reclassificação	$Augmentation$	$Cluster$	Parcelamento
Reclassificação	0.00089	-	-	-	-
$Augmentation$	0.96755	0.01723	-	-	-
$Cluster$	0.09895	0.72199	0.47055	-	-
Parcelamento	1	0.00089	0.96755	0.09895	-
DEC	0.16967	0	0.0195	0.00001	0.16967

Fonte: Autor

Na tabela 23, avaliando os p-valores das comparações duas a duas do teste de Nemenyi, podemos ver que o teste só revelou haver diferença significativa com um nível de 5% (α =0.05) entre os métodos (Caselas verdes-métrica KS):

- Tadicional e Reclassificação (p-valor=0.00089)
- Reclassificação e $Augmentation$ (p-valor=0.01723)
- Reclassificação e Parcelamento (p-valor=0.00089)
- Reclassificação e DEC (p-valor=0.00000)
- $Augmentation$ e DEC (p-valor=0.0195)
- $Cluster$ e DEC (p-valor=0.00001)

Avaliando num nível de significância de 10% (α =0.1) os métodos que revelaram diferença significativa foram (Caselas laranjas-métrica KS):

- Tadicional e $Cluster$ (p-valor=0.09895)
- $Cluster$ e Parcelamento (p-valor=0.09895)

O teste não apresentou evidências de diferença significativa entre os métodos para as demais combinações dois a dois (Caselas brancas-métrica KS).

Tabela 24 – Resultados do Teste de Nemenyi (p-valor) do Experimento III para a métrica ROC

Método	Tradicional	Reclassificação	$Augmentation$	$Cluster$	Parcelamento
Reclassificação	0.00044	-	-	-	-
$Augmentation$	0.89853	0.02293	-	-	-
$Cluster$	0.05988	0.73749	0.51275	-	-
Parcelamento	0.99993	0.0011	0.96032	0.10543	-
DEC	0.23955	0	0.01457	0	0.15143

Fonte: Autor

Na tabela 24, avaliando os p-valores das comparações duas a duas do teste de Nemenyi, podemos ver que o teste só revelou haver diferença significativa com um nível de 5% (α =0.05) entre os métodos (Caselas verdes-métrica ROC):

- Tadicional e Reclassificação (p-valor=0.00044)
- Reclassificação e $Augmentation$ (p-valor=0.02293)
- Reclassificação e Parcelamento (p-valor=0.0011)
- Reclassificação e DEC (p-valor=0.00000)
- $Augmentation$ e DEC (p-valor=0.01457)
- $Cluster$ e DEC (p-valor=0.00000)

Avaliando num nível de significância de 10% (α =0.1) os métodos que revelaram diferença significativa foram (Caselas laranjas-métrica ROC):

- Tadicional e $Cluster$ (p-valor=0.05988)

O teste não apresentou evidências de diferença significativa entre os métodos para as demais combinações dois a dois (Caselas brancas-métrica ROC).

Tabela 25 – Resultados dos métodos do Experimento III (%)

Método	Métrica			
	AUC	KS	ACC	F_1
Tradicional	58.39 ± 0.53	12.57 ± 0.97	58.42 ± 0.59	69.14 ± 0.52
Reclassificação	55.34 ± 0.29	8.09 ± 0.43	61.25 ± 0.37	73.80 ± 0.28
$Augmentation$	57.93 ± 0.70	12.02 ± 1.05	56.39 ± 0.46	66.79 ± 0.38
$Cluster$	56.69 ± 0.78	10.13 ± 1.19	52.31 ± 0.48	64.41 ± 0.45
Parcelamento	58.34 ± 0.53	12.63 ± 0.81	55.21 ± 0.45	65.23 ± 0.46
DEC	**60.66 ± 0.44**	**16.15 ± 0.67**	**65.77 ± 0.30**	**78.30 ± 0.22**

Fonte: Autor

Alguns métodos de inferência dos rejeitados têm resultados muito similares com o método tradicional, são eles: $Augmentation$ e Parcelamento. Já nos métodos de reclassificação e $Cluster$, suas performances ficaram abaixo, cujos resultados são mostrados na Tabela 25.

O método de DEC foi superior em todas as métricas de avaliação. Analisando o F_1, métrica que combina precision e recall, de modo a trazer um número único que indica a qualidade geral numa média harmônica, a performance foi superior em 13% e 6% comparadas com o método tradicional e reclassificação (método que obteve a segunda melhor performance). Quando é observado o KS, o DEC tem ganhos mais expressivos, o desempenho da métrica é superior a 25%, observando todos os métodos propostos.

Na Tabela 26, são apresentadas as taxas de inadimplência por classe de escore e a DTI para os respectivos métodos, obtidos a partir da diferença das classes dos intervalos extremos $DTI(1;5)$. Observa-se que o DTI foi de 19.7% para o método DEC, seguindo pelos métodos Tradional e Parcelamento (14.5%), e os demais métodos abaixo de 10%.

Tabela 26 – ExpI: Inadimplência por intervalo de quintil

Faixa Escore (Quintil)	Tradicional	Reclassificação	$Augmentation$	Cluster	Parcelamento	DEC
Q1	31,45%	22,99%	28,35%	26,93%	30,70%	32,30%
Q2	24,99%	22,74%	23,78%	25,30%	24,12%	26,25%
Q3	22,34%	22,48%	22,26%	23,00%	21,09%	19,06%
Q4	19,31%	22,23%	21,04%	20,56%	18,55%	14,79%
Q5	16,87%	21,97%	19,27%	19,18%	16,18%	12,60%
DTI(1;5)	14.59%	1.02%	9.08%	7.76%	14.52%	19.70%

Fonte: Autor

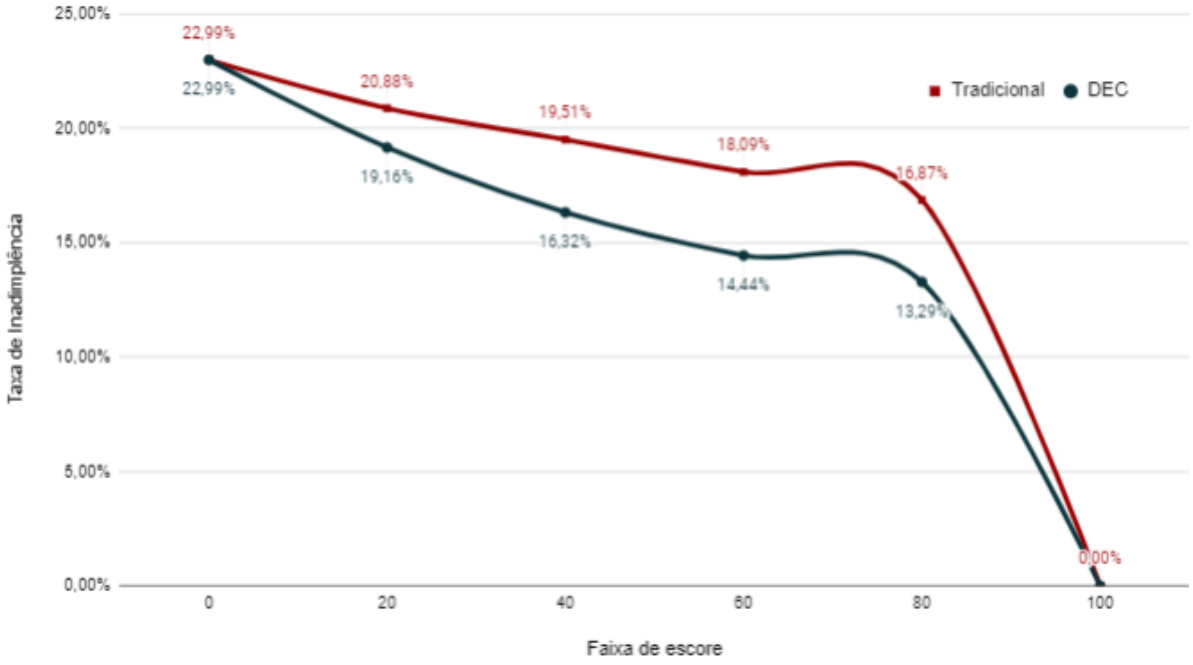
Na Figura 15, os métodos avaliados foram o tradicional e o DEC, uma vez que o método

DEC teve melhor performance comparados com os demais métodos.

Supondo um ponto de corte no escore 40, teríamos uma taxa de inadimplência para o método tradicional de 19,51% e para o DEC de 16,32%, ou seja, o DEC teria uma redução de inadimplência de 16% mantendo a mesma aprovação comparado com o tradicional.

Com o aumento do ponto de corte, podemos observar que o DEC aumenta o impacto na redução da inadimplência: 20% para o ponto de corte de 60, e 21% para o de 80.

Figura 15 – Simulação aprovação e inadimplência por ponto de corte - Escore percentil.

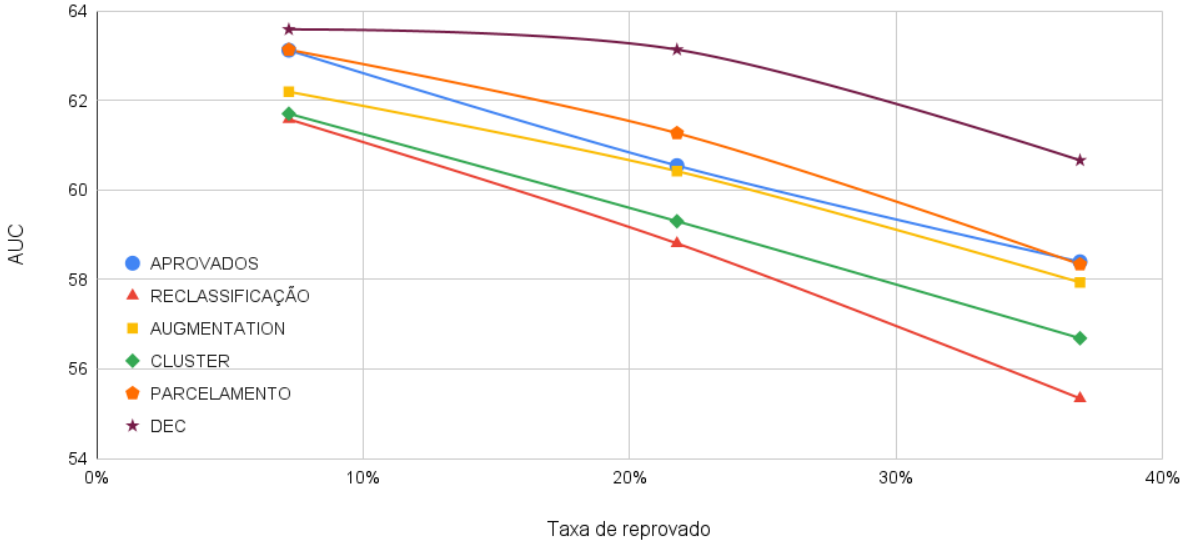


Fonte: Autor

Na Figura 16 mostra que o método de DEC tem melhor desempenho em todos os três experimentos, indicando que, independente da variação da taxa de reprovação, o método se destaca em relação aos demais. Também podemos observar que, com o aumento da taxa de rejeitados dos experimentos, a performance do método de reclassificação cai de forma mais acelerada quando comparamos com os demais.

Vale destacar também que quanto menor a taxa de rejeitados, menor a discrepância entre os métodos. O conjunto de dados I tem uma taxa de reprovação de 7,21% e a amplitude entre os métodos é de 2,71% Para o conjunto de dados II, a amplitude é de 4,33, entretanto, quando retiramos o método de DEC da análise, a amplitude entre os demais métodos é bem semelhante ao conjunto de dados I.

Figura 16 – Comparação dos métodos de inferência dos rejeitados para todos os experimentos.



Fonte: Autor

5 CONCLUSÃO E TRABALHOS FUTUROS

5.1 CONSIDERAÇÕES FINAIS

O problema da inferência dos rejeitados tem uma longa história na área de crédito, com grandes debates e divergências. Pode ser visto como um problema estatístico com dados faltantes, uma vez que os comportamentos dos proponentes rejeitados não estão disponíveis. Dependendo do tipo de falta, ou seja, se a falta for por $Missing\ completely\ at\ random$ (MCAR), $Missing\ not\ at\ random$ (MNAR) e $Missing\ at\ random$ (MAR), várias técnicas estatísticas são utilizadas. Em outra perspectiva, a inferência dos rejeitados também pode ser considerada como um problema de aprendizado de máquina, onde os algoritmos aprendem a usar as informações da população negada e otimizam a performance dos modelos gradativamente.

Nesta dissertação, foram apresentados cinco métodos de inferência dos rejeitados, utilizando-se das informações de clientes reprovados. Os métodos foram apresentados e aplicados em três bases de dados reais, contendo carteiras de clientes em diversos setores do mercado de crédito, entre 2018 e 2020, com perfis de reprovação diferentes, variando de 7,2% a 36,8%.

Para todas as bases, foram construídos seis modelos, aplicando os métodos tradicionais de modelagem e inferência dos rejeitados, sendo elas: Reclassificação, $Augmentation$, $Cluster$, Parcelamento e $Deep\ Embedded\ Clustering$ (DEC). A técnica utilizada para treinamento dos diversos modelos foi a regressão logística.

As estratégias adotadas para avaliar o desempenho dos métodos foram métricas comumente usadas pelo mercado de crédito e artigos relacionados a inferência dos rejeitados. São elas: Área da curva ROC (AUC), estatística do teste de Kolmogorov Smirnov (KS), acurácia, $F_1\ score$ e a diferença entre as taxas de inadimplência (DTI). E por fim foi realizado o teste de hipóteses de Kruskal-Wallispara verificar se existe diferença significativa entre os métodos nas diversas métricas. Como também o teste de comparação múltipla de Nemenyi, um teste $post-hoc$, que é usado para fazer comparações em pares com o intuito de verificar qual dos métodos diferem entre si.

Dadas as amplas aplicações de métodos de aprendizado de máquina e a crescente influência da $deep\ learning$ nos problemas atuais, esta dissertação propõe o uso de um algoritmo de $deep\ learning$ como uma solução para a inferência dos rejeitados.

Os conjuntos de dados mostram que os métodos propostos de DEC e parcelamento atin-

gem um desempenho superior ao método tradicional na maioria das métricas de avaliação estudadas, e, consequentemente, superiores às demais abordagens clássicas de inferência dos rejeitados em $credit\ scoring$.

Além disso, o método de DEC é muito mais eficiente quando é observado o F_1, ou seja, o modelo tende ter uma harmonização entre o $precision$ e o $recall$ e por consequencia uma melhor qualidade no modelo. De modo geral, os métodos de Reclassificação e $Cluster$ tiveram os piores indicadores de desempenho. Já o $Augmentation$ teve seus resultados mais similares ao método Tradicional.

Em geral, observando todos os conjuntos de dados, o método de DEC tem resultados melhores que os demais métodos nas diversas métricas de avaliação. Um ponto interessante é que, no teste de comparação múltipla de Nemenyi, o DEC é estatisticamente diferente dos métodos de Reclassificação, $Augmentation$ e $Cluster$ a um nível de 5% de confiança.

No conjunto de dados I, percebe-se que o DEC tem performance superior que os demais métodos, contudo, para os dois principais indicadores de avaliação, AUC e KS, o método sofre uma grande variação comparada com os demais métodos. Para a AUC o desvio-padrão foi de $\pm 2,75$ e para o KS $\pm 1,19$.

Devido a essa variação, os métodos tradicional e parcelamento devem ser observados com cuidado, pois seus indicadores de desempenho não estão tão longe do método DEC. Observou-se também que os métodos não são diferentes entre si a um nível de 5%. Todavia, quando observados pelo teste de Nemenyi, alguns dos métodos são diferentes ao fazer comparação em pares. Para a métrica de KS, 5 dos 15 pares de métodos, a métrica teve diferença significativa em um nível de 5%. Já para a ROC, o evento aconteceu em 4 dos 15 pares de métodos.

Nos indicadores de DTI, de modo geral, os métodos de reclassificação, tradicional, parcelamento e DEC tiveram suas performances bem similarescom um leve destaque para o DEC que tem DTI de 21,77%. Observando a métrica de simulação de ponte de corte o DEC, método de melhor desempenho, não é apresentado grande diferença quando comparamos com o método tradicional. Seus ganhos giram entre 2% e 5% de redução de inadimplência ao longo dos pontos de corte.

No Experimento II, é visto que os métodos que têm performances próximas ou superiores ao processo tradicional de modelagem de inferência dos rejeitados foram o parcelamento e o DEC, contudo, o DEC tem resultados superiores em todas as métricas: AUC de 63%, KS de 19,9%, ACC de 67,6% e F_1 de 78,9%. Isto representa incrementos de 4,3%, 24,6%, 14,9% e 18,9% comparados com o método tradicional, nas métricas citadas respectivamente.

No teste de Kruskal-Wallis, observou-se que os métodos são diferentes entre si a um nível de 5%. Quando observados pelo teste de Nemenyi alguns dos métodos são diferentes. Para ambas as métricas de KS e ROC 7 dos 15 pares de métodos a métrica teve diferença significativa em um nível de 5%, significa que 47% dos métodos são diferentes.

Já no Experimento III, o perfil dos dados era bem mais complexo para uma boa discriminação de bons e maus clientes. Dado esse cenário, o DEC conseguiu aprender melhor devido sua característica do algoritmo, onde transforma o espaço de dados X em um espaço de características latentes Z (usando um mapeamento não linear — DNN).

Por consequência, o DEC teve indicadores de performance superiores aos demais métodos. Quando comparamos o DEC com os métodos tradicional, augmentation e parcelamento, o desempenho do algoritmo de deep learning é superior em: $AUC =$ (3,9%, 4,7%, 4%), $KS =$ (28,5%, 34,4%, 27,9%) e $ACC =$ (12,6%, 16,6%, 19,1%).

Percebeu-se que os métodos não são diferentes entre si a um nível de 5%. Contudo, quando observados pelo teste de Nemenyi, alguns dos métodos são diferentes. Para as métricas de KS e ROC 6 dos 15 pares de métodos, a métrica teve diferença significativa em um nível de 5%.

Apenas o DTI do DEC e Parcelamento tiveram índice superior a 14%, 14,59% para o Tradicional e 14,5% para o DEC. Os demais métodos tiveram resultados muito baixos, em destaque para o método de Reclassificação 1,02%, isso significa que o modelo proposto pelo método é muito mal ajustado.

No incador de simulação de ponto de corte,o DEC aumenta o impacto na redução da inadimplência ao longo dos cenários de PC quando comparamos com método tradicional, as reduções seguem da seguinte forma: PC de 20 redução de 8%, PC de 40 redução de 16%, PC de 60 redução de 20% e PC de 80 redução de 21%. Ou seja, para qualquer cenário de PC o DEC consegue melhorar a estratégia de concessão de crédito usando modelos de $credit\ scoring$. Também foi explorado as possíveis distribuições de aprovados e rejeitados e descoberto que as amostras aceitas e rejeitadas tiveram distribuições diferentes, o que forneceu uma compreensão abrangente da inferência dos rejeitados inerente ao mecanismo de aprovação das carteiras.

No Experimento I, a taxa de reprovação da carteira é baixa, 7%. Dado esse cenário, os métodos tiveram resultados bem similares. Olhando as métricas de AUC e KS, a variação dos indicadores foi de 2,1% e 2,4% respectivamente.

Já nos conjuntos de dados II e III, com taxas de reprovação de 21,7% e 36,8%, com os métodos de Reclassificação e $Cluster$, seus indicadores de performance ficaram bem abaixo

quando se observa os demais métodos. O método de DEC, nessas bases de dados, tem performance superior a todos os métodos, contudo, não podemos desprezar e observar em futuros modelos desenvolvidos os métodos de Parcelamento e $Augmentation$.

Fica claro que, com o aumento da taxa de reprovação da carteira, os métodos propostos começam a auxiliar para um melhor desempenho e, por consequência, auxiliar as estratégias de pontuação de crédito. Portanto, deve-se observar a priori a taxa de reprovação da carteira para decidir se faz necessário a utilização de métodos de inferência dos rejeitados.

Deve-se ressaltar que o uso de métodos de inferência dos rejeitados precisa observar como foi a tomada de decisão de crédito anterior, como se deu o processo da estratégia de crédito.

5.2 CONTRIBUIÇÕES

A primeira contribuição foi a utilização de algoritmo de $deep\ learning$ no processo de inferência dos rejeitados. Esses algoritmos são fortemente usados em processamento de imagem, texto e áudio. Nos problemas de inferência dos rejeitados, tais algoritmos vêm sendo avaliados recentemente. Com esta análise, foi possível concluir que o algoritmo de $deep\ learning$ tem potencial para ser trabalhado.

Em segundo, a comparação dos métodos propostos em três problemas (bases de dados), com características completamente diferentes, também é uma importante contribuição deste trabalho. Pois, a variação da taxa de reprovação é um fator determinante na utilização e/ou definição do método de inferência dos rejeitados. Muitos outros trabalhos limitaram-se a um único problema, de modo que a robustez e adaptação dos métodos a diferentes situações não era englobada.

Além disso, aplicações em dados reais é raro dado a complexidade de ter acesso a esse tipo de informação. Outra grande contribuição deste trabalho é reforçar que métodos de inferência dos rejeitados são de suma importância para auxiliar na eliminação do viés de modelagem nos modelos de $credit\ scoring$ desenvolvidos atualmente. Este trabalho objetiva servir de inspiração para trabalhos futuros na área.

5.3 TRABALHOS FUTUROS

Diante das considerações anteriores, acreditamos que o presente trabalho pode ser continuado e aprimorado de diversas maneiras. Dentre essas possibilidades, podemos abranger um

maior número de métodos de inferência dos rejeitados e continuar pesquisando na literatura métodos de inferência que não foram contemplados neste trabalho, como: $Augementation\ 2^6$, $Fuzzy\ Augementation$, $Memory - BasedReasoning$, S_3VM, entre outras.Tais métodos não foram trabalhados, pois o objetivo era utilizar métodos fortemente utilizados na literatura e compará-los com o método DEC.

Aumentar o número de conjuntos de dados com perfis de taxa de reprovação diferentes, variando essa taxa entre 5% e 80% aproximadamente – com isso, é possível mapear diversos perfis de carteiras de crédito. Por exemplo, empresas de empréstimo pessoal tendem a ter uma taxa de reprovação alta, já produtos de consignado privado tendem a ter uma reprovação baixa.

Investigar no estado da arte novos algoritmos de $deep\ learning$, uma vez que o proposto na dissertação teve resultados promissores, como também realizar $ensemble$ de métodos de inferência para avaliar se a combinação propicia uma melhora nos resultados.

REFERÊNCIAS

Alves M.C. Estratégias para o desenvolvimento de modelos de credit score com inferência dos rejeitados. *Dissertação de Mestrado – Instituto de Matemática e Estatística, USP São Paulo.*, 2008.

Ash D., Meesters S. Best practices in reject inferencing. *Wharton Financial Institution Center*, 2002.

Banasik J., Crook J. Does reject inference really improve the performance of application scoring models? *Journal of Banking and Finance*, n. 28, p. 857874, 2004.

Banasik J., Crook J. Credit scoring, augmentation and lean models. *Journal of the Operational Research Society*, n. 56, p. 1072-1091, 2005.

Banasik J., Crook J. Reject inference, augmentation, and sample selection. *European Journal of Operational Research*, n. 183, p. 15821594, 2007.

Banasik J., Crook J. Reject inference in survival analysis by augmentation. *Journal of the Operational Research Society*, n. 61, p. 473485, 2010.

Banasik J., Crook J., Thomas L. Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, n. 54, p. 822832, 2003.

Bücker M.,Van Kampen M. Reject inference in consumer credit scoring with nonignorable missing data. *Journal of Banking e Finance*, n. 37, p. 10401045, 2013.

Chen G.G., Astebro T. Bound and collapse bayesian reject inference for credit scoring. *Journal of the Operational Research Society*, n. 63, p. 1374-1387, 2012.

Chen W., Liu Y., Xiang G., Liu Y. A three-stage data mining model for reject inference. *Fifth International Conference on Business Intelligence and Financial Engineering*, p. 34-38, 2012.

Connover W.J. Pratical nonparametric statistics. *John Wiley*, 1999.

Crook J., Banasik J. Does reject inference really improve the performance of application scoring models? *Journal of Banking e Finance*, v. 4, n. 28, p. 857-874, 2004.

Dempster A.P., Laird N.M. "maximum likelihood from incomplete data. *Journal of the Royal Statistical Society*, n. 39, p. 1-38, 1977.

Diniz C.A.R., Louzada F. Modelagem estatística para risco de crédito. *ABE - Associação Brasileira de Estatística*, v. 1, p. 1-178, 1977.

Diniz C.A.R., Louzada F. Métodos estatísticos para análise de dados de crédito. *6th Brazilian Conference on Statistical Modelling in Insurance and Finance*, 2013.

Dizaji K., Huang H. Deep clustering via joint convolutional autoencoder embedding and relative entropy minimization. *IEEE International Conference on Computer Vision (ICCV)*, p. 57475756, 2017.

Duda R., Hart P., Stork D. Pattern classification. *John Wiley*, 2001.

Durand D. Risk elements in consumer instalment financing. *New York: National Bureau of Economics*, 1941.

Ehrhardt A., Biernacki C. Reject inference methods in credit scoring: A rational review. *Journal of Applied Statistics*, p. 1-22, 2020.

Feelders A. Credit scoring and reject inference with mixture models. *Intelligent Systems in Accounting, Finance e Management*, v. 9, n. 1, p. 18, 2000.

Feelders A. An overview of model based reject inference for credit scoring. *Utrecht University, Institute for Information and Computing Sciences*, 2003.

Fisher R.A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, v. 1, n. 7, p. 179–188, 1936.

Guo X., Gao L. Improved deep embedded clustering with local structure preservation. *International Joint Conference on Artificial Intelligence (IJCAI)*, v. 26, p. 1753-1759, 2017.

Hand D. Reject inference in credit operations: theory and methods. *The Handbook of Credit Scoring*, 2001.

Hand D.J. ,Henley, W.E. Can reject inference ever work? *IMA Journal of Management Mathematics*, v. 5, n. 1, p. 45-55, 1993.

Hanley J.A. The robustness of the binormal assumptions used in fitting roc curves. *Medical Decision Making*, n. 8, p. 197203, 1988.

Hollander M., Wolfe D.A. Nonparametric statistical methods. *John Wiley Sons*, v. 3, p. 1-813, 2014.

Hsia D. C. Credit scoring and the equal credit opportunity. *Hastings LJ*, v. 30, n. 2, p. 1-79, 1978.

Hsia D.C. Credit scoring and the equal credit opportunity act. *Hastings Law Journal*, v. 30, n. 1, p. 371-448, 1978.

Hsieh N.C. Hybrid mining approach in the design of credit scoring models. *Expert Systems with Application*, v. 28, p. 655-665, 2005.

Jancey R. Multidimensional group analysis. *Australian Journal of Botany*, n. 14, p. 127-130, 1966.

Joanes D.N. Reject inference applied to logistic regression for credit scoring. *IMA Journal of Management Mathematics*, v. 5, n. 1, p. 35-43, 1993.

Kim Y., Sohn S.Y. Technology scoring model considering rejected applicants and effect of reject inference. *Journal of the Operational Research Society*, v. 1, n. 58, p. 13411347, 2007.

Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. *International joint Conference on artificial intelligence*, v. 1, n. 14, p. 11371145, 1995.

Kozodoi N., Katsas P. Shallow self-learning for reject inference in credit scoring. *ECML/PKDD*, p. 1-17, 2019.

Kruskal W.H., Wallis W.A. Use of ranks in one-criterion variance analysis. *Journal american statistical association*, v. 47, p. 583-621, 1952.

Lewis E. Introduction to credit scoring. *San Rafael: Fair Isaac and Co. Inc*, 1992.

Li F., Qiao H. Discriminatively boosted image clustering with fully convolutional auto-encoders. *Pattern Recognition*, n. 83, p. 161-173, 2017.

Li Z., Tian Y., Li K. Zhou F., Yang W. Reject inference in credit scoring using semi-supervised support vector machines. *Expert Systems With Applications*, n. 74, p. 105-114, 2017.

Lim M.K., Sohn S.Y. Cluster-based dynamic scoring model. *Expert Systems with Application*, v. 32, p. 427-431, 2007.

Lloyd. Least squares quantization in pcm. *IEEE Trans Inform Theory (Special Issue on Quantization)*, v. 28, n. 135, p. 129137, 1982.

Maaten L., Hinton G. Visualizing data using t-sne. *Journal of Machine Learning Research*, v. 1, n. 9, p. 25792605, 2008.

MacQueen J. Some methods for classification and analysis of multivariate observations. *Statistics and Probability*, v. 1, n. 1, p. 281-297, 1967.

MacQueen J. K-means-type algorithms: A generalized convergence theorem and characterization of local optimality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, n. 1, p. 81-87, 1984.

Metz C.E. "statistical analysis of roc data in evaluating diagnostic performance. *Multiple Regression Analysis: Applications in the Health Sciences*, v. 1, n. 13, p. 365384, 1986.

Nelder, J.A., Wedderburn, R. Generalized linear models. *J. R. Statist. Soc. A*, n. 135, p. 370-384, 1972.

Nemenyi P. Distribution-free multiple comparisons. *Princeton University*, 1963.

Nguyen H. T. Reject inference in application scorecards: evidence from france. *EconomiX Working Papers*, v. 1, 2016.

Parnitzke T. Credit scoring and sample selection bias. *Institute of Insurance Economic*, 2005.

Paula, G.A. Modelos de regressão com apoio computacional. *IME-USP*, v. 0, n. 1, 2004.

Pires I.R., Teixeira T.R., Souza J.B. Modelo estrategico para a tomada de decisão nas operações de crédito: Um estudo de caso utilizando redes neurais articiais. 2008.

Rubin D., Little R. Statistical analysis with missing data. *John Willey*, v. 3 Edition, 2020.

Rui X., Wunsch D. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, v. 16, n. 3, p. 645678, 2005.

Shearer C. The crisp-dm model: the new blueprint for data mining. *Journal of data warehousing*, v. 5, n. 4, p. 13-22, 2000.

Shevock A., Cumsille P., Graham J. Methods for handling missing data. *Research Methods in Psychology*, v. 2 Edition, 2012.

Siddiqi N. *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. New York: John Wiley Sons, 2006. ISBN 978–0-471–75451–0.

Silva J. *Gestão e análise de risco de crédito*. New York: Cengage Learning, 2016. ISBN 978-8522126743.

Thomas L., Edelman D., Crook J. Credit scoring and its applications. *SIAM-Monographs on mathematical modeling and computation*, 2002.

Tian Y., Yong Z., Luo J. A new approach for reject inference in credit scoring using kernel free fuzzy quadratic surface support vector machines. *Applied Soft Computing*, v. 17, p. 96-105, 2018.

Tomazela S.M.O. "avaliação de desempenho de modelos *credit score* ajustados por análise de sobrevivência. *Dissertação de Mestrado. Instituto de Matemática e Estatística - USP*, 2007.

Verstraeten G. The impact of sample bias on consumer credit scoring performance and probatability. *Journal of the operational research society*, v. 56, n. 8, p. 981-992, 2005.

Xie J., Girshick R. Unsupervised deep embedding for clustering analysis. *International Conference on Machine Learning (ICML)*, v. 33, p. 478487, 2016.

Yang B., Hong M. Towards k-means-friendly spaces: Simultaneous deep learning and clustering. *International Conference on Machine Learning (ICML)*, v. 70, p. 38613870, 2017.