SemeAd 2023
XXVI Seminários em Administração

07, 08, 09 e 10 de novembro de 2023
ISSN 2177-3866

Identificação de inadimplentes no setor de conexão e acesso à internet banda larga por fibra óptica.

JOSÉ NATANIEL CENTENO KLUG
ESCOLA SUPERIOR DE AGRICULTURA LUIZ DE QUEIROZ - ESALQ

FERNANDO FREIRE VASCONCELOS
FACULDADE DE ECONOMIA, ADMINISTRAÇÃO E CONTABILIDADE DA UNIVERSIDADE DE SÃO PAULO - FEA

Identificação de inadimplentes no setor de conexão e acesso à internet banda larga por fibra óptica.

Introdução

O mercado de telecomunicações passou pela maior turbulência durante os anos vividos de pandemia pela Covid-19. A pandemia com seu distanciamento social e os fechamentos generalizados gerou um surto inevitável de consumo de serviços digitais (Rahul, 2020).

O avanço rápido da doença e a movimentação de pessoas e seus trabalhos para o ambiente doméstico criou um ambiente de crescimento para o mercado de telecomunicações em especial para as tecnologias como a fibra óptica.

Neste mesmo ambiente de crescimento, com muita oportunidade, um problema se mostrou mais evidente do que antes: a inadimplência e o cancelamento de usuários. Mediante observação direta do gestor verificou-se que o setor trata este comportamento como natural e aceitam que é parte da dinâmica de mercado o cliente deste serviço ficar inadimplente e acabar por não manter o serviço, cancelando este, muitas vezes sem nem contatar ou informar a operadora.

Quando o valor excede o limite determinado pela empresa podem acontecer problemas de fluxo de caixa ou depreciação da qualidade da oferta o que poderá gerar uma nova onda de inadimplentes sustentando então um ciclo vicioso.

Para a definição de risco de inadimplência é comum encontrar modelos de regressão logística (Lopes, Ciribeli, Massardi, Mendes. 2017) que consiste na técnica de análise multivariada para verificar ocorrências e de elementos relacionados as variáveis (Fávero, Belfiore. 2017). No entanto ao pensar no risco de inadimplência deve-se incorrer sobre as condições para a concessão de crédito em que existem peculiaridades e características muito pontuais em cada crédito que podem impactar a inadimplência. (Jannuzzi, 2010)

No ramo de telecomunicações, não diferente de outros, alguns dados são muito comuns de serem coletados com os usuários como nome, data de nascimento, documentos, endereço, telefones e outros meios de contato. Além disso, claramente, são necessários dados relacionados ao produto ou serviço contratado como nome do serviço, data de vencimento, forma de pagamento, valor, condições de pagamento de taxas, tempo de contrato, formas de reajuste, descontos concedidos. A partir destes dados coletados, que não incluem variáveis comuns ao mercado financeiro como renda, estado civil e grau de instrução (Lima, Serra, Fávero, 2021), a modelagem para definição de concessão de crédito deve ser realizada e projetada para alinhar o risco de inadimplência com os objetivos da empresa. (Jiang, et al. 2021).

Seria possível, com base em dados coletados no processo de venda de um serviço de telecomunicações, em especial da banda larga fixa por fibra óptica, de identificar se este cliente se tornará um inadimplente? É possível classificar os clientes com base em idade, gênero, ticket do produto contratado, dia de vencimento, forma de pagamento escolhida, classificação de sistemas de crédito como Sistema de Proteção ao Crédito (SPC) e, a partir desta classificação, determinar quais seriam os inadimplentes para evitar que estes entrem na carteira da operadora?

O setor de telecomunicações

As telecomunicações, enquanto serviço à sociedade, existem desde a invenção do rádio, mas é na década de 1980 que aconteceram significativos movimentos de expansão pelo mundo tendo como principal vetor de crescimento novas tecnologias e o barateamento de tecnologias já existentes (Dilon, et al. 2020). Há ainda a mudança, em muitos países do mundo, do sistema público para o

privado o que gerou competitividade e influenciou diretamente a oferta de melhores e mais baratos serviços na área. (Silva, Perobelli. 2018)

Ao longo de sua evolução ficou claro que o serviço é dividido em três camadas: física – que incluí os componentes mecânicos, físicos e elétricos; a camada de transporte – que compreende os serviços de manipulação de dados a fim que o serviço transite entre um ponto e outro caracterizado por adicionar valor ao espectro físico e; a camada de aplicação – na qual estão os serviços consumidos pelos clientes. Uma operadora de telecomunicações está especialmente situada nos componentes físicos e de transporte. (Dilon, et al. 2020)

Atualmente as redes de telecomunicações são majoritariamente compostas por camadas físicas que se utilizam de tecnologias ópticas, em especial, a fibra óptica. Considerado um ambiente extremamente competitivo, coexistindo, em um mesmo mercado, diversas operadoras, as nuances de prestação de serviço, qualidade e mesmo suspensão dos serviços por inadimplência não dependem mais exclusivamente das operadoras, mas sim do mercado em que estão inseridas. (Anjos, et al. 2021)

As operadoras só podem atuar em duas frentes para garantir a sustentação de sua base de clientes: nos serviços da camada de transporte (Ribeiro, et. al. 2020) e na melhoria dos controles de entrada de clientes para reduzir a inadimplência e os custos (Freitas, et. al. 2020).

Material e Métodos

O desenvolvimento do trabalho se deu com dados de uma empresa de telecomunicações, focada no serviço de acesso e conexão fixa de banda larga por fibra óptica, do oeste do estado do Paraná, sendo uma das 80 maiores do Brasil. A metodologia do estudo foi a quantitativa.

A população alvo foi a de usuários (consumidores do serviço de telecomunicações), contratantes do serviço de acesso e conexão fixa de banda larga por fibra óptica na modalidade de serviço residencial independente do seu tempo de carteira junto a empresa.

Como instrumento da coleta de dados foram realizadas consultas ao banco de dados privado da empresa listando, ao final de cada mês, diversas variáveis independentes e a variável dependente (DUM_inadImediata) usada nos modelos. As variáveis bem como a interpretação de seus nomes constam na Tabela 1.

Tabela 1. Variáveis empregadas na modelagem e a explicação de sua aplicação e condição

Variável	Especificação
Variável dependente	
DUM_inadImediata	- Sem atraso ou pagamento até o dia de vencimento da conta (*dummy* = 0) - Inadimplente Imediato (*dummy* = 1)
Variáveis independentes	
MET_contaRecValor	Valor da parcela da mensalidade
VAR_spcClasse	Categórica com 6 classificações que contêm 1 letra de A à F, transformada em *dummy*.
VAR_diaVcto	Categórica com 4 classificações que contém o dia do vencimento escolhido pelo usuário sendo 10, 15, 20 ou ainda outros dias agrupados, transformada em *dummy*.
VAR_clienteIdadeAnos	Idade do usuário em anos
VAR_planoFormaPgtoPro	Categórica com 2 classificações, transformada em *dummy*. - Modalidade Cartão - Modalidade Tradicional

Variável	Especificação
MET_clienteCadastroMeses	Tempo desde o primeiro cadastro do usuário em meses (usuário pode ter retornado a operadora e ter um cadastro mais antigo que o contrato)
MET_planoContratoMeses	Tempo desde a primeira contratação do bloco de planos atual em meses (usuário pode ter contratado um plano, exemplo, em 2019 e mudado de plano, sem cancelar, até 2022, esta variável contempla a contratação inicial)
MET_planoAtivacaoMeses	Tempo deste a ativação do plano atual em meses
VAR_clienteGenero	- Masculino (*dummy* = 0) - Feminino (*dummy* = 1)
VAR_spcSituacao	Categórica com 3 classificações sendo novo, restrito e não restrito, transformada em *dummy*.
MET_spcProbabilidade	Valor de 0 a 100 oriundo da consulta do SPC que trata da probabilidade de inadimplência de um determinado cliente na janela de 12 meses seguintes a consulta
MET_spcScore	Valor de 0 a 1000 oriundo da consulta SPC que trata do nível de *credit scoring* dado ao cliente pelo sistema do SPC
VAR_unidadeNegocio	Categórica com 4 classificações representando as unidades de negócio da empresa transformada em *dummy*: - CASCAVEL - SANTATEREZADOOESTE - TOLEDO - TUPASSI

Fonte: Dados do autor. Base de dados utilizada para modelagem.

A decisão pelas variáveis escolhidas se deu por dois fatores principais: i) a facilidade de coleta delas no momento da contratação do serviço pelo usuário, possibilitando, com poucas interações que elas sejam utilizadas em uma predição futura e; ii) pelos diversos trabalhos no âmbito acadêmico e empresarial que as utilizam, a saber:

- Idade e Gênero. Albuquerque et al. (2017); Amorim Neto e Carmona (2004); Guimarães e Chaves Neto (2002); Jannuzzi (2010); Locatelli et al. (2015); Maciel e Maciel (2017); Ritta, et al. (2015); Souza, et al. (2018).
- Tempo de relacionamento com a empresa. Albuquerque et al. (2017); Amorim Neto e Carmona (2004); Guimarães e Chaves Neto (2002).
- Valor do produto ou serviço. Amorim Neto e Carmona (2004); Guimarães e Chaves Neto (2002); Jannuzzi (2010); Ritta, et al. (2015).

A coleta dos dados foi realizada de março de 2021 a agosto de 2022 totalizando 449.106 observações e 14 variáveis coletadas que levaram a 29 variáveis para os modelos a serem desenvolvidos. Devido a diversas observações estarem indisponíveis (*missing values*) foi necessária a remoção destas resultando em um *dataset* com 91.900 observações.

Para garantir a aplicabilidade do modelo foram criados *datasets* para treino (contendo 80% das observações) e para teste (contendo os 20% restantes). Foram feitas outras separações de dataset com diferentes dimensionamentos entre treino e teste, mas não se mostraram tão eficientes quanto a escolhida.

O método de classificação será de Árvores de Decisão, um método comum de distribuição de decisões entre ramos iniciando sempre de uma pergunta e, a partir da avaliação de impactos de outras variáveis sobre a pergunta, definem-se os rumos, binários, desta decisão. Larose, (2005); Brodley e Utgoff, (1995); Weinberg e Last (2019).

O modelo será comparado com outros modelos de *machine learning* como:

- Regressão logística, que estuda a probabilidade de ocorrência de um evento binário como a inadimplência, 1 para o inadimplente e 0 para o adimplente. Fávero, Belfiore, (2017); Gouvêa, et al. (2013).
- Árvore de Decisão utilizando o algoritmo de Random Forest, que geram diversos segmentos de dados de forma aleatória usando a técnica de *bootstrap*. Ibañez (2016)

Resultados e Discussão

A análise dos dados considera como ponto de interesse principal um modelo de predição que possibilite a minimização da perda financeira da empresa com relação aos inadimplentes, ou seja, procura-se que os possíveis inadimplentes sejam identificados com maior assertividade mesmo que para isso perca-se capacidade preditiva nos clientes adimplentes. Neste sentido é importante buscar acurácia acima de 65% e sensitividade acima de 80% (Lima, et al. 2021). Para tal serão gerados modelos diferentes e depois comparados entre si para a decisão do melhor modelo a ser adotado.

Estatística Descritiva

Consultando os dados, verifica-se que a variável de inadimplência exibe maior frequência (73,2%) de dados em adimplentes o que é esperado uma vez que o usuário do serviço deve pagar por ele em dia para evitar o corte, também se nota uma concentração de clientes na unidade de CASCAVEL (87,4%) das observações.

Há uma pequena distorção nas unidades de negócio em que a (CERTTOB4B), a classe (E) e a situação (NOVO) não tem dados suficientes, aparentemente, para ter uma boa explicação no modelo o que possivelmente resultará na exclusão da variável *dummy* desta classificação. A Tabela 3 mostra a estatística descritiva das variáveis discretas do modelo.

Tabela 3. Estatística descritiva para as variáveis discretas

Variável	**Mean (SD)**	**Median [Min, Max]**
MET_contaRecValor	114 (22.2)	120 [5.00, 500]
VAR_clienteIdadeAnos	39.1 (13.1)	37.0 [18.0, 89.0]
MET_planoAtivacaoMeses	9.52 (13.8)	5.00 [0, 176]
MET_clienteCadastroMeses	23.1 (42.7)	7.00 [0, 324]
MET_planoContratoMeses	7.37 (9.21)	5.00 [0, 485]
MET_spcProbabilidade	27.5 (40.8)	4.85 [0.220, 100]
MET_spcScore	531 (328)	640 [0, 998]

Fonte: dados do autor.

A estatística descritiva mostra que o valor médio das contas a receber é de R$ 114 com um desvio padrão de 22,2 o que é corroborado em dados da empresa para a base completa de clientes. Os clientes desta empresa têm em média 38,1 anos sendo a mediana em 37 anos com mínimos em 18 anos (maioridade legal) e máximos em 89 anos.

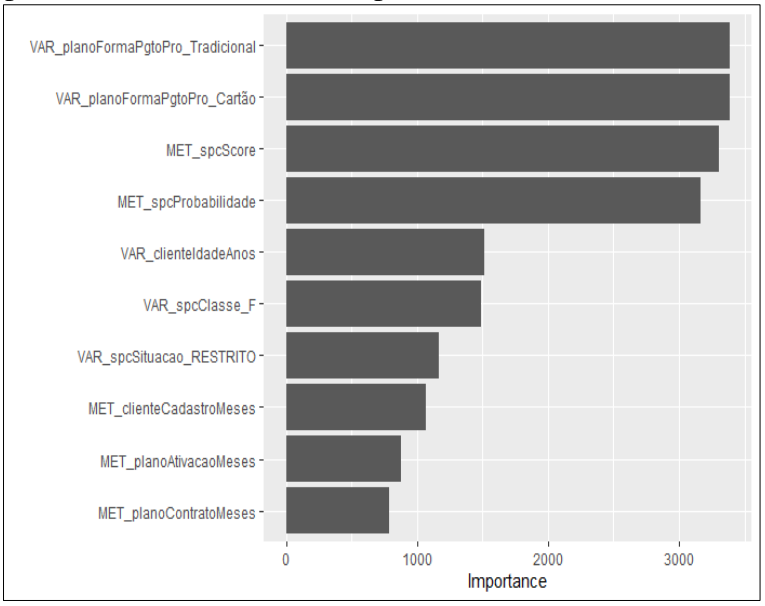
Nota-se ainda algumas curiosidades sobre os dados como o máximo tempo em meses de um cliente desta base é de 324, ou seja, 27 anos. A empresa foi fundada em 1995 (27 anos atrás) e por isso é plausível que existam clientes com cadastros tão antigos quanto este. A mediana no tempo de ativação dos planos é de 5 meses enquanto a média é de 9,52 meses ao mesmo tempo que a mediana nos contratos é de 5 meses e a média de 7,37 meses.

Árvore de Decisão (CART_model)

Uma árvore de decisão ou, do inglês, *Classification and Regression Trees* (CART) é uma técnica que visa encontrar quais variáveis possuem maior carga de informação para responder a melhor classificação das perguntas, desde que corretas, realizadas. Kelleher, et al. (2015).

Neste modelo utilizando-se de um custo de complexidade (cp) de 0,00001 para se encontrar o menor erro entre as predições chega-se um erro de 0.80629 com um cp de 0.000094300 que será usado para poda da árvore. A Figura 1 mostra a importância de cada variável do modelo.

Figura 1. Importâncias das variáveis independentes no modelo de árvore de decisão

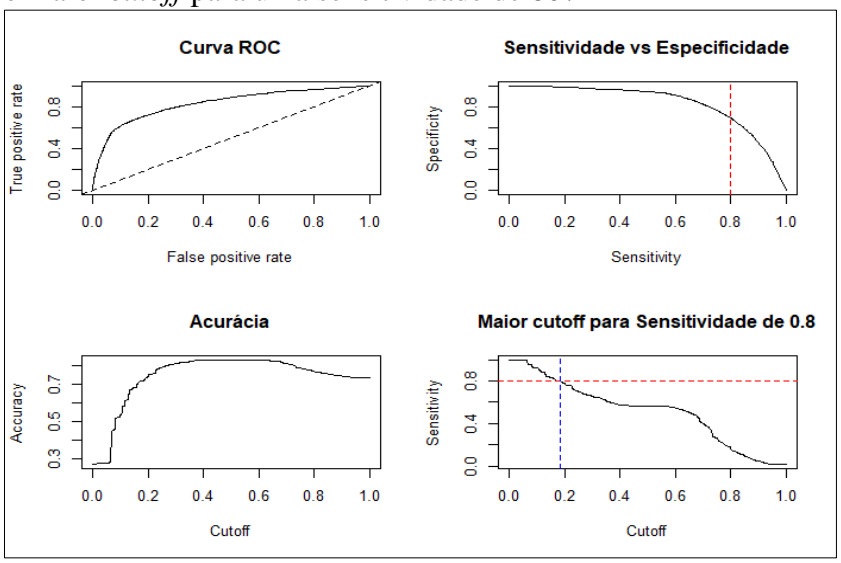


Fonte: Resultados da pesquisa

A forma de pagamento tradicional é a variável que mais influência no modelo de decisão, seguida pelo *credit scoring,* probabilidade e classe do tipo F, todos provenientes do SPC. Em seguida aparece a idade do cliente o que é plenamente corroborado pela literatura uma vez que se espera que quanto mais idade tenha o cliente menor seja a chance deste se tornar inadimplente (Januzzi, 2010). Também se nota que a partir deste ponto as variáveis, nas 10 mais importantes, que influenciam são aquelas de tempo de relacionamento com a empresa o que é aderente a outros modelos encontrados na literatura (Gouvêa, et al. 2013).

A árvore depois de podada foi usada para predição com base no *cutoff* que retornasse pelo menos 80% de sensitividade, conforme exibe a Figura 2.

Figura 2. Curva ROC, sensitividade, especificidade e acurácia para o modelo de árvore de decisão podada buscando o maior *cutoff* para uma sensitividade de 80%



Fonte: Resultados da Pesquisa

Com base no modelo o menor *cutoff* encontrado foi de 0.1829268 que gerou a matriz de classificação exibida na Tabela 4.

Tabela 4. Matriz de Classificação do Modelo de Árvore de Decisão na base de treino

Observado	Previsão do Modelo			
	Adimplente (0)	Inadimplente (1)	Observações	Nível de Acerto
Adimplente (0)	37195	16631	53826	Especificidade: 69,10%
Inadimplente (1)	3933	15761	19694	Sensitividade: 80,03%
				Acurácia: 72,03%

Fonte: resultados da pesquisa.

O resultado encontrado é satisfatório para o modelo com um índice de acurácia elevado, na casa dos 72% e com elevado índice de sensitividade (>80%). O modelo foi aplicado ao conjunto de dados de teste para aferir a funcionalidade e capacidade preditiva deste. Para a geração da informação foi utilizado o mesmo cutoff da base de treino sobre as probabilidades calculadas na base de teste. O resultado é exibido na Tabela 5.

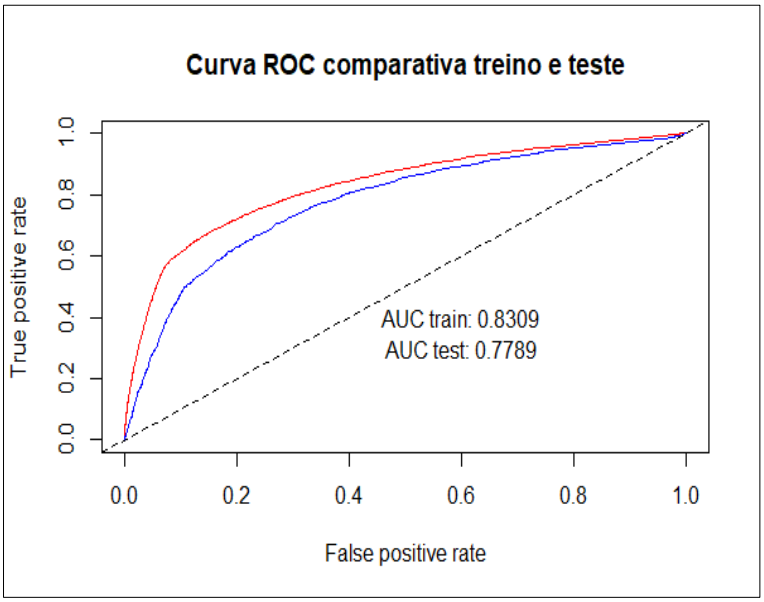
Tabela 5. Matriz de Classificação do Modelo de Árvore de Decisão na base de teste

Observado	Previsão do Modelo			
	Adimplente (0)	Inadimplente (1)	Observações	Nível de Acerto
Adimplente (0)	9148	4337	13485	Especificidade: 67,84%
Inadimplente (1)	1225	3670	4895	Sensitividade: 74,97%
				Acurácia: 69,74%

Fonte: resultados da pesquisa.

Apesar do resultado na base de testes ter sido inferior a base de treino ainda assim o modelo se mostrou promissor e viável para aplicação com uma acurácia maior que 65% e uma sensitividade bastante elevada. A Figura 3 mostra a capacidade preditiva do modelo com base na Curva ROC comparativa entre a base de treino e a base de teste.

Figura 3. Curva ROC das bases de treino e teste do modelo de Árvore de Decisão.



Fonte: resultados da pesquisa

Regressão Logística (GLM_model)

Os modelos de regressão logísticos binários são amplamente usados nas áreas de risco de crédito e consideram a probabilidade de um determinado cliente se tornar inadimplente. É um método de aprendizado de máquina supervisionado que dá melhor previsibilidade se comparado a outros métodos. Lima, et al. (2021).

Todas as variáveis disponíveis no banco de dados foram utilizadas para modelagem e, após isso, o procedimento de revisão das variáveis significativas (*stepwise*) foi aplicado resultando na Tabela 6.

Tabela 6. Modelo logístico binário integral e com procedimento *stepwise* para as variáveis independentes significativas

	GLM_model		**GLM_model_step**	
	Estimativa	**Erro Padrão**	**Estimativa**	**Erro Padrão**
(Intercept)	1.2664918 ***	0.1966265	1.3775198 ***	0.1880899
MET_contaRecValor	-0.0016195 ***	0.0004440	-0.0016584 ***	0.0004434
VAR_clienteIdadeAnos	-0.0103477 ***	0.0008011	-0.0104984 ***	0.0007954
MET_planoAtivacaoMeses	-0.0055955	0.0010109	-0.0057981 ***	0.0009773
MET_clienteCadastroMeses	-0.0002139 ***	0.0002587	NA ***	NA
MET_planoContratoMeses	-0.0083991 **	0.0015104	-0.0084091 ***	0.0015103
MET_spcProbabilidade	-0.0046623 ***	0.0015322	-0.0057406 ***	0.0014348

MET_spcScore	-0.0021173	0.0002443	-0.0020087 ***	0.0001689
VAR_spcClasse_A	0.1380383	0.1962536		
VAR_spcClasse_B	0.2098114	0.1732148		
VAR_spcClasse_C	0.2583540 *	0.1486262	0.0854201 *	0.0349724
VAR_spcClasse_D	0.3218694 **	0.1280580	0.1707298 **	0.0533749
VAR_spcClasse_E	0.3315230	0.1176103	0.2037269 **	0.0781706
VAR_spcClasse_F	NA ***	NA		
VAR_diaVcto_10	-0.2203126 ***	0.0243557	-0.2203637 ***	0.0243540
VAR_diaVcto_15	-0.2394162 ***	0.0282011	-0.2390775 ***	0.0281972
VAR_diaVcto_20	0.1327065	0.0279140	0.1332953 ***	0.0279039
VAR_diaVcto_Outros	NA **	NA		
VAR_unidadeNegocio_CASCAVEL	0.2922975 ***	0.1084510	0.2963727 **	0.1083133
VAR_unidadeNegocio_SANTATEREZADOOESTE	0.4909288 ***	0.1203412	0.4938608 ***	0.1202773
VAR_unidadeNegocio_TOLEDO	0.4209088	0.1111818	0.4238801 ***	0.1110640
VAR_unidadeNegocio_TUPASSI	NA ***	NA		
VAR_planoFormaPgtoPro_Cartão	-1.4369090	0.0195910	-1.4373104 ***	0.0195781
VAR_planoFormaPgtoPro_Tradicional	NA ***	NA		
VAR_clienteGenero_FEMININO	0.0668754	0.0186779	0.0672595 ***	0.0186422
VAR_clienteGenero_MASCULINO	NA ***	NA		
VAR_spcSituacao_NAO_RESTRITO	-0.2662401 ***	0.0272035	-0.2693352 ***	0.0271084
VAR_spcSituacao_NOVO	-0.3477890	0.0647287	-0.3449944 ***	0.0646324
VAR_spcSituacao_RESTRITO	NA	NA		

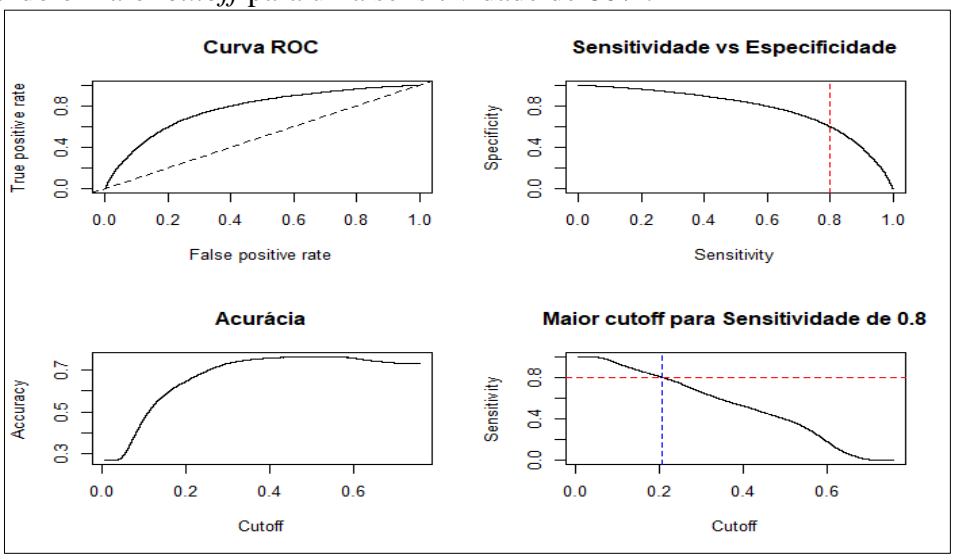
Fonte: Resultados da pesquisa
Obs: Nível de significância: *** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

Como os modelos foram estimados pelo método de Máxima Verossimilhança (Fávero, 2017) faz-se necessário a confirmação estatística de que os termos de erros estão menores entre os modelos e para isso compara-se o *Log Likelihood* em que no modelo completo tinha o valor de -35930 e no modelo com as variáveis significativas chegou a -35932 uma perda imperceptível de performance na qualidade da predição com um número muito menor de variáveis (9 a menos).

A retirada das variáveis pelo procedimento de *stepwise* reduz a necessidade de recursos computacionais para o processamento do modelo e facilita o trabalho de aplicação dele em dados posteriores, diminuindo o número de variáveis. Conforme consta na explicação das variáveis nota-se que (MET_clienteCadastroMeses) não tem significância estatística pois é possível que um cliente seja usuário do serviço desta operadora de telecom diversas vezes, não simultâneas, durante sua vida. Os estimadores e seus sinais também têm aderência ao modelo e a outros trabalhos (Gouvêa, et al. 2021) como, por exemplo, a variável (VAR_clienteIdadeAnos) em que é esperado que um consumidor de mais idade esteja financeiro estável e por isso tenha menor (sinal negativo) tendência a inadimplência. Bem como é esperado que um cliente de classe C, D ou E no SCP tenha um estimador positivo maior que um de classe A.

Foi aplicado também a modelagem de escolha do *cutoff* para garantir pelo menos 80% de Sensitividade no modelo, conforme exibe a Figura 4.

Figura 4. Curva ROC, sensitividade, especificidade e acurácia para o modelo de regressão logística binária buscando o maior *cutoff* para uma sensitividade de 80%.



Fonte: Resultados da pesquisa

Com base no modelo o menor *cutoff* encontrado foi de 0.2066074 que gerou a matriz de classificação exibida na Tabela 7.

Tabela 7. Matriz de Classificação do Modelo Logístico Binário na base de treino

Observado	Previsão do Modelo			
	Adimplente (0)	Inadimplente (1)	Observações	Nível de Acerto
Adimplente (0)	32181	21645	53826	Especificidade: 59,79%
Inadimplente (1)	3938	15756	19694	Sensitividade: 80,00%
				Acurácia: 65,20%

Fonte: Resultados da pesquisa

O resultado encontrado é satisfatório para o modelo com um índice de acurácia elevado, na casa dos 65% e com elevado índice de sensitividade (80%). O modelo foi aplicado ao conjunto de dados de teste para aferir a funcionalidade e capacidade preditiva deste. Para a geração da informação foi utilizado o mesmo cutoff da base de treino sobre as probabilidades calculadas na base de teste. O resultado é exibido na Tabela 8.

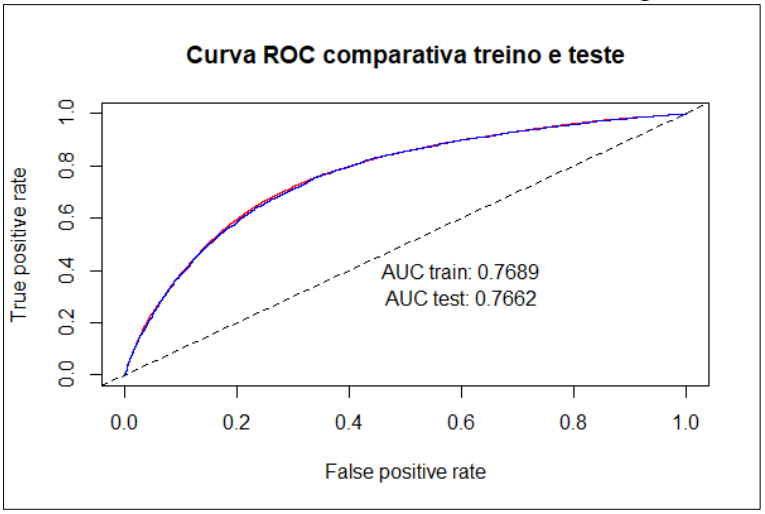
Tabela 8. Matriz de Classificação do Modelo Logístico Binário na base de teste

Observado	Previsão do Modelo			
	Adimplente (0)	Inadimplente (1)	Observações	Nível de Acerto
Adimplente (0)	7966	5519	13485	Especificidade: 59,07%
Inadimplente (1)	961	3934	4895	Sensitividade: 80,37%
				Acurácia: 64,74%

Fonte: Resultados da pesquisa

O resultado da base de teste foi bastante satisfatório e muito próximo ao encontrado na base de treino. A Figura 5 mostra a capacidade preditiva do modelo com base na Curva ROC comparativa entre a base de treino e a base de teste.

Figura 5. Curva ROC das bases de treino e teste do modelo de Regressão Logística Binária



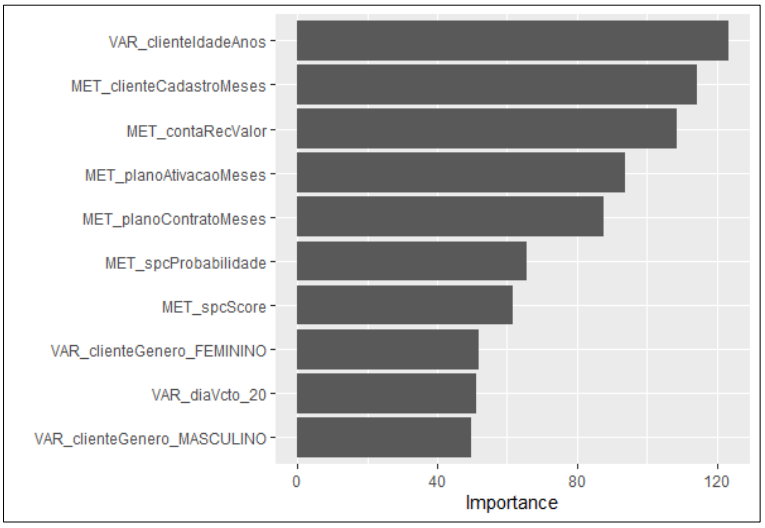
Fonte: Resultados da pesquisa

Árvore de Decisão: Random Forest (RF_model)

Uma variação da árvore de decisão, que utiliza um algoritmo de divisão da base de treino, conhecido como *bagging*, pode ser usado também para a definição de inadimplência. Em muitos casos esse método tem melhor capacidade preditiva que outros comumente utilizados. Madaan, et al. (2021).

Utilizou-se um total de 200 árvores, valor que se mostrou mais eficiente na acurácia (74,68%) que de forma aleatória utilizaram até 9 variáveis (mtry = 9). A Figura 6 mostra a importância de cada variável do modelo.

Figura 6. Importâncias das variáveis independentes no modelo de árvore de decisão utilizando o algoritmo de Random Forest

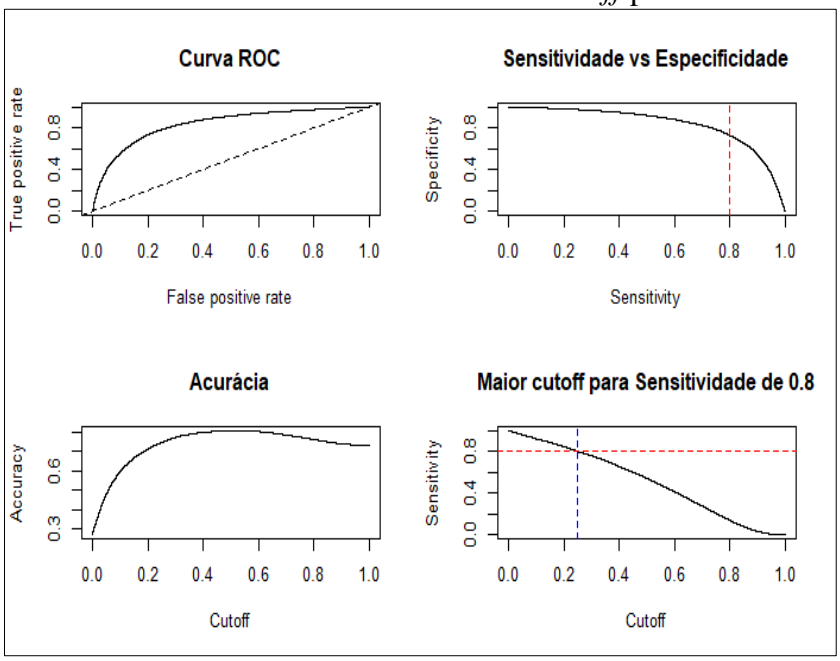


Fonte: Resultados da pesquisa

Há uma mudança considerável na ordenação de importância das variáveis em relação ao modelo de árvore de decisão tradicional, agora a variável mais importante é a idade do cliente seguida do tempo que este cliente se relaciona com a empresa que são amplamente discutidas na literatura (Januzzi, 2010), em seguida aparece o valor do serviço contratado, mais variáveis de tempo de relacionamento, que também são exploradas na literatura (Gouvêa, et al. 2013) e, por fim, variáveis de *credit scoring*. A forma de pagamento deixou de ter importância entre as top 10 neste modelo.

Foi analisada a curva ROC a partir da definição de um *cutoff* que retornasse pelo menos 80% de sensitividade, conforme exibe a Figura 7.

Figura 7. Curva ROC, sensitividade, especificidade e acurácia para o modelo de árvore de decisão utilizando o algoritmo de Random Forest buscando o maior *cutoff* para uma sensitividade de 80%



Fonte: Resultados da Pesquisa

Com base no modelo o menor *cutoff* encontrado foi de 0.2500985 que gerou a matriz de classificação exibida na Tabela 9.

Tabela 9. Matriz de Classificação do Modelo de Árvore de Decisão utilizando o algoritmo de Random Forest na base de treino

Observado	Previsão do Modelo			
	Adimplente (0)	Inadimplente (1)	Observações	Nível de Acerto
Adimplente (0)	39148	14678	53826	Especificidade: 72,73%
Inadimplente (1)	3938	15756	19694	Sensitividade: 80,00%
				Acurácia: 74,68%

Fonte: resultados da pesquisa.

O resultado encontrado é satisfatório para o modelo com um índice de acurácia elevado, na casa dos 74% e com elevado índice de sensitividade (80%). O modelo foi aplicado ao conjunto de

dados de teste para aferir a funcionalidade e capacidade preditiva deste. Para a geração da informação foi utilizado o mesmo cutoff da base de treino sobre as probabilidades calculadas na base de teste. O resultado é exibido na Tabela 10.

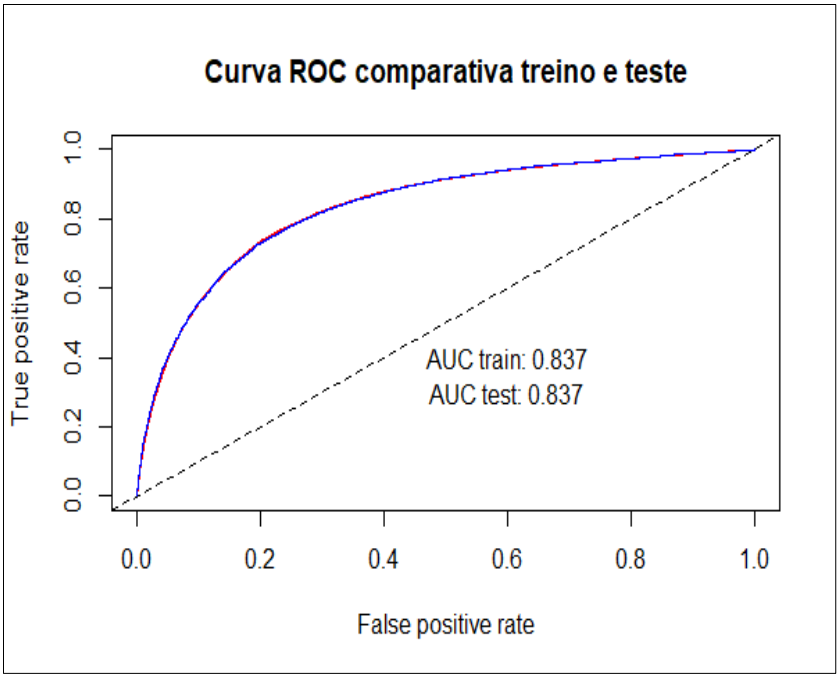
Tabela 10. Matriz de Classificação do Modelo de Árvore de Decisão utilizando o algoritmo de Random Forest na base de teste

Observado	Previsão do Modelo			
	Adimplente (0)	Inadimplente (1)	Observações	Nível de Acerto
Adimplente (0)	9753	3732	13485	Especificidade: 72,32%
Inadimplente (1)	970	3925	4895	Sensitividade: 80,18%
				Acurácia: 74,42%

Fonte: resultados da pesquisa.

O resultado na base de teste foi excelente e teve pouca perda de acurácia. A Figura 8 mostra a capacidade preditiva do modelo com base na Curva ROC comparativa entre a base de treino e a base de teste.

Figura 3. Curva ROC das bases de treino e teste do modelo de Árvore de Decisão utilizando o algoritmo de Random Forest.



Fonte: resultados da pesquisa.

Comparativo entre modelos

Para ter uma ampla visão sobre os modelos testados e assumir o melhor para o desempenho da função de predição é importante verificar os parâmetros comparáveis entre eles com base na aferição da base de teste. A Tabela 11 mostra estes parâmetros.

Tabela 11. Comparativo entre modelos utilizados

Modelo	CART_model	GLM_model_step	RF_model
Acurácia	69,74%	64,74%	74,42%
Especificidade	67,84%	59,07%	72,32%
Sensitividade	74,97%	80,37%	80,18%
AUC	0,7789	0,7662	0,8369

Fonte: resultados da pesquisa.

O desempenho do modelo de árvore de decisão utilizando o algoritmo de Random Forest é evidente e muito superior aos demais modelos tendo a maior acurácia, mantendo um ótimo balanço na sensibilidade e mostrando desempenho na área abaixo da curva ROC.

Comparativo com outros estudos

Com o intuito de perceber mudança ou vieses no presente estudo foram comparados os resultados obtidos, em especial dos estimadores do modelo logístico, com a literatura disponível:

- A idade do cliente demonstrou ter alta importância nas decisões e teve seu estimador com sinal negativo no modelo logístico significando que, quando mais idade tem o cliente, menor a chance dele se tornar inadimplente. Este comportamento está em linha Gouvêa, et. al. (2013), Lima, et. al. (2021) e Januzzi (2010) mas contrário a Locatelli et al. (2015) e Maciel e Maciel (2017).
- Gênero teve peso significativo no modelo logístico e participação na árvore de decisão. Comportamento semelhante foi encontrado em Maciel e Maciel (2017) e Ritta, et al. (2015).
- O tempo de relacionamento do cliente com a empresa, medido neste estudo por 3 variáveis, MET_clienteCadastroMeses, MET_planoAtivacaoMeses e MET_planoContratoMeses, se mostrou significativo em todos os modelos e teve seu estimador com sinal negativo no modelo logístico significando que, quanto mais tempo o cliente permanece com a empresa, menor a chance dele se tornar inadimplente em linha com Albuquerque et al. (2017), Amorim Neto e Carmona (2004), Guimarães e Chaves Neto (2002) e Lima, et. al. (2021).
- Por fim o valor do serviço ou do produto também teve significância nos modelos tendo seu estimador negativo no modelo logístico significando que quanto maior o valor da conta a pagar menor a chance de inadimplência. Diferente dos estudos de Amorim Neto e Carmona (2004), Guimarães e Chaves Neto (2002), Jannuzzi (2010), Lima, et. al. (2021) e Ritta, et al. (2015). A ressalva é que todos os estudos apresentados são de empréstimos bancários de alto valor e aqui, o valor é de um serviço de telecomunicações, mensal, muito inferior a uma parcela ou mesmo totalizador de um financiamento.

Considerações Finais

Este estudo teve como objetivo principal identificar a capacidade preditiva de inadimplentes utilizando modelos de aprendizado de máquina diferentes e medindo sua capacidade para a aplicação real dentro de uma empresa do ramo de telecomunicações de nível médio.

O objetivo foi alcançado em todos os modelos propostos, desde a árvore de regressão, ao modelo logístico geral e por fim no modelo utilizando a técnica de Random Forest. Neste sentido, sim, é possível identificar e prever um comportamento de inadimplência de clientes nestes dados com uma acurácia em nível adequado, maior que 65%, e ainda uma sensitividade, ou seja, a correta predição de positivos verdadeiros em excelente desempenho chegando à casa de 80%.

Também foi possível perceber a relação entre este estudo e outros estudos do mercado financeiro, com características de estudos diferentes, mas com variáveis independentes similares tendo comportamentos muito alinhados com a literatura o que engrandece a confiança na capacidade preditiva do modelo.

O modelo utilizando Randon Forest (RF_model) foi implantado em produção na empresa para monitoramento e melhorias das estratégias de aceitação de novos clientes, melhorando de fora sensível os problemas de caixa, pela baixa inadimplência, e os problemas de rescisão de contratos, conhecido por *churn*.

Referências

Albuquerque, P. H. M.; Medina, F. A. S.; Silva, A. R. da. (2017). Regressão Logística Geograficamente Ponderada Aplicada a Modelos de Credit Scoring. Revista Contabilidade e Finanças, 28(73), 93-112. doi:10.1590/1808-057x201703760.

Amorim Neto, A. A.; Carmona, C. U. D. M. (2004). Modelagem do risco de crédito: Um estudo do segmento de pessoas físicas em um banco de varejo. REAd-Revista Eletrônica de Administração, 10(40), 1-23.

Anjos, V. S.; Bortoletto, R. C.; Waldman, H. (2021). A influência do Comportamento dos Usuários e das Operadoras de Telecomunicações em uma Rede Óptica Elástica. Anais da XIX Escola Regional de Redes de Computadores, (pp. 13-18). Porto Alegre: SBC. doi:10.5753/errc.2021.18535

Brodley, C.E.; Utgoff, P.E. (1995). Multivariate decision trees. Machine learning, 19(1): 45-77. Disponível em: <https://link.springer.com/article/10.1007/BF00994660>. Acesso em: 02 agosto 2022.

Dilon, M. L. S.; Betarelli Jr., A. A.; Faria, W. R.; Montenegro, R. L. G. (2020). Produtividade e Influências Intersetoriais das Telecomunicações nas Economias Mundiais. Economia Ensaios, v. 36, n. 1, 2021. DOI: 10.14393/REE-v36n1a2021-46254

Fávero, L.P.; Belfiore, P. (2017). Manual de análise de dados: estatística e modelagem multivariada com Excel®, SPSS® e Stata® (1a ed.). Rio de Janeiro: Elsevier.

Freitas, L. C.; Prado, T. S.; Souza Filho, A. L.; Moura Filho, R. N.; Baigorri, C. M.; Morais, L. E. (2020). Economia do compartilhamento de infraestruturas no setor de telecomunicações brasileiro: inventário e o desenho de um mecanismo geral de compartilhamento. Revista Latinoamericana de Economia y Sociedad Digital, Issue 1, agosto 2020. DOI: 10.53857/DMTI9200

Gouvêa, M. A.; Gonçalves, E. B.; Mantovani, D. (2013). Análise de risco de crédito com o uso de regressão logística. Revista Contemporânea de Contabilidade. doi: 10.5007/2175-8069.2013v10n20p139.

Guimarães, I. A.; Chaves Neto, A. (2002). Reconhecimento de padrões: metodologias estatísticas em crédito ao consumidor. RAE Eletrônica, 1(2), 1-14. doi:10.1590/s1676-56482002000200006.

Ibañez, M. M. (2016). Uso de Redes Neurais Nebulosas e Florestas Aleatórias na Classificação de Imagens em um Projeto de Ciência Cidadã (Dissertação de Mestrado). Instituto Nacional de Pesquisas Espaciais, São José dos Campos, SP, Brasil.

Jannuzzi, F. C. K. (2010). Um estudo sobre as variáveis que impactam a inadimplência no crédito concedido para projetos imobiliários (Dissertação de Mestrado). Universidade Estácio de Sá, Rio de Janeiro, RJ, Brasil.

Jiang, J.; Liao, L.; Xi, L.; Wang, Z.; Xiang, H. (2021). Deciphering big data in consumer credit. Journal of Empirical Finance. doi:10.1016/j.jempfin.2021.01.009.

Kelleher, J. D.; Namee, B. M.; D'Arcy, A. (2015). Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies. MIT Press.

Larose, D. T. (2005). Data Mining: Concepts and Techniques, 2ed. Elsevier. San Francisco, CA, USA.

Lima, R. B.; Serra, R. G.; Fávero, L. P. L. (2021). Determinantes Hierárquicos da Inadimplência de Financiamento Imobiliário de Pessoa Física. Advances in Scientific and Applied Accouting (ASAA). doi: 10.14392/asaa.2021140202.

Locatelli, R. L.; Ramalho, W.; Silvério, R. A. de O.; Afonso, T. (2015). Determinantes da inadimplência no crédito habitacional direcionado a classe média emergente brasileira. Revista de Finanças Aplicadas, 1(1), 1-30.

Lopes, M. G.; Ciribeli, J. P.; Massardi, W. D. O.; Mendes, W. D. A. (2017). Análise dos indicadores de inadimplência nas linhas de crédito para pessoa física: um estudo utilizando modelo de regressão logística. Estudos Do CEPE. doi: 10.17058/cepe.v0i46.11099.

Maciel, H. M.; Maciel, W. M. (2017). Análise da Inadimplência Bancária: Um Estudo de Caso da Região Metropolitana de Fortaleza. Conexões - Ciência e Tecnologia, 11(3), 12–23. doi:10.21439/conexoes.v11i3.867.

Madaan, M.; Kumar, A.; Keshri, C.; Jain, R.; Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. IOP Conf. Ser. Mater. Sci. Eng. DOI: 10.1088/1757-899x/1022/1/012042

Ribeiro, D.; Reichardt, C.; Neves, M. V. (2020). Migração de servidores para nuvem: estudo de caso de provedor de internet Foxnet Telecomunicações. IGNIS: Periódico Científico de Arquitetura e Urbanismo, Engenharias e Tecnologia da Informação, 9(1). Disponível em: <https://periodicos.uniarp.edu.br/index.php/ignis/article/view/2345>. Acesso em: 09 de Setembro de 2022.

Ritta, C. de O.; Gorla, M. C.; Hein, N. (2015). Modelo de regressão logística para análise de risco de crédito em uma instituição de microcrédito produtivo orientado. Iberoamerican Journal of Industrial Engineering, 7(13), 103-122. doi:10.13084/2175-8018/ijie.v7n13p103-122.

Rahul, De; Neena, P.; Abhipsa, P. (2020). Impact of digital surge during Covid-19 pandemic: A viewpoint on research and practice. International Journal of Information Management, 55, 102171. doi: 10.1016/j.ijinfomgt.2020.102171.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. Neural Networks, 61(1): 85-117. Doi: 10.1016/j.neunet.2014.09.003.

Silva, G. D; Perobelli, F. S. (2018). Interconexões Setoriais e PIB per capita: há relação direta entre ambas as variáveis? Estud. Econ., São Paulo, vol.48 n.2, p. 251-282, abr.-jun. 2018. DOI: 10.1590/0101-41614823gdfp.

Sousa, Q. H.; Petri, S. M.; Anjos, E. A. (2018). Análise dos fatores preditivos de risco para inadimplência dos cooperados em uma cooperativa de crédito. III Congresso de Contabilidade da UFRGS e III Congresso de Iniciação Científica em Contabilidade da UFRG, Porto Alegre, RS, Brasil, 3.

Weinberg, A.I.; Last, M. (2019). Selecting a representative decision tree from an ensemble of decision-tree models for fast big data classification. Journal of Big Data, 6(1): 1-17. Disponível em: <https://link.springer.com/article/10.1186/s40537-019-0186-3>. Acesso em: 02 de agosto de 2022.