PREVENÇÃO PERDAS COM INADIMPLÊNCIA: Uma proposta prática para o uso de Inteligência Artificial na otimização de recursos públicos

Hector Honorio Santos Tomelin[(1)]
Atuação profissional como empresário desde 1987, (gestão acadêmica, de 84 a 98), Autor do Institucional do Estado de Minas Gerais para investidores (Minas an Outlook), Atuação *off-shore* para Jostens Learning Systems, CA, USA. Atuação em Sivicultura e psicultura, setor imobiliário (condomínio rural). Professor na UNA, SENAC, SENAI, Católica-SC..Coordenador robótica e Diretor na SUCESU-MG. Administrador, Mestre em Economia. Analista econômico SAMAE-JS, Presidente Conselho Fiscal do Instituto Seguridade Municipal.

Ricardo Alexandre de Mello Oliveira [(2)]
Graduado em Ciências Econômicas e Contábeis. Pós-graduação (*latu sensu*) Ciência de dados, Mestre em Engenharia da Produção. Doutorando em Engenharia e Gestão do Conhecimento. Analista Econômico da CIA Águas de Joinville.

Christiano Piccinin[(3)]
Engenheiro de Dados. Graduado em sistema de informação pela UNIVILLE. Experiência em criação de arquiteturas de sistemas de dados, otimização de processos e inovação tecnológica.

Alexandre Leopoldo Gonçalves(4)
Graduado em Ciência da Computação pela FURB em 1997, Mestrado e Doutorado em Engenharia de Produção pela UFSC em 2000 e 2006. É Professor Associado no Departamento de Computação/UFSC e Professor Permanente do Programa de Pós-Graduação em Engenharia e Gestão do Conhecimento/UFSC. Possui experiência em áreas como Ciência de Dados, Aprendizado de Máquina e Inteligência Artificial.

Sidney Marques de Oliveira Jr[(5)]
Graduado em Ciências Contábeis, Pós-graduação (*latu sensu*) em Finanças e Controladoria, Mestre em Gestão de Negócios. Mais de 30 anos de mercado, sendo 15 anos no setor de infraestrutura. Presidente da CIA Águas de Joinville.

Endereço[(1)]: Rua Carlos Frederico Ramthum 13.059 – Casa – Bairro Vila Chartres – Jaraguá do Sul – SC – CEP 89.267-000 – Brasil – Tel: +55 (47) 3275-1567 – celular: +55 (47) 98447-9567 – email: htomelin@outlook.com – https://orcid.org/0009-0004-3961-8059.

ABSTRACT

Sanitation is fundamental to human life. Brazil's sanitation services are tariffed. When a user does not pay the tariff, the others must pay for the defaulter. The penalty creates stress between the defaulters and service provider. The specific prediction of the defaults enables preventive actions to reduce the consequences of stress. To develop a prediction model, data from a city very close to the universalization of services, Jaraguá do Sul, SC, *was used. The analysis was based on the evolution of payment terms over 25 months, using AutoML tools. The best performing algorithm was LGBM with an assertiveness of 97.09%.*

*keywords***:** *Auto machine learning; AutoML; Sanitation; Default; Systemic costs*.

RESUMEM

El saneamiento es fundamental para la vida humana. En Brasil los servicios de saneamiento son tarifados. Cuando algún usuario no paga la tarifa, los demás tienen que costear el deudor. La penalización genera estrés entre los deudores y el proveedor de servicio. La capacidad de predecir quien se tornará deudor abre la posibilidad de actuar preventivamente para la reducción de consecuencias perjudiciales para esta situación. Para desarrollo del modelo de predicción se utilizaron los datos de una ciudad que está muy cerca de la universalización de los servicios, Jaraguá do Sul, SC. *El análisis se realizó sobre la evolución de los plazos del pago efectivo a lo largo de 25 meses, con herramienta de AutoML. El algoritmo con mejor rendimiento fue el LGBM con asertividad de 97,09%.*

Palabras clave: Aprendizaje automatizado; AutoML; Saneamiento; Deuda; Costos sistémicos.

RESUMO

O saneamento é fundamental para a vida humana. No Brasil os serviços de saneamento são tarifados, sendo as empresas envolvidas impactadas quando algum usuário não efetua o pagamento da tarifa. A capacidade de prever quem se tornará inadimplente, abre a possibilidade de agir preventivamente para a redução das consequências nefastas desta situação. Para desenvolver um modelo de predição, usou-se os dados de uma cidade bem próxima da universalização dos serviços, Jaraguá do Sul[1], SC. A análise foi realizada com base na evolução dos prazos de pagamento ao longo de 25 meses, com ferramentas de AutoML. O algoritmo com melhor desempenho foi o LGBM (Light Gradient Boosting Machine) com uma assertividade de 97,09%.

Palavras-chave: Aprendizado de máquina automatizada; Saneamento; Inadimplência; Custos sistêmicos.

1. INTRODUÇÃO

1.1. CONCEITUANDO SANEAMENTO BÁSICO

O saneamento básico é essencial para o desenvolvimento humano, conforme o 6º Objetivo de Desenvolvimento Sustentável das Nações Unidas (do inglês *Sustainable Development Goals* – SDG), prevendo salvaguardas e gestão sustentável para a universalização de serviços sanitários e disponibilidade de água potável (United Nations, 2015). A falta de serviços sanitários (recolher e afastar o esgoto) adequados diminui o nível de saúde da população, sendo associado a doenças como diarreia, desenvolvimento cognitivo e altas taxas de mortalidade. Por outro lado, a efetividade destes serviços reduz os índices de diversas doenças, trazendo algumas para a quase inexistência (Sinharoy et al., 2019).

O saneamento pode ser visto como um conjunto de ações e serviços que possam prover de forma adequada, organizada e equitativa higiene e água potável. (Capone, 2017).

No Brasil, o conjunto de serviços para prover o saneamento é feito por entidades públicas ou privadas. Estas organizações devem criar toda a estrutura operacional para atender estes serviços: captação, tratamento e distribuição da água potável; coleta, afastamento (transporte) e tratamento do esgoto; coleta, transporte e destinação dos resíduos sólidos, bem como os serviços de drenagem pluvial. Exceto a drenagem pluvial, os serviços são tarifados individualmente, ou seja, usuário por usuário (Borges et al., 2022). Os dados do **S**istema **N**acional de **I**nformações sobre **S**aneamento (SNIS[2]) demonstram que estas tarifas provêm os recursos para estes serviços na maioria casos. (Brasil, 2022).

1.2 MAGNITUDE DA INADIMPLÊNCIA EM JARAGUÁ DO SUL

Um aspecto de grande potencial de impacto é o comportamento da inadimplência (pagamento em data posterior à acordada previamente, ou não efetivação). A inadimplência de alguns clientes é uma questão que deve ser bem gerenciada pelos administradores do negócio, visto que esta pode conduzir à insolvência, e prejudica os usuários adimplentes do sistema que devem suportar o custo adicional gerado pelos inadimplentes (Cui, 2022).

Em Jaraguá do Sul, os serviços de saneamento são executados por uma autarquia municipal: Serviço Autônomo Municipal de Água e Esgoto (SAMAE). Ao se observar o SIG (Sistema de Informações Gerenciais) da autarquia e suas bases de dados, percebe-se que a inadimplência em situação de suspensão de serviços (faixa de corte) marcava em torno de 3,6% antes da COVID-19, subiu para 7% no auge da pandemia e retornou aos níveis anteriores (3,3% em dezembro/22).

No acumulado de 2021 e 2022, este montante representou 4,4%. Para fins de referência, este montante representaria ao longo de 4 anos, o investimento necessário para implantar toda a rede de esgoto para uma população de 14 mil pessoas, com um custo orçado de 18 milhões de reais (bairros Três Rios e Stº Antonio).

O procedimento adotado em relação aos inadimplentes durante a virada de exercício consiste na transferência do crédito para a prefeitura, deixando-o vinculado ao imóvel. Deste modo, a esperança é de que, quando o proprietário atualizar ou modificar o registro de seu imóvel, este venha a fazer a liquidação da dívida.

[1] População urbana atendida em quase 100%, esgoto em 91% e resíduos sólidos coletados em quase 100% da população total.
[2] http://app4.mdr.gov.br/serieHistorica/

Esta estratégia de ação reativa se apresenta muito demorada e pouco efetiva, o que onera os adimplentes do sistema.

Deste modo, pretende-se propor a adoção de uma nova estratégia que seja preventiva. O primeiro passo nesta nova estratégia é saber quem é o inadimplente antes que ele se torne inadimplente com corte de fornecimento da água (mais de 50 dias de atraso em alguma fatura), para que possam ser adotadas medidas mitigadoras e de menor potencial de conflito.

1.3 O PROBLEMA

O procedimento atual em os casos de inadimplência, consistem em 2 ações: comunicação de atraso e corte no fornecimento de água (atrasos maiores que 50 dias). A estratégia de ação é meramente reativa, contribuindo para um custo adicional quase de 4% sobre o sistema, pois se apresenta muito demorada e pouco efetiva, onerando duplamente os adimplentes do sistema (Cui, 2022).

Além do custo da perda da receita, essa ação reativa cria estresse entre o usuário e o SAMAE-JS (estresse psicológico entre os envolvidos diretos, e institucional que gera desgaste de imagem). O ato de cobrar cria tensão, e quando o resultado é o corte do fornecimento de água, esta tensão vai a níveis altíssimos (há relatos até de agressão contra o pessoal encarregado dessa ingrata tarefa).

Desse modo, pretende-se propor a adoção de estratégia preventiva e não reativa. O primeiro movimento é saber quem será o inadimplente antes dele se enquadrar na condição de corte de fornecimento da água (mais do que 50 dias de atraso em alguma fatura), a fim de procurar antecipadamente uma solução alternativa, com a participação do usuário, o que pode evitar uma tensão desnecessária.

2. METODOLOGIA

Nessa sessão, alguns conceitos básicos serão abordados para o entendimento e o embasamento deste trabalho, tais como: inadimplência, aprendizado de máquina e seus principais tipos, algoritmos de regressão linear, e métricas de avaliação dos algoritmos, bem como considerações sobre os dados utilizados neste trabalho.

2.1 INADIMPLÊNCIA

O conceito de inadimplência pode parecer óbvio, mas convém rever o seu conceito geral, e o seu significado no contexto do setor de saneamento básico no Brasil.

Seguindo os princípios de Varian (1996), uma definição prática de inadimplência: quando o consumidor faz o pagamento de suas obrigações pecuniárias após o prazo acordado, ou, não o faça. Isto transforma a carteira de recebíveis em um ativo de risco, e quanto maior o risco, maior a aversão e maior o custo de capital imposto ao produtor. É considerada um custo sistêmico[3], normalmente.

A medida da inadimplência é tomada dividindo o total de títulos de cobrança não pagos em uma determinada data pelo valor total faturado com vencimento na mesma data. Por exemplo: Os valores faturados em fevereiro vencem em março normalmente.

Conforme Cui (2022), deve-se reforçar o fato de que quando a inadimplência sobe, dois problemas são gerados: o usuário fica sem crédito e o prestador de serviços não obtém os recursos financeiros esperados.

Quando o consumidor fica sem crédito na praça, por exemplo, ao tentar adquirir um veículo ou imóvel, o crédito será negado. Isto afeta os segmentos da economia onde a concessão de crédito é essencial para as mais diversas cadeias produtivas. (Cui et al., 2021).

2.2 APRENDIZADO DE MÁQUINA

Segundo Russel e Norvig (2003), a Inteligência Artificial (IA) pode ser analisada sob diferentes dimensões, incluindo processos de pensamento, raciocínio e comportamento. Entre estas dimensões encontra-se a perspectiva de comportamento a partir de como o ser humano atua. Nesta perspectiva, a IA por ser entendida como “a arte de criar máquinas que executam funções que exigem inteligência quando executadas por pessoas” (Kurzweil, 1990). Ou como afirma McCarthy (2007), “é a ciência e a engenharia da criação de máquinas

[3] Custo não gerenciável pela organização. Depende mais do ambiente e contexto macro do que do microeconômico.

inteligentes, especialmente programas de computador inteligentes…". Desta forma, a IA confere às máquinas capacidades inerentes aos seres humanos, tais como, percepção e análise visual, processamento de linguagem natural, reconhecimento de padrões, tomada de decisão, entre outros vários exemplos.

De modo geral, engloba as capacidades de percepção, aprendizado, raciocínio e ação, inerentes aos seres humanos criando as bases para diferentes campos de estudo, entre eles reside o Aprendizado de Máquina (do inglês *Machine Learning* - *ML*). O ML trata-se de um ramo da Inteligência Artificial baseado na ideia de que sistemas podem aprender com dados, identificar padrões e tomar decisões com o mínimo de intervenção humana (He, 2021). Caracteriza-se assim, por técnicas que permitem aos computadores automatizarem a construção de modelos analíticos (baseados em modelos matemáticos) a partir de dados estruturados e não estruturados.

Ao contrário da programação tradicional, onde são programadas as regras que manipulam os dados, os algoritmos de ML podem aprender com a experiência passada, conseguem identificar os padrões em dados de entrada e gerar respostas para novos dados submetidos a determinado modelo. Neste sentido, a dimensão de aprendizagem é aspecto fundamental em ML podendo ser dividida em: aprendizado supervisionado onde as instâncias do conjunto de dados apresentando ao modelo possuem rótulos pré-definidos; aprendizado não supervisionado onde o conjunto de dados não está rotulado; semi-supervisionado onde o conjunto de dados possui instâncias rotuladas e não rotuladas; e aprendizagem por reforço, onde o computador tomará decisões que são recompensadas ou penalizadas (Karmaker, 2022).

É inegável o potencial do ML nas mais diversas áreas de aplicação, todavia, o processo para produzir modelos úteis e confiáveis não é trivial. Neste sentido, uma das áreas que têm ganhado relevância é o Aprendizado de Máquina Automatizado (do inglês *Automated Machine Learning* - AutoML). Segundo Prasad et al. (2021), os recursos do AutoML simplificam os diversos processos encapsulando dados e código em um fluxo que abstrai a complexidade, promovendo um caminho para a normatização da IA.

Ainda segundo os autores, apesar de algumas limitações do AutoML, a sua adoção em áreas em que os dados são estruturados e refinados facilita as diversas tarefas envolvidas em ML.

2.3 ALGORITMOS COM REGRESSÃO LINEAR

Diversos algoritmos usam como modelo matemático a regressão linear. De acordo com Hill (2011) e Neter (1996), a regressão linear consiste em construir uma interpretação de função complexa, em uma função simples tipo linear clássica, tal como: $Y = \alpha + a.X$, ou $Y = \alpha + a_1.X_1 + a_2.X_2 + ...+ a_n.X_n$.

Esta é uma modelagem de mais fácil compreensão, permitindo prever o comportamento de uma variável em função de outra, ou outras. A versão original das funções lineares acima, poderia ser derivada de uma função do tipo: $Y = \propto . X_1^{a1} . X_2^{a2} . \ldots . . X_n^{an}$ (função tipo Coob-Douglas).

Diversos algoritmos se utilizam deste modelo matemático e, em conjunto com o aprendizado de máquina supervisionado, podem criar modelos de predição com bom nível assertividade: destacam-se: *LGBM - Light Gradiente Boosting Machine e Random* Forest.

Segundo Alzamzami et al. (2020) Osman et al. (2021), o *LGBM* é um *Gradient Boosting Framework* que também se utiliza da lógica de aprendizado por Árvore de Decisão. A diferença principal em relação a outros modelos baseados em Árvore de Decisão é que, no *LGBM*, as árvores crescem "verticalmente" enquanto em outros algoritmos crescem "horizontalmente". O algoritmo usa duas técnicas principais: o *Gradient-based One-Side Sampling* (GOSS) e o *Exclusive Feature Bundling* (EFB). O *Random Forest*, por sua vez, é um algoritmo de aprendizagem supervisionado que usa o método de aprendizagem do conjunto para a regressão. O método de aprendizagem combina previsões de vários algoritmos de forma automática visando estabelecer previsões mais acuradas se comparadas à utilização de um único modelo.

Por fim, cabe mencionar que, para este trabalho, a biblioteca de *software* escolhida é o PyCaret[4] e, conforme Ali (2020), trata-se de uma biblioteca de código aberto em Python® voltada à automatização de tarefas de ML. Essencialmente, é um *wrapper* Python® em torno de várias bibliotecas e estruturas de aprendizado de máquina, como scikit-learn®, XGBoost®, LightGBM®, CatBoost®, spaCy®, Optuna®, Hyperopt®, Ray®, entre outras.

[4] PyCaret. An open source, low-code machine learning library in Python. april 2020. <https://pycaret.gitbook.io/docs/>. access 2023,march 8st.

2.4 MÉTRICAS PARA AVALIAÇÃO DE MODELOS DE REGRESSÃO

Ao utilizar recursos de AutoML supervisionado é importante conhecer as métricas utilizadas para avaliar o desempenho dos algoritmos utilizados e sua eficácia na construção de modelos de predição. Segundo Lira (2022), Hair (2006) e Padhma (2021), quando se trata de erros para regressão, têm-se como as principais métricas de avaliação o Erro Absoluto Médio (do inglês *Mean Absolute Error* - MAE**)**, o Erro Médio Quadrático (do inglês *Mean Squared Error* - MSE)**,** a Raiz Quadrada do Erro Médio (do inglês *Root Mean Squared Error* - RMSE), Raiz Quadrada do Logaritmo do erro médio (do inglês *Root Mean Logarithmic Error* - RMSLE), o Erro Percentual Médio Absoluto (do inglês *Mean Absolute Percentage Error* - MAPE). Khan Academy (2020) cita ainda o R^2, o coeficiente de determinação (do inglês *coeficient of determination*).

3 ANÁLISE E DISCUSSÃO DOS RESULTADOS

3.1 EXTRAÇÃO DE DADOS

Os dados extraídos a partir do sistema SANSYS (J-TECH, 2022) referem-se às cobranças (faturas) cujas datas de vencimento ficam entre maio de 2020 e maio de 2022 (ambos meses integralmente inclusos).

A partir disso, os dados primários foram armazenados em formato tipo csv (*Comma Separated Values*), utilizando-se o relatório: **FAT0022**, este relatório contém dados de cada fatura gerada para uma determinada matrícula em um determinado mês e ano, informando se foi paga (com a data de pagamento) ou se está inadimplente.

O pré-tratamento dos dados consistiu na eliminação de registros com comportamento diferenciado (órgãos de governo e grupos de fragilidade social), e dados incompatíveis com os filtros utilizados. A ferramenta utilizada foi o Knime[5] analytics platform nesta etapa. As faturas que serão objeto de análise são as que estão quitadas (pagas) ou vencidas sem quitação. O ponto focal para a extração dos dados foi o dia **31 de outubro de 2022.**

3.2 ANÁLISE DOS DADOS

O presente estudo pretende avaliar a evolução do comportamento de prazo de pagamento dos usuários visando identificar o padrão daqueles que, em algum momento, tiveram um período de atraso no pagamento superior a 50 dias (faixa de corte).

O hiato entre o pagamento e o vencimento de cada fatura, para os registros sem quitação, deve ser realizado comparando o vencimento com a data de extração dos dados (31/10/22). De fato, foram extraídos o total de 1.142.029 registros com a seguinte relação de "hiato" (vencimento e pagamento), conforme a Tabela 1.

Tabela 1 - Distribuição prazos pagamentos

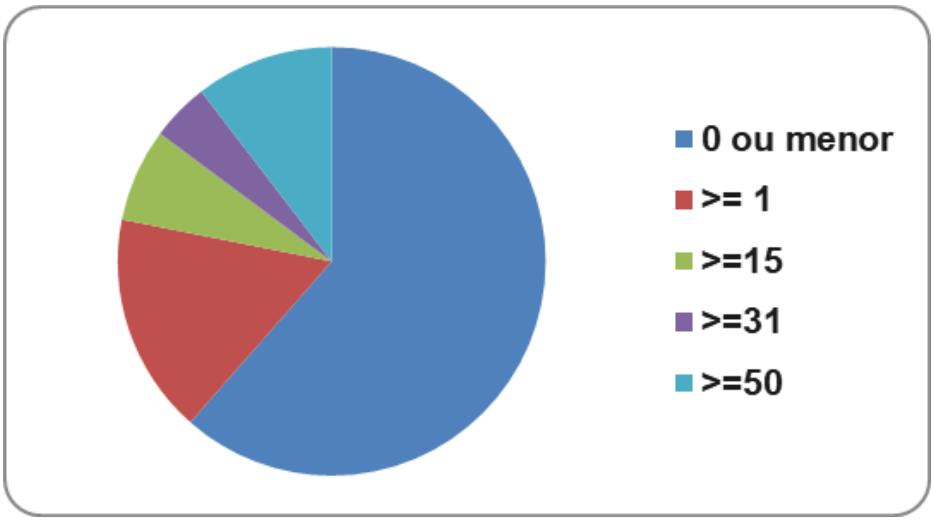
Dias atraso	Ocorrências	% total
0 ou menor	702.175	61,48
>= 1	190.200	16,65
>=15	80.374	7,04
>=31	50.200	4,40
>=50	119.080	10,43
Total	1.142.029	100,00

Fonte: Autores

Pode-se verificar que a maior parte dos pagamentos são efetuados dentro dos prazos, e os que são pagos em atraso, a maior parte (3/4) é com prazo inferior ou igual a 50 dias, conforme o Gráfico 1 e a Tabela 1.

[5] KNIME AG. **KNIME**. Version 4.7.8, January 2024. <https://www.knime.com>. Zurich, Switzerland. Acesso em 4 jan.2024

Figura 1 – Prazos de Pagamento



Fonte: Autores

3.3 CÁLCULO DO ATRASO (*DELAY*)

Para o cálculo e registro desta informação deve ser acrescentada uma coluna numérica, tipo inteiro, calculada diferentemente de acordo com a situação das faturas. Esta coluna deve ser denominada como "delay*"*.

Para as faturas quitadas (pagas), há uma data de pagamento efetivamente ocorrida a ser observada, ou seja, a diferença entre data de pagamento e data de vencimento.

No caso de pagamentos antecipados, os valores devem ser NEGATIVOS. Para pagamentos no vencimento, o valor deve ser ZERO, ao passo que pagamentos após vencimento produzem valores POSITIVOS.

Quando a coluna "delay" possuir um valor maior que 50, o usuário entra na faixa de corte, ou seja, é o ponto onde o SAMAE interrompe o fornecimento de serviços de água. É um ponto de desgaste para o usuário e o SAMAE, que é chamado de *default*.

3.4 REFORMATAÇÃO DOS DADOS

Pretende-se verificar nesta análise a evolução do prazo de atraso nos pagamentos. Ou seja, como este prazo de atraso se comportou. Se foi constante, ou se obedeceu a algum tipo de progressão e sob qual taxa ao definir uma função tendo como variável independente o tempo de atraso anterior (mês 25, mês 24, ..., mês 1). Para os usuários que tiveram inadimplência, espera-se obter uma função com 1ª derivada positiva, e a 2ª, ou igual a 0 (zero), ou positiva. Isto pode indicar uma progressão do tipo aritmética, geométrica ou até exponencial, para os usuários que evoluem para a situação de *default*.

Como o objetivo é determinar se um cliente está em vias de entrar na faixa de corte, ou seja, um atraso de pagamento maior que 50 dias, o resultado esperado é uma variável tipo *dummy* ou dicotômica (sim ou não). Em caso afirmativo o resultado será 1, caso contrário 0. O modelo matemático deve ser de regressão linear, mesmo que exija alguns ajustes, especialmente na variável de resultado que pode ser classificatória ou até de probabilidade (HILL et al., 2011).

Espera-se que a variação no atraso seja o determinante do cliente entrar em *default* ou não, sendo este o aspecto a ser testado. Entende-se que os dados relevantes para a previsão são: "matricula" (identificação do usuário no sistema); "referencia" (mês/ano a que se refere aquela fatura); "delay*"* (dias entre o "pagamento" e o vencimento); "default" (se aquele usuário, naquela referência está em situação de *default* ou não).

Deste modo, a tabela simplificada e "pivotada" em função das referências, estará pronta para os testes com o AutoML e o algoritmo selecionado (baseados em regressão linear).

Um dos principais pontos é o processo de "pivotagem" (*pivot node; group by* Matricula*; pivot by* Referencia*; Aggregation by* Delay, Default), que transforma a tabela fazendo com que para cada matrícula sejam criadas *n*

colunas com o *delay* para cada referência. Criou-se uma coluna "default*"* com o valor 1 para quando algum *delay* for maior que 50, e 0 quando nenhum *delay* for menor ou igual a 50 (menor processamento).

Sendo assim, a nova tabela (denominada NewDelay), "transposta" ou "pivotada", em torno da referência, com os dados de "delay*"* e "default*"* em uma coluna com a denominação equivalente à referência.

O total de linhas obtidas foi de 47.441. Destas, 18.486 matrículas estiveram na posição de *default* ao longo de 25 meses, o que equivale a 38,97% dos usuários. Ou seja, em determinado momento quase 30% dos usuários estavam em situação de corte ao longo dos 2 últimos anos.

3.5 AUTOML PARA MODELAGEM E PREDIÇÃO

Após a adequação dos dados foi obtida uma nova tabela, onde as referências de pagamento se tornaram colunas também, com apenas duas informações para cada uma das 25 referências: O *delay* para cada referência e se este determina uma situação de *default*. Ou seja, todas as ocorrências de um certo usuário que estavam em 25 linhas da tabela original, agora estão todas contidas em uma única linha, com 2 colunas para cada referência (50 colunas para cada usuário), nesta nova tabela (NewDelay).

Esta nova tabela será usada para determinação do melhor algoritmo para predição com uma ferramenta adequada de Auto ML, separada em 80% (37.952 registros) para treino e 20% (9.489 registros) para execução do algoritmo selecionado com a ferramenta de Auto ML.

Como se trata de modelagem com regressão linear, o aspecto mais importante a ser testado é a assertividade – R^2. Desta forma, o melhor resultado obtido foi de 0,9709 ou 97,09%, pelo algoritmo LGBM.

Na parte de AutoML para determinação do melhor algoritmo e aplicação do mesmo, foram utilizados ferramentas e pacotes na linguagem Python®, a saber, Jupiter Notebook® (Jupiter, 2022); Pycaret® e LGBM (LightGBM *classifier*). A execução do AutoML utilizou a tabela NewDelay, sendo obtidos os resultados apresentados na Tabela 2.

O modelo para predição a ser obtido deve ser considerado eficaz para a predição de quais matrículas entrarão ou não em *default*.

Deve-se observar que o algoritmo *Random Forest Regressor* teve um bom desempenho, (similar ao LGBM) no que diz respeito ao R^2, com 0,9703, e até superior em algumas métricas. Logo é uma alternativa que poderia ser considerada.

Table 2 - Resultado do teste de algoritmos por AutoML Pycaret®

Abrev.	Nome modelo	MAE	MSE	RMSE	R^2	RMSLE	MAPE	TT(sec)
lgbm	Light Gradient Boosting Machine	0,0315	0,0069	0,0829	0,9709	0,0581	0,0516	0,7710
rf	Random Forest Regressor	0,0207	0,0070	0,0837	0,9703	0,0618	0,0260	37,4510
et	Extra Trees Regressor	0,0405	0,0113	0,1064	0,9522	0,0774	0,0569	12,9140
dt	Decision Tree Regressor	0,0149	0,0149	0,1218	0,9372	0,0844	0,0174	0,6290
knn	K Neighbors Regressor	0,0696	0,0405	0,2011	0,8293	0,1311	0,1531	2,8800
gbr	Gradiente Boosting Regressor	0,1725	0,0589	0,2426	0,7516	0,1605	0,2626	7,9620
ada	AdaBoost Regressor	0,3403	0,1486	0,3851	0,3733	0,2849	0,3236	1,3590
lar	Least Angle Regression	0,3876	0,1825	0,4271	0,2304	0,2959	0,5315	0,0650
br	Bayesian Ridge	0,3878	0,1824	0,4271	0,2304	0,2959	0,5316	0,0740
ridge	Ridge Regression	0,3876	0,1825	0,4271	0,2304	0,2959	0,3150	0,0530
lr	Linear Regression	0,3876	0,1825	0,4271	0,2304	0,2959	0,5315	0,0750
en	Elastic Net	0,3926	0,1828	0,4275	0,2290	0,2970	0,5341	0,0700
lasso	Lasso Regression	0,3972	0,1837	0,4286	0,2251	0,2985	0,5366	0,0720
omp	Orthogonal Matching Pursuit	0,3891	0,1847	0,4298	0,2208	0,3000	0,5281	0,0600
huber	Huber Regressor	0,3485	0,2032	0,4507	0,1429	0,2878	0,6153	1,1370
llar	Lasso Least Angle Regression	0,4743	0,2372	0,4870	-0,0005	0,3428	0,6134	0,0580
dummy	Dummy Regressor	0,4743	0,2372	0,4870	-0,0005	0,3428	0,6134	0,0460
par	Passive Aggressive Regressor	0,8769	4,8899	2,0012	-19,6634	0,3967	1,9369	0,0990

Source: Authors

3.6 EXECUÇÃO DA PREDIÇÃO

A execução do algoritmo baseado em LGBM utilizou-se de um modelo matemático-computacional para determinar quais usuários entrarão em *default* na próxima referência de vencimento (sequência dos dados analisados).

Após a execução deste modelo, os resultados obtidos determinam quais registros entrarão em *default*. Esta informação encontra-se na coluna "default". Ao assumir o valor de 1, indica que o algoritmo prediz que determinado usuário deverá entrar em *default*; ao assumir o valor de 0, indica que o usuário **não** deverá entrar em *default*.

O $\mathbf{R^2}$ obtido foi de 0,9715, enquanto para MSE, o resultado foi de 0,0069 (Tabela 3). Estes dados são compatíveis com a fase de teste, com uma discreta melhoria na assertividade.

Tabela 3 - Resultado execução algoritmo baseado em LGBM

Sigla	MODELO	MAE	MSE	RMSE	R^2	RMSLE	MAPE
lightgbm	Light Gradient Boosting Machine	0,0317	0,0069	0,0828	0,9715	0,0578	0,0512

Fonte: Autores

A consistência percebida pela comparação das métricas de avaliação de modelos de regressão da fase de testes e execução aumentam a confiança nos resultados obtidos.

4. CONCLUSÃO

A água é essencial para a vida, logo a sua disponibilidade é essencial para toda e qualquer atividade humana, de modo que quando um cliente se permite entrar numa situação de indisponibilidade, aquela receita estará definitivamente perdida ou só será retornada em prazo muito longo, o que faz com que este ônus recaia sobre os adimplentes do sistema. O que lhes impõem um peso duplo, pois suportam os custos dos inadimplentes, além dos seus próprios custos. Socialmente, trata-se de uma injusta distribuição dos gastos.

A mudança no hiato de tempo entre o vencimento de uma fatura de um determinado cliente e a data do seu efetivo pagamento, podem e devem ser utilizados para prever quais clientes entrarão em *default*. Isto permite às empresas de saneamento preverem, antes que os próprios clientes saibam quem eles são e, assim, tomar medidas que possam minimizar os impactos.

O uso de Aprendizado de Máquina permite a criação de modelos de predição com uma velocidade e precisão inconcebíveis há pouco mais de 1 década. A prova disto é o uso de recursos de AutoML que revelam o algoritmo *LGBM* como de alta eficácia, atingindo os seguintes valores: R^2 de 0,9715; MAE de 0,0317; MSE de 0,0828; RMSLE de 0,0578; e MAPE de 0,0512.

O algoritmo de regressão linear LGBM pode e deve ser usado nos sistemas comerciais das empresas de saneamento, para redução dos atritos inerentes aos conflitos derivados da inadimplência (usuários e empresas de saneamento) pela ação preventiva e não meramente reativa.

O modelo produzido neste estudo com base no algoritmo *LGBM* tem como resultado final uma tabela com duas colunas: "matricula", sendo o identificador do usuário, e "default", onde 1 indica que o usuário deve entrar em *default* na próxima referência, e 0 indica que ele não deve entrar em *default* na próxima referência.

Estes dados possuem um formato simples, e podem ser importados e exportados para qualquer sistema comercial de modo eficaz, como forma de alimentar dados que permitam ações preventivas contra inadimplência pelo setor comercial das empresas de saneamento.

A lógica dita que o próximo passo dessa pesquisa é o Desenvolvimento e disponibilização de uma API (internet) que possa receber um vários registros e devolver a probabilidade de cada um entrar em Default.

Por fim, conclui-se que a utilização de AutoML e LGMB possui forte potencial para redução de conflitos e as consequências danosas dos atritos inerentes a situação de *default*, pois permite que ações corretivas, ou ao menos mitigadoras, sejam tomadas de forma antecipada pelas organizações responsáveis por este bem fundamental para toda a vida, a água.

AGRADECIMENTOS

Especialmente ao mentor, Alexandre Leopoldo Gonçalves, pelo encorajamento em perseguir os objetivos.

Ao SAMAE de Jaraguá do Sul, pela colaboração com sessão de dados.

Ao financiamento concedido ao estudante Ricardo Alexandre de Mello Oliveira pelo programa UNIEDU da Secretaria de Estado da Educação.

REFERÊNCIAS

1. Ali, M. (2020, April). An open source, low-code machine learning library in Python. Pycaret Docs. Retrieved April, 2020 from https://pycaret.gitbook.io/docs

2. Brasil (2022, December). Governo Federal do Brasil (Secretaria Nacional de Saneamento Ambiental - SNASA- do Ministério do Desenvolvimento Regional), Retrieved December 11, 2022, from http://app4.mdr.gov.br/serieHistorica

3. Alzamzami, F., Hoda, M., & Saddik, A.E. (2020). Light Gradient Boosting Machine for General Sentiment Classification on Short Texts: A Comparative Evaluation. IEEE Access, 101840-101858. https://doi.org/10.1109/ACCESS.2020.2997330

4. Borges, M.C.P., et al. (2022). The Brazilian National System for Water and Sanitation Data (SNIS): Providing information on a municipal level on water and sanitation services. Journal of Urban Management, 11, 4, 530-542.

5. Capone, D., Cumming, O., & Nichols, D. (2017). Water and sanitation in urban América. *American Journal Public Health Association*, *110*(10), 1445-1578. Retrieved October, 2020, from https://ajph.aphapublications.org/doi/epdf/10.2105/AJPH.2020.305833

6. Cui, W., & Kaas, L. (2021). Default cycles. *Journal of Monetary Economics*, *117*, 377-394. https://doi.org/10.1016/j.jmoneco.2020.02.001

7. Cui, W. (2022). Macroeconomic Effects of Delayed Capital Liquidation. *Journal of European Economic Association*, *20*(4), 1683-1742. https://doi.org/10.1093/jeea/jvac023

8. Hair Jr., J. F., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2009). *Análise Multivariada de Dados* (6th ed., p. 688). São Paulo: Bookman

9. He, X., Zhao, K., & Chu, X. (2021). AutoML: A survey of the state-of-the-art. *Knowledge-Based Systems*, *212*, 1-27. https://doi.org/10.1016/j.knosys.2020.106622

10. Hill, R. C., Griffiths, W. E., & Lim, G. C. (2011). *Principles of econometrics* (4th ed., p. 784). New Jersey, USA: John & Sons

11. J-Tech (2022, December). *SANSYS*. Retrieved April 21, 2023, from http://jtech.com.br

12. Jupiter (2022). *Jupiter Notebook*. Retrieved October 7, 2022, from https://jupyter.org

13. Karmaker, S., Hassan, M., Smith, M., Xu, L., Zhai, C., & Veeramachaneni, K. (2022). AutoML to Date and Beyond: Challenges and Opportunities. *Association for Computing Machinery*, *54*(8), 1-35. https://doi.org/10.1145/3470918

14. Khan Academy (2022, November). *R2 or coeficiente of determination*. Retrieved March 11, 2023, from https://en.khanacademy.org/math/ap-statistics/bivariate-data-ap/assessing-fit-least-squares-regression/v/r-squared-or-coefficient-of-determination

15. Kurzweil, R. (1990). *The Age of Intelligent Machines* (p. 580). MIT Press

16. Lira Jr., A. (2022). *Métricas para algoritmos de regressão*. IMasters Forum. Retrieved December 19, 2022, from https://imasters.com.br/data/metricas-para-algoritimos-de-regressao

17. McCarthy, J. (2004). *What Is Artificial Intelligence?* Retrieved June 26, 2023, from https://www-formal.stanford.edu/jmc/whatisai.pdf

18. Neter, J., Kutner, M., Nachtshein, C., & Wasserman, W. (1996). *Applied linear statistical models* (1st ed., p. 1408). McGraw-Hill/Irwin

19. Osman, M., He, J., Mokbal, F.M.M., Zhu, N., Qureshi, S. (2021). *ML-LGBM: A Machine Learning Model based on Light Gradient Boosting Machine for the Detection of Version Number Attacks in RPL-Based Networks*. Retrieved July 27, 2023, from https://ieeexplore.ieee.org/document/9448047

20. Padhma, M. (2021). *End-to-End Introduction to Evaluating Regression Models*. Retrieved April 17, 2023, from https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/

21. Prasad, V.V.P. Senthil Kumar, Lokeswari Y. Venkataramana, G. Prasannamedha, S. Harshana, S. Jahnavi Srividya, K. Harrinei, Sravya Indraganti. Automating water quality analysis using ML and auto ML techniques. **Environmental Research**. 2021. Volume 202 ,2021, 111720, ISSN 0013-9351. <https://doi.org/10.1016/j.envres.2021.111720>

22. Russell, S., & Norvig, P. (2003). *Artificial Intelligence: A Modern Approach* (2nd Ed., P. 1132). New Jersey: Prentice Hall

23. Sinharoy, S. S., Pittluck, R., & Clasen, T. (2019). Review of drivers and barriers of water and sanitation policies for urban informal settlements in low-income and middle-income countries. *Utilities Policy Journal*, *60*, 2-8. https://doi.org/10.1016/j.jup.2019.100957

24. United Nations. (2015). General Assembly, Transforming our world: the 2030 Agenda for Sustainable Development, 21 October 2015, A/RES/70/1. Retrieved July, 22, 2023, from https://www.refworld.org/docid/57b6e3e44.html

25. Varian, H.R. (1996). Intermediate Microeconomics**:** A Modern approach (4th ed., p. 650). USA: W.W. Norton & Company Incorporation

APÊNDICES

"jupiter_aplicacao_automl_samae.ipynb" – conjunto de scripts para execução sob o Jupiter.

"modelo_lightgbm.pkl" – Pacote para implementação de LightGBM para Python.

"SAMAE JS.knwf" – Workflow das atividades feitas sob Knime (análise exploratória).

"SAMAE JS.json" – Workflow das atividades feitas sob Knime (análise exploratória) em formato json.

Disponíveis no link de compartilhamento abaixo:

https://drive.google.com/drive/folders/1IeZNCoi09tuLmG5HaYBKOqqEg8vo4fe4?usp=sharing