Predição de Pagamentos Atrasados Através de Algoritmos Baseados em Árvore de Decisão

Prediction of Late Payments Using Decision Tree-Based Algorithms

Arthur Flor de Sousa Neto[1]
orcid.org/0000-0002-0522-2150

José Fernando Guilhermino da Silva[1]
orcid.org/0000-0001-6439-2435

Glauber Nascimento de Oliveira[1]
orcid.org/0000-0002-2472-9811

[1]Escola Politécnica de Pernambuco, Universidade de Pernambuco, Recife, Brasil. E-mail: afsn@ecomp.poli.br

DOI: 10.25286/repa.v6i5.1746

Como citar este artigo pela NBR 6023/2018: NETO, A. F. S.; SILVA, J. F. G.; OLIVEIRA, G. N. Predição de Pagamentos Atrasados Através de Algoritmos Baseados em Árvore de Decisão. Revista de Engenharia e Pesquisa Aplicada, Recife, v.6, n. 5, p. 1-10, Novembro, 2021.

RESUMO

O processo *Invoice-to-Cash* é essencial para a estabilidade financeira de qualquer empresa, tendo em vista a coleta de contas a receber como sua principal atividade. No entanto, apesar da sua importância, a etapa de cobrança é geralmente processada manualmente, no qual ocasiona no contato a todos os clientes em intervalos fixos, mesmo que alguns sempre tenham pago em dia. Assim, o trabalho explora técnicas de mineração de dados com aprendizado de máquina, visando otimizar o processo de coleta através de predições dos pagamentos das faturas. Para isto, foi utilizado oito algoritmos baseados em árvore de decisão, aplicados em três etapas: (i) identificar as faturas com pagamento no prazo ou atrasado; (ii) identificar entre as faturas atrasadas, pagamento no mês de vencimento ou posterior; e (iii) prever entre as faturas atrasadas, quantos dias de atraso terão além do mês de vencimento. Por fim, através dos resultados obtidos dos melhores modelos para cada etapa, foi obtido uma precisão média de 81,85%, 85,63% e 73,98%, respectivamente.

PALAVRAS-CHAVE: Contas a Receber; Coleta de Pagamento; Mineração de Dados; Aprendizado de Máquina; Árvore de Decisão;

ABSTACT

The Invoice-to-Cash process is essential for the financial stability of any company, in view of the collection of accounts receivable as its main activity. However, despite its importance, the billing stage is usually processed manually, which causes contact with all customers at fixed intervals, even if some have always paid on time. Thus, the work explores data mining techniques with machine learning, aiming to optimize the collection process through prediction of invoice payments. For this, eight algorithms based on decision tree were used, applied in three steps: (i) identify invoices with payment on time or late; (ii) identify among the overdue invoices, payment in the due date month or later; and (iii) predict among the overdue invoices, how many days of delay they will have beyond the due month. Finally, through the results obtained from the best models for each step, was obtained an average precision of 81.85%, 85.63% and 73.98%, respectively.

KEY-WORDS: Accounts Receivable, Payment Collection; Data Mining; Machine Learning; Decision Tree;

1 INTRODUÇÃO

O *Business Analytics* é a prática exploratória dos dados de uma organização com ênfase na análise estatística. Dessa forma, é utilizado por empresas através de procedimento sistemático de coleta de dados para tomar decisões baseadas em dados **[1][2]**. Nas últimas décadas, o volume de dados disponíveis aumentou e esse processo analítico passou por uma grande transição. Logo, métodos automatizados tornaram-se necessários para a análise de dados. Isso abriu espaço para o aprendizado de máquina; conjunto de métodos de detecção de padrões nos dados para, em seguida, utilizar o conhecimento adquirido na predição de novos dados.

Neste contexto, empresas possuem o processo de *Order-to-Cash*, no qual se refere ao recebimento e processamento de pedidos dos clientes **[2]**. Embora seu número de etapas possa variar de empresa para empresa (dependendo do seu segmento e tamanho), o conjunto de atividades do setor financeiro, conhecido como *Invoice-to-Cash*, são essenciais para o funcionamento de qualquer negócio **[1][2][3]**.

O processo *Invoice-to-Cash*, por sua vez, lida com a priorização de contas, atividades de contato com o cliente, chamadas de cobrança, escalonamento e resolução de disputas **[3]**. Na maioria das vezes, essas etapas são processadas manualmente e, portanto, lentas, caras e imprecisas, apesar de sua importância para os negócios. Além disso, as ações de cobrança são tipicamente genéricas e não consideram as especificidades do cliente. Nesse sentido, todos os clientes são contatados em intervalos fixos, embora alguns sempre tenham pago em dia; e geralmente quanto mais tarde o contato com um cliente, menor a probabilidade de as faturas serem pagas a tempo **[2][3]**. Além disso, o contato repetido a bons clientes pode diminuir a sua satisfação. Tais ineficiências nas práticas atuais levam a atrasos significativos nas coletas ou até mesmo à falha na cobrança antes dos prazos **[3][4]**.

De fato, é de grande interesse administrar de forma eficaz a coleta de contas a receber, no qual só a indústria de construção canadense em 2012 gerou 111 bilhões de dólares, todos em forma de fatura. Além disso, as empresas atuais acumulam grandes volumes de dados sobre seus clientes, o que torna possível a eficácia da coleta em grande escala **[1][4][5][6]**.

Assim, este trabalho propõe uma modelagem para o processo de identificação de pagamentos, visando otimizar a etapa de coleta de contas a receber através de informações sobre pagamentos com potencial atraso. Desta forma, técnicas de mineração de dados em conjunto a modelos baseados em árvore de decisão, foram utilizados para alcançar o objetivo. Além disso, o projeto proposto está disponível em: https://github.com/arthurflor23/invoice-payment-prediction.

Por fim, este trabalho está organizado da seguinte maneira: a seção 2 fala sobre a fundamentação teórica, no qual pode-se entender a área do negócio, mineração de dados e os trabalhos relacionados; a seção 3 descreve a base de dados utilizada, assim como o pré-processamento aplicado; a seção 4 apresenta o experimento realizado, bem como os resultados alcançados; e a seção 5 conclui o trabalho realizado, trazendo as considerações finais.

2 FUNDAMENTAÇÃO TEÓRICA

Nesta seção, são apresentados os temas que constituem a base teórica deste trabalho.

2.1 *ACCOUNTS RECEIVABLE*

Accounts Receivable, ou contas a receber, são consideradas uma das partes essenciais das finanças na cadeia de abastecimento e da estabilidade financeira das empresas **[4]**.

Existem muitas métricas usadas para medir a eficácia da cobrança de uma empresa **[4]**. Uma das medidas mais utilizadas é o de Dias de Vendas Pendentes (*Days Sales Outstanding*, DSO). Ela expressa o tempo médio em dias em que as contas a receber estão pendentes e é definido como:

$$DSO = \frac{ET \times NDP}{CSP} \quad \textbf{(1)}$$

onde NDP corresponde ao número de dias no período analisado (*Number of Days in Period Analyzed*) e CSP ao número de vendas a crédito para o período analisado (*Credit Sales for Period Analyzed*) **[2]**. Esse tipo de métrica auxilia no tempo necessário para cobrar as faturas. Ou seja, se for possível prever o resultado de uma fatura, pode-se usar essas informações para direcionar o processo de cobrança **[2]**.

Segundo Zeng et al. **[2]**, normalmente o setor de cobrança espera até que as faturas estejam inadimplentes para iniciar as ações de cobrança. No entanto, é possível se beneficiar do contato preventivo com potenciais contas inadimplentes. Além disso, mesmo após o vencimento de uma fatura, é benéfico saber quais faturas provavelmente serão pagas mais cedo ou mais tarde, caso nenhuma ação seja tomada. Assim, o contato pode ser priorizado com base na estimativa do atraso.

2.2 MINERAÇÃO DE DADOS

Devido ao grande volume de dados atualmente e sua disponibilidade, o campo de estudo de mineração de dados vem atraindo cada vez mais a atenção no meio acadêmico e industrial **[7]**.

De certa maneira, a mineração de dados explora identificar padrões válidos nos dados, com objetivo de extrair conhecimento útil através de técnicas de pré-processamento. Assim, pode ser utilizado em conjunto a abordagens de Inteligência Artificial, servindo como base de entrada aos modelos de aprendizado de máquina **[5][7]**.

2.3 APRENDIZADO DE MÁQUINA

A partir do reconhecimento de padrões e Inteligência Artificial, o Aprendizado de Máquina (*Machine Learning*), explora a construção de algoritmos que aprendem e fazem previsões em grandes volumes de dados **[3]**. Dessa forma, o aprendizado de máquina pode ser aproveitado para aprimorar o processo de cobrança, pois permite previsões de datas de pagamento mais precisas, utilizando dados históricos **[1][2][6]**.

De acordo com Hu **[1]** e Nanda **[3]**, as abordagens comumente utilizadas para previsão de pagamento no processo de cobrança, são:

- **Modelo de Classificação Binária:** tarefa de classificar os elementos, dado dois grupos (classes), como: (i) pagamentos no prazo; e (ii) pagamento atrasado **[3]**;
- **Modelo de Classificação Multiclasse:** tarefa de classificar os elementos, dado três ou mais grupos. Neste contexto, o modelo classifica as faturas utilizando intervalos (*buckets*) de dias de atraso pré-definidos **[5]**;
- **Modelo de Regressão:** tarefa de prever um valor contínuo referente ao número de dias de atraso. Neste cenário, o modelo prevê os dias de atraso, em que pode ser classificado posteriormente, ou não **[2]**.

Diante das abordagens, os principais modelos utilizados são árvores de decisão **[1]**. Dessa forma, não somente atendem as três abordagens, como também trazem bons resultados através do treinamento em atributos relevantes **[3]**.

2.3.1 Árvore de Decisão

Algoritmos baseados em árvore de decisão são utilizados devido sua capacidade de construção com base em atributos mais relevantes. Além disso, podem ser utilizados tanto para tarefas de classificação, quanto para regressão **[8]**.

Para predição de pagamentos e coleta de contas a receber, alguns métodos de construção podem ser utilizados no algoritmo de árvore. O primeiro método, *Bootstrap Aggregating* (*Bagging*), consiste em gerar subconjuntos de exemplos através de um sorteio simples com reposição, sobre o conjunto de dados de treinamento original, chamado de "*bags*". Cada subconjunto é utilizado para a construção de uma nova árvore, tendo como resultado, a combinação das decisões de cada uma **[8]**.

No entanto, através do método *Bagging*, a estrutura da árvore pode ter semelhanças entre os seus subconjuntos, e por sua vez, alta correção em suas previsões. Para evitar a combinação de estruturas semelhantes e obter uma previsão fraca, o *Random Forest* altera a maneira de como os subconjuntos são aprendidos, de modo que as previsões resultantes tenham menos correlação através da aleatoriedade nas subdivisões **[9]**.

Outro método de construção, o *Boosting*, explora o processo de combinar e complementar as árvores ao longo do aprendizado. Dessa forma, os modelos não são mais treinados separados, mas sim de forma sequencial, a partir de ajustes dos modelos treinados previamente. Além disso, variações no aprendizado do modelo, podem melhorar o desempenho **[10]**.

Neste contexto, o *Adaptive Boosting* (AdaBoost) **[10]** foi proposto para classificações binárias, em que o próprio método ajusta seus parâmetros iterativamente (modelo aditivo). Por outro lado, o *Gradient Boosting* **[11]**, propõe ser mais genérico e atender problemas mais complexos. Além disso, permite a otimização de

DOI: 10.25286/repa.v6i5.1746

duas ou mais funções de perda, o que o torna robusto a *outliers* quando comparado ao AdaBoost. Recentemente, foi proposto o método *Random Undersampling Boosting* (*RUSBoost*) **[12]**, no qual se utiliza do AdaBoost, mas traz o foco no balanceamento das classes. Esse problema de desbalanceamento é aliviado durante o aprendizado, através da subamostragem aleatória da amostra em cada iteração do algoritmo. Por fim, o *eXtreme Gradient Boosting* (*XGBoost*) **[13]**, traz o foco para a construção do modelo através do *Gradient Boosting* através do aprendizado sequencial. O *XGBoost* visa a adição sequencial das árvores no treino do modelo como uma forma de otimizar o resultado.

2.4 TRABALHOS RELACIONADOS

As técnicas de mineração de dados fornecem grande ajuda na otimização do *Invoice-to-Cash*, tendo em vista os desafios ao lidar com grandes volumes de dados financeiros **[1]**.

Zeng et al. **[2]**, aborda o problema da redução de contas a receber pendentes através de melhorias na estratégia de cobrança. Para isso, foi demonstrado como o aprendizado supervisionado pode ser usado para desenvolver modelos que preveem os pagamentos de faturas recém-criadas. Isso permite ações de cobrança mais assertivas para cada cliente. Os algoritmos utilizados por Zeng et al. **[1]** foram baseados em árvore de decisão e delimitaram as classes em *ranges* de 30 dias, compondo 5 *buckets*.

O trabalho de Hu **[1]**, por sua vez, apresenta previsões precisas sobre os pagamentos das faturas com base em dados históricos do cliente. Além da criação dos dados históricos, foi realizado um extenso estudo sobre atributos mais relevantes e técnicas com melhor aproveitamento neste cenário. Hu **[1]** conclui que *Random Forest* obteve os melhores resultados, assim como uma boa flexibilidade entre o desbalanceamento das classes (4 *buckets* de 30 dias de intervalo). Além disso, também reforça o uso do algoritmo *Support Vector Machine* (SVM) como segunda abordagem.

No trabalho de Nanda **[3]**, os algoritmos baseados em árvore tiveram três abordagens: (i) regressão, para o modelo prever um número contínuo referente ao número de dias em atraso; (ii) classificação binária, para apenas determinar se a fatura será paga em dia ou atrasada; e (iii) classificação multiclasse, referente aos *buckets*, assim como os trabalhos anteriores. Neste cenário, Nanda **[3]** alcançou bons resultados com *Random Forest* e *Gradient Boosting*.

Por fim, Shah **[5]** apresentou outra forma de criar os dados históricos. Em seu processo, foi utilizado cada *bucket* definido para delimitar os dados históricos de cada cliente. Desse modo, foi possível criar para cada intervalo de dias de atrasado, seu respectivo histórico. Ainda segundo Shah **[5]**, a desvantagem dessa abordagem foi sua limitação aos próprios intervalos, já que para qualquer alteração, seja por regra de negócio, ou adaptação aos dados, o processo do histórico terá que ser refeito. Além disso, o algoritmo *Adaptive Boosting* (AdaBoost) foi o único utilizado no trabalho.

3 MATERIAIS E MÉTODOS

Esta seção consiste em descrever a base de dados do estudo, assim como apresentar as técnicas adotadas para o pré-processamento e criação de atributos no experimento realizado.

3.1 DESCRIÇÃO DA BASE DE DADOS

A base de dados utilizada neste trabalho foi fornecida por uma empresa privada cuja atividade é a produção e o processamento de alimentos. O conjunto comporta registros de todo o processo financeiro interno, no período de janeiro de 2018 até fevereiro de 2021.

O conjunto de dados oferece informações da fatura, cliente e empresa em todo o processo financeiro. Além disso, possui cerca de 11 milhões de registros e 131 atributos, numéricos.

3.2 ANÁLISE DESCRITIVA DOS DADOS

A análise descritiva dos dados é considerada uma etapa fundamental para o processo de descoberta de conhecimento. Tendo em vista a interpretação prévia dos dados, as técnicas de mineração são aplicadas de maneira mais assertiva. Nesta etapa, os dados podem ser organizados em distribuição de frequência e visualizados através de gráficos.

As análises indicam um grande distanciamento entre o limite inferior de faturas pagas em dia (368 dias antes da data de vencimento) e do limite máximo de faturas pagas com atraso (511 dias após a data de vencimento). Além disso, o

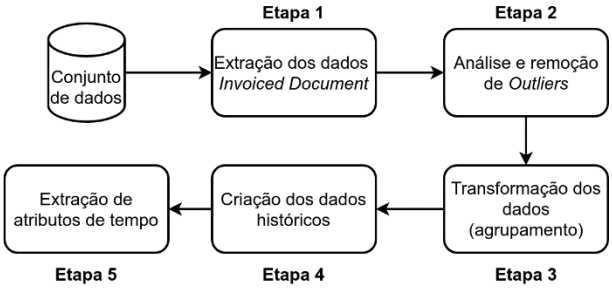
número de faturas atrasadas é cerca de 34% menor do que faturas pagas no prazo.

A distribuição do valor total das faturas entre os dados, também é um fator importante para análise. Isso pois, há grande dispersão referente aos valores, sendo o mínimo de 0,01 *cents*, e máximo de 8.5 milhões de dólares.

3.3 PRÉ-PROCESSAMENTO DOS DADOS

A etapa de pré-processamento é essencial para a organização e preparação da base, visando a boa qualidade dos dados para a etapa de modelagem. A Figura 1 mostra as etapas do pré-processamento realizadas neste trabalho.

Figura 1 – Etapas do pré-processamento realizadas



Fonte: Os Autores.

Tendo em vista que o conjunto de dados bruto consiste em todos os registros do processo financeiro interno, a etapa 1 visa a seleção dos dados de interesse do trabalho, ou seja, do tipo *Invoiced Document*. Também foram removidos registros com datas de processamento inconsistentes, como faturas com data de criação maior que o próprio vencimento e pagamento. Além disso, como regra de negócio, foi desconsiderado faturas com valor menor a mil dólares, por serem consideradas vendas internas. Ao fim da etapa 1, a base de dados passa a ter cerca de 2.7 milhões de registros.

A etapa 2, análise e remoção de *outliers*, foi aplicada visando minimizar a dispersão dos dados com respeito ao tempo de vencimento e ao valor da fatura. Assim, foi utilizado o método *Robust Z-score* **[14]**, já que utiliza o Desvio Mediano Absoluto (*Median Absolute Deviation*), em vez do desvio padrão. O MAD é calculado através da diferença absoluta entre cada valor com a mediana da amostra (x̃), para então calcular a mediana dessas diferenças:

$$MAD = median\{|x_i - \tilde{x}|\} \quad \textbf{(2)}$$

Por sua vez, o *Robust Z-score* é calculado a seguir, sendo a constante 0,6745, o 0,75 quartil da distribuição, para o qual o MAD converge.

$$M_i = \frac{0.6745(x_i - \tilde{x})}{MAD} \quad \textbf{(3)}$$

Além disso, como sugerido em Leys et al. **[14]**, o valor de corte adotado foi de 3. Isso significa que todo valor acima do ponto de corte é considerado um *outlier* e é então removido.

A etapa 3, consiste na transformação dos dados, no qual foi realizado o agrupamento de faturas de mesmo cliente com mesma data de vencimento. Esse processo foi necessário, pois foi observado a existência de diversas faturas para um mesmo cliente em uma mesma data. Assim, caso uma única fatura tenha chance de atraso, a cobrança ainda será feita. O Quadro 1 descreve os atributos criados na etapa 3.

Quadro 1 - Atributos criados na etapa de agrupamento para cada cliente

#	DESCRIÇÃO DOS ATRIBUTOS
1	Número de faturas na mesma data de vencimento
2	Número de faturas pendentes na mesma data de vencimento
3	Proporção de faturas pendentes pelo total de faturas na mesma data de vencimento
4	Soma dos valores das faturas na mesma data de vencimento
5	Soma dos valores das faturas pendentes na mesma data de vencimento
6	Proporção do valor das faturas pendentes pelo valor total das faturas na mesma data de vencimento

Fonte: Os Autores.

A etapa 4 consiste na criação de novos atributos baseado nos dados atuais, no qual visa complementar os dados com o histórico do cliente na data de cada fatura **[1][2][3]**. Assim, criou-se três grupos para os dados históricos. O primeiro é referente à mediana de dias de atraso do cliente, tanto para faturas pagas, quanto para

DOI: 10.25286/repa.v6i5.1746

os pendentes. O Quadro 2 apresenta os atributos criados com a mediana de dias de atraso.

Quadro 2 - Atributos criados com base na mediana histórica de dias de atraso do cliente.

#	DESCRIÇÃO DOS ATRIBUTOS
7	Mediana dos dias de atraso
8	Desvio absoluto dos dias de atraso
9	Mediana dos dias de atraso, posterior ao mês de vencimento
10	Desvio absoluto dos dias de atraso, posterior ao mês de vencimento
11	Mediana dos dias de atraso das faturas pendentes
12	Desvio absoluto dos dias de atraso das faturas pendentes
13	Mediana dos dias de atraso pendentes, posterior ao mês de vencimento
14	Desvio absoluto dos dias de atraso pendentes, posterior ao mês de vencimento

Fonte: Os Autores.

O segundo grupo consiste na quantidade total de faturas pagas em relação à quantidade de faturas pagas com atraso. O Quadro 3 descreve os atributos criados em relação a faturas pagas.

Quadro 3 - Atributos com relação à quantidade de faturas pagas e pagas com atraso do cliente.

#	DESCRIÇÃO DOS ATRIBUTOS
15	Total de faturas pagas
16	Total de faturas pagas com atraso
17	Total de faturas pagas com atraso, posterior ao mês de vencimento
18	Proporção de faturas pagas com atraso pelo total de faturas pagas
19	Proporção de faturas pagas com atraso posterior ao mês de vencimento, pelo total de faturas pagas
20	Valor total das faturas pagas
21	Valor total das faturas pagas com atraso
22	Valor total das faturas pagas com atraso, posterior ao mês de vencimento
23	Proporção do valor total das faturas pagas com atraso pelo valor total de faturas pagas
24	Proporção do valor total das faturas pagas com atraso posterior ao mês de vencimento, pelo valor total de faturas pagas

Fonte: Os Autores.

Por último, o terceiro grupo é referente a quantidade de faturas pendentes e a quantidade de faturas pendentes com atraso. O Quadro 4 mostra os atributos históricos criados em relação à quantidade de faturas pendentes.

Quadro 4 - Atributos com relação à quantidade de faturas pendentes e pendentes com atraso do cliente.

#	DESCRIÇÃO DOS ATRIBUTOS
25	Total de faturas pendentes
26	Total de faturas pendentes com atraso
27	Total de faturas pendentes com atraso, posterior ao mês de vencimento
28	Proporção de faturas pendentes com atraso pelo total de faturas pendentes
29	Proporção de faturas pendentes com atraso posterior ao mês de vencimento, pelo total de faturas pendentes
30	Valor total das faturas pendentes
31	Valor total faturas pendentes com atraso
32	Valor total das faturas pendentes com atraso, posterior ao mês de vencimento
33	Proporção do valor total das faturas pendentes com atraso pelo valor total de faturas pendentes
34	Proporção do total das faturas pendentes com atraso posterior ao mês de vencimento, pelo total de pendentes

Fonte: Os Autores.

A etapa 5 visa a criação de atributos relacionados com informações de tempo. Dessa forma, foi possível extrair o dia da semana em que a fatura será vencida, a quantidade de dias desde a criação até o vencimento da fatura, e quantidade de dias do vencimento até o fim do mês. O Quadro 5 mostra os atributos criados a partir da data de criação e vencimento.

Quadro 5 - Atributos criados com base nas datas de criação e vencimento de cada fatura.

#	DESCRIÇÃO DOS ATRIBUTOS
35	Quantidade de dias até o vencimento
36	Quantidade de dias até o fim do mês, a partir da data de vencimento
37	Dia da semana do último dia do mês

Fonte: Os Autores.

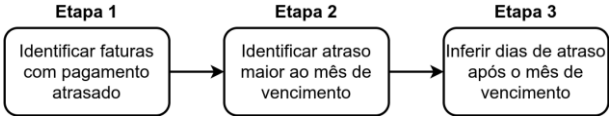
Por fim, após as etapas do pré-processamento serem aplicadas, a base de dados passa a ter cerca de 580 mil registros e 37 atributos.

3.4 METODOLOGIA EXPERIMENTAL

As etapas da mineração de dados podem variar segundo o modelo de estrutura definido. No entanto, o modelo comumente utilizado é o *Cross-Industry Standard Process of Data Mining* (CRISP-DM) **[7]**, no qual possui cerca de seis fases bidirecionais e organizadas de forma cíclica. As fases do processo CRISP-DM são: (i) Entendimento do Negócio, no qual busca entender o objetivo do negócio que se deseja atingir; (ii) Entendimento dos Dados, visa conhecer os dados, identificando os mais relevantes para solucionar o problema; (iii) Preparação dos Dados, possui o objetivo de tratar e melhorar a qualidade dos dados com técnicas de pré-processamento; (iv) Modelagem, define as técnicas e algoritmos de aprendizado de máquina que serão aplicadas, de modo a alcançar o objetivo definido; (v) Avaliação, testa e valida o modelo desenvolvido, visando obter a confiabilidade do processo; e (vi) *Deployment*, no qual é responsável pela implantação do modelo.

Com isso, a realização do experimento, assim como as etapas de mineração de dados, foi aplicada seguindo a metodologia CRISP-DM. De tal forma, foi possível definir quais etapas são necessárias para alcançar o objetivo do trabalho, que é de otimizar o processo *Invoice-to-Cash* no setor de cobrança. As três etapas são compostas por: (i) identificar as faturas que serão pagas no prazo ou com atraso; (ii) identificar se as faturas com atraso, serão pagas ainda no mês de vencimento ou não; e (iii) inferir o número de dias de atraso, para as faturas que serão atrasadas além do mês de vencimento. A Figura 2 sintetiza as etapas definidas.

Figura 2 – Etapas definidas para o experimento.



Fonte: Os Autores.

A abordagem através de árvore de decisão foi escolhida por ser um modelo que oferece transparência da decisão, podendo ser fornecida ao usuário. Assim, a implementação dos modelos de classificação e regressão se deu pelos algoritmos: (i) *Bagging*; (ii) *Balanced Bagging*; (iii) *Random Forest*; (iv) *Balanced Random Forest;* (v) *AdaBoost*; (vi) *Gradient Boosting*; (vii) *RUSBoost*; e (viii) *XGBoost.*

Para a execução e validação dos modelos, o conjunto de dados foi particionado em duas partes: (i) dados de teste, composto por toda fatura com data de vencimento em fevereiro de 2021; e (ii) dados de treino, composto por toda fatura com data de vencimento anterior a fevereiro de 2021.

Além disso, são adotados subconjuntos de dados para cada etapa conforme o seu objetivo. Isso significa que, na etapa 1 são utilizados todos os dados para treino (570.270) e teste (13.123), para identificação das faturas pagas com atraso. Já na etapa 2, para identificar os atrasos além do mês de vencimento, são utilizadas as faturas atrasadas, que consiste em 227.048 dados para treino e 5.693 para teste. E por fim, para predição de dias de atraso na etapa 3, também são utilizadas as faturas atrasadas, porém, testadas apenas nas faturas com atraso posterior ao mês de vencimento, que é 1.493.

Na fase de treinamento, foi aplicado o procedimento de validação cruzada *Repeated Stratified k-Fold* (*k=10*, n=3), no qual consiste em dividir o conjunto de dados em *k* partições, de forma aleatória e mantendo a proporção original das classes. Assim, cada partição *k* é utilizada como um conjunto de validação, enquanto todas as outras partições são usadas como um conjunto de treinamento. Esse processo é executado *n* vezes. Por fim, o desempenho médio do modelo é validado.

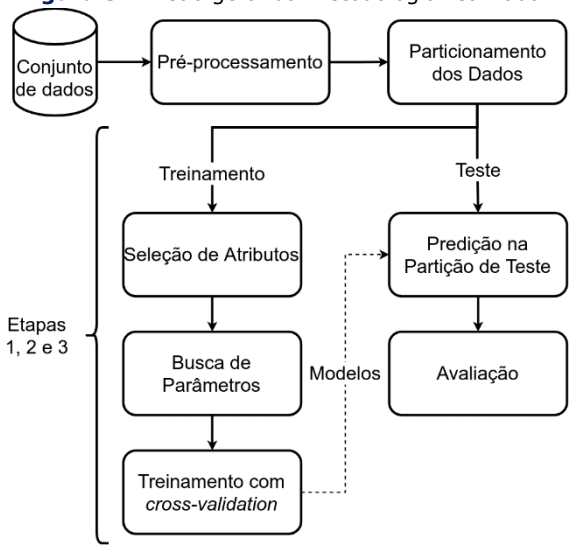
Antes de cada treinamento dos modelos, foram utilizadas as técnicas *Recursive Feature Elimination* (RFE) e *Grid Search* com *cross-validation*, para selecionar os melhores atributos e as melhores combinações de parâmetros, respectivamente. Para análise dos modelos nas etapas 1 e 2, a métrica *f1-score* com média *macro* foi utilizada **[16]**.

Para análise dos modelos na etapa 3, foi utilizada a métrica *Root Mean Squared Error* (RMSE) **[16]**. Essa métrica consiste na raiz quadrática média dos erros entre os valores reais e as predições referente aos dias de atraso. Além disso, é analisada a acurácia do modelo referente ao número de dias de atraso posterior ao mês de vencimento. Neste contexto, é considerado as predições de dias, somado a uma tolerância de

DOI: 10.25286/repa.v6i5.1746

±0 até ±3 dias. A Figura 3 apresenta a visão geral da metodologia realizada.

Figura 3 – Visão geral da metodologia realizada.



Fonte: Os Autores.

4 ANÁLISE E DISCUSSÃO DOS RESULTADOS

Nesta seção, os experimentos realizados são apresentados e os resultados obtidos em cada etapa são detalhados nas subseções seguintes. Por fim, os pontos mais relevantes são discutidos.

4.1 IDENTIFICAÇÃO DE PAGAMENTOS NO PRAZO E COM ATRASO

Na etapa 1, o objetivo consiste em distinguir as faturas com pagamentos no prazo e com atraso. Assim, além de modelos tradicionais de classificação, também foi utilizado modelos focados em lidar com o desbalanceamento de classes (Balanced).

Inicialmente, através do RFE, foi possível remover 3 atributos nesta etapa e ainda manter bons resultados entre os modelos. Estes atributos estavam relacionados à quantidade de faturas pagas e pendentes no dia. Além disso, o *Grid Search* proporcionou as melhores combinações de parâmetros. Assim, foi utilizado o ganho de informação, calculado pela função de qualidade da árvore e seus atributos (entropia). Também foi utilizado, na construção da maioria dos modelos, o valor de 512 estimadores. Para as modelos *Balanced*, a melhor estratégia de reamostragem foi de 70% da classe minoritária sobre a classe majoritária.

Assim, como pode ser observado na Tabela 1, os modelos *XGBoost* e *Random Forest* tiveram os melhores resultados, com *f1-score* de 81,85% e 81,76%, respectivamente. Já o *AdaBoost* obteve o pior resultado, com 79,51% em *f1-score*.

Tabela 1 - Resultados obtidos em 13.123 dados, com classificação de faturas pagas no prazo e com atraso (P = *Precision*, R = *Recall*, F = *F1-Score*).

Modelo	P (%)	R (%)	F (%)
AdaBoost	79,73	79,37	79,51
Bagging	81,35	81,53	81,43
B. Bagging	81,40	81,76	81,52
B. Random Forest	81,59	81,97	81,71
Gradient Boosting	81,49	81,01	81,20
Random Forest	81,64	**82,00**	81,76
RUSBoost	80,03	80,30	80,13
XGBoost	**81,75**	81,99	**81,85**

Fonte: Os Autores.

Através dos resultados obtidos, também foi analisada a quantidade dos dados identificados e identificados corretamente. O modelo *XGBoost*, por exemplo, obteve por volta de 82,09% no reconhecimento dos dados, e cerca de 81,20% no reconhecimento de faturas atrasadas.

4.2 IDENTIFICAÇÃO DE PAGAMENTOS ATRASADOS DENTRO E ALÉM DO MÊS DE VENCIMENTO

Para a etapa 2, é considerado apenas as faturas em atraso para treinamento e teste. Neste cenário, os modelos visam identificar os pagamentos atrasados no mês de vencimento ou posterior.

Como na etapa anterior, o RFE resultou na exclusão de 4 atributos, também relacionados à quantidade de faturas pagas e pendentes no dia. Os parâmetros obtidos pelo *Grid Search*, por sua vez, variaram no número de estimadores utilizados na construção, agora 256, e no valor da reamostragem (modelos *Balanced*), agora com 75% da classe minoritária sobre a majoritária.

Os resultados alcançados pelos modelos foram bem diversos. O *AdaBoost* obteve o melhor resultado em *precision*, com 89,70%. Já no *recall* e *f1-score*, o *Balanced Random Forest* alcançou 85,36% e 85,63%, respectivamente. Como pior resultado, o *RUSBoost* obteve 81,43%. A Tabela 2 apresenta os melhores resultados alcançados por cada modelo.

Tabela 2 - Resultados obtidos em 5.693 dados, com classificação de faturas atrasadas no mês de vencimento e posterior (P = *Precision*, R = *Recall*, F = *F1-Score*).

Modelo	P (%)	R (%)	F (%)
AdaBoost	**89,70**	79,90	83,22
Bagging	89,16	81,42	84,27
B. Bagging	86,36	84,93	85,63
B. Random Forest	85,92	**85,36**	**85,63**
Gradient Boosting	87,99	81,05	83,66
Random Forest	87,67	83,23	85,09
RUSBoost	84,65	79,35	81,43
XGBoost	85,90	81,49	83,31

Fonte: Os Autores.

Diante dos resultados, o Balanced Random Forest obteve 88,96% no reconhecimento geral dos dados e 77,76% no reconhecimento de faturas atrasadas além do mês de vencimento.

4.3 PREDIÇÃO DOS DIAS DE ATRASO ALÉM DO MÊS DE VENCIMENTO

Na etapa 3, as faturas atrasadas são utilizadas para treinamento, tendo o conjunto de teste as faturas com atraso além do mês de vencimento.

Diferente das etapas anteriores, através do RFE foi possível excluir cerca de 10 atributos e ainda manter bons resultados. Estes atributos removidos eram relacionados a data e histórico de faturas pagas e pendentes. Além disso, como parâmetros obtidos pelo *Grid Search*, os modelos em geral alcançaram bons resultados com 512 estimadores na construção.Nesse contexto, o modelo *Random Forest* obteve o menor valor de RMSE, com cerca de 3,5792. Porém, quanto a acurácia, considerando os valores de tolerância (±0 até ±3 dias), o *AdaBoost* obteve os melhores resultados, indo de 60,23% (sem variação) até 84,16% (com 3 dias de variação). A Tabela 3 detalha os resultados.

Tabela 3 - Resultados obtidos em 1.493 dados, com a predição do número de dias de atraso posterior ao mês de vencimento (A = Acurácia).

Modelo	RSME	A ±0 (%)	A ±1 (%)	A ±2 (%)	A ±3 (%)
AdaBoost	3,6164	**60,23**	**71,91**	**79,62**	**84,16**
Bagging	3,5786	54,24	70,12	78,46	84,12
Gradient Boosting	6,1018	47,53	61,04	69,31	73,70
Random Forest	**3,5792**	54,29	70,10	78,41	**84,16**
XGBoost	4,7917	53,86	61,08	66,43	71,47

Fonte: Os Autores.

4.4 DISCUSSÕES

Os resultados alcançados por cada modelo e em cada etapa, de certo modo, possibilita o suporte à tomada de decisão na coleta de contas a receber. Comparado aos trabalhos relacionados, a abordagem da modelagem, definida em 3 etapas, traz melhor detalhamento a decisão.

Assim, as 3 etapas em conjunto oferecem 81,85% na identificação de faturas com pagamento atrasado, 85,63% na identificação desse pagamento atrasado ser após o mês de vencimento, e por fim, de 60,23% até 84,16% na acurácia de predição desses dias de atraso. Assim, os modelos *XGBoost*, *Balanced Random Forest* e *AdaBoost* compõem cada etapa do processo definido, respectivamente.

Em geral, os modelos utilizaram o ganho de informação através da entropia como medida de qualidade para as divisões dos nós das árvores. Dessa forma, foi possível lidar melhor com os dados desbalanceados e alcançar resultados ligeiramente melhores, quando comparado a utilização do índice de impureza (Gini). Também vale mencionar os diferentes desempenhos dos modelos, conforme o objetivo de cada etapa. Na etapa 1, por exemplo, foi observado o aumento da precisão dos modelos no início da semana, enquanto na etapa 2, a melhor precisão ficou no fim do mês. Isto pois a possibilidade de pagamento dentro do mês de vencimento diminui. Por fim, a acurácia da etapa 3, decai para todos os modelos, conforme passam os dias do mês de vencimento, indicando maior variação dos dias de atraso, ao longo do mês.

5 CONCLUSÕES

Tendo em vista a otimização no processo de coleta de contas a receber pelo setor financeiro de uma empresa, este trabalho apresenta uma abordagem de predição de pagamentos atrasados através de algoritmos de árvore de decisão. O processo de identificação foi modelado em 3 etapas: (i) faturas pagas com atraso; (ii) faturas pagas com atraso além do mês de vencimento; e (iii) dias de atraso além do mês de vencimento.

Para alcançar o objetivo proposto, foi realizado um experimento, no qual considera desde a etapa de coleta de dados até a avaliação dos modelos de aprendizado de máquina. Como proposta, foi utilizado técnicas de pré-processamento para remoção de *outliers* e criação de atributos

DOI: 10.25286/repa.v6i5.1746

baseado em histórico, assim como 8 algoritmos de árvore de decisão parametrizados.

Através dos resultados obtidos em modelo, foi identificado em média 80% das faturas atrasadas, faturas atrasadas com pagamento além do mês de vencimento, e por fim a quantidade de dias de atraso. Portanto, o trabalho proposto possibilita o auxílio no processo de coleta de contas a receber, oferecendo até 3 etapas de detalhamento para a tomada decisão. Vale ressaltar, que a modelagem proposta também é generalizada para outras empresas, com diferentes setores de atuação, tendo em vista a utilização de dados históricos.

Como trabalhos futuros, o refino das três etapas será aplicado. Ou seja, uma melhor seleção de atributos, busca por parâmetros e até modelos mais robustos, como redes neurais, podem incrementar nos resultados do experimento. Além disso, a predição dos dias de atraso (etapa 3), será modelada através de intervalos de dias (*buckets*).

REFERÊNCIAS

[1] HU, P. **Predicting and improving invoice-to-cash collection through machine learning**. Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2015.

[2] ZENG, S.; MELVILLE, P.; LANG, C.A.; BOIER-MARTIN, I.M.; MURPHY, C. **Using predictive analysis to improve invoice-to-cash collection**. Proceedings of the 14nd international conference on Knowledge Discovery and Data mining, 2008.

[3] NANDA, S. **Proactive Collections Management: Using Artificial Intelligence to Predict Invoice Payment Dates**. The Credit & Financial Management Review. Credit Research Foundation, 2018.

[4] KOUVELIS, P.; ZHAO, W. **Supply chain finance**. The Handbook of Integrated Risk Management in Global Supply Chains, 2011.

[5] SHAH, H. **Customer Payment Prediction in Account Receivable**. International Journal of Science and Research (IJSR), vol. 8, 2019.

[6] PLACENCIA, J. O.; HALLO, M.; LUJÁN-MORA S. **Detection of Taxpayers with High Probability of Non-payment: An Implementation of a Data Mining Framework**. 15th Iberian Conference on Information Systems and Technologies, 2020.

[7] SUBASI, A.; CANKURT, S. **Prediction of default payment of credit card clients using Data Mining Techniques**. International Engineering Conference, 2019.

[8] ÇETINKAYA, Z.; HORASAN, F. **Decision Trees in Large Data Sets**. Uluslararası Muhendislik Arastirma ve Gelistirme Dergisi, 2021.

[9] SHAIK, A. B.; SRINIVASAN, S. **A Brief Survey on Random Forest Ensembles in Classification Model**. International Conference on Innovative Computing and Communications, p. 253–260, 2018.

[10] FERREIRA, A. J.; FIGUEIREDO, M. A. T. **Boosting Algorithms: A Review of Methods, Theory, and Applications**. Ensemble Machine Learning, 2012.

[11] ZHANG, C.; ZHANG, Y.; SHI, X.; ALMPANIDIS, G.; FAN, G.; SHEN, X. **On Incremental Learning for Gradient Boosting Decision Trees**. Neural Processing Letters, 2019.

[12] SEIFFERT, C.; KHOSHGOFTAAR, T. M.; VAN HULSE, J.; NAPOLITANO, A. **RUSBoost: A Hybrid Approach to Alleviating Class Imbalance**. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans, p. 185-197, 2010.

[13] CHEN, T.; GUESTRIN, C. **XGBoost: A Scalable Tree Boosting System**. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'16), 2016.

[14] LEYS, C.; LEY, C.; KLEIN, O.; BERNARD, P.; LICATA, L. **Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median**. Journal of Experimental Social Psychology, p. 764–766, 2013.

[15] WIRTH, R.; HIPP, J. **CRISP-DM: Towards a Standard Process Model for Data Mining**. Proceedings of the Fourth International Conference on the Practical Application of Knowledge Discovery and Data Mining, 2000.

[16] STRECHT, P.; CRUZ, L.; SOARES, C.; MOREIRA, J.; ABREU, R. **A Comparative Study of Classification and Regression Algorithms for Modelling Students' Academic Performance**. 8th International Conference on Educational Data Mining, 2015.