A IMPORTÂNCIA DO TRATAMENTO DE OUTLIERS NAS TAREFAS DE MINERAÇÃO DE DADOS

João Ubirajara Gomes Oliveira & Gabriel Ludvig Laskos
Orientação: Prof. Dr. Rubens Lopes de Oliveira

RESUMO

Este trabalho de pesquisa ressalta a importância crítica da gestão de outliers (valores atípicos) na mineração de dados, dada a sua capacidade de distorcer análises e comprometer a precisão de modelos preditivos. O estudo aborda a natureza dos outliers, classificando-os e enfatizando a necessidade de compreender sua origem antes do tratamento. A identificação desses valores não só garante a integridade dos dados, mas também revela insights valiosos, como fraudes ou falhas de sistema, que seriam ignorados sem uma análise cuidadosa.

Para a detecção de outliers, o artigo explora diversos métodos, incluindo abordagens estatísticas como o Teste de Grubbs, Z-Score, Intervalo Interquartil (IQR) com Boxplot e Mediana Absoluta do Desvio (MAD). A escolha do método é crucial e depende da natureza e volume dos dados. Após a identificação, o tratamento pode envolver a remoção, transformação de dados (e.g., logaritmos) ou substituição de valores (e.g., por média ou mediana). A decisão sobre a estratégia de tratamento é contextual, visando sempre a robustez e a qualidade dos resultados.

No contexto da avaliação de desempenho de modelos de regressão, o trabalho empregou métricas como o Erro Médio Absoluto (MAE), a Raiz do Erro Quadrático Médio (RMSE) e o Coeficiente de Determinação (R2). O MAE, em particular, foi enfatizado como um parâmetro crucial para medir a precisão dos modelos, indicando a magnitude média dos erros sem considerar sua direção. Os resultados apresentados nas conclusões do trabalho demonstram a eficácia das estratégias de tratamento de outliers na melhoria dessas métricas de desempenho, especialmente o MAE, evidenciando a redução do erro médio e, consequentemente, a maior acurácia dos modelos após a devida gestão dos dados atípicos.

Como resultado final, conclui-se que a gestão eficaz de outliers é fundamental para a confiabilidade dos resultados na mineração de dados. A aplicação de métodos de deteção e tratamento adequados, juntamente com a avaliação rigorosa do desempenho dos modelos por meio de métricas como o MAE, é essencial para construir modelos robustos e embasar decisões precisas, transformando dados brutos em informações acionáveis.

Palavras-chave: Outliers; Mineração de Dados; Deteção de Anomalias; Pré-processamento de Dados; Qualidade de Dados.

1. INTRODUÇÃO

A crescente digitalização de processos e a proliferação de dispositivos conectados têm gerado um volume sem precedentes de dados em diversas áreas do conhecimento. Nesse cenário, a mineração de dados emerge como uma disciplina fundamental, cujo objetivo principal é extrair conhecimento, padrões e informações úteis a partir de grandes e complexos conjuntos de dados [1]. A capacidade de transformar dados brutos em insights acionáveis é crucial para a tomada de decisões estratégicas em setores que vão desde finanças e saúde até marketing e manufatura.

No entanto, a qualidade dos dados é um fator determinante para o sucesso de qualquer processo de mineração de dados. Dados imprecisos, incompletos ou inconsistentes podem levar a análises distorcidas e conclusões errôneas, comprometendo a validade dos modelos e a eficácia das decisões. Entre os diversos desafios relacionados à qualidade dos dados, a presença de outliers (valores atípicos ou discrepantes) se destaca como um dos mais críticos. Outliers são observações que se desviam significativamente das demais observações em um conjunto de dados, podendo ser resultado de erros de medição, erros de entrada de dados, variabilidade natural ou eventos anômalos [2].

A identificação e o tratamento adequado de outliers são etapas essenciais no pré-processamento de dados, pois esses valores podem exercer uma influência desproporcional sobre as análises estatísticas e os algoritmos de aprendizado de máquina. Por exemplo, a presença de outliers pode afetar drasticamente a média e o desvio padrão de um conjunto de dados, distorcendo a representação central e a dispersão dos dados. Em modelos preditivos, outliers podem levar a um ajuste inadequado, resultando em menor precisão e capacidade de generalização [3].

Este artigo tem como objetivo explorar a importância da identificação e eliminação de outliers nas tarefas de mineração de dados. Para tanto, será apresentada uma fundamentação teórica sobre o que são outliers e por que sua deteção é crucial. Em seguida, serão detalhados os principais métodos de deteção, incluindo abordagens estatísticas, baseadas em distância, em densidade e em modelos. Posteriormente, serão discutidas as estratégias de tratamento de outliers, como remoção, transformação de dados, substituição de valores e utilização de métodos robustos. O impacto dos outliers na mineração de dados será analisado em profundidade, e estudos de caso e aplicações práticas serão apresentados para ilustrar a relevância do tema. Por fim, serão apresentadas as conclusões e considerações finais sobre a gestão de outliers para garantir a qualidade e a confiabilidade dos resultados na mineração de dados.

2. FUNDAMENTAÇÃO TEÓRICA

No contexto da mineração de dados e da análise estatística, um outlier é uma observação que se distancia significativamente das demais observações

em um conjunto de dados. Essa discrepância pode ser tão grande que levanta a suspeita de que o outlier foi gerado por um mecanismo diferente ou que representa um evento raro e incomum [4]. A identificação de outliers é um campo de estudo importante, pois esses valores podem ter um impacto substancial na análise de dados, distorcendo resultados e levando a conclusões errôneas se não forem devidamente tratados.

Os outliers podem se manifestar de diversas formas e por diferentes razões. Em alguns casos, são o resultado de erros de medição ou de entrada de dados, como um erro de digitação que registra uma idade de 200 anos para uma pessoa. Em outros, podem ser anomalias genuínas que representam eventos raros, mas importantes, como uma transação financeira fraudulenta ou um pico inesperado no tráfego de rede que indica um ataque cibernético. A distinção entre um erro e uma anomalia real é crucial e muitas vezes depende do contexto do problema e do domínio de aplicação [5].

Existem diferentes tipos de outliers, que podem ser classificados com base em sua natureza e no contexto em que ocorrem:

- **Outliers Pontuais (Point Outliers):** São observações individuais que se desviam significativamente do restante dos dados. São os tipos mais comuns e geralmente os primeiros a serem identificados em análises exploratórias. Por exemplo, um único valor de temperatura extremamente alto em um conjunto de dados de temperaturas diárias.
- **Outliers Contextuais (Contextual Outliers):** Uma observação é considerada um outlier contextual se seu valor é anômalo em um contexto específico, mas não em outro. Por exemplo, uma temperatura de 30°C em um dia de verão no Brasil pode ser normal, mas a mesma temperatura em um dia de inverno na Antártida seria um outlier contextual. A deteção desses outliers requer a consideração de atributos contextuais, como tempo ou localização [6].
- **Outliers Coletivos (Collective Outliers):** Um conjunto de observações relacionadas é considerado um outlier coletivo se, quando consideradas em conjunto, elas se desviam significativamente do restante dos dados, mesmo que as observações individuais não sejam outliers. Por exemplo, uma sequência de transações de baixo valor que, em conjunto, podem indicar um padrão de lavagem de dinheiro [7].

A compreensão da natureza e da origem dos outliers é fundamental para decidir a melhor abordagem para sua deteção e tratamento. A simples remoção de um outlier sem entender sua causa pode levar à perda de informações valiosas ou à ocultação de problemas reais nos dados.

2.1 Importância da identificação de Outliers

A identificação de outliers é uma etapa crítica e indispensável no processo de mineração de dados, com implicações diretas

na qualidade, precisão e confiabilidade das análises e modelos construídos. Ignorar a presença de outliers pode levar a uma série de problemas que comprometem a validade dos resultados e a eficácia das decisões baseadas nos dados [2].

Primeiramente, outliers podem distorcer significativamente as estatísticas descritivas de um conjunto de dados. Medidas como a média e o desvio padrão são particularmente sensíveis a valores extremos. Por exemplo, em um conjunto de dados de salários, a presença de um ou dois salários extremamente altos pode elevar artificialmente a média, dando uma falsa impressão da remuneração típica. A mediana, por ser uma medida de tendência central mais robusta a outliers, é frequentemente utilizada em conjunto com a média para fornecer uma visão mais completa da distribuição dos dados [3].

Em segundo lugar, a presença de outliers pode impactar negativamente o desempenho de algoritmos de aprendizado de máquina. Muitos algoritmos, como regressão linear e redes neurais, são baseados em suposições sobre a distribuição dos dados e são sensíveis a valores extremos. Outliers podem atuar como pontos de alavancagem, puxando a linha de regressão ou as fronteiras de decisão em direções incorretas, resultando em modelos com menor poder preditivo e maior erro de generalização. O modelo pode se sair bem nos dados de treinamento, mas ter um desempenho ruim em dados novos [3].

Além disso, a identificação de outliers pode revelar informações valiosas e insights importantes que, de outra forma, passariam despercebidos. Em alguns contextos, os outliers não são meros erros, mas sim anomalias que indicam eventos incomuns, mas significativos. Por exemplo, em sistemas de deteção de fraude, transações que são consideradas outliers podem ser indicadores de atividades fraudulentas. Em monitoramento de sistemas, um outlier no consumo de recursos pode sinalizar um ataque cibernético ou uma falha de hardware. Nesses casos, a deteção de outliers é o objetivo principal da análise, e não um obstáculo a ser removido [7].

Finalmente, a identificação de outliers é fundamental para garantir a integridade e a qualidade dos dados. Ao detetar valores atípicos, é possível investigar suas causas, que podem incluir erros de medição, falhas de sensores, erros de entrada manual ou problemas no processo de coleta de dados. Corrigir esses erros na fonte não apenas melhora a qualidade do conjunto de dados atual, mas também previne a ocorrência de problemas semelhantes em futuras coletas de dados, contribuindo para a construção de bases de dados mais limpas e confiáveis a longo prazo [2].

Em suma, a identificação de outliers não é apenas uma etapa técnica no pré-processamento de dados, mas uma prática essencial que garante a validade das análises, a robustez dos modelos e a descoberta de insights críticos, transformando dados brutos em informações valiosas para a tomada de decisões informadas

3. MÉTODOS DE DETECÇÃO DE OUTLIERS

A deteção de outliers é um campo de pesquisa ativo e crucial na mineração de dados, com uma vasta gama de técnicas desenvolvidas para identificar observações que se desviam significativamente do padrão esperado. A escolha do método adequado depende de diversos fatores, como a natureza dos dados, o volume, a dimensionalidade e o conhecimento prévio sobre a existência e o tipo de outliers. Nesta seção, exploraremos os principais métodos de deteção de outliers, categorizando-os para facilitar a compreensão.

3.1 Métodos estatísticos

Os métodos estatísticos são amplamente utilizados para a deteção de outliers, especialmente quando a distribuição dos dados é conhecida ou pode ser aproximada por uma distribuição paramétrica. Esses métodos baseiam-se em modelos estatísticos para identificar observações que se encontram nas caudas da distribuição, ou seja, que são improváveis de ocorrer sob o modelo assumido. A seguir, alguns dos métodos estatísticos mais comuns.

3.2 Testes de hipóteses

Testes de hipóteses, como o Teste de Grubbs (para um único outlier em uma distribuição normal) ou o Teste de Dixon (para um ou dois outliers em pequenas amostras), são utilizados para verificar se uma observação específica pode ser considerada um outlier. Esses testes calculam uma estatística de teste e a comparam com um valor crítico para determinar se a observação deve ser rejeitada como parte da distribuição principal. Embora eficazes para um número limitado de outliers e distribuições específicas, sua aplicabilidade é restrita em grandes conjuntos de dados ou em distribuições não-normais [8].

3.3 Desvio padrão e Z-Score

Para dados que seguem uma distribuição normal, o desvio padrão é uma medida de dispersão que pode ser utilizada para identificar outliers. Observações que estão a uma certa quantidade de desvios padrão de distância da média são consideradas outliers. O Z-Score (ou escore padrão) quantifica essa distância em termos de desvios padrão, sendo calculado pela fórmula:

$$Z = (x - \mu) / \sigma,$$

onde x é a observação, μ é a média e σ é o desvio padrão. Valores com Z-Score absoluto superior a um determinado limiar (geralmente 2, 2.5 ou 3) são frequentemente classificados como outliers. Este método é simples e intuitivo, mas é sensível à presença de outliers na própria estimativa da média e do desvio padrão, o que pode mascarar a deteção de outros outliers [3].

3.4 Intervalo interquartil (IQR) e boxplot

O método do Intervalo Interquartil (IQR) é uma abordagem não paramétrica e mais robusta a outliers do que os métodos baseados na média e desvio padrão, pois utiliza quartis em vez da média. O IQR é a diferença entre

o terceiro quartil (Q3) e o primeiro quartil (Q1) dos dados (IQR = Q3 - Q1). Outliers são definidos como valores que estão abaixo de Q1 - 1.5 * IQR ou acima de Q3 + 1.5 * IQR. Essa técnica é visualmente representada pelo boxplot, que exibe a mediana, os quartis e os valores extremos, com os outliers plotados como pontos individuais fora dos 'bigodes' do gráfico. O boxplot é uma ferramenta poderosa para a visualização e identificação rápida de outliers em distribuições assimétricas ou com valores extremos [3].

3.5 Mediana absoluta do desvio (MAD)

A Mediana Absoluta do Desvio (MAD) é outra medida de dispersão robusta a outliers, utilizada para detetar outliers em dados univariados. A MAD é calculada como a mediana dos desvios absolutos de cada ponto de dados em relação à mediana do conjunto de dados. Observações que se desviam da mediana por um múltiplo da MAD (geralmente 2 ou 3 vezes) são consideradas outliers. A MAD é menos sensível a valores extremos do que o desvio padrão, tornando-a mais adequada para conjuntos de dados com outliers significativos [9].

4. TRATAMENTO DE OUTLIERS

Após a identificação de outliers, é essencial aplicar técnicas apropriadas para seu tratamento, visando garantir a qualidade dos dados e a robustez das análises subsequentes. A escolha da técnica mais adequada depende do contexto, da natureza dos dados e do objetivo da análise. Nesta seção, são abordadas as principais estratégias de tratamento de outliers, discutindo suas vantagens, limitações e aplicabilidades.

4.1 Remoção de outliers

A remoção de outliers é uma das abordagens mais diretas e comumente utilizadas. Consiste em simplesmente eliminar do conjunto de dados as observações que foram identificadas como discrepantes. Essa técnica é particularmente eficaz quando os outliers são claramente erros de medição, entrada de dados ou registros que não representam o fenômeno estudado.

Por exemplo, em um conjunto de dados clínicos, uma idade registrada como 999 anos pode ser evidentemente um erro e, portanto, justificaria a exclusão. Métodos como o Z-score e o IQR são frequentemente utilizados para identificar esses valores extremos que podem ser removidos de forma segura [8].

Contudo, a remoção indiscriminada pode resultar na perda de informações valiosas, especialmente em domínios onde os outliers representam casos raros, porém relevantes — como fraudes financeiras ou diagnósticos médicos incomuns. Portanto, é fundamental analisar o contexto antes de optar pela exclusão dos dados discrepantes [6].

4.2 Transformação de dados

Transformar os dados é uma alternativa que visa reduzir o impacto dos outliers, ao invés de removê-los. Técnicas de transformação estatística, como logaritmos,

raiz quadrada, ou Box-Cox, podem comprimir a escala dos dados e suavizar a influência de valores extremos.

Por exemplo, aplicar uma transformação logarítmica em variáveis com distribuição assimétrica pode melhorar a normalidade dos dados e reduzir o peso dos outliers sobre os modelos estatísticos [10].

A principal vantagem dessa abordagem é manter todas as observações no conjunto de dados, evitando a perda de informações. No entanto, nem todas as variáveis são adequadas para transformação (ex.: dados com valores negativos ou nulos), e é necessário avaliar cuidadosamente os efeitos da transformação sobre a interpretação dos resultados.

4.3 Substituição de Valores

A substituição consiste em manter os registros outliers, mas substituindo seus valores por uma estimativa mais plausível. Essa estimativa pode ser baseada em medidas centrais, como a média, mediana ou moda, ou ser resultado de algoritmos de imputação, como k-Nearest Neighbors (kNN), Regressão ou Árvores de Decisão [11].

A imputação por mediana é especialmente útil em distribuições assimétricas, pois é mais robusta a outliers do que a média. Já técnicas mais complexas, como imputação por modelos preditivos, consideram múltiplas variáveis para estimar o valor substituído com maior precisão.

Apesar de ser útil para manter a integridade do conjunto de dados, a substituição de valores pode introduzir viés se for aplicada indiscriminadamente, especialmente em grandes proporções de dados discrepantes.

4.4 Métodos Robustos

Em algumas situações, em vez de modificar os dados, é mais apropriado empregar algoritmos e técnicas estatísticas robustas, que sejam menos sensíveis à presença de outliers. Modelos robustos são projetados para minimizar o impacto dos valores extremos sem a necessidade de pré-processamento específico.

Por exemplo, a Regressão de Mínimos Quadrados Ordinários (OLS) é fortemente influenciada por outliers, enquanto alternativas como a Regressão de Mínimos Quadrados Ponderados ou a Regressão Quantílica oferecem maior resistência a esses efeitos [12].

Além disso, algoritmos de aprendizado de máquina, como Árvores de Decisão, Random Forests e Redes Neurais Profundas, apresentam maior tolerância a outliers por sua própria natureza ou por meio de regularizações específicas.

O uso de métodos robustos é particularmente valioso em contextos onde os outliers contêm informações significativas e não devem ser removidos nem transformados. Nesses casos, manter a integridade dos dados originais pode ser essencial para a descoberta de padrões raros, mas relevantes.

5. IMPACTO DOS OUTLIERS NA MINERAÇÃO DE DADOS

A presença de outliers em um conjunto de dados pode ter um impacto profundo e multifacetado em todas as etapas do processo de mineração de dados, desde a análise exploratória até a construção e avaliação de modelos. Compreender esses impactos é fundamental para justificar a necessidade de deteção e tratamento de outliers, garantindo a validade e a confiabilidade dos resultados obtidos.

5.1. Distorção de estatísticas descritivas

Os outliers são capazes de distorcer significativamente as estatísticas descritivas básicas de um conjunto de dados. A média aritmética, por exemplo, é extremamente sensível a valores extremos. Um único outlier muito grande ou muito pequeno pode puxar a média para longe do centro da maioria dos dados, fornecendo uma representação enganosa da tendência central. Da mesma forma, o desvio padrão e a variância, que medem a dispersão dos dados em torno da média, são inflacionados pela presença de outliers, sugerindo uma variabilidade maior do que a real. Isso pode levar a interpretações incorretas sobre a distribuição dos dados e a tomada de decisões baseadas em informações imprecisas [3].

5.2. Prejuízo à qualidade dos modelos preditivos

Um dos impactos mais críticos dos outliers é o prejuízo à qualidade e ao desempenho dos modelos preditivos e de classificação. Muitos algoritmos de aprendizado de máquina, especialmente aqueles baseados em otimização de funções de custo que utilizam a soma dos erros quadráticos (como a regressão linear e as redes neurais tradicionais), são altamente sensíveis a outliers. Esses valores extremos podem atuar como pontos de alavancagem, influenciando desproporcionalmente o ajuste do modelo. O resultado é um modelo que se ajusta mal aos dados normais, apresentando alta variância e baixa capacidade de generalização para novos dados. Isso se manifesta em:

- **Redução da Precisão:** O modelo pode ter um desempenho inferior em tarefas de regressão ou classificação, resultando em previsões imprecisas ou classificações erradas.
- **Overfitting:** Em alguns casos, o modelo pode tentar se ajustar aos outliers, tornando-se excessivamente complexo e perdendo a capacidade de generalizar para dados não vistos. Isso é particularmente problemático em modelos que buscam minimizar o erro global, onde um grande erro em um outlier pode dominar a função de custo [13].
- **Interpretabilidade Comprometida:** A presença de outliers pode ocultar as verdadeiras relações entre as variáveis, tornando a interpretação dos parâmetros do modelo mais difícil e menos confiável. Por exemplo, em um modelo de regressão, os coeficientes podem ser distorcidos, levando a conclusões

errôneas sobre a influência de certas características [14].

5.3. Impacto em algoritmos de agrupamento (Clustering)

Algoritmos de agrupamento, como K-Means, são baseados na distância entre os pontos de dados. Outliers podem afetar significativamente a formação de clusters, puxando os centroides para longe dos agrupamentos reais ou formando clusters espúrios compostos por apenas alguns outliers. Isso pode levar a uma segmentação incorreta dos dados, prejudicando a descoberta de padrões e a identificação de grupos homogêneos [15].

5.4. Dificuldade na descoberta de padrões

Outliers podem mascarar padrões e relações importantes nos dados. Ao introduzir ruído ou desviar a atenção para valores extremos, eles podem dificultar a identificação de tendências, correlações e regras de associação que seriam evidentes em um conjunto de dados limpo. Isso é especialmente relevante em tarefas de descoberta de conhecimento, onde o objetivo é encontrar informações implícitas e previamente desconhecidas [1].

5.5. Conclusões enganosas e decisões inadequadas

Em última análise, o impacto mais grave dos outliers não tratados é a possibilidade de levar a conclusões enganosas e, consequentemente, a decisões inadequadas. Se as análises e os modelos são construídos sobre dados distorcidos por outliers, as estratégias e ações baseadas nesses resultados podem ser ineficazes, custosas ou até mesmo prejudiciais. Por exemplo, em uma análise de risco de crédito, a falha em identificar outliers fraudulentos pode levar a perdas financeiras significativas para uma instituição [2].

Em resumo, a gestão eficaz de outliers é um pilar fundamental para garantir a integridade e a validade de qualquer projeto de mineração de dados. A negligência nesse especto pode comprometer a precisão das análises, a robustez dos modelos e a confiabilidade das decisões, transformando um processo promissor em uma fonte de informações distorcidas.

6. ESTUDOS DE CASOS

Os estudos de caso apresentados a seguir compartilham uma estrutura comum de análise. Em todos os casos, os dados foram submetidos a um processo sistemático que incluiu: (i) pré-processamento com remoção de colunas irrelevantes, tratamento de valores ausentes e codificação de variáveis categóricas; (ii) deteção e tratamento de outliers com a técnica do Intervalo Interquartil (IQR); (iii) normalização dos dados numéricos utilizando o método StandardScaler; (iv) treinamento de um modelo de regressão supervisionada, com destaque para o uso do algoritmo XGBoost (ou Regressão Linear, conforme o caso); e (v) avaliação da performance dos modelos por

meio de métricas como RMSE (Root Mean Squared Error), MAE (Mean Absolute Error) e R^2 (coeficiente de determinação). Esses procedimentos garantiram maior robustez às análises e contribuíram significativamente para a melhoria na qualidade das previsões. A seguir, detalha-se cada um dos estudos de caso.

6.1 Estudo de caso: deteção e tratamento de outliers na base air quality

Este estudo analisou dados ambientais coletados por sensores automáticos para medir a qualidade do ar, incluindo monóxido de carbono (CO), dióxido de nitrogênio (NO2), temperatura, umidade e outros gases. Após uma inspeção inicial, variáveis como CO e NO2 apresentaram concentrações extremas em horários específicos. Esses valores foram identificados como outliers utilizando boxplots e a técnica do intervalo interquartil (IQR). O tratamento incluiu a remoção de registros com múltiplos valores discrepantes e aplicação de transformação logarítmica em variáveis como CO.

A base foi normalizada com StandardScaler e usada em um modelo XGBoost para prever concentrações de poluentes. O desempenho do modelo melhorou consideravelmente após o tratamento dos outliers. Além da redução no erro quadrático médio (MAE), foi observada maior estabilidade nas previsões. Este caso demonstra a importância do pré-processamento em conjuntos de dados ambientais, frequentemente sujeitos a medições anômalas.

Adotando-se o processo sistemático descrito na introdução desse capítulo, verificou-se os seguintes resultados de desempenho (MAE):

(i) tratamento de valores ausentes e codificação de variáveis categóricas; **5.68**

(ii) tratamento de outliers com a técnica do Intervalo Interquartil (IQR); **3.87**

(iii) normalização dos dados numéricos: **0.17**

6.2 Estudo de caso: análise de outliers em dados de criminalidade

Neste estudo, uma base de dados com indicadores sociais e estatísticas criminais dos EUA foi usada para prever a taxa de crimes violentos por população. A análise inicial revelou forte correlação entre variáveis socioeconômicas, como desemprego, pobreza e educação, e a criminalidade. Outliers foram detetados com Z-score e boxplots, com destaque para registros com taxas de crimes violentos anormalmente altas ou zero em variáveis relevantes.

O tratamento incluiu remoção seletiva de registros com múltiplas inconsistências e imputação pela mediana. Após normalização dos dados, o modelo XGBoost foi treinado e apresentou melhoria significativa na performance, evidenciada pela redução do RMSE. As variáveis mais importantes incluíram nível educacional e estrutura familiar. Este caso reforça o impacto dos outliers na modelagem preditiva em contextos sociais complexos.

Adotando-se o processo sistemático descrito na introdução desse capítulo, verificou-se os seguintes resultados de desempenho (MAE):

(i) tratamento de valores ausentes e codificação de variáveis categóricas; **27.00**

(ii) tratamento de outliers com a técnica do Intervalo Interquartil (IQR); **19.71**

(iii) normalização dos dados numéricos: **19.71**

6.3 Estudo de caso: identificação e tratamento de outliers em dados automotivos para predição de consumo urbano

Com foco na eficiência de veículos, este estudo utilizou uma base automotiva contendo variáveis como potência (horsepower), peso (curb-weight), tamanho do motor (engine-size) e tipo de tração. O objetivo foi prever o consumo urbano (city-mpg) com Regressão Linear. Foram detetados outliers em variáveis como horsepower e peso utilizando histogramas, Z-score e boxplots.

O tratamento incluiu remoção de registros extremos, substituição de valores pela mediana e transformação logarítmica. Após a normalização dos dados, o modelo apresentou melhor desempenho, com aumento no R^2 e redução no MAE. Além disso, os coeficientes do modelo tornaram-se mais interpretáveis, demonstrando relações coerentes entre potência, peso e consumo. O estudo ilustra o impacto positivo da preparação dos dados em problemas de engenharia.

(i) tratamento de valores ausentes e codificação de variáveis categóricas; **0.83**

(ii) tratamento de outliers com a técnica do Intervalo Interquartil (IQR); **0.83**

(iii) normalização dos dados numéricos: **0.04**

6.4 Estudo de caso: Predição do nível de vício em celular entre adolescentes

A base de dados analisada continha respostas de adolescentes sobre hábitos digitais, escolaridade e contexto familiar. A variável alvo foi o Addiction Level, representando o nível de dependência de celular. As variáveis preditoras incluíam gênero, horas no celular, tipo de escola e uso de redes sociais. Categóricas foram codificadas com LabelEncoder e variáveis numéricas normalizadas.

Outliers foram detetados com IQR e removidos para evitar distorções. O modelo XGBoost foi então treinado com validação por métricas de regressão. O resultado foi um R^2 superior a 0.80, com RMSE e MAE dentro de limites aceitáveis. As variáveis com maior importância foram tempo de uso diário, desempenho escolar e uso de redes sociais. Este caso evidencia a eficácia da modelagem supervisionada com dados comportamentais e a necessidade de cuidado no tratamento de dados sensíveis.

(i) tratamento de valores ausentes e codificação de variáveis categóricas; **0.35**

(ii) tratamento de outliers com a técnica do Intervalo Interquartil (IQR); **0.34**

(iii) normalização dos dados numéricos: **0.34**

6.5 Estudo de caso: análise da qualidade de vinhos

Este estudo explorou a base de dados Wine Quality, que contém informações físico-químicas de amostras de vinhos tintos e brancos, associadas à respectiva nota de qualidade (variável alvo) atribuída por especialistas. As variáveis disponíveis incluem, entre outras, acidez fixa, acidez volátil, teor de álcool, pH, cloretos e densidade, que são determinantes para a avaliação sensorial da bebida.

Após a etapa inicial de exploração dos dados, verificou-se a presença de valores extremos em variáveis como acidez volátil e teor alcoólico, que poderiam distorcer o treinamento dos modelos de previsão da qualidade. Para tratar esses casos, foram utilizados boxplots e a técnica do Intervalo Interquartil (IQR), possibilitando a deteção de outliers em faixas discrepantes. Em situações em que os valores extremos não apresentavam justificativa química plausível, os registros foram removidos. Em outras variáveis, como o teor alcoólico, foi aplicada transformação logarítmica para reduzir assimetrias.

Na sequência, as variáveis numéricas foram normalizadas com o método StandardScaler, assegurando que atributos com escalas diferentes (por exemplo, pH e sulfatos) não dominassem o processo de modelagem. O conjunto de dados resultante foi dividido em treino e teste, sendo treinado um modelo de regressão supervisionada com o algoritmo XGBoost, escolhido pela sua robustez em problemas de previsão com múltiplas variáveis correlacionadas.

Os resultados indicaram melhora significativa no desempenho após o tratamento de outliers e normalização. O modelo apresentou redução nos erros médios (MAE e RMSE) e incremento no R^2, refletindo maior capacidade preditiva. Além disso, foi observada maior consistência nas previsões em faixas intermediárias de qualidade do vinho, que tendem a ser mais frequentes na base.

Empregando os parâmetros de desempenho selecionados para nosso estudo observamos:

(i) tratamento de valores ausentes e codificação de variáveis categóricas; **0.30**

(ii) tratamento de outliers com a técnica do Intervalo Interquartil (IQR); **0.27**

(iii) normalização dos dados numéricos: **0.27**

6.6 Estudo de Caso: Análise da relação entre Comunidades e Crimes

Este estudo de caso utilizou a base Communities and Crime Unnormalized, que reúne indicadores socioeconômicos, demográficos e policiais de comunidades dos Estados Unidos, com o objetivo de prever a taxa de criminalidade. As variáveis incluem fatores como renda média, taxa de desemprego, nível educacional, proporção

de jovens, investimentos em policiamento, entre outros.

Na inspeção inicial, verificou-se a presença de valores faltantes em variáveis relacionadas a estatísticas socioeconômicas e policiais. Esses casos foram tratados por meio da remoção de atributos com excesso de dados ausentes e imputação da média para valores isolados. Em seguida, a análise exploratória identificou outliers em indicadores como taxa de desemprego, densidade populacional e investimentos policiais, que apresentavam concentrações extremamente elevadas em algumas comunidades específicas.

A deteção foi realizada por meio de boxplots e aplicação da técnica do Intervalo Interquartil (IQR). O tratamento incluiu a remoção de registros com múltiplos outliers severos e, em variáveis com forte assimetria, como densidade populacional, foi aplicada transformação logarítmica.

Com o objetivo de uniformizar a escala das variáveis, foi utilizada a normalização com StandardScaler, assegurando que atributos como renda média e percentual de analfabetismo não dominassem o processo de aprendizado do modelo.

Na etapa de modelagem, a base foi dividida em treino e teste, e um modelo de regressão supervisionada com XGBoost foi empregado para prever a taxa de crimes violentos. O desempenho do modelo foi avaliado por métricas como RMSE, MAE e R^2. Após o tratamento de outliers e normalização, observou-se uma redução significativa nos erros médios e maior estabilidade nas previsões, especialmente para comunidades de perfil socioeconômico intermediário.

(i) tratamento de valores ausentes e codificação de variáveis categóricas; **1549.00**

(ii) tratamento de outliers com a técnica do Intervalo Interquartil (IQR); **1502.30**

(iii) normalização dos dados numéricos: **0.14**

6.7 Estudo de Caso: Análise de Desempenho de notas de Estudantes

Este estudo utilizou a base Student Performance, que contém informações demográficas, sociais e acadêmicas de estudantes do ensino médio em Portugal, com foco no desempenho escolar em disciplinas como Matemática e Português. As variáveis incluem dados sobre idade, sexo, escolaridade dos pais, hábitos de estudo, consumo de álcool, tempo de lazer e notas finais (variável alvo).

Na etapa inicial, foi conduzido um processo de pré-processamento que envolveu: remoção de atributos redundantes (como identificadores não relevantes), tratamento de valores ausentes e codificação de variáveis categóricas (por exemplo, sexo, profissão dos pais e status de relacionamento familiar) utilizando a técnica de one-hot encoding.

A análise exploratória identificou outliers em variáveis relacionadas às notas parciais (G1 e G2) e ao consumo de álcool, que em alguns registros apresentavam valores incompatíveis com o padrão da amostra. A deteção foi realizada por meio de boxplots

e da técnica do Intervalo Interquartil (IQR). O tratamento incluiu a remoção de registros com múltiplas inconsistências e a aplicação de transformação logarítmica em variáveis assimétricas, como o tempo de estudo semanal.

Em seguida, os dados numéricos foram normalizados com o método StandardScaler, assegurando que atributos de escalas diferentes, como idade e horas de estudo, tivessem impacto equilibrado no processo de aprendizado.

Seguindo a metodologia de registro de desempenho do tratamento de outliers adotada para os casos de uso em estudo neste capítulo, obteve-se os resultados a seguir:

(i) tratamento de valores ausentes e codificação de variáveis categóricas; **2,33**

(ii) tratamento de outliers com a técnica do Intervalo Interquartil (IQR); **2,20**

(iii) normalização dos dados numéricos: **0.15**

7. CONCLUSÕES

Os gráficos de linha a seguir (figura 1) apresentam a evolução do MAE em cada estudo de caso. Cada curva representa um caso estudado e os três pontos indicam:

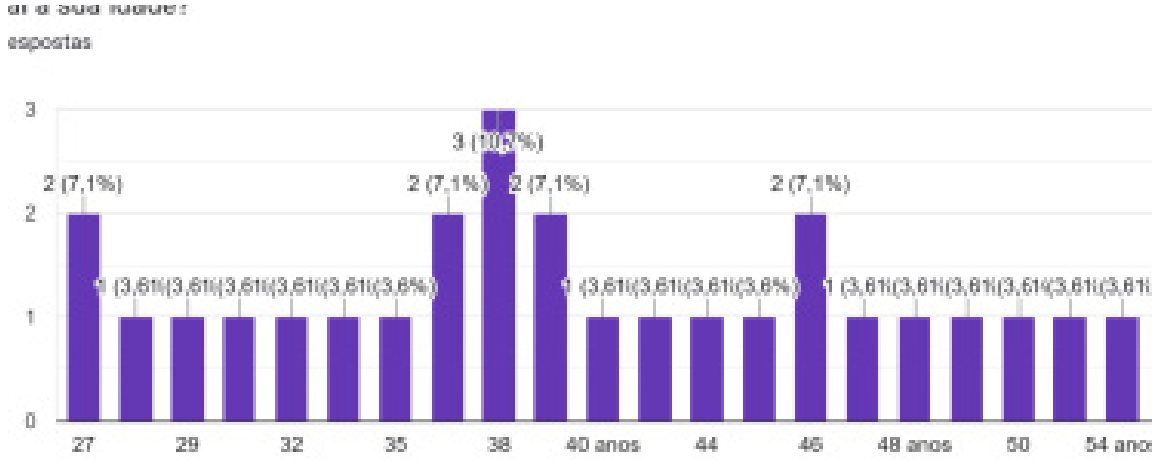
1. Tratamento apenas de valores faltantes

2. Tratamento de valores faltantes + outliers

3. Tratamento de valores faltantes + outliers + normalização

Esses formatos evidenciam, claramente, a tendência de redução do erro ao longo das etapas de pré-processamento, destacando a importância do tratamento de outliers e da normalização para melhorar o desempenho dos modelos de regressão.

Figura 1- Gráfico de linha (MAE)



O tratamento de possíveis outliers, seguido de uma etapa de normalização dos resultados evidencia um padrão consistente.

7.1 Impacto inicial dos outliers

Nos estudos Air Quality, Automotivo, Communities and Crime e Desempenho de Estudantes, a diferença entre considerar apenas dados faltantes e aplicar também o tratamento de outliers foi significativa. Isso mostra que os outliers estavam distorcendo a distribuição dos dados e comprometendo a capacidade preditiva do modelo.

7.2 Efeito da normalização após tratamento de outliers

A normalização potencializou ainda mais os ganhos. No caso Air Quality e Automotivo, o MAE caiu drasticamente após essa etapa, evidenciando que a uniformização das escalas das variáveis é um passo crucial quando lidamos com atributos heterogêneos.

7.3 Bases menos sensíveis a outliers

Em bases como Vício em Celular e Wine Quality, os ganhos com o tratamento de outliers foram pequenos, mas a normalização trouxe estabilidade e consistência, evitando que variáveis em escalas diferentes dominassem o aprendizado.

7.4 Casos complexos (Criminalidade e Communities and Crime)

Apesar de a redução no MAE não ser tão expressiva no caso de Criminalidade, a remoção de outliers garante maior robustez, evitando que registros extremos influenciem excessivamente o modelo. Já em Communities and Crime, o impacto foi enorme após normalização, mostrando que dados socioeconômicos tendem a apresentar valores extremos e desbalanceados, tornando indispensável o pré-processamento.

7.5 Considerações finais

A análise realizada ao longo dos estudos de caso evidencia que o tratamento de outliers é uma etapa fundamental no desenvolvimento de modelos de regressão. Sua adequada identificação e gestão não apenas mitigam distorções estatísticas, mas também aumentam a estabilidade dos modelos e a confiabilidade das previsões, contribuindo para uma interpretação mais consistente das relações entre as variáveis.

A etapa subsequente de normalização dos dados atua de forma complementar, ao equalizar a escala das variáveis numéricas. Isso evita que atributos com magnitudes maiores dominem o processo de aprendizado, garantindo que todas as variáveis contribuam de maneira justa e proporcional para o resultado final.

Assim, a combinação entre o tratamento de outliers e a normalização dos dados revelou-se a estratégia mais eficaz para o pré-processamento em mineração de dados. Essa abordagem integrada demonstrou ser robusta e generalizável, promovendo ganhos significativos em desempenho — expressos pela redução do MAE e RMSE e pelo aumento

do R^2 — em conjuntos de dados de naturezas distintas, sejam eles ambientais, sociais, educacionais ou de engenharia. Em suma, essa prática sistemática é essencial para transformar dados brutos em conhecimento confiável e acionável.

Referências

[1] Fayyad, U. M., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, 17(3), 37-54.

[2] Awari. (2023). Outliers na mineração de dados: identificação e tratamento de valores atípicos. Disponível em: https://awari.com.br/outliers-na-mineracao-de-dados-identificacao-e-tratamento-de-valores-atipicos/

[3] DataGeeks. (2024). Outliers: O que são Pontos Fora da Curva e Como Tratá-los? Disponível em: https://www.datageeks.com.br/outliers/

[4] Han, J., Kamber, M., & Pei, J. (2011). Data Mining: Concepts and Techniques. Morgan Kaufmann.

[5] IBM. (s.d.). Análise de valor discrepante. Disponível em: https://www.ibm.com/docs/pt-br/planning-analytics/2.0.0?topic=ai-outlier-analysis

[6] Aggarwal, C. C. (2016). Outlier Analysis. Springer.

[7] Trend Micro. (s.d.). O que é Mineração de Dados?. Disponível em: https://www.trendmicro.com/pt_br/what- is/machine-learning/data-mining.html

[8] Barnett, V., & Lewis, T. (1994). Outliers in Statistical Data. John Wiley & Sons.

[9] Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. Journal of Experimental Social Psychology, 49(4), 764-766.

[10] Osborne, J. W. (2010). Improving your data transformations: Applying the Box-Cox transformation. Practical Assessment, Research, and Evaluation, 15(12).

[11] Batista, G. E. A. P. A., & Monard, M. C. (2003). An analysis of four missing data treatment methods for supervised learning. Applied Artificial Intelligence, 17(5-6), 519-533.

[12] Rousseau, P. J., & Leroy, A. M. (2005). Robust Regression and Outlier Detection. Wiley.

[13] Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

[14] Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). Introduction to Linear Regression Analysis. John Wiley & Sons.

[15] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), 226-231.