MODELAGEM PREDITIVA E INTERPRETÁVEL PARA PREVISÃO DE LUCRO NO COMÉRCIO ELETRÔNICO: UMA ANÁLISE COMPARATIVA ENTRE MÉTODOS ESTATÍSTICOS E ENSEMBLE DE APRENDIZADO DE MÁQUINA

Alexandre Bezerra de Souza[1], Romário de Sousa Almeida[1,2], Arthur Victor Ramos Vidal Negreiros[1], Maria Helena Ramos de Macedo[1], Eduarda da Silva Brito[1], Ítalo Gabriel da Silva[1]

[1]Universidade Estadual da Paraíba - UEPB, Campina Grande, Brasil
(alexandre.souza@aluno.uepb.edu.br)
[2]Universidade Federal de Lavras - UFLA, Lavras, Brasil

Resumo: O crescimento do comércio eletrônico tem intensificado a necessidade de modelos preditivos capazes de extrair conhecimento a partir de grandes volumes de dados transacionais. Este estudo avalia o desempenho de modelos estatísticos e de aprendizado de máquina na previsão do lucro, incluindo Regressão Linear, Random Forest, XGBoost e Stacked Ensemble. Foi implementado um pipeline completo em R, contemplando pré-processamento, engenharia de atributos, validação cruzada com 10 folds e otimização de hiperparâmetros. O desempenho foi avaliado por métricas como RMSE, MAE, MAPE e R^2. Os resultados indicaram desempenho semelhante entre os modelos, com leve vantagem para o Random Forest, sugerindo uma relação predominantemente linear entre as variáveis. A análise de interpretabilidade, baseada em importância de variáveis e valores SHAP, evidenciou a variável Sales como principal preditora do lucro. Os achados indicam que modelos mais simples podem ser suficientes em cenários com baixa complexidade estrutural, contribuindo para aplicações práticas no suporte à tomada de decisão no comércio eletrônico.

Palavras-chave: Análise preditiva; Inteligência de negócios; Modelagem regressiva; Interpretabilidade de modelos; Ciência de dados.

INTRODUÇÃO

O avanço das tecnologias digitais tem impulsionado significativamente o crescimento do comércio eletrônico, resultando em um aumento expressivo das transações online em escala global. Segundo projeções de mercado, esse setor apresenta crescimento acelerado, impulsionado pela digitalização e pela maior acessibilidade às plataformas de compra (SIKDER; ROLFE, 2023).

O mercado global de comércio eletrônico atingiu aproximadamente US$ 13 trilhões em 2021, com expectativa de alcançar US$ 55,6 trilhões até 2027, apresentando uma taxa de crescimento anual composta de 27,4% (RESEARCH AND MARKETS, 2022).

Nesse contexto, a expansão do volume de dados exige abordagens analíticas eficientes para extrair informações relevantes. Modelos preditivos tornam-se essenciais para antecipar tendências, otimizar estratégias e apoiar a tomada de decisão (DRITSAS; TRIGKA, 2025).

Apesar da disponibilidade de dados, muitas empresas ainda enfrentam dificuldades em extrair valor estratégico dessas informações, o que pode resultar em perda de competitividade. Assim, técnicas de modelagem estatística e aprendizado de máquina são fundamentais para identificar padrões e relações entre variáveis.

A regressão linear, embora amplamente utilizada, pode apresentar limitações na captura de relações complexas. Por outro lado, algoritmos de aprendizado de máquina conseguem modelar padrões não lineares, oferecendo potencial para melhorar a precisão das previsões.

Diante disso, o objetivo deste estudo é analisar e prever o lucro em um comércio eletrônico utilizando

modelos estatísticos e de aprendizado de máquina, identificando o modelo com melhor desempenho e fornecendo suporte à tomada de decisão baseada em dados.

MATERIAL E MÉTODOS

Conjunto de dados e variáveis

O conjunto de dados utilizado neste estudo foi obtido na plataforma *Kaggle* e contém dados transacionais de uma empresa de comércio eletrônico. A variável resposta é o Profit (Lucro obtido), enquanto os preditores incluem: *Sales* (vendas totais), *Quantity* (quantidade de produtos), *Category* (categoria do produto), *Region* (região), *Month* (mês) e *Year* (ano).

Tabela 1. Descrição das variáveis.

Variável	Descrição
Profit (Y)	Lucro obtido pela empresa, variável que o modelo busca prever.
Sales	Total de vendas realizadas.
Quantity	Quantidade de produtos vendidos em cada transação.
Category	Categoria do produto vendido.
Region	Região de onde foi feita a venda.
Month	Mês em que a venda ocorreu.
Year	Ano em que a venda ocorreu.

Inicialmente, foi realizada uma análise exploratória dos dados (EAD), com o objetivo de compreender a estrutura do conjunto de dados, identificar possíveis inconsistências e avaliar a distribuição das variáveis. Foram criadas variáveis derivadas a partir da data, especificamente mês e ano, permitindo capturar possíveis padrões temporais. Não foram identificados valores ausentes ou duplicados, indicando consistência do conjunto de dados.

Pré-processamento

Na etapa de pré-processamento, as variáveis categóricas foram transformadas em representações numéricas por meio da criação de variáveis *dummy*, possibilitando sua utilização nos modelos de aprendizado de máquina. Além disso, as variáveis numéricas foram normalizadas com o objetivo de padronizar as escalas e melhorar o desempenho dos algoritmos. Todo o pipeline de preparação dos dados foi implementado na linguagem R, utilizando o framework *tidymodels*.

Divisão dos Dados e Validação

Para a construção dos modelos, o conjunto de dados foi dividido em duas partes: 80% dos dados foram utilizados para treinamento e 20% para teste. Essa divisão permite avaliar o desempenho dos modelos em dados não observados, garantindo maior capacidade de generalização.

Adicionalmente, foi empregada validação cruzada do tipo *k-fold*, com k = 10, na qual os dados de treinamento são subdivididos em 10 partes, sendo cada uma utilizada como conjunto de validação enquanto as demais são utilizadas para treinamento, reduzindo o risco de sobreajuste.

Modelos Avaliados

Foram avaliados modelo estatístico clássico, a Regressão Linear Múltipla, e modelos de aprendizado de máquina, o Random Forest, o *Extreme Gradient Boosting* (XGBoost) e o Stacked Ensemble.

A Regressão Linear Múltipla é descrita pela seguinte equação:

$$y = \beta_0 + \sum_{j=1}^{n} \beta_j x_j + \varepsilon$$

Onde y representa a variável dependente (lucro), β_0 é o intercepto do modelo, β_j são os coeficientes associados às variáveis explicativas, x_j são as variáveis independentes (como vendas, quantidade, categoria e região), e ε é o termo de erro aleatório que representa a parte não explicada pelo modelo (REDDY, 2023).

O modelo Random Forest consiste em um conjunto de árvores de decisão construídas a partir de subconjuntos aleatórios dos dados e das variáveis. A previsão final é obtida pela média das previsões individuais de cada árvore, reduzindo a variância e aumentando a robustez do modelo. Conforme equação descrita:

$$\hat{y}(x) = \frac{1}{K} \sum_{k=1}^{K} h_k(x)$$

Onde $\hat{y}(x)$ representa o valor previsto para a entrada x, K é o número total de árvores na floresta e h_k representa a previsão da k-iésima árvore de decisão.

Foi ajustado um modelo preditivo baseado no algoritmo XGBoost. pode ser representado como a soma de funções de árvores de decisão adicionadas sequencialmente:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \ , \quad f_k \in \mathcal{F} \qquad (1)$$

onde $\hat{y}_i$ é o valor predito para a i-ésima observação, x_i representa o vetor de variáveis explicativas, K é o número total de árvores no modelo e f_k. corresponde à k-ésima árvore de decisão pertencente ao espaço de funções $\mathcal{F}$.

A função objetivo otimizada pelo XGBoost é dada por:

$$\mathcal{L} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \qquad (2)$$

onde $\mathcal{L}$ representa a função objetivo, $l(y_i, \hat{y}i)$ é a função de perda que mede a discrepância entre os valores observados y_i e preditos $\hat{y}_i$, e $\Omega(f_k)$ é o termo de regularização que penaliza a complexidade do modelo, onde é dada por:

$$\Omega(\mathrm{f_k}) = \gamma \mathrm{T} + \frac{1}{2}\lambda \sum_{\mathrm{j}=1}^{\mathrm{T}} \mathrm{w_j^2} \qquad (3)$$

Onde γ é o parâmetro que penaliza o número de folhas, T representa o número de folhas da árvore, λ é o parâmetro de regularização associado aos pesos das folhas, e w_j corresponde ao peso da j-ésima folha.

O modelo Stacked Ensemble combina as previsões de diferentes modelos base, como Regressão Linear, Random Forest e Gradient Boosting, com o objetivo de explorar os pontos fortes de cada abordagem e reduzir os erros de previsão.

A otimização dos modelos Random Forest e XGBoost foi realizada por meio de busca aleatória de hiperparâmetros, com seleção baseada na minimização do erro quadrático médio (RMSE).

Métricas de Avaliação

A avaliação do desempenho dos modelos foi realizada utilizando diversas métricas estatísticas.
O erro absoluto médio (MAE): Métrica muito utilizada em modelos de regressão, que mede a média das diferenças absolutas entre e os valores previstos e valores reais, fazendo uma indicação de forma direta da precisão do modelo.

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \hat{y}_i|$$

Onde y_i é o valor observado, $\hat{y}_i$ o valor previsto e n o número total de observações.

A raiz do erro quadrático médio (RMSE): Métrica usada para avaliar a performance de modelos de regressão.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

Onde y_i é o valor observado, $\hat{y}_i$ o valor previsto e n o número total de observações.

O erro percentual absoluto médio (MAPE): Mede o erro médio absoluto como uma porcentagem dos valores reais.

$$MAPE = \frac{100}{n}\sum_{i=1}^{n} \frac{|y_i - \hat{y}_i|}{y_i}$$

Onde y_i é o valor observado, $\hat{y}_i$ o valor previsto e n o número total de observações.

O erro médio de viés (Bias) é definido como:

$$Bias = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)$$

Onde y_i é o valor observado, $\hat{y}_i$ o valor previsto e n o número total de observações. Ele indica a tendência do modelo em superestimar ou subestimar os valores observados.

O erro quadrático médio (MSE): Mede a média dos erros ao quadrado entre os valores previstos e os valores reais.

$$MSE = \frac{1}{n}\sum_{i=1}^{n} (\hat{y}_i - y_i)^2$$

Onde y_i é o valor observado, $\hat{y}_i$ o valor previsto e n o número total de observações.

O coeficiente de determinação (R^2): Mede a proporção da variabilidade dos valores observados que é explicada pelo modelo de regressão. Valores próximos de 1 indica que o modelo explica bem os dados, enquanto valores próximos de 0 indicam baixo poder explicativo.

$$R^2 = 1 - \frac{SQ_{res}}{SQ_{tot}}$$

Onde SQ_{res} é a soma dos quadrados dos resíduos e SQ_{tot} soma total dos quadrados.

Interpretabilidade dos Modelos

Por fim, para garantir a interpretabilidade dos modelos, foram aplicadas técnicas como análise de importância de variáveis, importância por permutação, valores SHAP e gráficos de dependência parcial. Essas abordagens permitem compreender a contribuição de cada variável na previsão do lucro, fornecendo maior transparência ao processo de modelagem.

Todas as análises foram realizadas utilizando o pacote de *software* R versão 4.4.1 (R CORE TEAM, 2025) com o RStudio IDE.

RESULTADOS E DISCUSSÃO

Análise Exploratória dos Dados

A análise exploratória inicial revelou padrões importantes na distribuição e relacionamento das variáveis. A Figura 1 apresenta a distribuição do lucro (*Profit*), evidenciando uma assimetria positiva, com maior concentração de valores em faixas inferiores e uma cauda longa à direita. Esse comportamento indica a presença de poucos valores elevados de lucro, o que pode influenciar métricas sensíveis a outliers, como o RMSE.

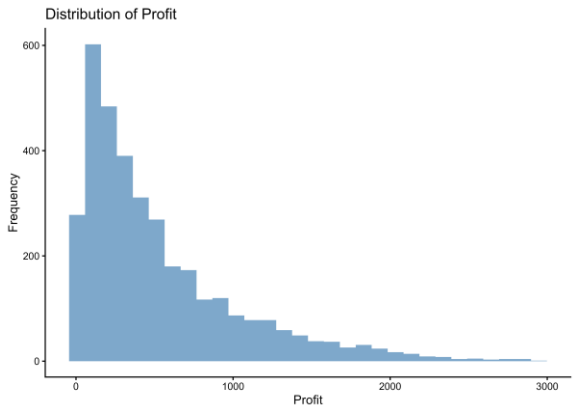

Figura 1. Distribuição do lucro (Profit).

A Figura 2 mostra a matriz de correlação entre as variáveis quantitativas. Observa-se uma correlação moderada a forte entre *Sales* e *Profit*, indicando que o volume de vendas é um dos principais fatores explicativos do lucro. A variável *Quantity* também apresenta correlação positiva, porém de menor magnitude.

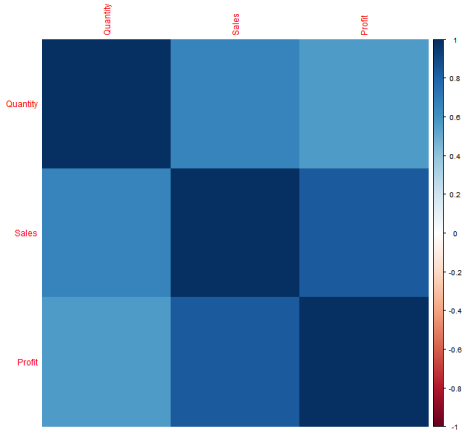

Figura 2. Matriz de correlação entre as variáveis quantitativas.

Desempenho dos Modelos

A comparação quantitativa dos modelos é apresentada na Tabela 2.

Tabela 2. Métricas de desempenho dos modelos para previsão do lucro.

Modelo	**R^2**	**RMSE**	**MSE**
LM	0,692	278,04	77306,54
RF	0,696	276,52	76466,04
XGB	0,680	283,01	80092,33
STACK	0,689	279,54	78144,17
Modelo	**MAE**	**MAPE**	**Bias**
LM	185,74	0,525	3,02
RF	186,01	0,569	5,67
XGB	189,79	0,525	0,29
STACK	189,14	0,627	2,41

R^2 = coeficiente de determinação; RMSE = raiz do erro quadrático médio; MSE = erro quadrático médio; MAE = erro absoluto médio; MAPE = erro percentual absoluto médio; Bias = erro médio de viés.

Os resultados indicam que todos os modelos apresentaram desempenho semelhante, com valores próximos de R^2, RMSE e MAE. O modelo Random Forest apresentou ligeira vantagem em termos de RMSE (276,52) e R^2 (0,696), porém a diferença em relação aos demais modelos é pequena.

A Figura 3 apresenta a comparação do RMSE entre os modelos, reforçando que não há diferenças expressivas de desempenho. Esse comportamento sugere que a estrutura dos dados é predominantemente linear, limitando o ganho obtido com modelos mais complexos.

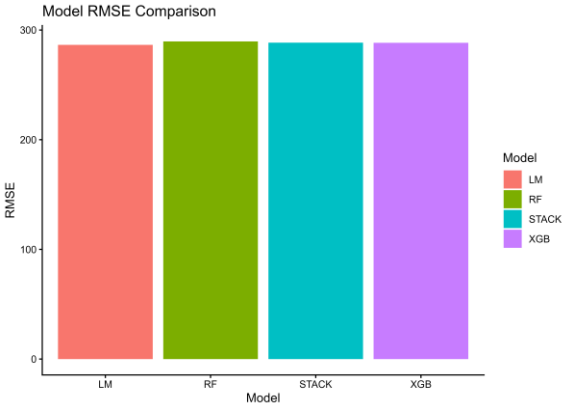

Figura 3. Comparação do erro RMSE entre os modelos de regressão.

Análise dos Valores Observados vs Preditos

A Figura 4 apresenta a relação entre valores observados e preditos para os modelos avaliados. Observa-se que os pontos estão concentrados próximos à linha 1:1, indicando boa capacidade preditiva.

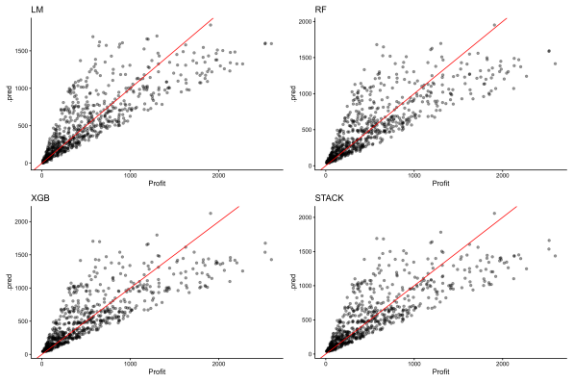

Figura 4. Valores observados versus preditos para os modelos Linear Regression (LM), Random Forest (RF), XGBoost (XGB) e Stacked Ensemble (STACK).

Entretanto, nota-se aumento da dispersão para valores mais elevados de lucro, sugerindo que os modelos apresentam maior dificuldade em capturar extremos. Esse comportamento é comum em dados com distribuição assimétrica.

Distribuição dos Erros e Calibração

A análise da distribuição dos resíduos é apresentada na Figura 5, que mostra que os erros estão centrados próximos de zero, indicando ausência de viés significativo e boa qualidade das previsões.

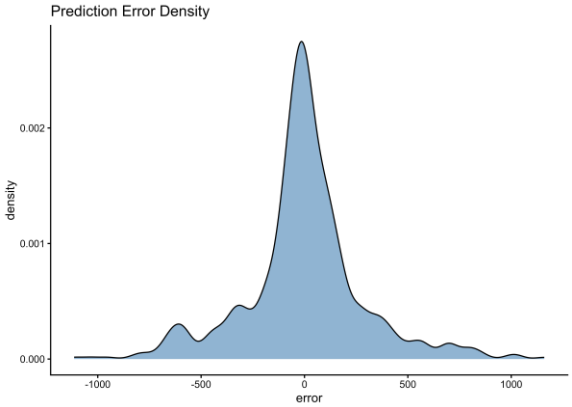

Figura 5. Densidade dos erros de previsão.

A Figura 6 apresenta o gráfico de calibração do modelo, evidenciando forte alinhamento entre valores observados e preditos. A proximidade da linha de regressão com a linha ideal confirma a consistência do modelo.

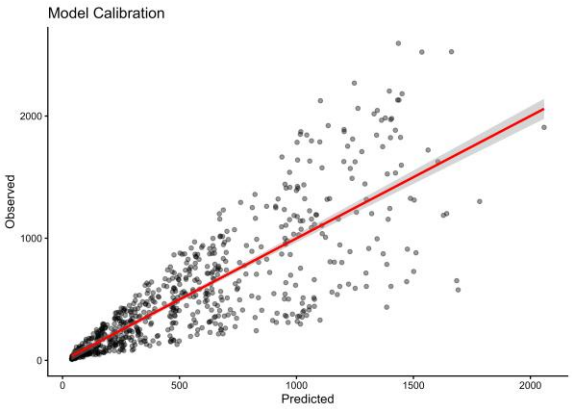

Figura 6. Calibração do modelo preditivo.

Importância das variáveis e Interpretação via SHAP

A Figura 7 apresenta a importância das variáveis obtida via SHAP. Observa-se que a variável Sales possui importância significativamente superior às demais, indicando seu papel dominante na previsão do lucro. A variável *Quantity* apresenta importância intermediária, enquanto *Month* e *Year* apresentam baixa contribuição.

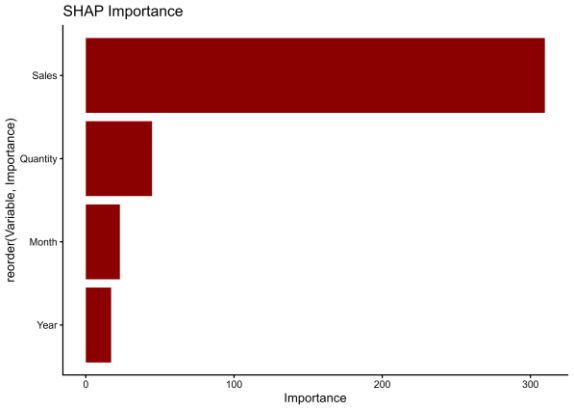

Figura 7. Importância das variáveis baseada em valores SHAP.

A Figura 8 apresenta o gráfico SHAP Beeswarm, que permite analisar o impacto das variáveis em diferentes observações. Nota-se que valores elevados de *Sales* estão associados a maiores valores de lucro, confirmando sua influência positiva.

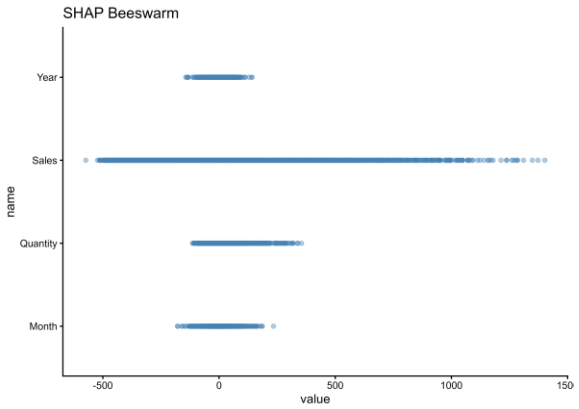

Figura 8. Gráfico SHAP Beeswarm para interpretação do modelo.

Relação Parcial entre Variáveis

A Figura 9 apresenta o gráfico de dependência parcial para a variável *Sales*, evidenciando uma relação aproximadamente linear com o lucro. Esse resultado reforça a conclusão de que a estrutura dos dados é predominantemente linear.

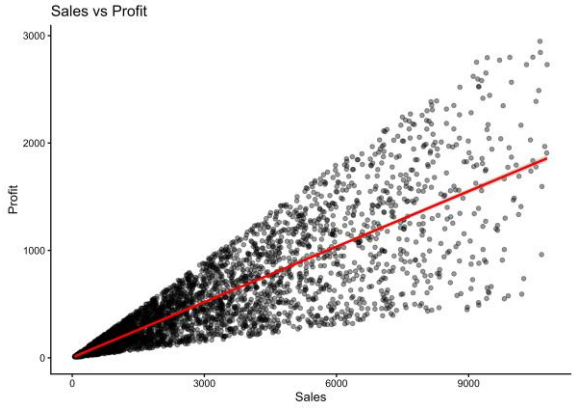

Figura 9. Gráfico de dependência parcial (*Partial Dependence Plot)* para a variável Sales.

Discussão Geral

Os resultados demonstram que todos os modelos apresentam desempenho semelhante, indicando que a relação entre as variáveis explicativas e o lucro é majoritariamente linear. Esse comportamento explica por que modelos mais complexos, como Random Forest e XGBoost, não apresentaram ganhos significativos em relação à regressão linear.

A predominância da variável Sales como principal preditor está alinhada com o contexto do comércio eletrônico, onde o volume de vendas está diretamente relacionado ao lucro. A baixa importância das variáveis temporais sugere que não há forte dependência sazonal no conjunto de dados analisado. A análise dos erros e da calibração indica que os modelos são estáveis e confiáveis, embora apresentem maior incerteza em valores extremos.

De forma geral, os resultados indicam que modelos simples e interpretáveis podem ser suficientes para este tipo de problema, reduzindo complexidade computacional sem perda significativa de desempenho.

CONCLUSÃO

Os resultados demonstram que os modelos avaliados apresentam desempenho semelhante, indicando uma relação predominantemente linear entre as variáveis explicativas e o lucro. Nesse contexto, a regressão linear mostrou-se eficiente, mesmo quando comparada a métodos mais complexos. A variável Sales destacou-se como principal fator explicativo, reforçando sua relevância estratégica.

Assim, conclui-se que modelos simples e interpretáveis podem ser preferíveis em cenários onde não há evidência significativa de não linearidade, permitindo redução de complexidade sem perda de desempenho.

AGRADECIMENTOS

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão de bolsas de pesquisa.

À Universidade Estadual da Paraíba (UEPB), pelo suporte institucional.

REFERÊNCIAS

AL RAHIB, Md Abdullah et al. Customer data prediction and analysis in e-commerce using machine learning. Bulletin of Electrical Engineering and Informatics, [S.l.], v. 13, n. 4, p. 2624–2633, ago. 2024. Disponível em: https://beei2.org/index.php/EEI/article/view/6420/3806.

DRITSAS, Elias; TRIGKA, Maria. International Journal of Information Systems and New Technologies, [s.l.], 2025. Disponível em: https://www.ijisnt.com/journal/index.php/public_html/article/view/3/13

PANGA, Naresh Kumar Reddy. Forecasting e-commerce trends utilizing linear regression, polynomial regression, random forest and gradient boosting for accurate sales and demand prediction. 2023. Disponível em: https://www.researchgate.net/profile/Naresh-Panga/publication/382571653

RESEARCH AND MARKETS. Global e-commerce market growth opportunities and forecasts 2022–2027. BusinessWire, 2022. Disponível em: https://www.businesswire.com/news/home/20220222005665/en/Global-E-Commerce-Market-Growth-Opportunities-and-Forecasts-2022-2027-A-US%24-55.6-Trillion-Market-by-2027---ResearchAndMarkets.com

SIKDER, Abu Sayed; ROLFE, Stephen. International Journal of Information Systems and New Technologies, 2023. Disponível em: https://www.ijisnt.com/journal/index.php/public_html/article/view/3/13