Análise de sentimentos em avaliações de clientes do *e-commerce* nacional e comparação de métodos tradicionais de *machine learning* com Redes Neurais *Long Short Term Memory* (LSTM)

Carlos Magno Santos Ribeiro de Brito

Universidade Federal da Bahia
Instituto de Matemática e Estatística

Introdução

- O *e-commerce* anualmente movimenta bilhões de dolares no mundo e está em franca expansão;
- Os empreendimentos estão cada vez mais adotando tecnologias diversas em suas plataforma, inclusive lançando mão da ciência de dados para isso;
- Para se manter competitiva e sólida neste mercado, uma empresa precisa de planejamento, inovação e, principalmente, entender sobre as necessidades dos clientes e como fidelizá-los;

Introdução

- Uso de métodos do *machine learning* e redes neurais surgem como opção para análise dos dados;
- A depender do contexto, faz-se necessário utilizar métodos que sejam mais eficazes e consigam ter um *tradeoff* entre tempo de processamento/acurácia satisfatório;
- Foram utilizados quatro métodos tradicionais de aprendizado de máquina (Regressão logística, Naive Bayes, XGBoost e Florestas Aleatórias) e um modelo de rede neural artificial (LSTM) para fins comparativos.

Objetivos

Objetivo principal

Realizar análise com base de dados real sobre melhores modelos para se utilizar em estudo de satisfação dos consumidores em um comércio eletrônico

Objetivo secundário

- Ponderar com a relação de eficiência *versus* custo;
- Aplicar conceitos diversos da ciência/engenharia de dados, como por exemplo a mineração de texto, pré processamento e pós processamento de dados;
- Apresentar uma visão geral sobre o comércio eletrônico (e-commerce), sua evolução e importância no mercado brasileiro.

Justificativa

Dentre as principais justificativas para o estudo dos *e-commerces*, tem-se:

- A análise de dados com essas técnicas pode ser feita de forma automatizada, reduzindo custos e tempo de processamento em comparação com análises manuais.;
- Com essas técnicas é possível prever o comportamento dos consumidores, ajudando as empresas a tomarem decisões estratégicas;
- A análise de dados com *machine learning* e redes neurais pode ser aplicada em diferentes áreas do e-commerce, como marketing digital, o seu uso pode ajudar o negócio a otimizar o seu sistema logístico, reduzindo os custos de entrega e aumentando a satisfação dos clientes.

Tipo de pesquisa e descrição dos dados

- A Olist é uma startup brasileira que atua no segmento de E-commerce por meio de marketplace.
- A empresa concentra vendedores que desejam anunciar em marketplaces como Mercado Livre, B2W, Via Varejo, Amazon, entre outros.
- A Olist concentra os produtos de todos os vendedores em uma loja única que fica visível ao consumidor final.
- Atualmente, a empresa reúne mais de 800 colaboradores e mais de 9 mil lojistas, além de 2 milhões de consumidores únicos.
- A base de dados escolhida para análise descreve a rotina de compra de um E-commerce e contém diversas informações sobre os produtos, como nome, preço, descrição, nota atribuída, comentários e local de compra.
- A análise se concentra especificamente nos comentários avaliativos (*review comment message*) e nas notas dadas pelos clientes (*review score*).

Tipo de pesquisa e descrição dos dados

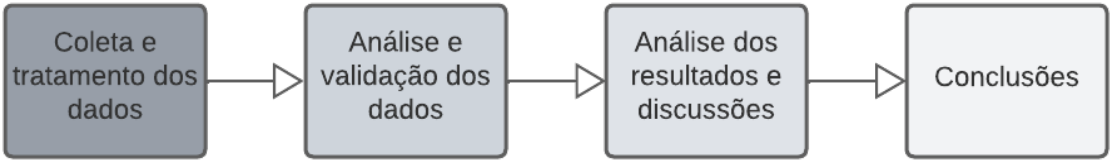


Figura 1: Fluxograma do trabalho

Exemplo da tabela usada

	review_id	order_id	review_comment_title	review_comment_message	...
type	hash	hash	string—null	string—null	
ex	da79b0a377eb	df73dbeba33	bom, mas	atende às expectativas	

	review_score	review_creation_date	review_answer_timestamp
type	number	datestring	datestring
ex	3	2018-01-18 00:00:00	2018-01-18 21:00:00

Tabela 1: Esquema da tabela *olist_order_review*

Dataset utilizado

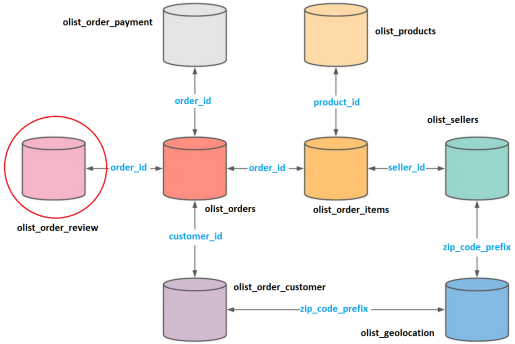


Figura 2: Esquemas do dataset publicado pelo Olist, adaptado pelo autor (2022)

Dataset transformado para uso

	index	text	label
type	bigint	string	boolean
exemplo	1	produto muito ruim	0

Tabela 2: Esquema de geração do *data_bin*

	index	text	label
type	bigint	string	number
exemplo	2	produto muito bom	4

Tabela 3: Esquema de geração do *data_gen*

- Os databases em formato CSV foram extraídos e exportados para leitura e armazenamento em diferentes variáveis.
- Itens duplicados, com valores nulos e valores discrepantes foram removidos.
- Separou-se dois *datasets* contendo informações dos comentários dos clientes (*review_comment_message*) e da sua nota de avaliação (*review_score*).
- Para o *data_bin* tem-se comentários associados categoricamente com valores binários (0 e 1) a partir das notas, com o valor atribuído de 0 para o intervalo $(0, 2]$ e de 1 para o intervalo de $(2, 5]$.
- No *data_cat* tem-se apenas os comentários e os reviews numéricos de 1 a 5.
- Ambas as bases são removidos os valores nulos.

Máquina para processamento

Todos os métodos foram executados em fila, sequencialmente entre eles, onde cada um deles foi executado em ordem, um de cada vez e de maneira procedural.
A máquina utilizada para todos eles foi o laptop Dell G3 com processador Intel Core i7 de 10ª geração, 16GB de RAM, 512GB de SSD e placa de vídeo NVIDIA RTX 2060 com 6GB de memória dedicada, incluindo o sistema operacional Windows 11 com WSL 2 instalado e Ubuntu 20.04 como ambiente de execução para as ferramentas de teste de software usadas.

Regressão Logística

- A regressão logística é um modelo estatístico robusto e eficiente, que permite a previsão da probabilidade de um evento binário de forma precisa e confiável:
  1. Este tipo de evento é aquele que pode ocorrer ou não, ou que pode ser classificado em duas categorias distintas.
  2. Por exemplo, em uma análise de crédito, o cliente pode ser aprovado ou não. Na medicina, o evento pode ser a cura ou não de uma doença.
- Ela apresenta um modelo linear generalizado que utiliza a função logística para modelar a relação entre as variáveis independentes e a variável dependente;
- A regressão logística utiliza a técnica de máxima verossimilhança para estimar os parâmetros do modelo a partir dos dados observados;
- A regressão logística tem sido amplamente utilizada em diversas áreas, como na análise de dados de sobrevivência, na análise de dados de saúde, na análise de dados financeiros, etc.

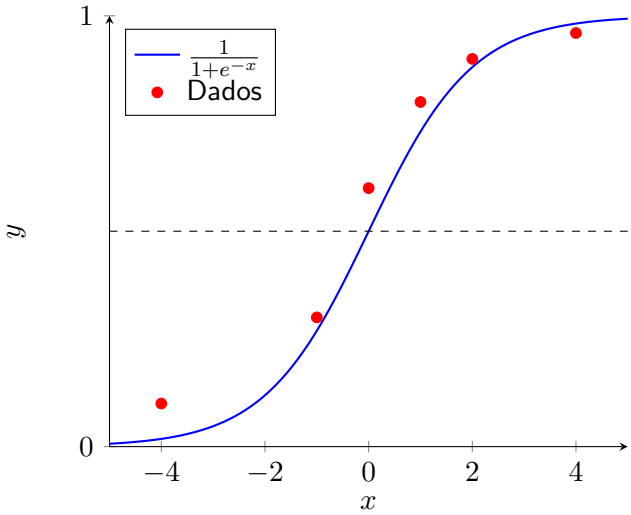


Figura 3: Modelo gráfico da regressão logística (curva sigmoide)

Naive Bayes

- A classificação Naive Bayes é um algoritmo de aprendizado de máquina supervisionado que utiliza o teorema de Bayes para classificar instâncias em classes discretas;
- A principal vantagem do algoritmo Naive Bayes é a sua simplicidade e eficiência computacional, o que o torna uma escolha popular para problemas de classificação em grande escala;
- O algoritmo Naive Bayes é baseado no teorema de Bayes, que fornece uma maneira de calcular a probabilidade condicional de uma hipótese, dado um conjunto de evidências;
- A principal suposição por trás do algoritmo Naive Bayes é a independência condicional das características;
- O algoritmo Naive Bayes tem sido aplicado em muitas áreas, incluindo reconhecimento de fala, processamento de texto e detecção de spam de e-mail;
- Embora o algoritmo Naive Bayes seja uma técnica de classificação simples e eficiente, ele também tem algumas limitações.

Florestas Aleatórias

- As florestas aleatórias são uma técnica de aprendizado de máquina que combina várias árvores de decisão para construir um modelo de classificação ou regressão.
- Cada árvore de decisão é construída a partir de um subconjunto aleatório dos dados de treinamento e um subconjunto aleatório dos recursos (também conhecidos como características ou atributos).
- A construção de uma árvore de decisão é feita por meio de uma série de etapas, onde a árvore começa com um único nó que representa todo o conjunto de dados de treinamento e é dividida em nós menores usando uma função de divisão que escolhe um recurso e um ponto de divisão que minimiza a impureza dos dados.
- Durante a fase de teste, a floresta aleatória retorna a classe mais comum ou a média das saídas das árvores individuais, dependendo se o problema é de classificação ou regressão, respectivamente.
- As florestas aleatórias apresentam várias vantagens em relação a outras técnicas de aprendizado de máquina, como bom desempenho em dados de alta dimensão, insensibilidade a outliers e dados ausentes e facilidade de paralelização.

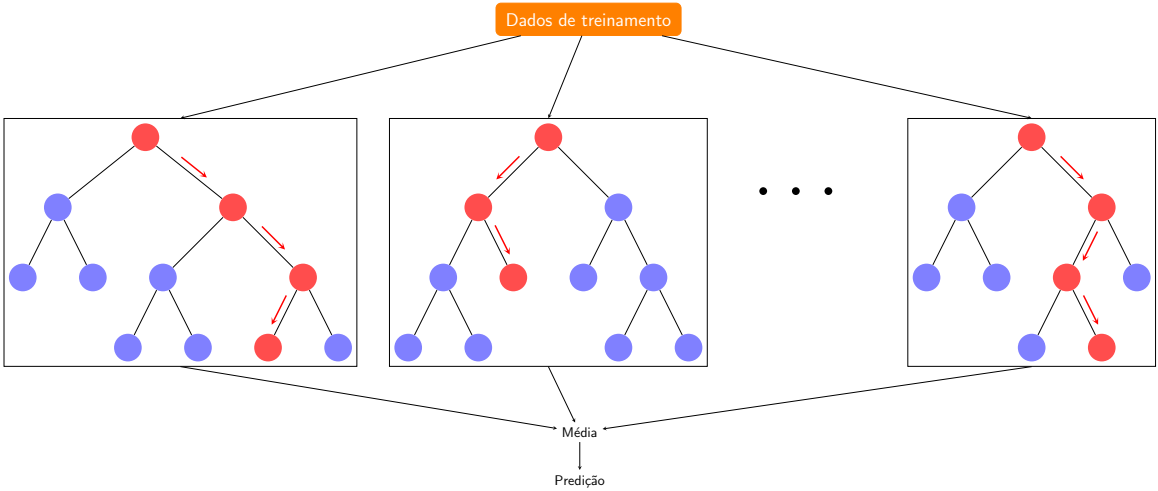


Figura 4: Esquema de funcionamento de uma floresta aleatória

XGBoost

- O XGBoost (*Extreme Gradient Boosting*) é um método de aprendizado de máquina baseado em árvores de decisão, assim como as florestas aleatórias, mas com algumas diferenças importantes.
- Enquanto as florestas aleatórias usam um conjunto de árvores de decisão independentes para fazer uma previsão, o XGBoost usa um conjunto de árvores de decisão sequenciais que são criadas iterativamente. Cada nova árvore é ajustada aos resíduos do modelo anterior, tentando corrigir os erros cometidos pelo modelo atual.
- O algoritmo XGBoost foi desenvolvido por Tianqi Chen e Carlos Guestrin em 2016 e é baseado na biblioteca de código aberto de mesmo nome. O XGBoost se tornou um dos algoritmos de aprendizado de máquina mais populares em competições de ciência de dados e é amplamente utilizado na indústria.
- Para construir o modelo XGBoost, o algoritmo usa um processo iterativo de adição de árvores, onde cada nova árvore é ajustada aos resíduos do modelo anterior, tentando corrigir os erros cometidos pelo modelo atual.

Redes Neurais LSTM

- Uma rede neural artificial *Long Short-Term Memory* (LSTM) é um tipo especializado de rede neural recorrente (RNN) projetada para lidar com o problema de dependência a longo prazo em sequências de dados.
- A principal característica da LSTM é a presença de unidades de memória, que são capazes de lembrar informações relevantes por um período prolongado de tempo.
- Essas unidades são compostas por três portões principais: o portão de entrada, o portão de esquecimento e o portão de saída.

Redes Neurais LSTM

- Durante o treinamento, a LSTM é capaz de aprender a modificar seus portões de entrada, esquecimento e saída, a fim de manter as informações relevantes na memória e descartar as informações irrelevantes.
- Os pesos das conexões entre as unidades de memória são atualizados usando o algoritmo de retropropagação através do tempo (BPTT).
- A arquitetura LSTM foi amplamente utilizada em diversas aplicações, incluindo processamento de linguagem natural, reconhecimento de fala e análise de séries temporais.
- Além disso, posteriormente foi proposta uma abordagem de aprendizado contínuo com LSTMs, chamada de "learning to forget", que permite que a rede esqueça informações irrelevantes à medida que recebe novos dados.

Distribuição

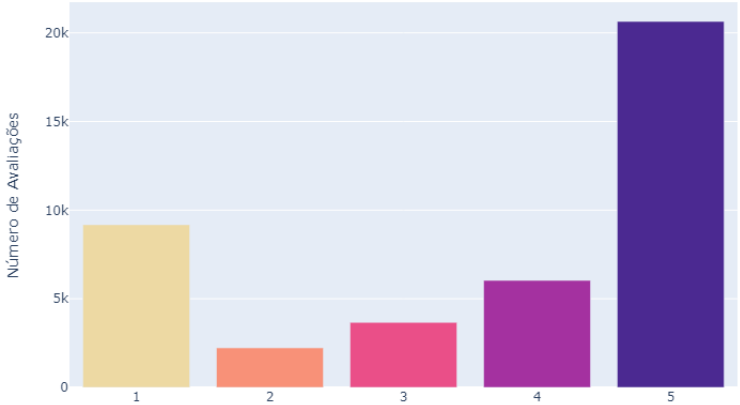


Figura 5: Distribuição das avaliações

Distribuição

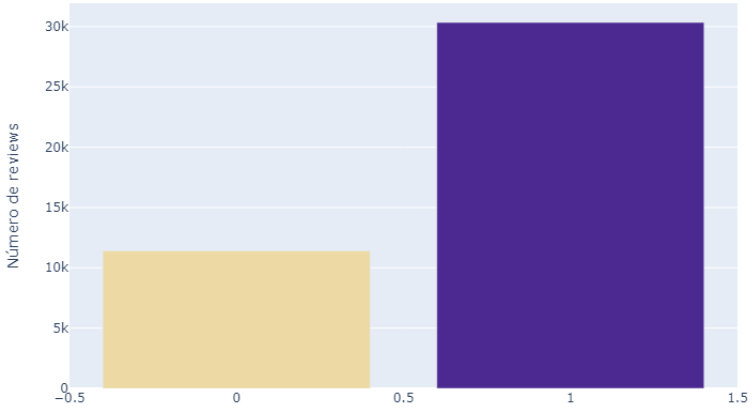


Figura 6: Distribuição das avaliações binárias

Word Cloud

Figura 7: Nuvem de palavras destacando os principais termos utilizados

Acurácias

modelo	Reg. Logística	F. Aleatórias	XGBoost	Naive Bayes
treino (%)	73.9	99.6	93.5	74.0
teste (%)	73.3	78.2	82.1	74.0

Acurácias

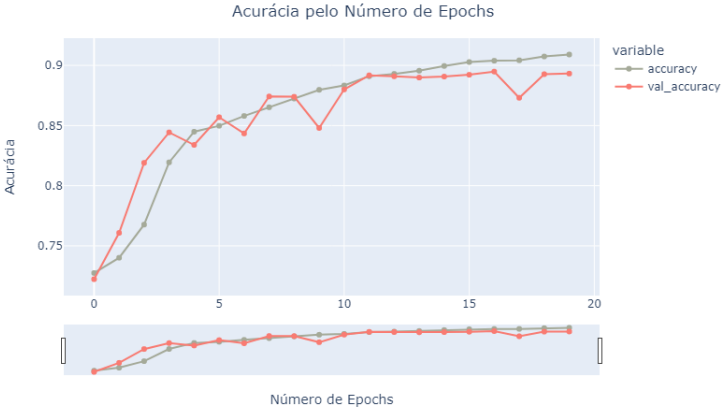


Figura 8: Relação acurácia por *epochs*

Acurácias

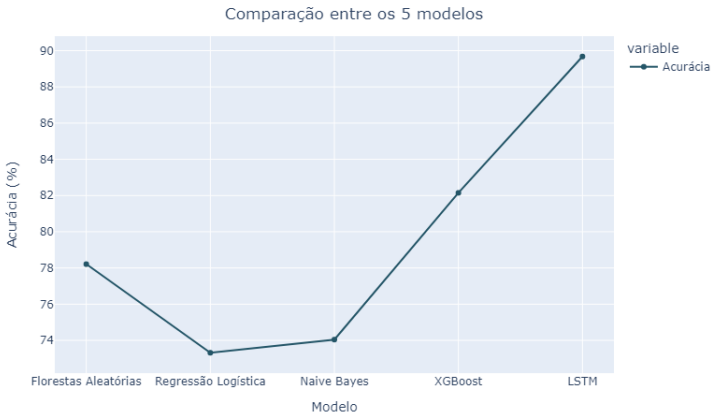


Figura 9: Comparação de acurácia entre modelos

Tradeoff

modelo	Reg. Logística	F. Aleatórias	XGBoost	Naive Bayes	LSTM
Tempo (s)	21.5	1.5	3.0	0.1	1500
Acurácia (%)	73.3	78.2	82.1	74.0	90.0

Tabela 4: Tempo de execução/Acurácia dos modelos avaliados

Matriz de confusão

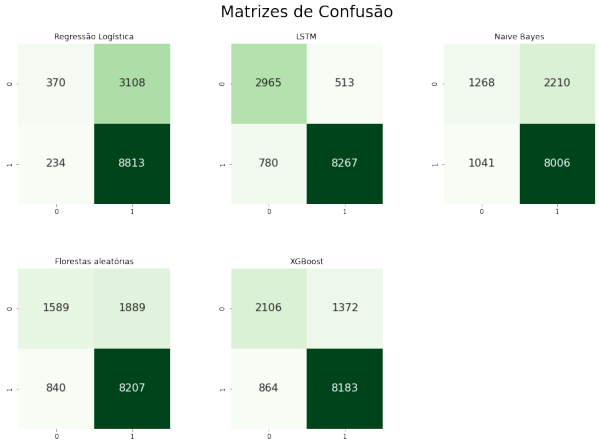
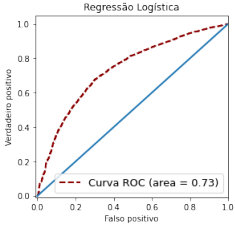
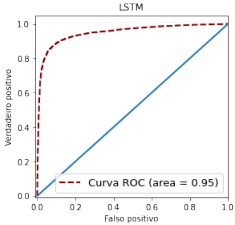
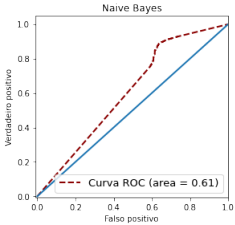


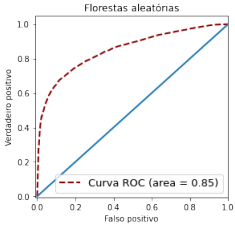
Figura 10: Curvas ROC dos modelos utilizados

Curva HOC









XGBoost

Conclusões

- A Rede neural LSTM apresentou a melhor performance entre os algoritmos de machine learning avaliados, com acurácia de 90% e maior capacidade de distinguir entre as classes, indicado pela área sob a curva ROC.
- XGBoost também apresentou resultados promissores, com acurácia de 82% e área sob a curva ROC de 0,89, mas cometeu mais erros na classificação de algumas amostras do que a LSTM.
- Regressão logística e Naive Bayes tiveram as piores performances, com acurácias de 73% e 61%, respectivamente, e áreas sob a curva ROC menores do que os outros modelos avaliados.

Conclusões

- É necessário encontrar um equilíbrio entre a rapidez da resposta e a precisão do modelo em muitas aplicações em tempo real.
- A escolha do modelo ideal depende de vários fatores, como o tamanho dos dados, a complexidade do problema e a disponibilidade de recursos de computação.
- Modelos mais simples, como regressão logística ou Naive Bayes, têm tempos de processamento menores, mas podem ter uma acurácia menor, enquanto modelos mais complexos, como redes neurais, podem ter uma acurácia muito alta, mas exigem uma grande quantidade de tempo de processamento. As Florestas Aleatórias e o XGBoost são modelos intermediários que podem ser mais adequados para muitas aplicações.

Conclusões

- O XGBoost possui vantagens em relação a redes neurais, como a capacidade de lidar com dados heterogêneos e faltantes de forma eficiente, e a simplicidade e rapidez no treinamento.
- O XGBoost também é interpretável, permitindo a identificação das variáveis mais importantes para a classificação dos dados.
- A LSTM, por sua vez, é capaz de lidar com dados sequenciais e com dependências de longo prazo, o que pode ser um desafio para algoritmos tradicionais de aprendizado de máquina.

Conclusões

- A LSTM também é capaz de aprender padrões complexos em dados sequenciais sem a necessidade de engenharia manual de características, tornando-se uma escolha popular em tarefas de processamento de linguagem natural e análise de séries temporais.
- A LSTM é capaz de lidar com dados de entrada de diferentes tipos e tamanhos, como sequências de palavras, imagens e dados numéricos.
- Para análise de sentimentos em reviews de usuários, a LSTM pode ser mais vantajosa do que o XGBoost devido à sua capacidade de capturar dependências de longo prazo nos dados sequenciais e trabalhar eficientemente com dados sequenciais.