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;
- Uso de métodos do *machine learning* e redes neurais surgem como opção para análise dos dados;

Introdução

- Usuários plenamente satisfeitos e/ou insatisfeitos tendem a avaliar corretamente o produto, porém, fora dessa faixa é encontrada muita inconsistência;

- 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 cinco métodos tradicionais de aprendizado de máquina (Regressão logística, Naive Bayes, XGBoost, LightGBM 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;
- Obter melhor precisão na classificação dos reviews com base nos comentários;
- 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.

Tipo de pesquisa e descrição dos dados

- A base utilizada é uma base pública disponível no Kaggle e publicada pela Olist;
- A análise se concentra especificamente nos comentários avaliativos e nas notas dadas pelos clientes;
- A tabela com possui 100 mil linhas depois de tratada com as informações pertinentes ao comentário avaliativo e à nota referente;
- Os atributos originais utilizados são *review_score* e *review_comment_message* da tabela *olist_order_review*

Exemplo

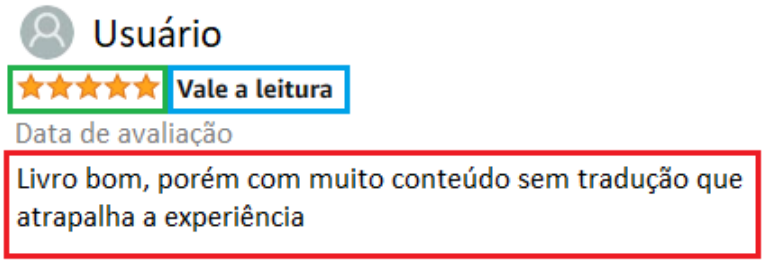


Figura 1: Exemplo de como é efetuada a avaliação

Tipo de pesquisa e descrição dos dados

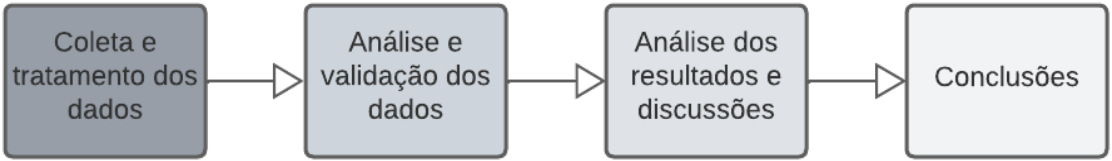


Figura 2: Fluxograma do trabalho

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

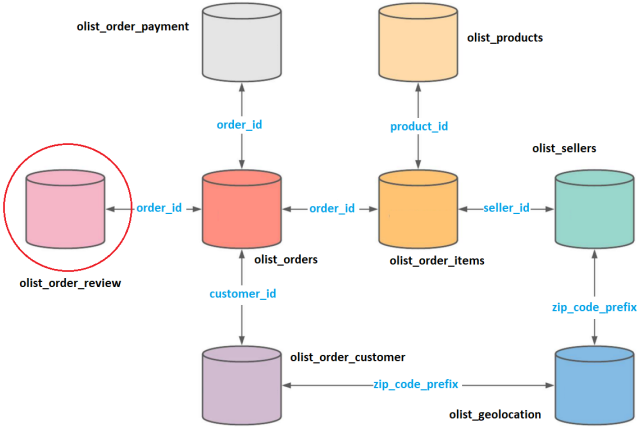


Figura 3: Esquemas do dataset publicado pelo Olist, adaptado pelo autor (2022)

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.
- Os dados de entrada (comentários) foram transformados em uma lista de números por meio de um processo de tokenização;
- 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.

Comparação dos métodos utilizados

Com base nos resultados e literatura, a tabela a seguir foi elaborada com as principais vantagens/características de cada método utilizado:

Método	Tempo de treinamento	Acurácia	Desempenho preditivo	Tipo de modelo	Uso
Regressão logística	Médio	Média	Baixo	Linear	Classificação binária ou multiclasse em dados com poucas variáveis.
Naive Bayes classifier	Muito Baixo	Média	Baixo	Probabilístico	Classificação binária ou multiclasse em dados com poucas variáveis e baixa correlação entre as variáveis.
Florestas aleatórias	Médio	Média	Médio	Árvore de decisão	Classificação binária ou multiclasse e regressão em dados com muitas variáveis e não-lineares.

Comparação dos métodos utilizados

Com base nos resultados e literatura, a tabela a seguir foi elaborada com as principais vantagens/características de cada método utilizado:

Método	Tempo de treinamento	Acurácia	Desempenho preditivo	Tipo de modelo	Uso
XGBoost	Baixo	Alta	Alto	Árvore de decisão	Classificação binária ou multiclasse e regressão em dados com muitas variáveis e não-lineares.
LightGBM	Muito baixo	Alta	alto	Árvore de decisão	Classificação binária ou multiclasse e regressão em dados com muitas variáveis e não-lineares.
LSTM	Muito alto	Muito alta	Muito alto	Rede neural	Modelagem de sequências temporais em dados com dependências de longo prazo.

Distribuição

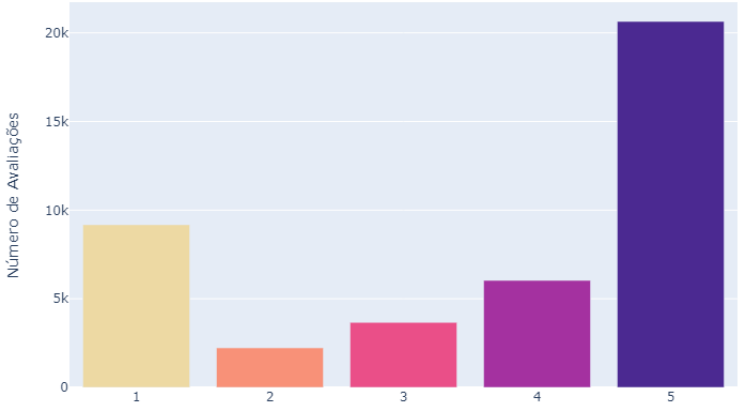


Figura 4: Distribuição das avaliações

- Nota 1 corresponde às piores avaliações, nota 5 às melhores;
- A distribuição das avaliações apresenta uma forma de "J", com grande quantidade de notas 5, 4 e 1, e pequena quantidade de notas 2 e 3;
- Isso pode ocorrer porque clientes extremamente satisfeitos ou insatisfeitos são mais propensos a deixar avaliações;
- Além disso, clientes que tiveram uma experiência neutra podem não se sentir motivados a deixar uma avaliação, o que resulta em uma concentração menor de avaliações com notas médias;
- A distribuição em forma de "J" pode fornecer informações importantes sobre a satisfação do cliente e a qualidade do produto.

Distribuição

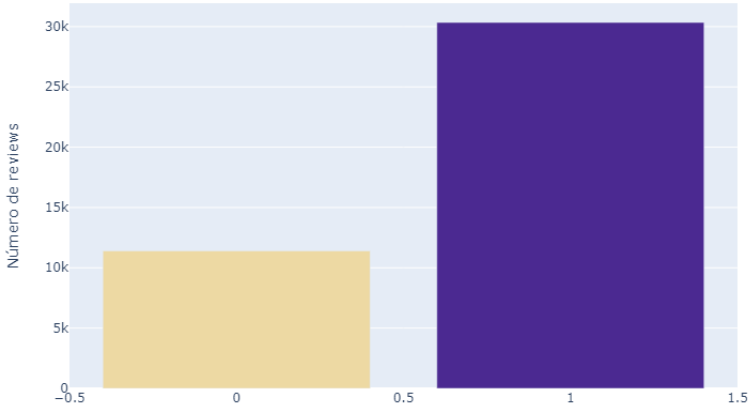


Figura 5: Distribuição das avaliações binárias

Word Cloud

Figura 6: Nuvem de palavras destacando os principais termos utilizados

- Nuvem de palavras é feita com as entradas de dados, que são os principais termos usados nos comentários, excluindo as *stopwords*;

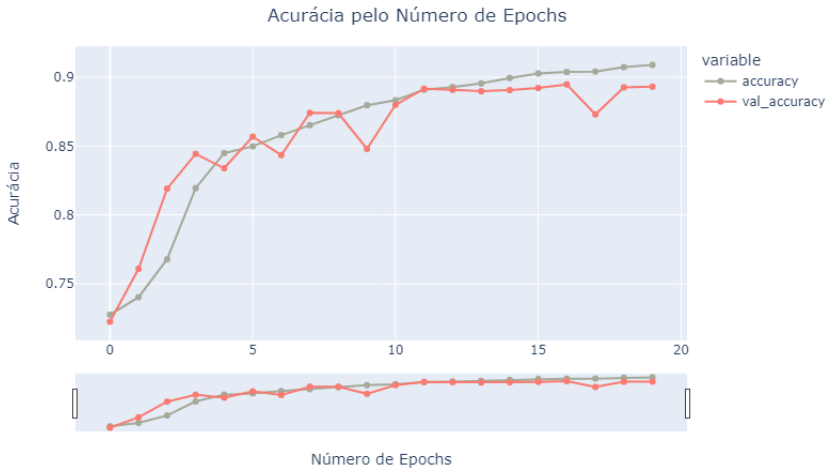
- Essa técnica visual pode ser usada em vários campos, incluindo análise de sentimentos, mineração de opiniões, análise de redes sociais, pesquisa de mercado e análise de feedback de clientes como o caso em específico

- Além disso, por ser uma forma visualmente atraente de resumir informações e destacar pontos importantes, é bastante útil para relatórios e análises.

Acurácias dos modelos de ML

modelo	Reg. Logística	F. Aleatórias	XGBoost	Naive Bayes	LightGBM
treino (%)	73.9	99.6	93.5	74.0	86.7
teste (%)	73.3	78.2	82.1	74.0	81.3

Acurácia da rede neural LSTM vs Epochs



Acurácias comparadas

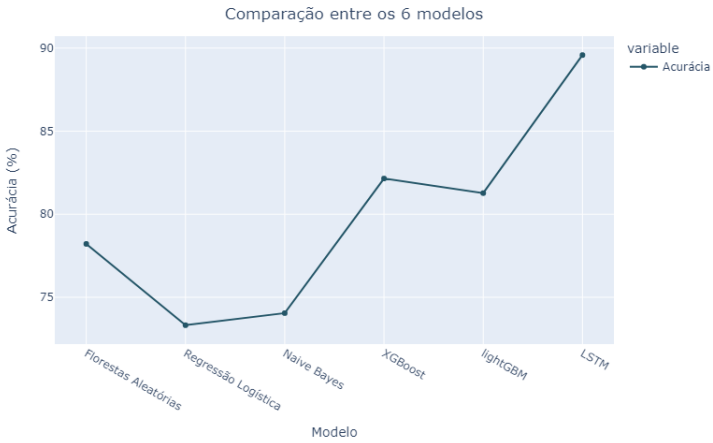


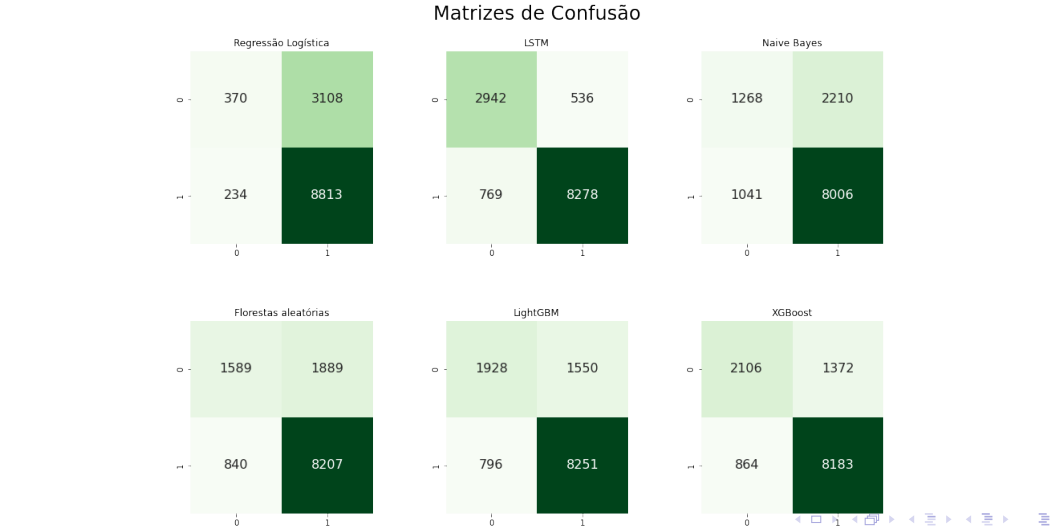
Figura 8: Comparação de acurácia entre modelos

Tradeoff

modelo	Reg. Logística	F. Aleatórias	XGBoost	Naive Bayes	LightGBM	LSTM
Tempo (s)	21.5	1.5	3.0	0.1	1.5	1500
Acurácia (%)	73.3	78.2	82.1	74.0	81.3	90.0

Tabela 4: Tempo de execução/Acurácia dos modelos avaliados em teste

Matriz de confusão



Curva HOC

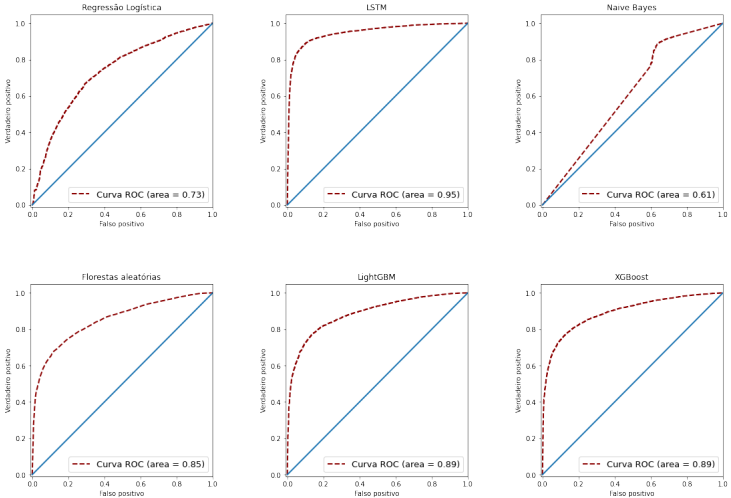


Figura 10: Curvas ROC dos modelos utilizados

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 e LightGBM também apresentaram resultados promissores, com acurácia de $\approx 82\%$ 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 74%, 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, LightGBM 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.
- O LightGBM é mais rápido que o XGBoost, usando uma técnica de otimização chamada "histogram-based"para encontrar os melhores pontos de divisão de árvore, permitindo o processamento de conjuntos de dados maiores e mais complexos em menos tempo.

Conclusões

- O LightGBM usa menos memória do que o XGBoost, utilizando uma técnica de amostragem chamada "leaf-wise"em vez de "level-wise", permitindo criar árvores mais profundas com menos nós.
- O LightGBM tem melhor desempenho em conjuntos de dados esparsos, sendo melhor em lidar com eles devido à técnica de exclusão de zeros para reduzir a sobrecarga computacional em recursos esparsos.
- 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.