Universidade Federal do Ceará - Campus Crateús
Disciplina: Ciência dos Dados
Alunos:
Francisca Maria Rodrigues Andrade - 564538
Gabriel Silva Nascimento - 571484
João Paulo Sousa Menezes - 557497
Maria Fernanda Vasconcelos Nunes - 554599

Aprendizado de Máquina e Processamento de Linguagem Natural

Para ter acesso ao nosso repositório do GitHub, clique aqui

1 Introdução

1.1 Contextualização

A competitividade no setor de e-commerce exige que empresas transformem grandes volumes de dados operacionais e textuais em decisões estratégicas para garantir a retenção de clientes. O problema central reside na complexidade de processar informações multidimensionais de forma escalável, o que demanda a integração de diferentes aplicações do **Aprendizado de Máquina**. A partir disso, nosso objetivo é automatizar a compreensão dos fatores que impactam a satisfação do consumidor e a eficiência operacional em mercados digitais.

1.2 Descrição do DataSet

Utilizamos o conjunto de dados real do ecossistema de e-commerce brasileiro da Olist, a maior loja de departamentos dentro dos marketplaces no Brasil. A Olist atua como um hub, conectando pequenos lojistas de todo o país a grandes canais de venda.

Para o nosso trabalho, partes específicas do conjunto de dados foram selecionadas:

Tabela 1: Descrição dos subconjuntos de dados do ecossistema Olist

Arquivo (.csv)	**Descrição Técnica**
`olist_orders`	Armazena o ciclo de vida do pedido, desde a compra até a entrega final.
`olist_order_items`	Contém dados sobre os itens de cada pedido, incluindo preços e valores de frete.
`olist_order_reviews`	Reúne as avaliações dos clientes, com notas numéricas e comentários textuais.
`olist_products`	Metadados sobre os produtos, como categoria, peso e dimensões físicas.
`olist_customers`	Informações sobre os compradores, focando na localização geográfica (cidade/estado).
`olist_sellers`	Dados cadastrais dos vendedores que utilizam a plataforma para comercializar itens.
`olist_geolocation`	Mapeamento de códigos postais para coordenadas geográficas, útil para análises de frete.
`olist_order_payments`	Registra os detalhes financeiros, como método de pagamento e número de parcelas.

2 Aprendizado de Máquina

A primeira parte do trabalho, dividida em duas abordagens, visa utilizar as informações tabulares para prever dois importantes pontos: Satisfação do cliente e Preço de frete.

i. Classificação em Dados Tabulares:

Objetivo: Prever a satisfação dos clientes com base em características das vendas.

2.1 Técnicas Utilizadas

- Imputação de valores ausentes: Utilizando mediana para maior robustez contra outliers.
- Escalonamento: Usando StandardScaler para normalização de grandezas.
- Aplicação de SMOTE: Balanceamento gerando dados sintéticos da classe minoritária.
- GridSearchCV: Usado com validação cruzada (5-fold).
- Treinamento de 5 modelos: KNN, Regressão Logística, Árvore de Decisão, Random Forest e MLP.

2.2 Resultados e Conclusões

Modelo	Acurácia	F1-Score	Recall	Precision
Regressão Logística	**0.7853**	**0.8716**	0.9473	0.8070
Random Forest	0.7708	0.8525	0.8612	0.8440
MLP (Rede Neural)	0.7592	0.8502	0.8889	0.8148
KNN	0.7307	0.8204	0.7999	0.8420
Árvore de Decisão	0.7078	0.8019	0.7689	0.8379

Figura 1: A Regressão Logística, com um F1-Score de 0.8716, foi dada como melhor modelo. Embora seja o modelo mais simples entre os testados, apresentou o melhor equilíbrio entre precisão e sensibilidade (Recall) para este conjunto de dados. O Recall da Regressão Logística (0.9473) indica que o modelo é extremamente eficiente em identificar clientes satisfeitos, embora tenha uma taxa de erro moderada em prever insatisfações (Falsos Positivos).

ii. Problema de Regressão:

Objetivo: Utilizar fatores logisticos e geográficos para prever a taxa de frete.

2.3 Técnicas Utilizadas

- Engenharia de feature para criar novos atributos, como peso cubado e distância.
- Pré-processamento com StandardScaler para padronizar números e OneHotEncoder para transformar categorias.
- Comparação entre três modelos: Regressão Linear, Random Forest e LightGBM usando pipelines e validação cruzada.

2.4 Resultados e Conclusões

Melhor modelo no treino: LightGBM com RSME médio de 7.521810. Sua performance no conjunto de teste:

Tabela 2: Métricas de Desempenho do Modelo de Regressão

Tabela 3: O modelo acerta bastante nos fretes de valor mais baixo (que são a maioria), mas perde precisão e começa a errar mais conforme o valor do frete aumenta.

Métrica	**Valor**
R^2 (Coeficiente de Determinação)	76,90%
MAE (Erro Absoluto Médio)	R$ 3,41
RMSE (Raiz do Erro Quadrático Médio)	R$ 7,49

3 Processamento de Linguagem Natural

Para a segunda parte do trabalho, nosso objetivo é utilizar os comentários de texto e classificá-los a fim de entender o nível de satisfação do cliente em insatisfeito (comentários de texto com 1 a 3 estrelas) ou satisfeitos (de 4 a 5 estrelas).

3.1 Técnicas Utilizadas

- Stemming: Uso do algoritmo RSLP (Removedor de Sufixos da Língua Portuguesa) para reduzir palavras aos seus radicais.
- Utilização de 5 modelos: Naive Bayes, Logistic Regression, Linear SVC (Support Vector Machine), Random Forest, Stochastic Gradient Descent (SGD).

3.2 Resultados e Conclusões

Figura 2: A nuvem de palavras destaca que radicais relacionados ao recebimento são muito frequentes em ambos os casos, o que pode significar que a qualidade da entrega é de grande influência.

4 Limitações Gerais

- Desbalanceamento de dados: Por se tratar de um DataSet real, existem informações de muitas pessoas e lojas diferentes, o que contribui para o desbalanceamento.
- O cálculo da distância usando Haversine considera apenas uma linha reta, o que é uma aproximação e gera imprecisões em rotas reais mais complexas.
- O modelo foi muito punido nas previsões para os estados do Norte. A principal suposição é que nessas rotas é necessário transporte fluvial e aéreo.
- Binarização Simplista: Ao transformar notas 1, 2 e 3 em "Insatisfeito", o modelo trata reclamações leves da mesma forma que críticas severas, perdendo a gradação do sentimento.
- Stemming e Semântica: O uso do radical (stemming) simplifica o vocabulário, mas pode perder nuances gramaticais importantes para o sentimento.