Relatório Técnico: Implementação e Análise do Algoritmo k-Nearest Neighbors (kNN) Aplicado ao Instagram

Grupo 99:
Gabriel Marçal de Oliveira
Leon Santana Barbosa

Data de entrega: 17/11/24

Resumo

Síntese do Projeto

Objetivo

Desenvolver um modelo preditivo utilizando o algoritmo k-Nearest Neighbors (kNN) para analisar e prever o impacto de influenciadores do Instagram com base em métricas como `influence_score`, `followers`, `avg_likes`, e outros atributos.

Metodologia

1. Coleta e Pré-processamento de Dados:
   - Conversão de valores textuais em numéricos (sufixos `k`, `m`, `%`).
   - Atribuição de códigos numéricos para países, organizados em faixas representando continentes.
   - Remoção de valores ausentes e normalização de variáveis importantes.

2. Análise Exploratória:
   - Visualização de correlações entre variáveis.
   - Análises gráficas (distribuições, scatterplots, boxplots).

3. Implementação do Modelo:
   - Construção de um modelo kNN básico.
   - Otimização de hiperparâmetros com **GridSearchCV** (melhores valores de `k` e métricas de distância).

4. Avaliação do Modelo:
   - Métricas de desempenho calculadas: MAE, MSE e RMSE.
   - Comparação gráfica entre valores reais e preditos.

Principais Resultados

1. Correlação:
   - Identificadas relações significativas entre `followers`, `avg_likes`, e `60_day_eng_rate`.

2. Desempenho do Modelo:
  - Com Hiperparâmetros Otimizados:
    - MAE: 0.0044
    - MSE: 6.52e-05
    - RMSE: 0.0081

3. Visualizações Relevantes:
  - Gráficos de dispersão e boxplots destacaram insights sobre engajamento e influência por continentes e outros atributos.

‘

Introdução

Contextualização do Problema

Com o crescimento das redes sociais, especialmente o Instagram, os influenciadores digitais desempenham um papel crucial no marketing e na comunicação global. Analisar e prever métricas de desempenho desses influenciadores, como engajamento e influência, é essencial para empresas e marcas que buscam parcerias estratégicas. Contudo, lidar com grandes volumes de dados e identificar padrões significativos pode ser desafiador.

O algoritmo k-Nearest Neighbors (kNN), conhecido por sua simplicidade e eficiência em prever valores com base em proximidades, é uma escolha apropriada para este problema. Ele permite explorar relações entre variáveis como número de seguidores, curtidas médias e taxa de engajamento, ajudando a identificar influenciadores com maior potencial de impacto.

Descrição do Conjunto de Dados

O conjunto de dados utilizado neste projeto contém informações detalhadas sobre influenciadores do Instagram. Os principais atributos incluem:

- rank: Posição do influenciador em um ranking.
- channel_info: Informações sobre o canal ou perfil do influenciador.
- influence_score: Pontuação que mede o impacto do influenciador.
- followers: Número de seguidores do perfil.
- avg_likes: Média de curtidas em suas postagens.
- 60_day_eng_rate: Taxa de engajamento nos últimos 60 dias.
- new_post_avg_like: Média de curtidas em postagens recentes.

- total_likes: Total de curtidas recebidas no perfil.
- country: País de origem do influenciador.

Transformações nos Dados

- Conversão de valores: Colunas contendo valores com sufixos (`k`, `m`, `b`, `%`) foram transformadas em números reais.

- Codificação por continente: Países foram classificados em faixas numéricas que representam continentes, permitindo análises regionais e facilitando o uso em modelos numéricos.

Metodologia

Análise Exploratória

A análise inicial identificamos variáveis-chave, como `followers`, `avg_likes` e `60_day_eng_rate`, que possuem forte correlação com o `influence_score`. Gráficos de dispersão e boxplots revelaram padrões importantes, incluindo a relação entre número de seguidores e engajamento médio, além de diferenças regionais no impacto de influenciadores.

Implementação do Algoritmo kNN

O algoritmo kNN foi configurado utilizando diferentes valores de `k` e métricas de distância (Euclidiana e Manhattan). A variável `country` foi transformada em faixas numéricas representando continentes, otimizando sua utilização no modelo preditivo.

Validação e Ajuste de Hiperparâmetros

A validação cruzada foi realizada com o GridSearchCV para testar combinações de parâmetros e encontrar a configuração ideal. Os melhores resultados foram obtidos com valores de `k=5` e métrica Euclidiana, melhorando a precisão do modelo.

Resultados

Métricas de Avaliação

As métricas de desempenho para o modelo kNN com os melhores parâmetros (k=5 e métrica Euclidiana) foram:

- **MAE:** 0.0044
- **MSE:** 6.52e-05
- **RMSE:** 0.0081

Esses resultados indicam uma boa precisão na previsão do **influence_score**, com erros absolutos e quadráticos pequenos.

Visualizações

- **Gráficos de dispersão** mostraram uma relação positiva entre `followers` e `avg_likes`, evidenciando o impacto do engajamento.
- **Boxplots** revelaram diferenças de engajamento entre continentes.
- **Gráfico de barras** comparando o `rank` com o `influence_score` permitiu visualizar como a posição no ranking se correlaciona com a influência.

Essas visualizações confirmam a eficácia do modelo e facilitam a compreensão das variáveis que afetam o desempenho dos influenciadores.

Discussão

Os resultados mostram que o modelo kNN com os melhores parâmetros teve bom desempenho, mas apresentou limitações, como valores faltantes e variáveis não exploradas. A escolha de k=5 e a normalização melhoraram o desempenho, mas o modelo pode ser sensível a diferentes hiperparâmetros e à distribuição dos dados

Conclusão e Trabalhos Futuros

O projeto mostrou a eficácia do kNN na previsão do influence_score, com bons resultados. Para melhorias futuras, seria interessante explorar modelos mais complexos, como regressão linear ou redes neurais, e incluir mais variáveis, como a categoria de conteúdo dos influenciadores. O tratamento de valores ausentes e a otimização do pré-processamento também são pontos a serem aprimorados.

Referências

https://neo4j.com/docs/graph-data-science/current/algorithms/knn/?utm_source=GSearch&utm_medium=PaidSearch&utm_campaign=Evergreen&utm_content=AMS-Search-SEMCE-DSA-None-SEM-SEM-NonABM&utm_term=&utm_adgroup=DSA&gad_source=1&gclid=CjwKCAiAxea5BhBeEiwAh4t5K3Q3H3dWDYd_a-yWxJ-RDi9ISsiv3SNHPYBT5fqBGhcFMB4zIUt0TxoC-NcQAvD_BwE

https://medium.com/brasil-ai/knn-k-nearest-neighbors-1-e140c82e9c4e