Jeferson Braga Luz
Rafael Rodrigues Souza

Relatório Técnico

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

Vitória da Conquista - BA
17/11/2024

Resumo

Este relatório detalha a implementação e análise do algoritmo k-Nearest Neighbors (kNN) aplicado ao conjunto de dados de influenciadores do Instagram. O objetivo foi explorar as variáveis relevantes para entender a relação entre fatores como seguidores, engajamento e influência. O processo incluiu a análise exploratória dos dados, a implementação do algoritmo kNN, a otimização de hiperparâmetros com GridSearchCV, e a validação cruzada para avaliar a consistência do modelo. Como resultados, foram calculadas métricas de desempenho, como MAE, MSE e RMSE, e gráficos foram gerados para ilustrar as principais descobertas.

Introdução

O crescimento do Instagram como uma plataforma de marketing social gerou uma demanda significativa por modelos que possam prever o impacto de influenciadores em suas audiências. O k-Nearest Neighbors (kNN) é um algoritmo simples, mas eficaz, para prever a influência de um influenciador com base em suas características. Neste projeto, utilizamos um conjunto de dados de influenciadores do Instagram, com variáveis como número de seguidores, engajamento e outros parâmetros, para aplicar o kNN e analisar as relações entre essas variáveis.

Descrição do Conjunto de Dados

O conjunto de dados contém as seguintes variáveis:

- **rank**: Ranking baseado no número de seguidores.
- **channel_info**: Nome de usuário do Instagram.
- **influence_score**: Pontuação de influência, considerando menções e popularidade.
- **posts**: Número de postagens.
- **followers**: Número de seguidores.
- **avg_likes**: Média de curtidas por postagem.
- **60_day_eng_rate**: Taxa de engajamento nos últimos 60 dias.
- **new_post_avg_like**: Média de curtidas nas novas postagens.
- **total_likes**: Total de curtidas recebidas em bilhões.
- **country**: País ou região do influenciador, transformado em faixas numéricas representando continentes.

Metodologia

A análise exploratória foi conduzida para compreender melhor o conjunto de dados e as variáveis mais relevantes para a análise. Algumas relações importantes foram observadas, como:

- A correlação entre **followers** e **avg_likes** sugere que influenciadores com mais seguidores tendem a ter mais curtidas, mas a relação não é linear.
- A variável **60_day_eng_rate** mostrou uma correlação positiva com a **taxa de engajamento**, o que indica que influenciadores com melhor desempenho recente mantêm um bom nível de engajamento.

A implementação do kNN foi feita utilizando o Scikit-Learn, com a distância Euclidiana como métrica para avaliar a proximidade entre os dados. A transformação da variável **country** foi realizada, atribuindo valores numéricos de acordo com os continentes para simplificar a análise.

O código foi estruturado para permitir o teste de diferentes valores de **k** e a avaliação do modelo com validação cruzada. Também foram avaliados os efeitos da normalização das variáveis, como **followers**, **avg_likes**, e **total_likes**, para melhorar a performance do modelo.

O processo de validação cruzada foi implementado para garantir que o modelo fosse robusto e não sofresse de overfitting. O ajuste de hiperparâmetros foi realizado utilizando **GridSearchCV**, buscando o melhor valor para **k** e a melhor configuração para a métrica de distância.

Resultados

O modelo kNN foi avaliado utilizando três métricas de erro: MAE, MSE e RMSE. Os resultados obtidos foram:

- **MAE (Mean Absolute Error)**: 2.80
  O MAE indica que, em média, a diferença entre as previsões e os valores reais

é de 2.80. Isso sugere que o modelo é capaz de fazer previsões razoavelmente próximas dos valores reais.

- **MSE (Mean Squared Error)**: 13.82
  O MSE é mais sensível a erros grandes devido ao efeito do erro ao quadrado. O valor de 13.82 indica que ainda existem alguns pontos em que a previsão diverge significativamente dos valores reais, mas o impacto desses erros não é excessivamente alto, dada a escala dos dados.
- **RMSE (Root Mean Squared Error)**: 3.71
  O RMSE, com valor de 3.71, reflete a magnitude dos erros de forma similar ao MSE, mas traz a métrica para a mesma escala das variáveis de interesse. O valor obtido é aceitável, embora haja espaço para melhorias.

Essas métricas sugerem que o modelo tem um desempenho sólido, com erros moderados e uma boa capacidade de generalização.

Foram gerados os seguintes gráficos para visualizar as relações entre as variáveis:

1. Gráfico de Dispersão (Scatter Plot) entre "followers" e "avg_likes":

   A relação entre **followers** e **avg_likes** mostra uma tendência positiva, confirmando que influenciadores com mais seguidores tendem a receber mais curtidas. No entanto, há uma variação significativa em influenciadores de menor porte, indicando a presença de fatores externos não modelados.

1. Gráfico de Dispersão (Scatter Plot) entre "rank" e "influence_score":

   A comparação entre o **rank** e o **influence_score** mostrou uma relação coerente, com influenciadores de ranking mais alto geralmente apresentando pontuações de influência mais elevadas, o que valida a lógica por trás da pontuação.

Discussão

Os resultados obtidos indicam que o modelo kNN foi capaz de capturar padrões significativos no conjunto de dados, especialmente ao prever a influência de um perfil baseada em variáveis como seguidores e curtidas. O MAE de 2.80 e o RMSE de 3.71 demonstram que o modelo possui um bom nível de precisão, mas os valores de MSE sugerem que alguns pontos possuem erros maiores, possivelmente devido a outliers ou a variabilidade não capturada por um algoritmo baseado em distância.

O ajuste de hiperparâmetros, como o valor de **k**, e a normalização das variáveis, contribuíram positivamente para o desempenho do modelo. No entanto, a simplicidade do kNN também pode ser uma limitação, pois ele não leva em conta relações complexas entre variáveis e pode ser afetado pela alta dimensionalidade dos dados.

Adicionalmente, a transformação da variável **country** para representar continentes simplificou a análise, mas pode ter perdido nuances regionais importantes. A inclusão de variáveis que capturem diferenças culturais e de comportamento entre regiões poderia potencialmente aumentar a precisão do modelo.

Conclusão

O projeto demonstrou que o algoritmo kNN é uma abordagem eficaz para a análise de influenciadores do Instagram, especialmente para dados quantitativos que apresentam padrões lineares ou próximos disso. O desempenho medido pelas métricas de erro foi satisfatório, embora existam áreas a serem melhoradas.

Como trabalhos futuros, recomenda-se:

1. **Explorar Modelos Mais Avançados**: Considerar o uso de algoritmos como árvores de decisão, random forest ou redes neurais, que podem capturar relações não lineares e complexas entre as variáveis.
2. **Incluir Variáveis Qualitativas**: Incorporar informações sobre o conteúdo das postagens, hashtags utilizadas e tipo de interação com os seguidores, o que pode fornecer uma visão mais detalhada do impacto e engajamento.
3. **Tratar Outliers**: Implementar técnicas de detecção e tratamento de outliers para reduzir o impacto de erros extremos no MSE e RMSE.

Em resumo, o kNN mostrou-se uma boa escolha inicial para o problema proposto, mas há oportunidades para melhorar o modelo e sua capacidade preditiva com a inclusão de mais variáveis e técnicas de modelagem avançadas.

Referências

GITHUB. *GitHub Repository for Project kNN Analysis*. [s. l.]: GitHub Inc., 2024. Disponível em: https://github.com/RafGuiro/PojetoKNNTIC36. Acesso em: 17 nov. 2024.

JHA, Suraj. *Top Instagram Influencers Data Cleaned*. Kaggle, 2023. Disponível em: https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned. Acesso em: 17 nov. 2024.

KAGGLE. *Python Data Analysis*. [s. l.]: Kaggle Inc., 2024. Disponível em: https://www.kaggle.com/learn/python. Acesso em: 17 nov. 2024.

PEDREGOSA, F.; VAARO, G.; GRAMFORT, A.; et al. *Scikit-learn: Machine Learning in Python*. [s. l.]: Scikit-Learn, 2011. Disponível em: https://scikit-learn.org/stable/user_guide.html. Acesso em: 17 nov. 2024.

WORLD BANK. *World Bank Country and Lending Groups*. Washington, D.C.: World Bank, 2024. Disponível em: https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups . Acesso em: 17 nov. 2024.