Relatório Técnico: Implementação e Análise do Algoritmo KNN

Cauã Tavares Nunes

Vinícius Levi dos Santos

JUAZEIRO - BA

17/11/2024

RESUMO

Este relatório apresenta um estudo técnico sobre a análise de influenciadores do Instagram utilizando o algoritmo k-Nearest Neighbors (kNN). O principal objetivo foi explorar e prever a taxa de engajamento dos influenciadores a partir de um conjunto de dados específico. A metodologia envolveu uma análise exploratória detalhada dos dados, conversão e normalização de variáveis, e implementação e validação do modelo kNN. Os resultados evidenciaram um modelo eficaz com um R2 score de 0.96, demonstrando a importância da normalização dos dados para melhorar o desempenho do modelo.

INTRODUÇÃO

A análise de influenciadores do Instagram é essencial para entender o impacto e o alcance das personalidades nas redes sociais. O uso do kNN é justificado pela sua simplicidade e eficácia em lidar com problemas de regressão neste contexto. O conjunto de dados utilizado compreende informações sobre influenciadores do Instagram, incluindo fatores como número de seguidores, média de likes, e taxa de engajamento.

METODOLOGIA

Análise Exploratória:

A análise inicial dos dados envolveu a conversão de valores expressos em formas como "1k" ou em percentagens, além de um mapeamento das localidades dos influenciadores para seus respectivos continentes, simplificando a análise geodemográfica. A visualização dos dados incorporou gráficos de distribuição por continente e análise da relação entre o número de seguidores e a média de likes.

Implementação do Algoritmo:

Para implementação do kNN, a variável país foi convertida em códigos de continente, o que facilitou a categorização para o modelo. As features escolhidas para a implementação foram normalizadas utilizando o MinMaxScaler, assegurando que todas as variáveis tivessem uma contribuição balanceada no modelo.

Validação e Ajuste de Hiperparâmetros:

O ajuste de hiperparâmetros foi realizado com a técnica de Grid Search, identificando que o melhor número de vizinhos para o kNN foi 3. A validação cruzada foi utilizada para garantir a robustez do modelo, com média de resultados consistentemente elevados.

RESULTADOS

Métricas de Avaliação:

O modelo apresentou uma performance substancial com MAE de 0.0397, MSE de 0.0042, RMSE de 0.0648 e um R2 Score de 0.96, indicando previsões precisas da taxa de engajamento.

Visualizações:

Foram gerados gráficos que mostram a precisão do modelo, comparando os valores reais e previstos de taxa de engajamento, além de visualizações que destacam a relação entre números de seguidores e likes.

DISCUSSÃO

Os resultados obtidos comprovam a eficácia da normalização dos dados e da escolha correta de hiperparâmetros para melhorar o desempenho do modelo. No entanto, algumas limitações incluem a dependência de dados históricos e a possibilidade de sobreajuste devido à escolha de atributos de entrada. O impacto das conversões de dados e do mapeamento para continentes provou-se benéfico.

CONCLUSÃO E TRABALHOS FUTUROS

Este projeto reforçou a importância de uma análise de dados minuciosa e da escolha de técnicas de modelagem apropriadas. Trabalhos futuros poderiam incluir a integração de dados temporais para prever tendências futuras e o uso de técnicas de aprendizado profundo para potencialmente melhorar o desempenho do modelo.

REFERÊNCIAS

Scikit-learn: Scikit-learn developers. Scikit-learn: Machine Learning in Python. Disponível em: https://scikit-learn.org/. Acesso em: 20/11/2024

Pandas: The Pandas development team. Pandas: Powerful data analysis tools for Python. Disponível em: https://pandas.pydata.org/. Acesso em: 20/11/2024.

Matplotlib: Hunter, J. D. Matplotlib: A 2D graphics environment. Computing in Science & Engineering, 9(3), 90-95, 2007.

Seaborn: Waskom, M. L. seaborn: Statistical data visualization. Journal of Open Source Software, 1(0.010), 218, 2017.

AMARAL, Fernando. Formação Cientista de Dados: O Curso Completo. Udemy. Disponível em: https://www.udemy.com/course/cientista-de-dados/.

GEEKSFORGEEKS. K-Nearest Neighbors (KNN) Regression with scikit-learn. Disponível em: <https://www.geeksforgeeks.org/k-nearest-neighbors-knn-regression-with-scikit-learn/>. Acesso em: 17 nov. 2024.