Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Residente 1: Guilherme Melo dos Santos

Residente 2: Gleidson de Meireles Costa

Data: 17/11/2024

Resumo

O trabalho desenvolvido teve como principal objetivo criar um modelo preditivo capaz de estimar a taxa de engajamento de influenciadores digitais utilizando informações como número de seguidores, postagens e médias de curtidas. Após a aplicação de técnicas de análise exploratória e tratamento dos dados, utilizamos a Regressão Linear como base para a modelagem. O desempenho do modelo foi avaliado por métricas como R^2, MSE e MAE, alcançando resultados satisfatórios com R^2 de 0.92 no conjunto de teste. Além disso, foram aplicadas técnicas de regularização (Ridge e Lasso) e validação cruzada para explorar a generalização e ajustar os hiperparâmetros. Este estudo destaca a relevância de um pré-processamento robusto e de escolhas bem fundamentadas na construção de modelos preditivos.

Introdução

A taxa de engajamento é um indicador crítico na avaliação da performance de influenciadores digitais, especialmente no Instagram, onde métricas como curtidas, comentários e visualizações desempenham um papel fundamental para marcas que buscam parcerias estratégicas. Dado esse contexto, foi elaborado um projeto para estimar, com base em variáveis facilmente observáveis, a taxa de engajamento de influenciadores. A utilização da Regressão Linear foi justificada pela sua capacidade de identificar padrões e relações lineares entre variáveis independentes e a variável dependente.

O conjunto de dados utilizado neste estudo foi extraído de um levantamento que abrange os principais influenciadores do Instagram. Ele contém informações sobre número de postagens, seguidores, curtidas médias, entre outras métricas. Composto por 200 registros inicialmente, o ataset foi ajustado para conter apenas 199 registros devido à necessidade de tratar dados ausentes. Este conjunto de informações foi o ponto de partida para a aplicação de diversas etapas metodológicas, como análise exploratória, tratamento e modelagem preditiva.

Metodologia

Para garantir que os dados fossem adequados ao processo de modelagem, foi realizado um pré-processamento abrangente. Em primeiro lugar, as variáveis categóricas, como os nomes dos influenciadores e países, foram descartadas, uma vez que não contribuíam diretamente para o objetivo preditivo. Dados ausentes ou inconsistentes foram tratados, e as variáveis restantes passaram por normalização utilizando o método `StandardScaler` da biblioteca Scikit-learn.

A análise exploratória revelou importantes insights sobre as relações entre as variáveis. A partir de um mapa de correlação, foi possível identificar dependências lineares que embasaram a seleção das variáveis independentes para o modelo, como `followers`, `avg_likes` e `new_post_avg_like`. Essas relações também ajudaram a confirmar a adequação do uso da Regressão Linear.

O modelo foi treinado utilizando a biblioteca Scikit-learn, sendo a variável dependente a taxa de engajamento registrada como `60_day_eng_rate`. Para melhorar a capacidade do modelo em lidar com possíveis multicolinearidades, aplicamos técnicas de regularização, como Ridge (L2) e Lasso (L1). Além disso, foi realizada validação cruzada utilizando o método K-Fold com 5 divisões, garantindo uma estimativa mais robusta do desempenho do modelo.

Resultados

Os resultados obtidos após a implementação do modelo foram promissores. A Regressão Linear base apresentou um R^2 de 0.92, indicando que o modelo foi capaz de explicar grande parte da variabilidade da taxa de engajamento. O erro quadrático médio (MSE) foi próximo de zero, enquanto o erro absoluto médio (MAE) ficou em torno de 0.005. Esses números reforçam a precisão do modelo na tarefa preditiva.

O uso de regularização trouxe resultados mistos. O modelo Ridge teve desempenho próximo ao modelo base, com um R^2 de 0.92, enquanto o Lasso apresentou dificuldades em ajustar os coeficientes, resultando em um R^2 insatisfatório de -0.0018. Isso aponta para a necessidade de ajustes mais refinados nos hiperparâmetros para essa técnica.

A validação cruzada, por outro lado, revelou uma grande variabilidade nos resultados, com um R^2 médio de -0.0595 ± 1.9182. Essa dispersão sugere que o modelo pode ser sensível ao conjunto de dados utilizado, refletindo a necessidade de ampliar a base de dados ou de explorar técnicas mais robustas.

Os coeficientes das variáveis fornecem insights interessantes sobre sua relação com a taxa de engajamento. Por exemplo, new_post_avg_like apresentou um coeficiente positivo significativo, indicando que a média de curtidas em postagens recentes tem grande impacto na taxa de engajamento. Em contrapartida, o número de seguidores (followers) apresentou um coeficiente negativo, possivelmente sugerindo que influenciadores com grandes audiências podem ter uma taxa de engajamento mais diluída.

Discussão

Os resultados obtidos demonstram que a Regressão Linear é uma técnica viável para prever a taxa de engajamento de influenciadores digitais, desde que aplicada a um conjunto de dados devidamente tratado. No entanto, algumas limitações foram identificadas durante o estudo. O tamanho relativamente pequeno da amostra e a ausência de variáveis qualitativas, como categoria do influenciador ou país, podem ter reduzido o potencial preditivo do modelo.

Outro ponto a considerar foi a alta variabilidade observada na validação cruzada. Isso pode indicar que o modelo apresenta limitações em generalizar para novos dados, sugerindo que métodos alternativos, como árvores de decisão ou redes neurais, podem ser explorados no futuro.

Apesar dessas limitações, o modelo é funcional e fornece insights valiosos sobre os fatores que influenciam a taxa de engajamento. Técnicas como regularização Ridge mostraram-se úteis, enquanto o Lasso pode necessitar de ajustes mais específicos.

Conclusão e Trabalhos Futuros

Este estudo confirmou a eficácia da Regressão Linear para tarefas de previsão no contexto de marketing digital. O modelo implementado apresentou métricas sólidas, como um R^2 de 0.92, e forneceu insights claros sobre a relevância de variáveis como média de curtidas e seguidores na taxa de engajamento.

Como propostas para trabalhos futuros, recomenda-se a inclusão de mais registros na base de dados, bem como o acréscimo de variáveis qualitativas que possam enriquecer o modelo. Além disso, a comparação com algoritmos mais avançados pode ser útil para avaliar se métodos não lineares oferecem vantagens neste contexto.

Referências

1. Documentação do Scikit -learn: https://scikit-learn.org
2. Documentação do Pandas: https://pandas.pydata.org
3. Dataset fictício adaptado para fins didáticos.
4. Conjunto de Dados: https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned