Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Anna Giulia Gomes Miranda

Victoria Beatriz Silva de Azevedo Reis

17 de novembro de 2024

Resumo

Este projeto tem como objetivo aplicar regressão linear para inferir taxas de engajamento de influenciadores no Instagram, usando um conjunto de dados contendo informações sobre métricas da conta, como número de seguidores, média de curtidas e taxa de engajamento nos últimos 60 dias. A abordagem envolve a preparação de dados, análise exploratória, implementação de modelos de regressão linear e avaliação do desempenho do modelo usando diversas métricas. O modelo final apresentou um R^2 de 0.92 e um RMSE de 0.05, indicando uma boa capacidade preditiva.

Palavras-chave: regressão linear, instagram, análise exploratória

Introdução

Um dos principais indicadores de desempenho dos influenciadores no Instagram é a taxa de engajamento. As marcas levam em conta essa taxa para determinar a eficácia de suas campanhas de marketing. O projeto foi justificado pela necessidade de prever com precisão esta métrica, com o objetivo de aprimorar a seleção de influencers pelas marcas. A regressão linear foi escolhida devido à simplicidade do algoritmo e à sua capacidade de capturar relações entre variáveis numéricas, além de ser rápida e eficiente mesmo com diversos dados, ajudando a compreender como duas ou mais variáveis estão relacionadas.

Os dados foram coletados de uma amostra de 200 influenciadores e possuem as seguintes variáveis: pontuação de influência, postagens, seguidores, média de curtidas, média de curtidas em novas postagens e total de curtidas.

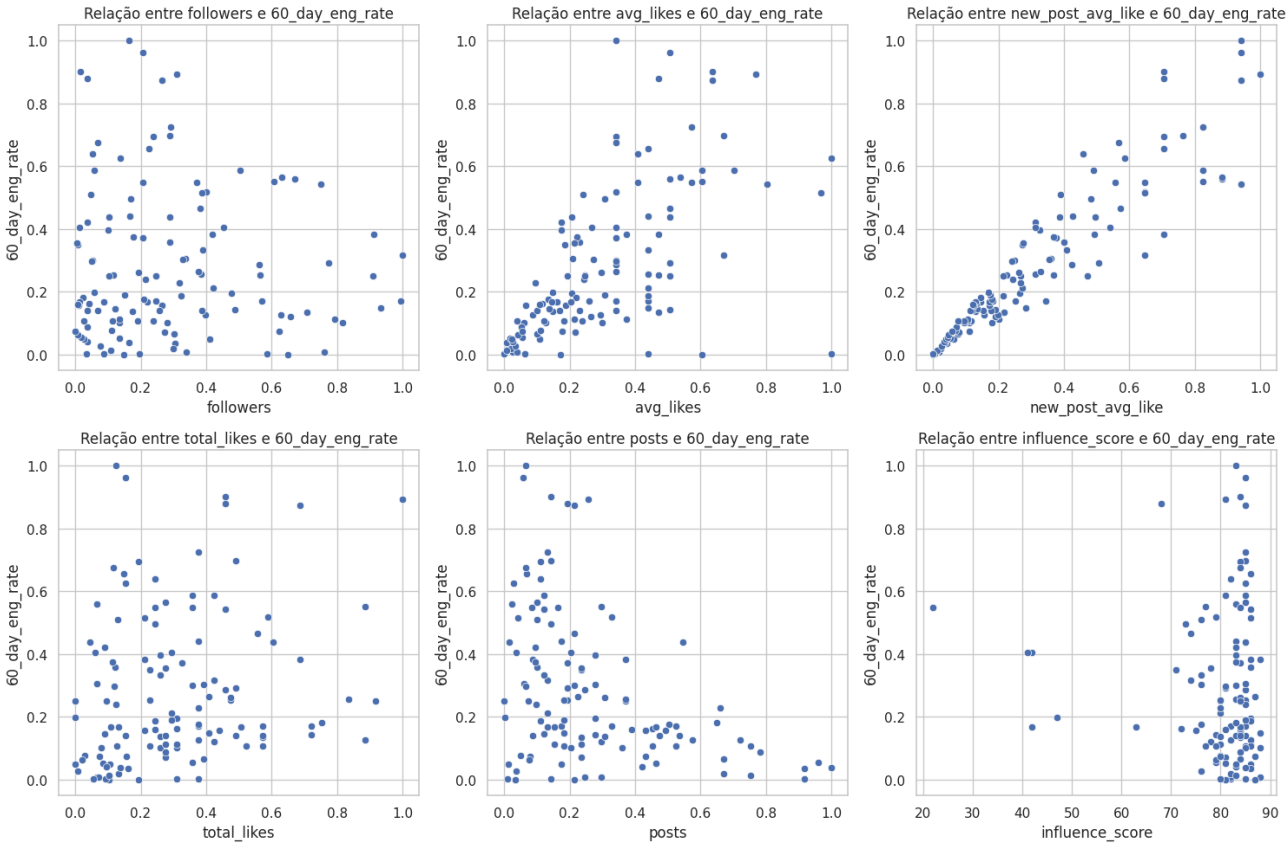
Metodologia

Análise Exploratória

Na análise inicial dos dados, foi observado que algumas variáveis apresentam correlações relevantes com a taxa de engajamento em 60 dias, como o número médio de curtidas que possui uma correlação de 0.61 e a média de curtidas em novas postagens com correlação de 0.93, sugerindo que essa variável em específico é um bom preditor da taxa de engajamento. Por outro lado, variáveis como postagens e pontuação de influência mostraram correlações negativas, com pouca ou nenhuma influência.

Ao analisar os gráficos de dispersão é possível concluir a relação que já foi afirmada acima, entre taxa de engajamento e número médio de curtidas. À medida que o número médio de curtidas aumenta, a taxa de engajamento também aumenta; no entanto, o gráfico indica que outros elementos também afetam a taxa. Além disso, a variável que representa a média de curtidas em novas postagens demonstra a maior correlação, sendo quase linear, sugerindo que essa variável desempenha o melhor papel na previsão da taxa de engajamento.

Correlações



Implementação do Algoritmo

Antes de implementar o algoritmo é necessário preparar os dados, convertendo eles para valores numéricos, permitindo que os mesmos possam ser utilizados em cálculos futuramente. O próximo passo, é a remoção de outliers, que podem distorcer os resultados da regressão linear.

Para garantir que as variáveis possam ser comparadas entre si, as colunas foram normalizadas entre 0 e 1 utilizando a técnica Min-Max. Em seguida, o conjunto de dados foi dividido em um conjunto de treinamento e um de teste, sendo 70% para treino e 30% para teste. A regressão linear foi implementada e o modelo foi treinado utilizando os dados de treino `X_train, y_train`.

Validação e Ajuste de Hiperparâmetros

As variáveis independentes foram postagens, seguidores, média de curtidas, média de curtidas em novas postagens e total de curtidas. Sendo escolhidas por conta do resultado da correlação dessas variáveis com a variável dependente taxa de engajamento de 60 dias.

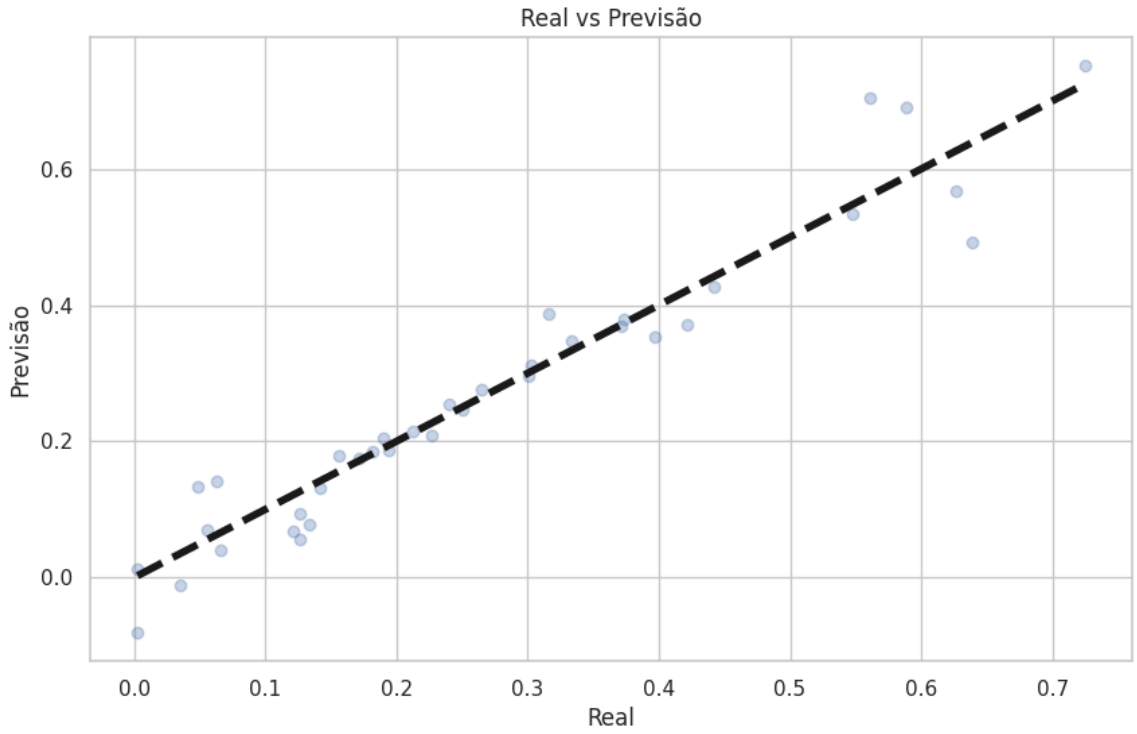
O processo de validação cruzada foi utilizado para garantir a generalização do modelo e evitar overfitting, dividindo o modelo em diferentes subconjuntos para treinar e testar o mesmo, garantindo a redução na variabilidade do resultado e por usar todo o conjunto de dados, mesmo que dividido, também garante uma avaliação mais confiável do modelo.

Resultados

Métricas de Avaliação

As principais métricas de avaliação do modelo utilizadas foram R^2 e erro quadrático médio (MSE). O modelo teve um bom desempenho em ambas as métricas, um MSE de 0.05 que mostra que há um baixo erro nas previsões e um coeficiente de determinação (R^2) de 0.92 que indica um bom ajuste entre os dados usados e a regressão linear, indicando que o modelo consegue capturar as tendências da taxa de engajamento com um nível razoável de precisão.

Visualizações



Discussão

Os resultados obtidos no modelo de Regressão Linear mostraram uma correlação entre algumas das variáveis independentes e a taxa de engajamento de 60 dias. Algumas variáveis do dataset não demonstraram ter uma correlação relevante, no entanto variáveis como a média de curtidas em novos posts demonstrou ter uma correlação expressiva com a variável dependente utilizada (taxa de engajamento).

Ao analisar a lógica da ferramenta Instagram como um todo, esse resultado pode ser explicado pelo fato de que uma grande quantidade de seguidores não assegura a existência de uma alta taxa de engajamento. Outras variáveis podem possuir uma maior influência no engajamento final.

As limitações encontradas durante a construção do modelo, envolvem o fato de que mesmo sendo conhecida como simples e de fácil compreensão, a regressão linear não é capaz de identificar relações não lineares complexas presentes nos dados. Além disso, a remoção dos outliers pode ter deixado de lado dados importantes sobre influenciadores excepcionais que não seguem uma lógica linear esperada. Embora a normalização também seja necessária, para um resultado mais condizente com o esperado, esta pode ter atenuado algum dos efeitos mais significativos que alguma dessas variáveis independentes exerciam sobre a taxa de engajamento.

Outra limitação, dessa vez ligada ao próprio dataset e não ao código, é a quantidade de dados limitada. Apesar de apresentar um resultado aceitável com base no exposto, a taxa de engajamento que foi analisada pode sofrer influência de outras variáveis, que podem até ser mais importantes que as apresentadas neste projeto.

Conclusão e Trabalhos Futuros

O projeto proporcionou uma compreensão precisa da conexão entre as variáveis quantitativas dos influenciadores do Instagram e suas taxas de engajamento. A Regressão Linear mostrou ser uma ferramenta eficaz para compreender essas relações, apesar de apresentar restrições em sua capacidade de prever com alta precisão o engajamento baseado em um conjunto simples de variáveis.

Considerando futuras melhorias no projeto, é possível dizer que explorar modelos mais complexos, como árvores de decisão ou modelos de redes neurais, é uma melhor opção quando comparado a regressão linear, pois podem capturar interações não lineares. Além de incorporar mais variáveis, incluindo dados

qualitativos, como o uso de vídeos, uso de hashtags populares, ou colaboração com marcas.

Referências

MIRKO STOJILJKOVIC. Linear Regression in Python. Disponível em: <https://realpython.com/linear-regression-in-python/>.
Linear Regression. Disponível em: <https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LinearRegression.html>.

DEBROY, S. Simple Linear Regression in Python. Disponível em: <https://medium.com/@shuv.sdr/simple-linear-regression-in-python-a0069b325bf8>.

Top Instagram Influencers Data (Cleaned). Disponível em: <https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned?resource=download>.
JÚNIOR, C. DE O. Prevendo Números: Entendendo as métricas R2, MAE, MAPE, MSE e RMSE. Disponível em: <https://medium.com/data-hackers/prevendo-n%C3%BAmeros-entendendo-m%C3%A9tricas-de-regress%C3%A3o-35545e011e70>.