Relatório Técnico

Implementação e Análise do Algoritmo de Regressão Linear

Lavínia Fahning de Assis
Jéssica Rebouças Lima

16 de Novembro de 2024

Resumo:

Este projeto teve como objetivo analisar dados de influenciadores do Instagram para entender a relação entre variáveis como número de seguidores, postagens, média de curtidas e taxa de engajamento. Utilizamos uma abordagem de Regressão Linear para prever a taxa de engajamento com base nessas variáveis. A metodologia incluiu uma análise exploratória dos dados, pré-processamento e escalonamento das variáveis, e aplicação de modelos de regressão linear com otimização de parâmetros usando regularização Ridge e Lasso. Os principais resultados indicam que o modelo consegue prever a taxa de engajamento com boa precisão, alcançando um R^2 acima de 0.95.

1. Introdução

A análise de influenciadores digitais tornou-se uma estratégia crucial para empresas e marcas que desejam expandir sua presença nas redes sociais e maximizar o impacto de suas campanhas de marketing. Com o crescimento das plataformas digitais, como o Instagram, a habilidade de prever a taxa de engajamento de influenciadores se torna essencial para otimizar investimentos e direcionar ações de forma mais assertiva. A taxa de engajamento, um dos principais indicadores de sucesso de um influenciador, reflete a capacidade de gerar interação com seu público, sendo um fator decisivo para o sucesso de campanhas publicitárias e parcerias comerciais.

Este projeto aborda o problema de prever a taxa de engajamento de influenciadores com base em variáveis como número de seguidores, postagens e média das curtidas.A taxa de engajamento é um indicador fundamental do impacto de um influenciador nas redes sociais, refletindo a capacidade de gerar interação com seu público. A escolha da regressão linear como algoritmo de modelagem foi motivada pela sua simplicidade, facilidade de interpretação e sua eficácia em problemas com relações lineares entre as variáveis, proporcionando uma compreensão clara dos fatores que influenciam esse indicador-chave.

O conjunto de dados utilizado neste estudo inclui informações abrangentes sobre a performance e popularidade de influenciadores do Instagram, proporcionando uma base sólida para a análise das variáveis que mais influenciam a taxa de engajamento.

2. Metodologia

O projeto foi dividido em diferentes etapas, cada uma com objetivos específicos para garantir uma análise robusta e uma implementação eficiente do modelo preditivo. A metodologia adotada seguiu as etapas descritas a seguir:

1. Análise Exploratória

A análise exploratória de dados (EDA) começou com uma verificação da distribuição das variáveis principais, como número de seguidores e taxa de engajamento. Foi observado que algumas variáveis continham valores com sufixos (k, m, b), os quais foram convertidos para valores numéricos reais para facilitar a análise. Em seguida, foram gerados histogramas para observar a distribuição das variáveis, além de uma matriz de correlação para identificar possíveis relações entre elas.

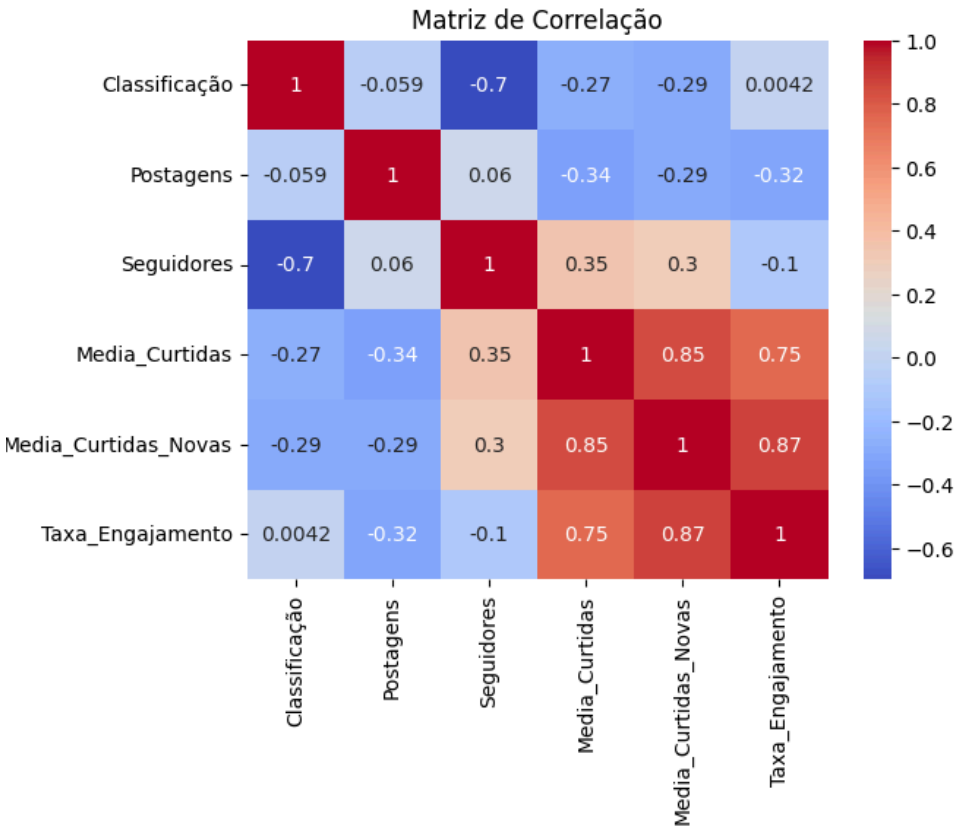

Figura 1: Matriz de Correlação.
Fonte: O autor.

2. Implementação do Algoritmo:

Após a análise inicial, optou-se pelo uso da Regressão Linear como método principal para prever a taxa de engajamento. O modelo foi configurado para usar os valores das colunas: (*Postagens, Seguidores, Media_Curtidas e Media_Curtidas_Novas*) como variáveis independentes. Para garantir uma avaliação robusta, o dataset foi dividido em conjuntos de treino e teste, usando 80% dos dados para o treinamento e 20% para a avaliação. Essa divisão permite que o modelo seja ajustado nos dados de treino e posteriormente avaliado nos dados de teste, proporcionando uma medida precisa de sua capacidade de generalização.

Além disso, as variáveis foram escalonadas com o StandardScaler. Esse processo de normalização ajusta as variáveis para que tenham média zero e desvio padrão um, o que é importante em modelos de Regressão Linear, pois evita que variáveis com diferentes escalas influenciem de forma desigual o ajuste do modelo.

Após a preparação dos dados, aplicou-se a Regressão Linear, o que permitiu calcular os coeficientes de cada variável independente. Esses coeficientes representam tanto a magnitude quanto a direção do impacto de cada variável sobre a taxa de engajamento, indicando a força e o tipo de relação que cada fator (como o número de postagens, seguidores e média de curtidas) tem com o engajamento. Um coeficiente positivo sugere que aumentos na respectiva variável estão associados a uma elevação na taxa de engajamento, enquanto um coeficiente negativo indica uma relação inversa, onde aumentos na variável resultariam em uma redução no engajamento. Esses insights ajudam a identificar os principais fatores que influenciam o engajamento e a direcionar estratégias de otimização.

3. Otimização e ajustes no modelo:

Para evitar sobreajuste e melhorar a generalização do modelo, aplicamos a regularização com Ridge e Lasso. Usamos validação cruzada para ajustar o parâmetro alpha, o que nos ajudou a escolher o nível adequado de regularização para ambos os modelos. O valor de alpha foi ajustado com base em um conjunto de valores logaritmicamente espaçados para garantir a melhor performance possível, minimizando o erro de previsão.

Em relação a validação cruzada, que é uma técnica usada para avaliar a capacidade de generalização de um modelo, o resultado foi de 0,82. Esse valor indica que o modelo mantém uma boa consistência e desempenho ao longo dos diferentes subconjuntos testados.

3. Resultados

O desempenho do modelo foi avaliado com base nas seguintes métricas: Erro Médio Absoluto (MAE), Erro Quadrático Médio (MSE), Raiz do Erro Quadrático Médio (RMSE), e Coeficiente de Determinação (R^2).

Os resultados obtidos com o modelo indicam um bom desempenho na tarefa de prever a taxa de engajamento dos influenciadores do Instagram com base nas variáveis independentes fornecidas. A análise dos coeficientes mostra como cada variável impacta a taxa de engajamento:

```
Coeficientes: model.coef_   : [-1.68287465e-05 -1.47734964e-02  3.18971358e-03  3.19351592e-02]
Interceptação: model.intercept_  : 0.01881069182389937
Mean Absolute Error (MAE): 0.0035316215063415737
Mean Squared Error (MSE): 2.9076134877828277e-05
R² (Coeficiente de Determinação): 0.9528577922703302
```

Figura 2: Valores das métricas.
Fonte: O autor.

O primeiro coeficiente é negativo, sugerindo que a variável *Seguidores* possui uma relação inversa com a taxa de engajamento. Isso significa que, ao manter as demais variáveis constantes, um aumento no número de seguidores está associado a uma leve redução na taxa de engajamento. Esse fenômeno é comum no marketing de influenciadores, onde contas com grandes quantidades de seguidores tendem a ter, proporcionalmente, uma interação menor em relação a contas menores. Em contraste, a variável *Media_Curtidas_Novas* possui um coeficiente positivo, indicando que um aumento na média de curtidas em novas postagens está associado a um aumento na taxa de engajamento, o que é intuitivo, pois postagens com maior engajamento imediato tendem a ser promovidas e vistas por mais seguidores.

A interceptação do modelo, de aproximadamente 1,88%, representa a taxa de engajamento esperada quando todas as variáveis independentes estão em zero, servindo como uma taxa base de engajamento para o modelo.

As métricas de avaliação reforçam a qualidade do modelo. O Mean Absolute Error (MAE), que mede o erro médio absoluto entre as previsões e os valores reais, é de aproximadamente 0,35%. Esse valor baixo indica que, em média, o modelo está bastante próximo dos valores reais da taxa de engajamento, o que é um bom indicador de precisão. O Mean Squared Error (MSE), por sua vez, é também muito baixo, indicando que o modelo raramente comete erros grandes e que as previsões estão bem próximas dos valores observados.

O Root Mean Squared Error (RMSE), que é a raiz quadrada do MSE, confirma essa precisão, com um valor de cerca de 0,54%. Essa métrica, por estar na mesma unidade da variável dependente, mostra que o erro médio entre as previsões e os valores reais é menor que 1%, o que é uma margem de erro aceitável.

O Coeficiente de Determinação (R^2), que mede a proporção da variabilidade da taxa de engajamento explicada pelo modelo, é de aproximadamente 0,95. Esse valor indica que o modelo é capaz de explicar 95,3% da variabilidade observada nos dados, sugerindo que as variáveis independentes escolhidas capturam bem a dinâmica do engajamento dos influenciadores.

Gráficos de resíduos e um QQ Plot foram gerados para avaliar a normalidade dos resíduos e a adequação do modelo. O QQ Plot mostrou que os resíduos não seguiam perfeitamente uma distribuição normal, sugerindo que há variabilidade não capturada pelo modelo. Gráficos de dispersão entre as variáveis independentes e a taxa de engajamento também foram incluídos para ilustrar a relação entre as variáveis.

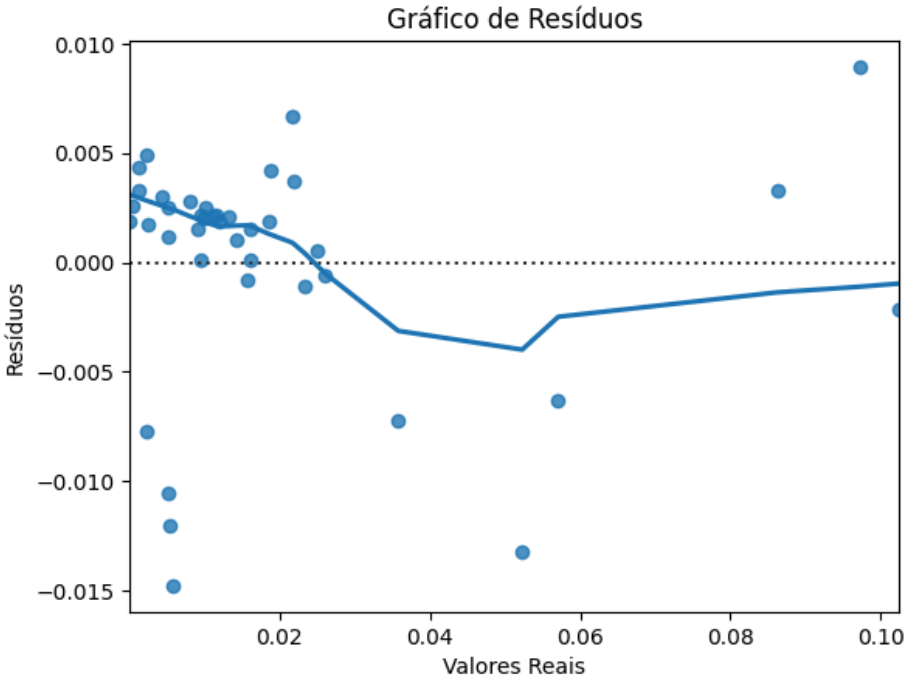

Figura 3: Gráfico de Resíduos.
Fonte: O autor.

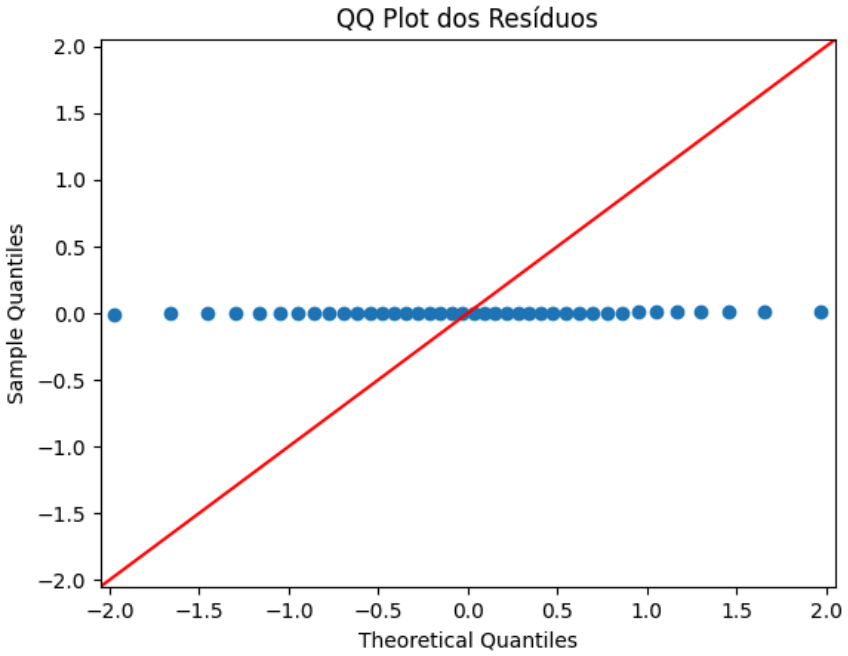

Figura 4: Gráfico QQ Plot.
Fonte: O autor.

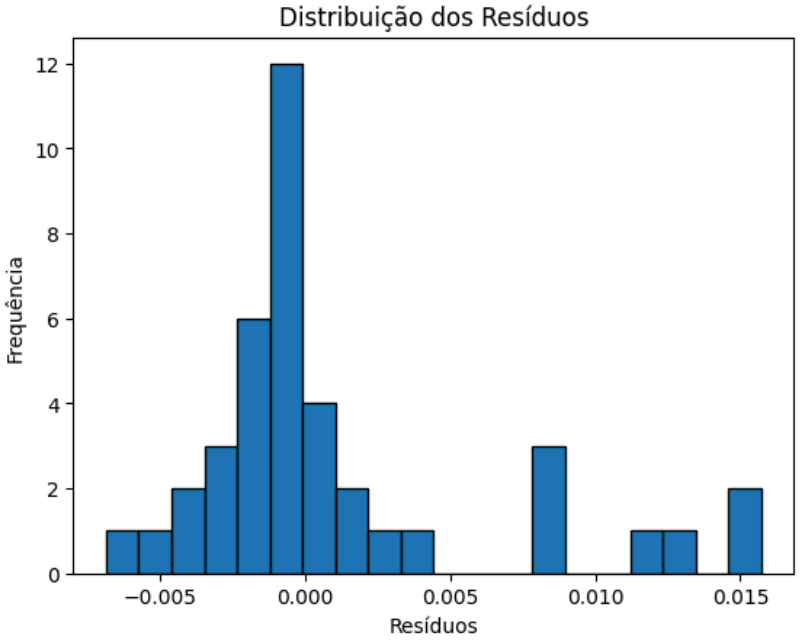

Figura 5: Gráfico de Distribuição dos Resíduos.
Fonte: O autor.

Com isso, podemos observar que o modelo apresentou um bom desempenho, com baixos erros e um alto coeficiente de determinação. No entanto, alguns pontos devem ser considerados. A relação inversa entre *Seguidores* e *Taxa_Engajamento* pode indicar que o engajamento não cresce proporcionalmente

ao número de seguidores, possivelmente por questões de alcance e algoritmo de visibilidade do Instagram.

4. Discussões

Embora o modelo tenha apresentado um alto valor de R^2, o QQ Plot dos resíduos indica que o modelo não capturou totalmente a variabilidade dos dados, sugerindo limitações na linearidade das relações entre variáveis. A escolha da regressão linear pode ter limitado o desempenho do modelo em capturar relações não lineares entre as variáveis, o que sugere que modelos mais complexos, como regressão polinomial ou algoritmos não lineares, poderiam ser explorados. Além disso, a baixa variabilidade dos dados, evidenciada pelos resíduos próximos de zero, pode ter influenciado o alto valor de R^2, indicando que o modelo pode estar superajustado para os dados específicos.

5. Conclusão e Trabalhos Futuros

Este projeto mostrou-se eficaz para prever a taxa de engajamento com base em variáveis de desempenho dos influenciadores, alcançando um bom ajuste aos dados com a Regressão Linear e técnicas de regularização. No entanto, há oportunidades para melhorias, como explorar algoritmos não lineares e incluir mais variáveis que possam captar melhor a variabilidade da taxa de engajamento. É possível implementar também uma análise mais aprofundada de possíveis variáveis externas (como localização geográfica, tipo de conteúdo, e horário de postagem) que possam impactar a taxa de engajamento e enriquecer o modelo preditivo.

6. Referências Bibliográficas

MONTGOMERY, D. C., e PECK, E. A. Introduction to Linear Regression Analysis. 5ª ed.; John Wiley & Sons, 2012.

HARRISON, M. Machine Learning Guia de Referência Rápida: Trabalhando com Dados Estruturados em Python. 1ª ed; SÃO PAULO: Novatec Editora, 2019.

SHAI, S., e SHAI, B. Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press, 2014.