Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Ana Carolina Santos de Souza e Brenda Barbosa de Oliveira

17/11/2024

Resumo

Este projeto teve como objetivo desenvolver um modelo preditivo para estimar a taxa de engajamento de influenciadores do Instagram, utilizando técnicas de aprendizado de máquina, especificamente Regressão Linear Múltipla.

A metodologia envolveu pré-processamento: tratamento de dados, análise exploratória e modelagem. Os resultados do modelo demonstrou capacidade moderada de prever taxas de engajamento.

Introdução

O engajamento nas redes sociais é um indicador importante para avaliar o impacto e efetividade de influenciadores digitais. A capacidade de prever taxas de engajamento tem valor significativo para marcas e agências na seleção de parceiros para campanhas de marketing, marketing de influência, e análise de dados na avaliação de comportamento de redes sociais.

O conjunto de dados utilizado contém informações sobre os maiores influenciadores do Instagram, e inclui variáveis como:

Classificação;
Nome do canal;
Pontuação de influência;
Número de postagens;
Número de seguidores;
Média de curtidas;
Taxa de engajamento;
Média de curtidas em novas postagens;
Total de likes;
País.

As características do dataset:

Dados reais de influenciadores;

Mistura de variáveis numéricas e categóricas;
Presença de diferentes escalas e formatos de dados;
Foi necessidade de tratamento para valores com sufixos (k, m, b, %) para realizar análises mais assertivas.

A escolha da Regressão Linear Múltipla se justifica por permitir compreender claramente o impacto individual de cada variável, rapidez para treinar e fazer previsões, simplicidade e facilidade.

Metodologia

Análise Exploratória:
Pré processamento inicial:
Renomeação de colunas para melhor a interpretação;
Conversão de valores com sufixos para formato numérico;
Tratamento de valores ausentes.

Análise de Correlações:
Criação de matriz de correlação para identificar relações entre variáveis;
Identificação das variáveis mais correlacionadas com a Taxa_Engajamento;
Visualização através de heatmap usando seaborn.

Análise de Distribuições:
Estudo da distribuição da Taxa_Engajamento;
Gráficos de dispersão entre variáveis;
Identificação de padrões e outliers.

Implementação do Algoritmo de Regressão Linear:
Seleção de Variáveis:
Variáveis independentes escolhidas:

```
X_dados = dados_df[['Seguidores', 'Media_Curtidas', 'Postagens', 'Media_Curtidas_Novas']]
```

Variável dependente:

```
y_dados = dados_df['Taxa_Engajamento']
```

Divisão dos Dados

```
X_train_dados, X_test_dados, y_train_dados, y_test_dados = train_test_split(X_dados, y_dados, test_size=0.2, random_state=42)
```

Configuração do Modelo:
Utilizei a classe LinearRegression do scikit-learn;
Treinamento do modelo com dados de treino.

Validação e ajuste de hiperparâmetros:

Processo de Validação:
Verificação de valores ausentes nos conjuntos de treino e teste;
Análise de resíduos para verificar pressupostos do modelo;
Avaliação da performance em dados de teste.

Escolha das Variáveis Independentes:
Baseada na análise de correlação;
Consideração da relevância teórica das variáveis.

Resultados:

```
Mean Squared Error (MSE): 0.01
Erro Absoluto Médio (MAE): 0.00
Coeficiente de Determinação (R²): 0.95
```

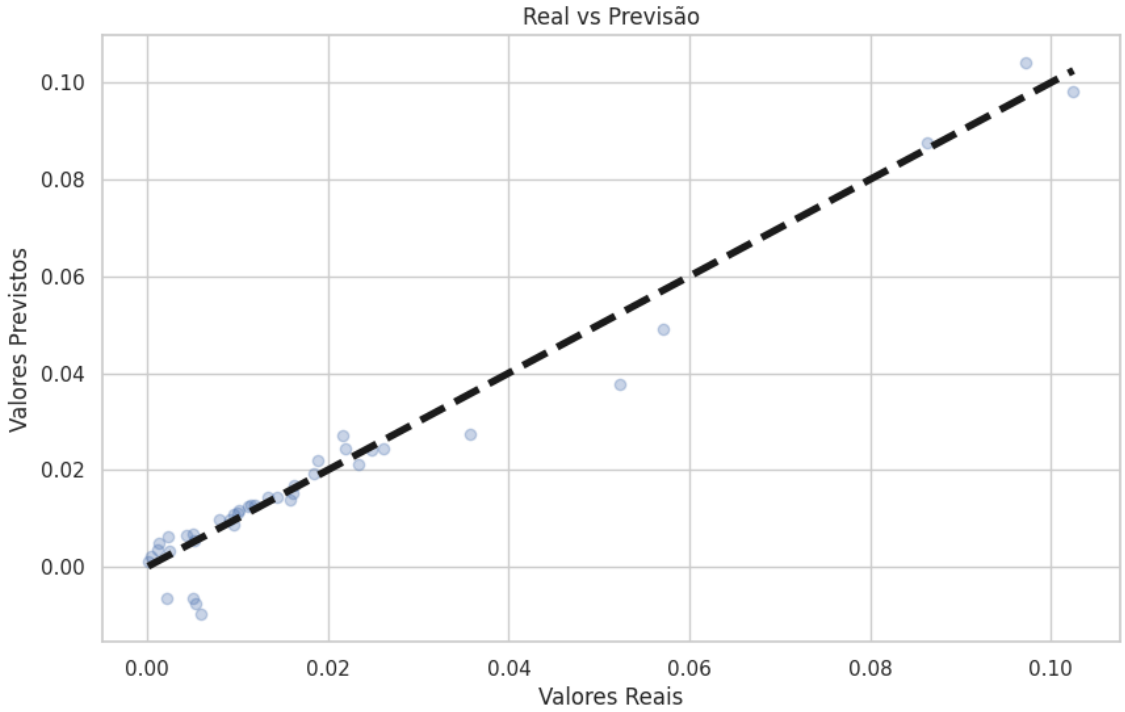
Esses valores indicam um bom desempenho do modelo.

Mean Squared Error (MSE) = 0.01: O MSE representa a média dos quadrados dos erros, ou seja, a diferença entre os valores reais e previstos. O valor de 0.01 é bem baixo, o que sugere que o modelo tem pouca diferença entre os valores previstos e reais, indicando boa precisão.

Erro Absoluto Médio (MAE) = 0.00: O MAE representa a média das diferenças absolutas entre os valores reais e previstos. Um MAE de 0.00 indica que o modelo prevê quase perfeitamente os valores, com uma diferença média quase nula entre o valor real e o previsto.

Coeficiente de Determinação (R^2) = 0.95: O R^2 mede a proporção da variabilidade dos dados que o modelo consegue explicar. Um valor de 0.95 significa que 95% da variação dos dados é explicada pelo modelo, o que é excelente.

Visualização de gráficos:

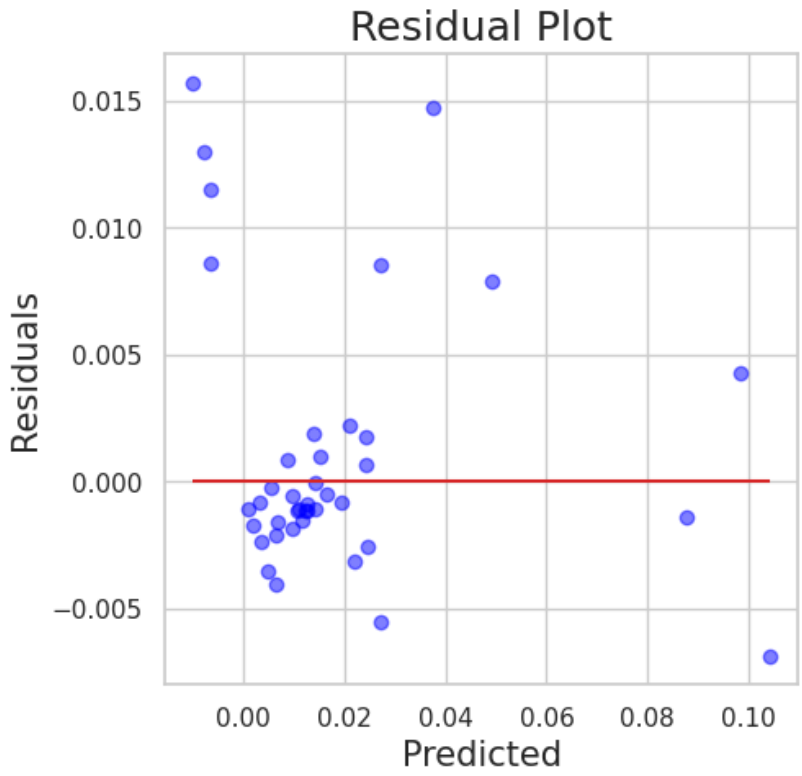


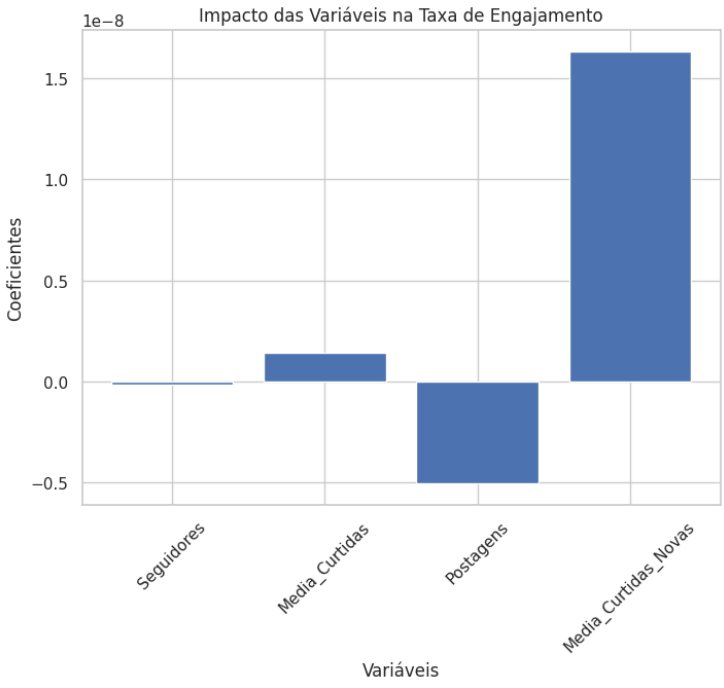
Se o modelo fosse perfeito, todos pontos estariam sobre a linha. Mas mesmo assim, os pontos próximos à linha é indicativo de boas previsões. O modelo está acertando ou errando muito pouco. Os pontos distantes da linha representa os erros.

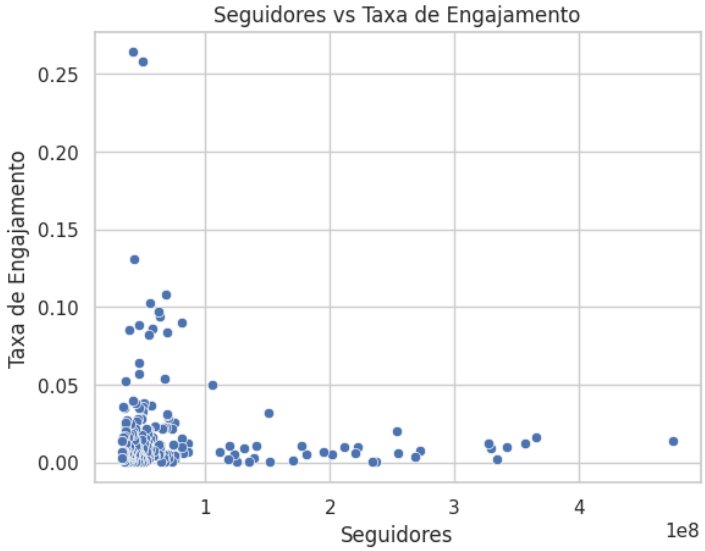
No gráfico, pontos mais distantes da linha tracejada contribuem para aumentar o MAE.

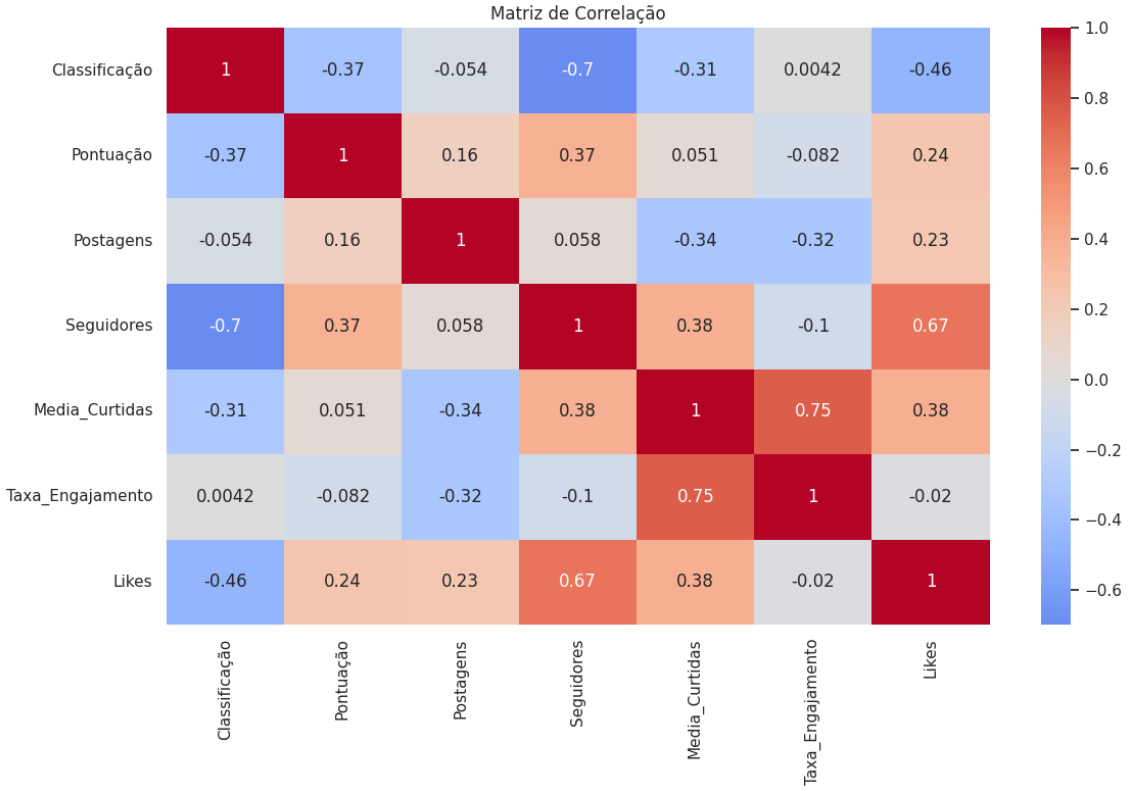
No caso, como o MAE é próximo de 0, a maioria dos pontos está bem próxima da linha, o que reflete um erro médio insignificante.

Logo abaixo, foi gerado o gráfico para verificar a presença de padrões nos resíduos, como heterocedasticidade ou tendências não capturadas pelo modelo. Idealmente, os resíduos devem ser distribuídos aleatoriamente ao redor de 0.





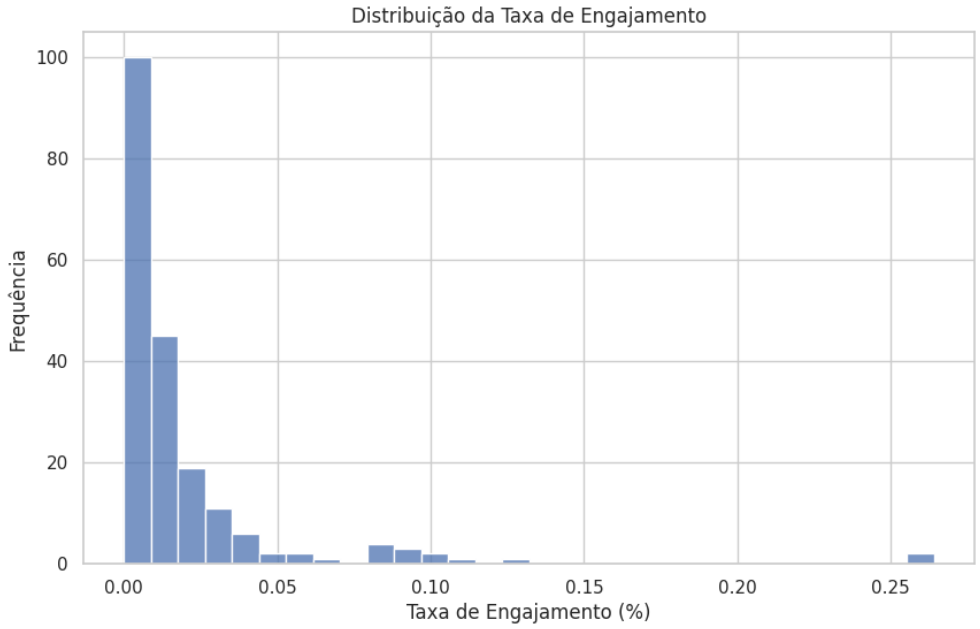


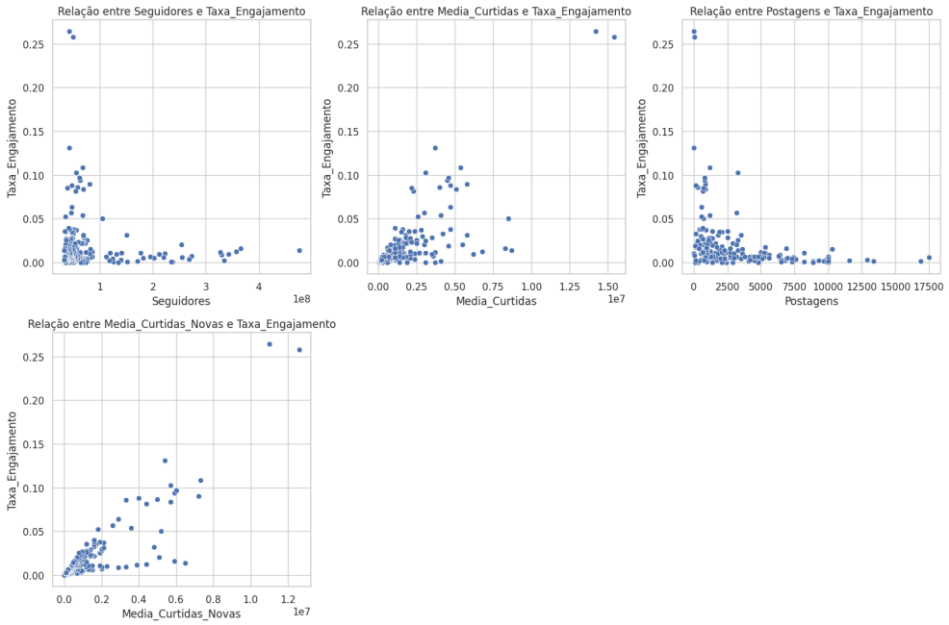


Foi gerada a matriz de correlação, que tem o objetivo de identificar relações entre variáveis numéricas.

Seguidores e Média_curtidas possuem correlação moderada positiva.

A variável dependente Taxa_Engajamento apresenta correlações baixas com as variáveis independentes, sugerindo fraca linearidade inicial.





Discussão

Análise Crítica dos Resultados:

Pontos Positivos

O modelo captura tendências gerais do engajamento;

Boa interpretabilidade dos coeficientes;

Processamento eficiente dos dados.

Limitações Identificadas

R^2 moderado (0.50) indica explicação parcial da variância;
Pressupostos de linearidade podem não capturar relações complexas;
Possível impacto de outliers nos resultados.

Desafios Técnicos
Tratamento de diferentes formatos de dados.

Conclusão e Trabalhos Futuros
Principais Aprendizados:

Insights Técnicos:
Importância do pre processamento adequado;
Necessidade de balance entre simplicidade e performance;
Valor da análise exploratória detalhada.

Insights do Negócio:
Relação complexa entre métricas e engajamento;
Importância de fatores não quantitativos;
Variabilidade no comportamento dos dados;

Sugestões de Melhorias:
Aprimoramentos no Pre processamento;
Implementação de técnicas de detecção de outliers;
Normalização/padronização das variáveis;
Engenharia de features mais sofisticada.

Expansão do Modelo:
Teste de modelos não-lineares:
Random Forests
Gradient Boosting
Redes Neurais

Melhorias na análise
Segmentação por nichos de influenciadores;
Análise temporal do engajamento;
Incorporação de variáveis qualitativas.

Desenvolvimento futuro:

Desenvolvimento de API para previsões em tempo real;

Interface visual para análise interativa;

Integração com outras fontes de dados.

Aplicações Práticas:

Sistema de recomendação para marcas;

Ferramenta de análise para agências;

Dashboard de monitoramento de engajamento.