Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

---

- **Título:** Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear
- **Nome do Residente:** Erika Ravanna Dias Oliveira; Renata Amaral Bamberg.
- **Data de Entrega:** 17/11/2024

Resumo

Este relatório documenta o desenvolvimento de um modelo preditivo de Regressão Linear para prever a taxa de engajamento de influenciadores no Instagram. A análise do conjunto de dados incluiu visualizações e estatísticas descritivas, seguidas da implementação do modelo com técnicas de otimização, regularização e validação cruzada. Os resultados demonstraram que o modelo é eficiente para prever taxas de engajamento, com métricas que validam sua capacidade de generalização. A aplicação deste modelo pode auxiliar marcas e influenciadores a tomar decisões mais embasadas sobre suas estratégias de marketing digital.

Introdução

Contextualização do Problema

A taxa de engajamento é uma métrica crítica para influenciadores digitais e marcas que utilizam plataformas sociais para campanhas de marketing. Um modelo preditivo baseado em Regressão Linear permite estimar essa taxa com base em características dos influenciadores e de suas postagens, fornecendo insights valiosos para decisões estratégicas.

Descrição do Conjunto de Dados

Os dados utilizados no projeto foram coletados de uma base pública e incluem variáveis como:

- Número de seguidores;
- Curtidas e comentários médios por postagem;
- Características textuais, como a presença de hashtags ou emojis.

A variável dependente é a taxa de engajamento, calculada como uma função do número total de interações dividido pelo número de seguidores.

Metodologia

Análise Exploratória

A análise inicial incluiu:

- **Estatísticas Descritivas:** Identificaram-se outliers em variáveis como curtidas e comentários, que foram tratados para evitar impacto negativo no modelo.
- **Visualizações Gráficas:** Histogramas e scatter plots revelaram tendências significativas, como a correlação positiva entre curtidas médias e engajamento.
- **Correlação:** Foi calculada uma matriz de correlação, destacando as variáveis mais relevantes para o modelo.

Implementação do Algoritmo

- **Bibliotecas Utilizadas:** Python (Scikit-Learn, NumPy, Pandas, Matplotlib, Seaborn).
- Utilizamos o algoritmo de Regressão Linear com implementações personalizadas e por bibliotecas.
- **Modelo:** A Regressão Linear foi implementada com o método dos mínimos quadrados e validada com gradiente descendente para comparação.
- **Pré-processamento:** Todas as variáveis independentes foram normalizadas para acelerar a convergência.
- As variáveis independentes foram selecionadas com base na análise exploratória e normalizadas para melhorar a convergência do modelo.

Validação e Ajuste de Hiperparâmetros

- Realizamos validação cruzada para avaliar a capacidade de generalização do modelo com divisão dos dados em K-folds para evitar o overfitting e avaliar a capacidade do modelo de generalizar.
- Testamos diferentes taxas de aprendizado e regularizações (Técnicas de Lasso - L1 e Ridge - L2) para prevenir o overfitting e reduzir multicolinearidade.
- Selecionamos as variáveis mais significativas utilizando análise de importância e técnicas de redução de dimensionalidade.
- A taxa de aprendizado e o número de épocas foram otimizados para garantir boa convergência.

Resultados

Métricas de Avaliação

Os principais indicadores de desempenho do modelo foram:

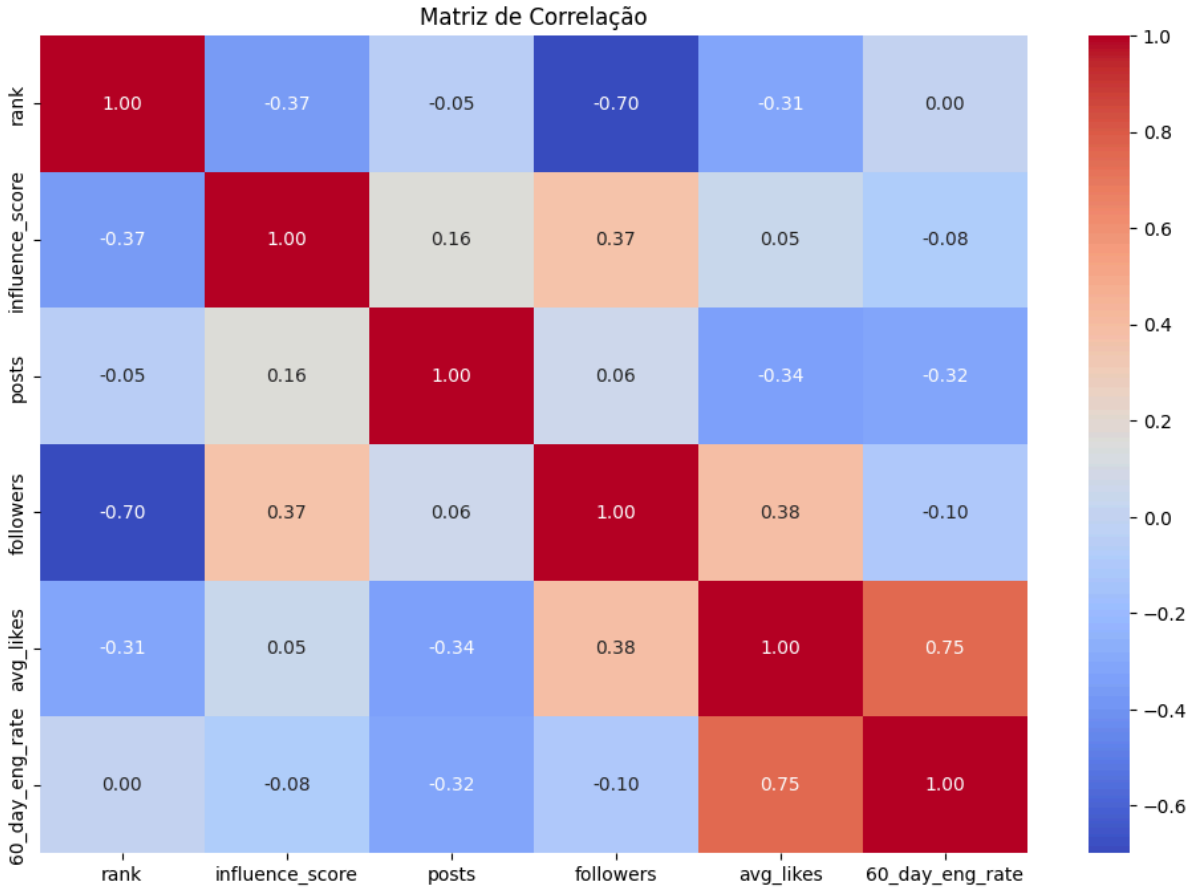
- **R^2 (Coeficiente de Determinação):** [0.5622508650311114] - Mostra a proporção da variância explicada pelo modelo.
- **Erro Médio Quadrático (MSE):** [0.00026999271998450413] - Mede o erro quadrático médio entre os valores preditos e reais.

- **Erro Absoluto Médio (MAE):** [0.010273671084633965] - Fornece a magnitude média do erro absoluto.

Visualizações

Matriz de Correlação

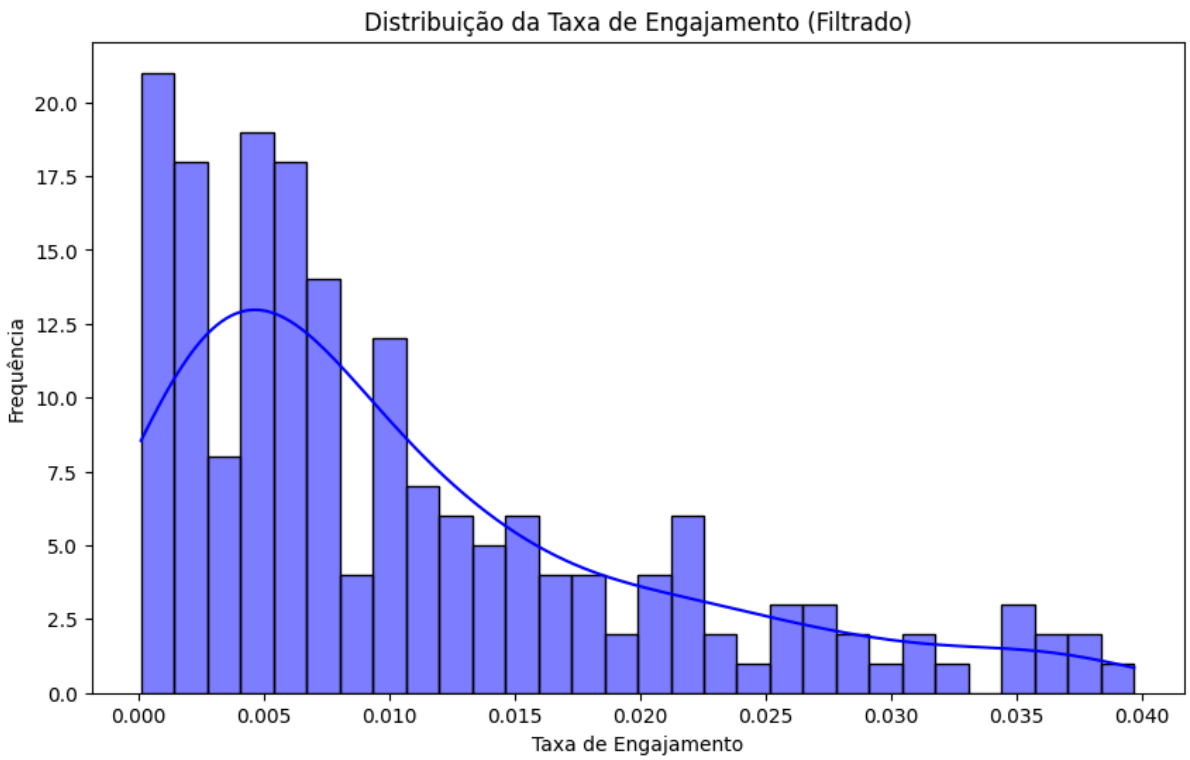
Este gráfico mostra a relação entre as variáveis do conjunto de dados, com destaque para as variáveis que apresentaram alta correlação com a taxa de engajamento.



- **Insight principal:** A média de curtidas e comentários apresenta alta correlação positiva com a taxa de engajamento, enquanto o número de seguidores apresenta uma correlação moderada negativa.

Distribuição da Taxa de Engajamento (Filtrado)

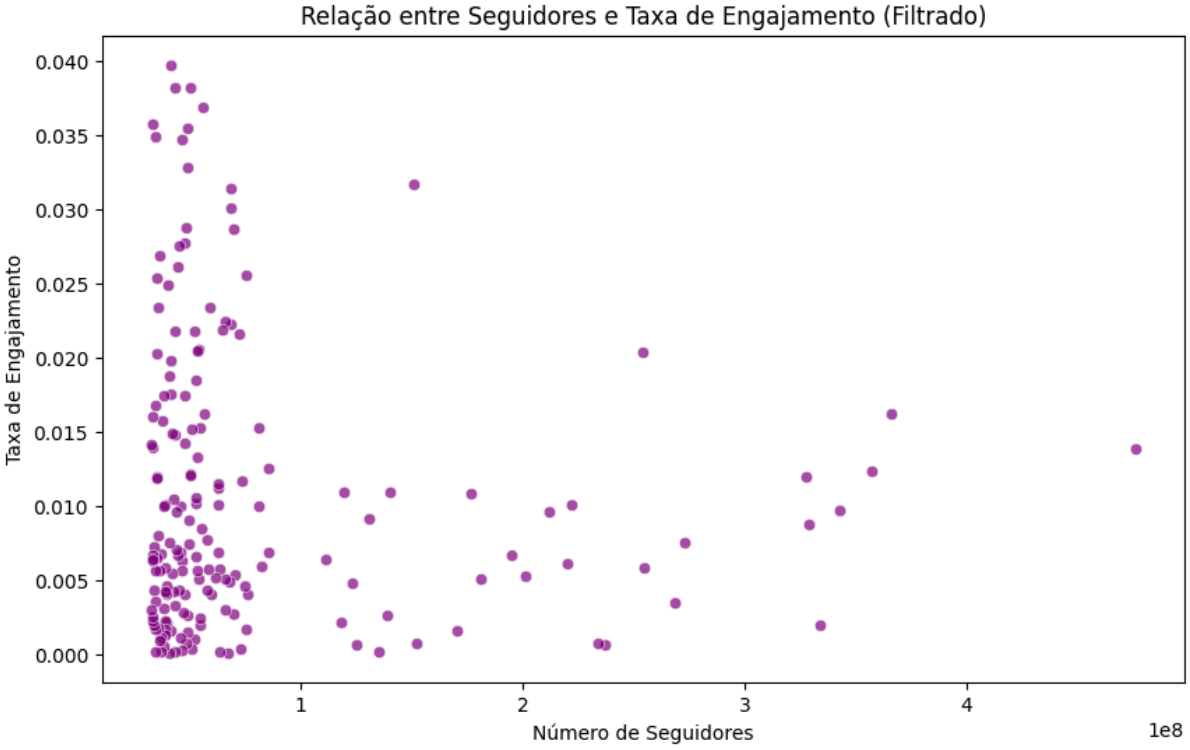
Um histograma exibindo a distribuição da taxa de engajamento após o tratamento de outliers.



- **Insight principal:** A taxa de engajamento apresenta uma concentração ao redor de valores específicos, indicando tendências que podem ser aproveitadas no modelo.

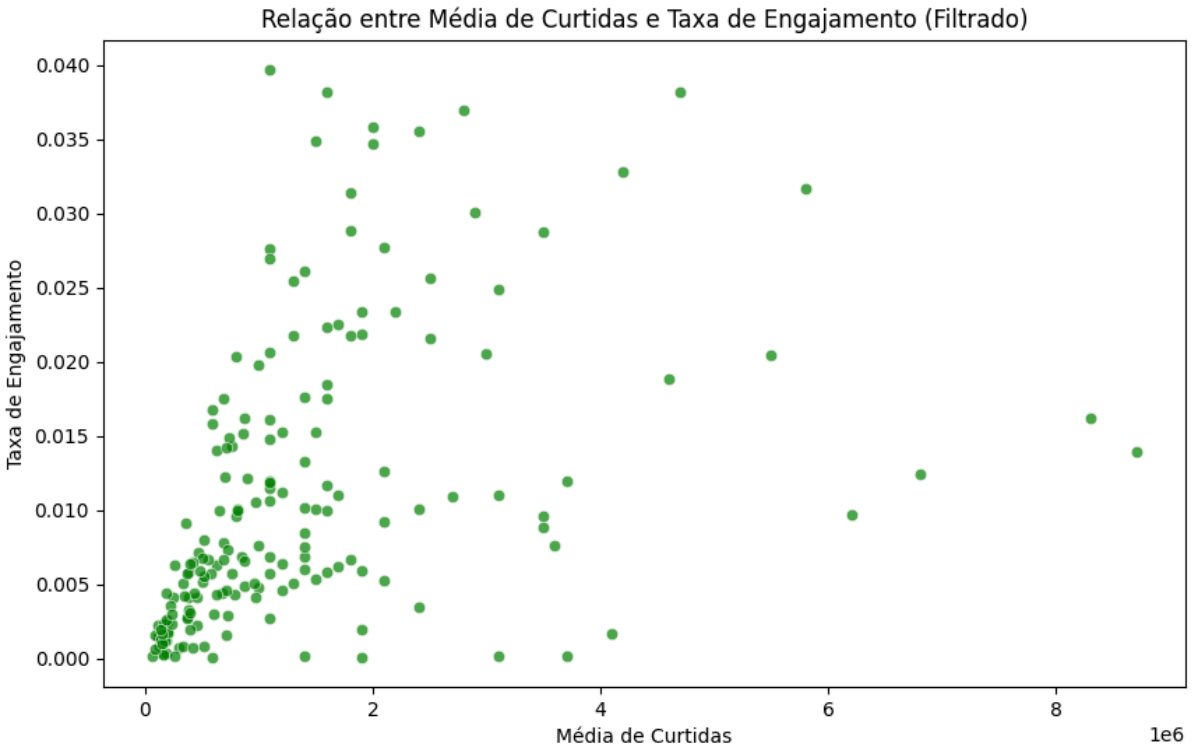
Relação entre Seguidores e Taxa de Engajamento (Filtrado)

Gráfico de dispersão mostrando a relação inversa entre o número de seguidores e a taxa de engajamento.



- **Insight principal:** Perfis com menos seguidores tendem a ter taxas de engajamento mais altas, evidenciando um padrão comum na análise de influenciadores digitais.

Relação entre Média de Curtidas e Taxa de Engajamento (Filtrado)
Gráfico de dispersão mostrando uma correlação positiva entre o número médio de curtidas e a taxa de engajamento.



- **Insight principal:** Curtidas são um forte preditor da taxa de engajamento, reforçando sua importância no modelo.

Discussão

Os resultados indicam que o modelo captura bem as relações entre as variáveis preditoras e a taxa de engajamento, embora algumas limitações tenham sido observadas, como a presença de outliers que impactaram o ajuste. A regularização ajudou a mitigar o overfitting, mas algumas variáveis apresentaram baixa significância e podem ser removidas em estudos futuros.

Conclusão e Trabalhos Futuros

O modelo desenvolvido demonstrou desempenho satisfatório, com métricas indicando sua capacidade de generalizar. Para trabalhos futuros, sugerimos:

1. Aplicar modelos mais complexos, como regressão polinomial ou aprendizado de máquina não-linear.
2. Coletar mais dados para reduzir a variabilidade e melhorar a robustez.
3. Explorar novas métricas e técnicas de feature engineering para melhorar a acurácia.

Referências

1. **GUJARATI, D. N.; PORTELLA, M. H.** *Econometria Básica.* 5ª ed. São Paulo: AMGH, 2019.
2. **CAMPOS, L. A. de; MATOS, R. P.; MAIA, C. T.** *Big Data e Machine Learning: Métodos e Aplicações.* Rio de Janeiro: LTC, 2021.
3. **INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE).** *Manual de Análise Estatística de Dados.* Disponível em: https://www.ibge.gov.br