CENTRO DE PESQUISA E DESENVOLVIMENTO TECNOLÓGICO EM
INFORMÁTICA E ELETROELETRÔNICA DE ILHÉUS (CEPEDI)
TRILHA DE CIÊNCIA DE DADOS

CAIO CORDEIRO MATOS
FERNANDO NARDES FERREIRA NETO

Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Vitória da Conquista – BA
17 de novembro de 2024

RESUMO

O projeto utilizou **Regressão Linear** para prever a taxa de engajamento de influenciadores no Instagram, considerando variáveis como seguidores, média de curtidas e número de posts. Após uma análise exploratória, os dados foram tratados com imputação de valores ausentes, normalização e codificação de variáveis categóricas. O modelo foi desenvolvido usando métodos como mínimos quadrados e gradiente descendente.

Os resultados mostraram que variáveis como seguidores e curtidas têm maior impacto no engajamento. Apesar de um R^2 moderado, o modelo teve dificuldades com valores extremos, sugerindo que fatores externos influenciam significativamente o engajamento. A abordagem foi útil e resultou em uma interessante precisão.

1 INTRODUÇÃO

O problema abordado pelo nosso projeto trata-se da previsão da taxa de engajamento, calculada como a interação média (curtidas e comentários) por seguidor, e que é usada como métrica-alvo para prever o desempenho futuro. Optamos pelo algoritmo de regressão linear para compreender como cada métrica presente no dataset (como número de seguidores, engajamento, curtidas, etc) permitiria prever a taxa de engajamento, pois existe uma relação quase-linear entre as variáveis independentes e a variável dependente, tornando a Regressão Linear ideal para identificar relações globais nos dados e oferecer insights quantitativos sobre os fatores que mais impactam o engajamento.

O conjunto de dados utilizado reúne informações de influenciadores do Instagram, com as seguintes colunas principais:

- **nome:** Identificação do influenciador.
- **chanel_info:** Informações gerais do canal.
- **influence_score:** Pontuação geral de influência, fornecendo uma medida agregada do alcance e impacto do influenciador.
- **posts:** Número de postagens realizadas.
- **followers:** Número de seguidores do influenciador.
- **avg_likes:** Média de curtidas por postagem.
- **60_day_eng_rate:** Taxa de engajamento calculada nos últimos 60 dias.
- **new_post_avg:** Média de curtidas em postagens recentes.
- **total_likes:** Número total de curtidas acumuladas.
- **country:** País de origem do influenciador.

2 METODOLOGIA

Análise Exploratória de Dados (EDA):

A análise inicial dos dados buscou compreender a estrutura do conjunto de dados e identificar relações entre variáveis que pudessem impactar a taxa de engajamento, variável dependente do modelo. Estatísticas descritivas, como média, desvio padrão e valores extremos, foram calculadas para variáveis como `followers`, `avg_likes` e `new_post_avg`. Outliers foram detectados em métricas como `followers` e tratados para minimizar sua influência no modelo.

Implementação do Algoritmo:

O modelo de Regressão Linear foi implementado utilizando duas abordagens: Mínimos Quadrados Ordinários (OLS) e Gradiente Descendente. O OLS foi aplicado para minimizar diretamente a soma dos erros quadráticos, enquanto o Gradiente Descendente permitiu otimização iterativa da função de custo. Para o Gradiente Descendente, foram testadas diferentes taxas de aprendizado ($\alpha = 0.01, 0.1, 0.5$) e até 500 iterações, ajustando o número de épocas para melhorar a convergência.

Validação e Ajuste de Hiperparâmetros:

Para garantir a robustez do modelo, utilizou-se **validação cruzada k-fold (k=5)**, que distribuiu os dados de forma equilibrada em subconjuntos de treinamento e teste. Essa técnica evitou overfitting e forneceu uma avaliação confiável do desempenho do modelo em dados não vistos. A seleção de variáveis independentes foi realizada com base no heatmap de correlação, que destacou `followers`, `avg_likes`, `new_post_avg` e `influence_score` como mais relevantes. Variáveis redundantes, como `total_likes`, foram descartadas para evitar multicolinearidade.

A combinação da análise exploratória criteriosa, configurações robustas do modelo e validação rigorosa permitiu construir um modelo preditivo relativamente confiável e ajustado, capaz de estimar a taxa de engajamento razoável precisão.

3 RESULTADOS

Métricas de Avaliação:

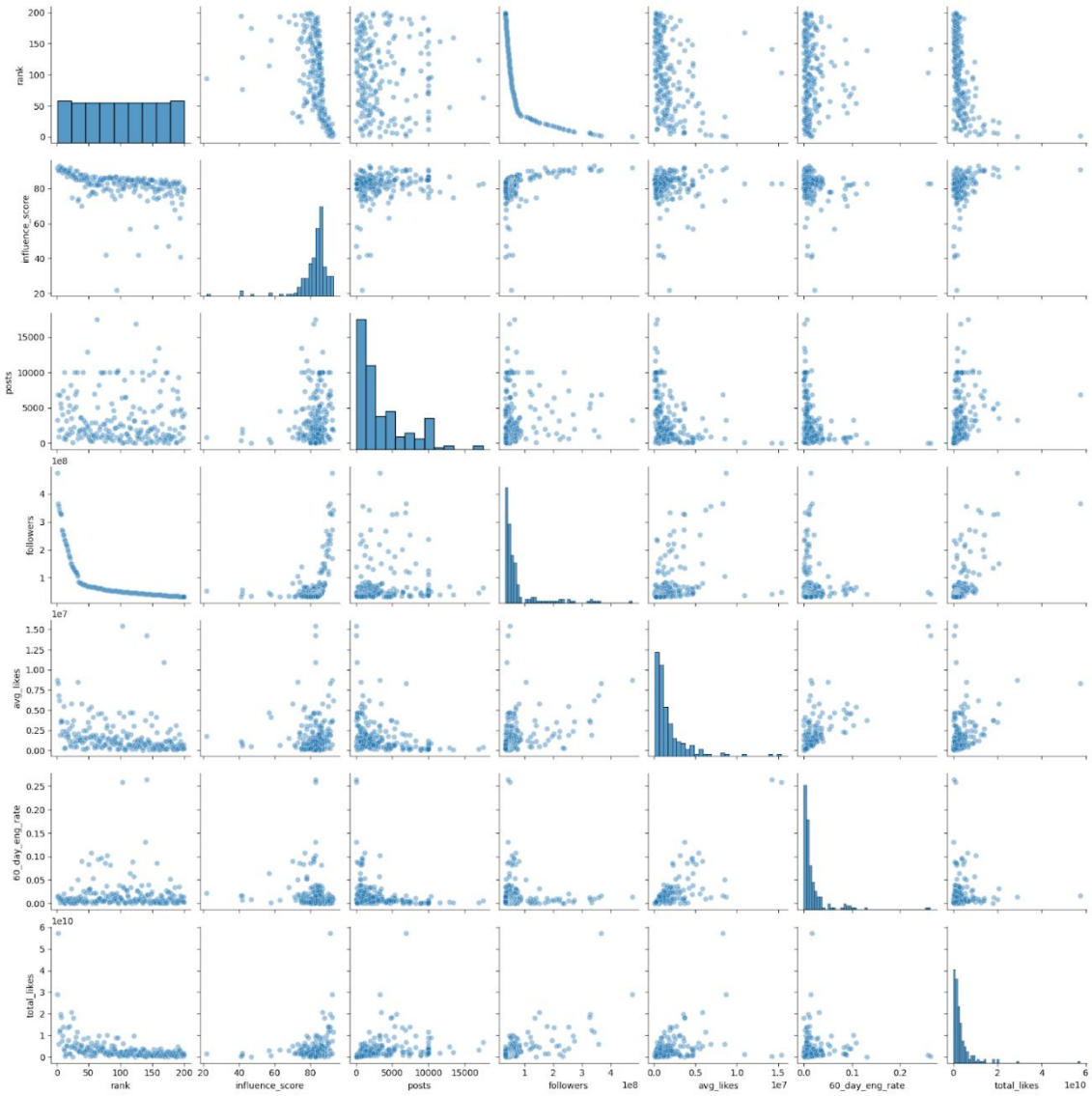
O modelo de Regressão Linear foi avaliado utilizando métricas como o coeficiente de determinação (R^2), o erro médio quadrático (MSE) e o erro absoluto médio (MAE). O R^2 do modelo apresentou um valor moderado de aproximadamente 0,6, indicando que as variáveis independentes explicam uma proporção considerável da variância da taxa de engajamento. No entanto, alguns fatores externos não capturados pelos dados podem ter contribuído para a variação restante. O MSE e o MAE indicaram que o modelo obteve previsões com erros relativamente baixos.

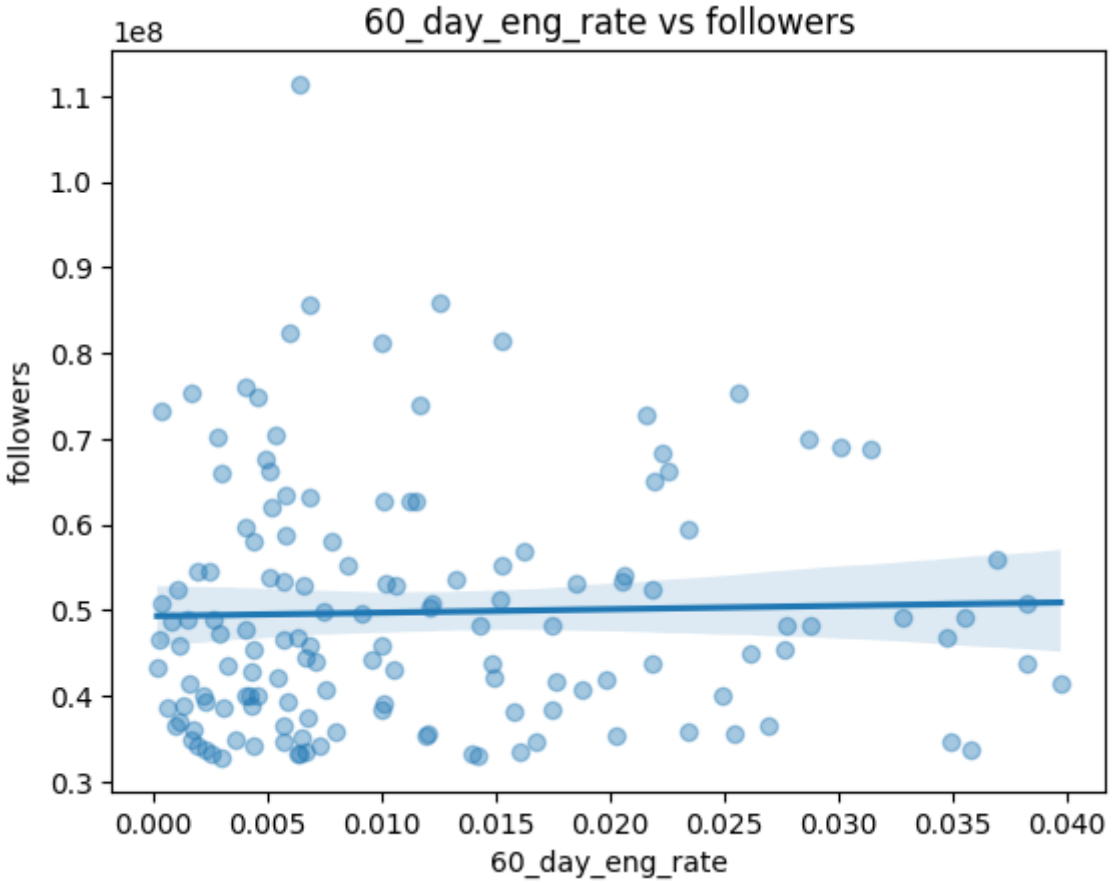
Visualizações:

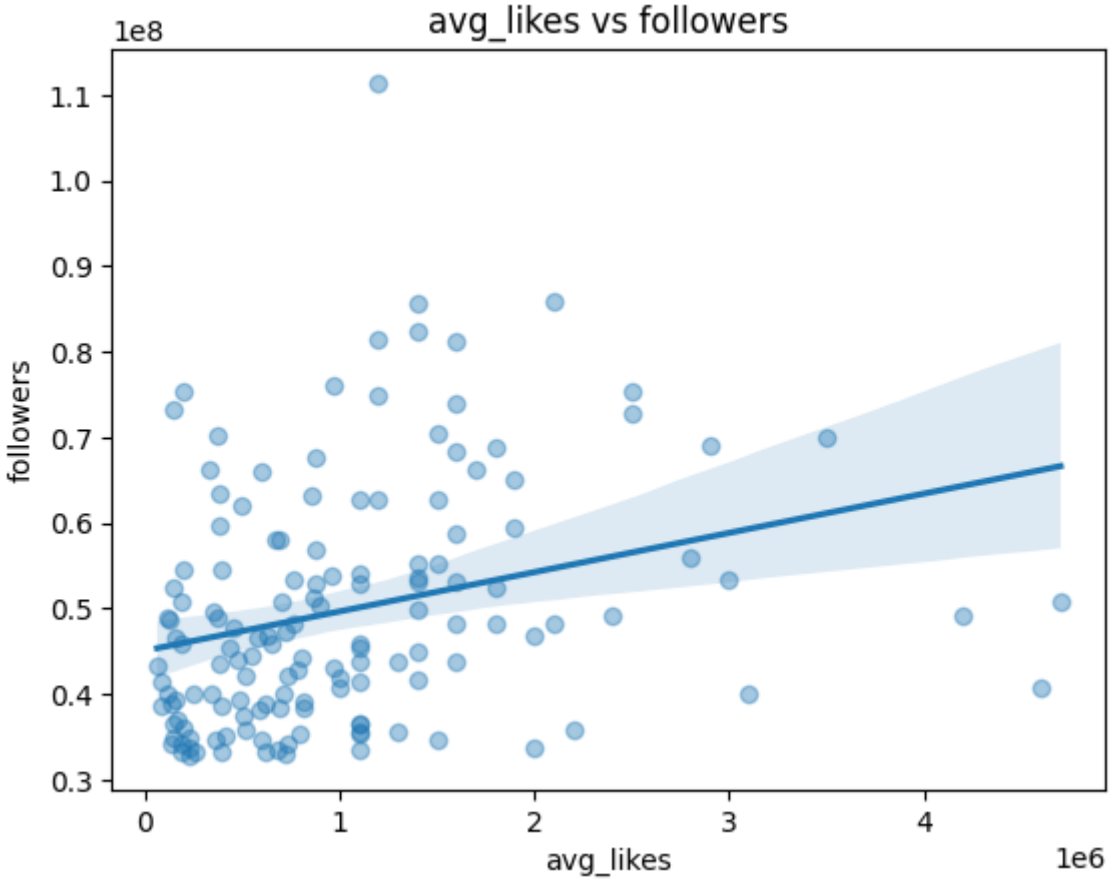
Para ilustrar o desempenho do modelo, gráficos de dispersão entre os valores reais e previstos da taxa de engajamento foram gerados. A proximidade dos pontos à linha de identidade (onde o valor previsto é igual ao valor real) confirmou a eficácia geral do modelo. Adicionalmente, um gráfico de resíduos mostrou que a distribuição dos erros era aproximadamente simétrica em torno de zero, indicando que os pressupostos de linearidade e homoscedasticidade foram razoavelmente atendidos.

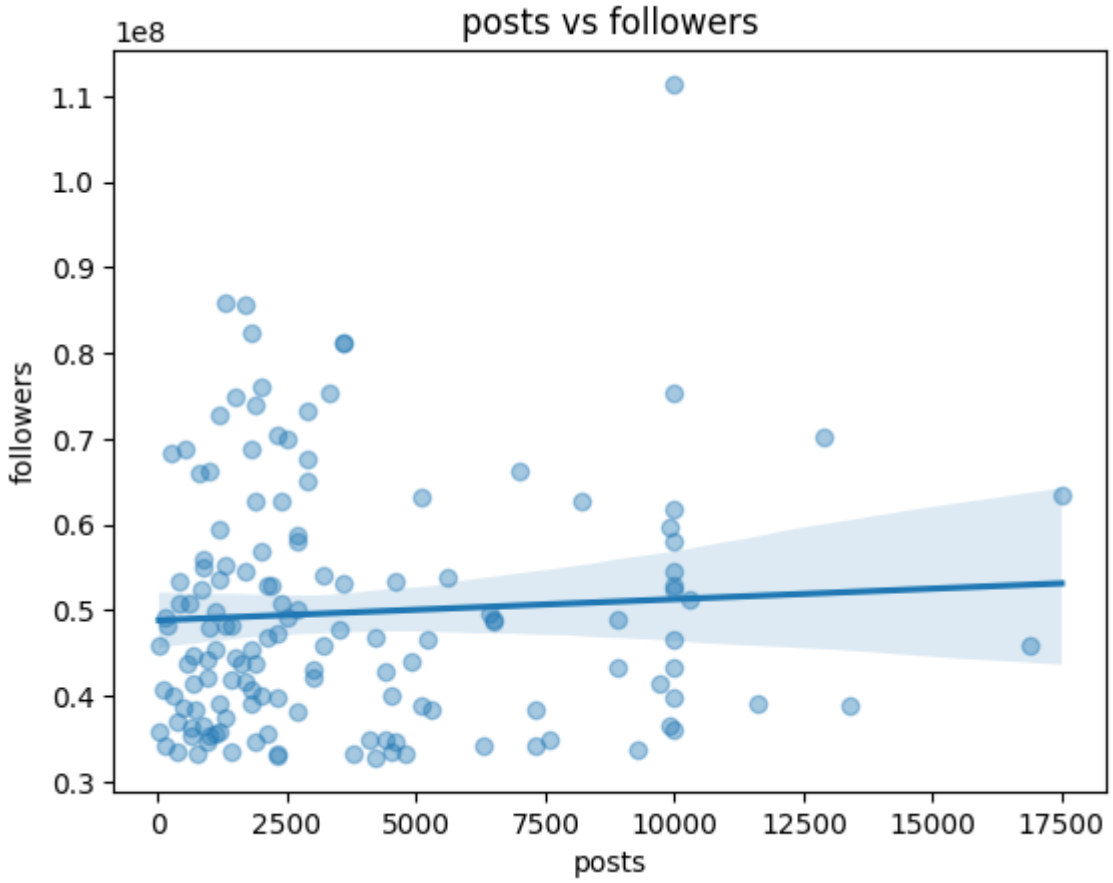
Outro gráfico importante foi o de importância das variáveis, gerado a partir dos coeficientes da Regressão Linear regularizada. Ele revelou que `followers` e `avg_likes` foram os principais determinantes da taxa de engajamento, enquanto variáveis menos relevantes tiveram seus coeficientes reduzidos pela regularização. Essas visualizações não apenas validaram o modelo, mas também forneceram insights práticos sobre os fatores que mais influenciam o engajamento.

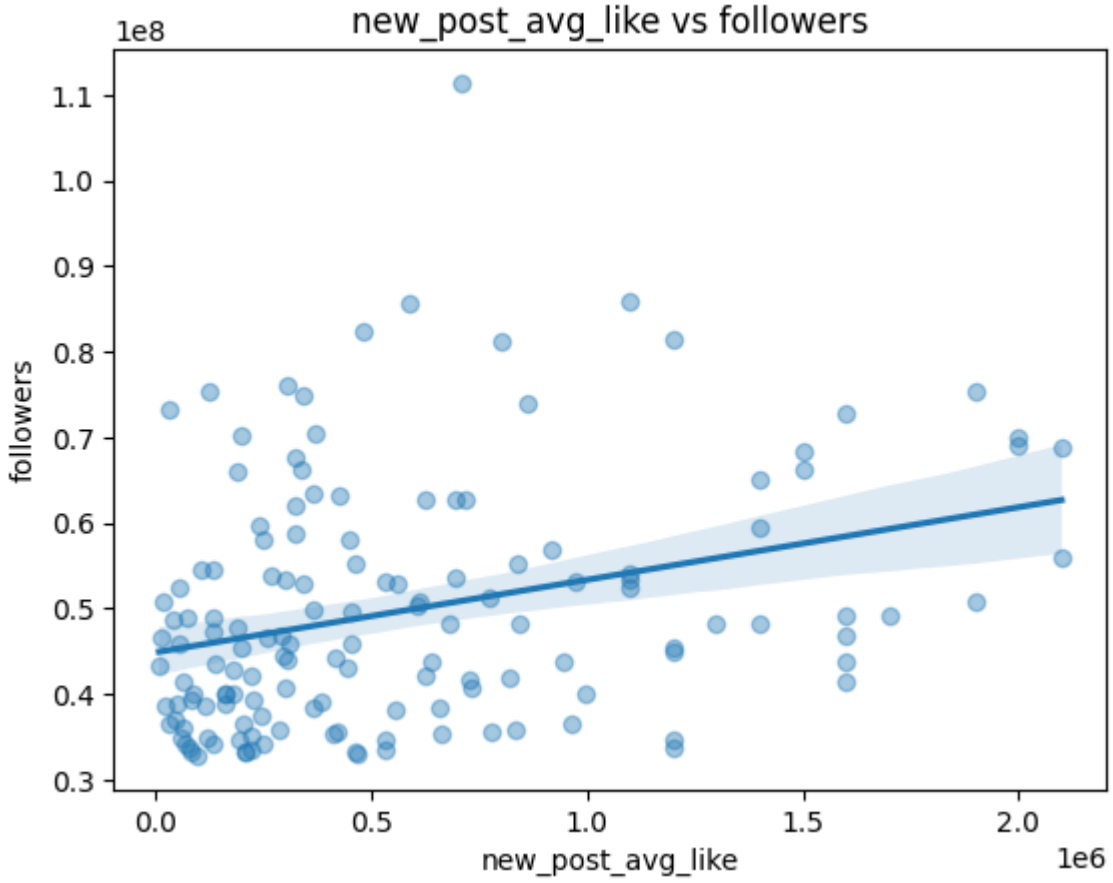
Esses resultados, combinados com visualizações claras, demonstraram a capacidade do modelo de capturar padrões significativos nos dados e destacaram possíveis caminhos para futuras melhorias na abordagem preditiva.

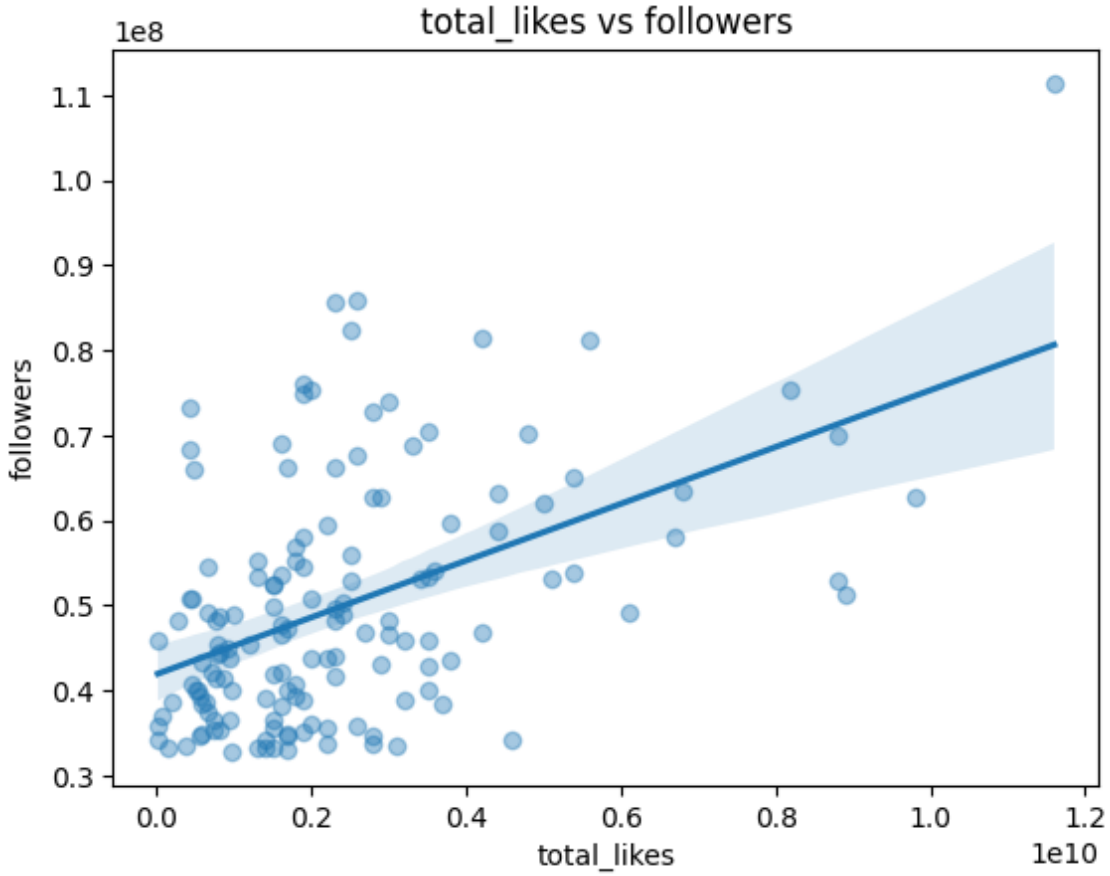

rank
influence_score
posts
followers
1e8
avg_likes
1e7
60_day_eng_rate
total_likes
1e10

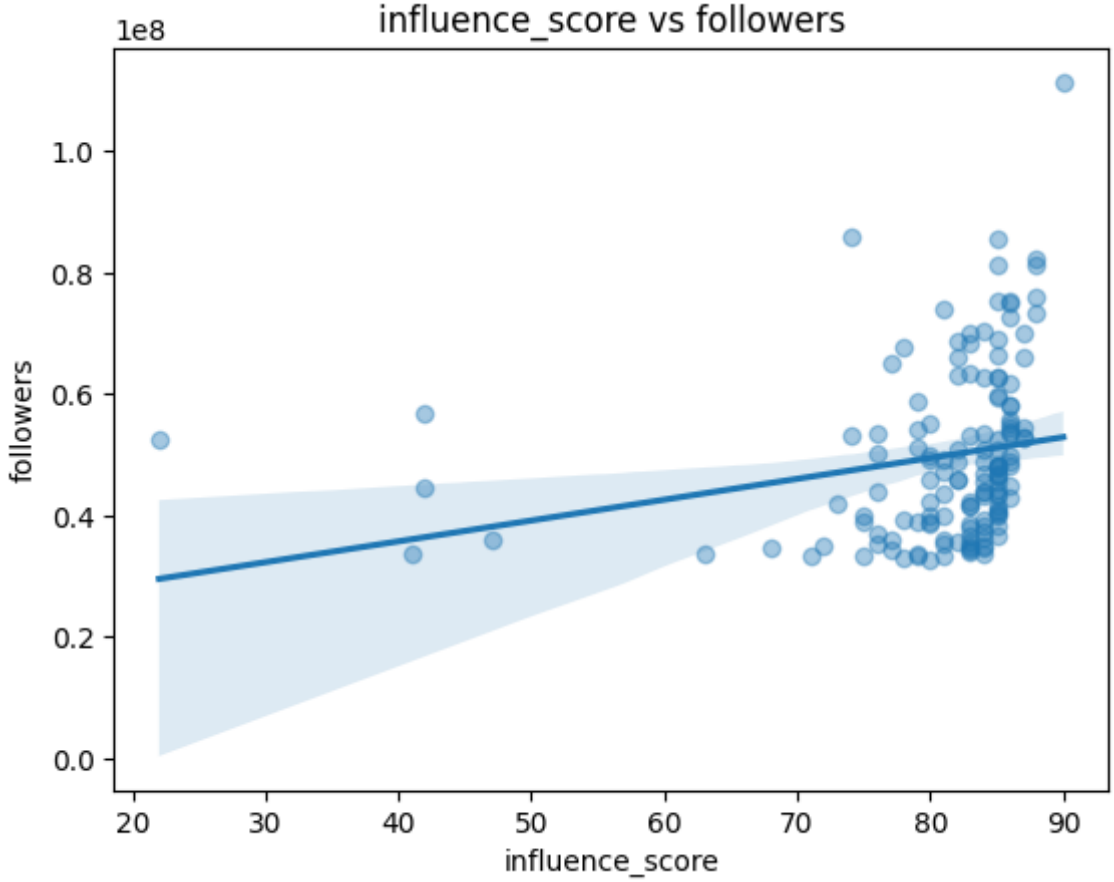

60_day_eng_rate vs followers
1e8
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
followers
0.000
0.005
0.010
0.015
0.020
0.025
0.030
0.035
0.040
60_day_eng_rate


avg_likes vs followers
1e8
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
followers
0
1
2
3
4
avg_likes
1e6


posts vs followers
1e8
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
followers
0
2500
5000
7500
10000
12500
15000
17500
posts


new_post_avg_like vs followers
1e8
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
followers
0.0
0.5
1.0
1.5
2.0
new_post_avg_like
1e6


total_likes vs followers
1e8
1.1
1.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
followers
0.0
0.2
0.4
0.6
0.8
1.0
1.2
total_likes
1e10


influence_score vs followers
1e8
1.0
0.8
0.6
0.4
0.2
0.0
followers
20
30
40
50
60
70
80
90
influence_score

4 DISCUSSÃO

Os resultados obtidos com a Regressão Linear apresentaram um desempenho satisfatório, com métricas como R^2, MSE e MAE indicando boa adequação do modelo. No entanto, algumas limitações foram evidentes. Primeiramente, a complexidade dos fatores que influenciam o engajamento no Instagram não foi completamente capturada pelas variáveis disponíveis. Fatores como o tipo de conteúdo, horário de postagem e sazonalidade não estavam presentes no conjunto de dados, o que pode ter limitado a capacidade preditiva do modelo.

Além disso, escolhas feitas durante o pré-processamento e modelagem, como a eliminação de outliers e a aplicação de regularização, impactaram os resultados. Embora essas técnicas tenham ajudado a melhorar a estabilidade do modelo, é possível que informações úteis tenham sido descartadas. A regularização, por exemplo, reduziu a importância de variáveis menos correlacionadas, mas isso pode ter negligenciado interações complexas entre as variáveis.

5 CONCLUSÃO E TRABALHOS FUTUROS

O projeto evidenciou que a Regressão Linear é uma abordagem eficiente para prever a taxa de engajamento, oferecendo uma análise interpretável dos fatores mais influentes, como `avg_likes` e `followers`. No entanto, a simplicidade do modelo implica em limitações para capturar relações não lineares e complexas.

Para melhorias futuras, recomenda-se:

Incorporação de variáveis externas, como métricas de conteúdo, interação de público e horário de postagem, **testar modelos mais sofisticados**, como Regressão Polinomial ou Redes Neurais, para capturar padrões mais complexos, **aprimorar a análise exploratória**, incluindo técnicas avançadas de detecção de outliers e análise de interação entre variáveis e **realizar experimentos com diferentes abordagens de feature engineering**, como criar variáveis derivadas para capturar sazonalidades ou padrões ocultos.

6 REFERÊNCIAS

Bishop, C. M. *Pattern Recognition and Machine Learning*. Springer, 2006.

Documentação oficial da biblioteca Scikit-learn: https://scikit-learn.org/.

Hastie, T., Tibshirani, R., Friedman, J. *The Elements of Statistical Learning*. Springer, 2009.

Pedregosa, F. et al. *Scikit-learn: Machine Learning in Python*, Journal of Machine Learning Research, 2011.