Fase 1 – Capacitação Tecnológica

Trilha 3 — Ciência de Dados

Relatório Técnico:

Implementação e Análise do Algoritmo de Regressão Linear

Luana Oliveira da Silva - polo Juazeiro
Elai Emylle Matos de Lima - polo Vitória da Conquista

17 de novembro de 2024

RESIDÊNCIA EM SOFTWARE
BAHIA + TECNOLOGIA + EMPREENDEDORISMO

INSTITUIÇÃO EXECUTORA cepedi
PARCEIRA GOVERNO DO ESTADO BAHIA SECRETARIA DO TRABALHO, EMPREGO, RENDA E ESPORTE
COORDENADORA Softex
APOIO MINISTÉRIO DA CIÊNCIA, TECNOLOGIA E INOVAÇÃO
GOVERNO FEDERAL BRASIL UNIÃO E RECONSTRUÇÃO

Resumo

Este relatório apresenta o desenvolvimento e análise de um modelo de Regressão Linear para prever a taxa de engajamento dos principais influenciadores no Instagram. O estudo utilizou o conjunto de dados "Top Instagram Influencers Data", que contém 200 registros e 10 atributos, como número de seguidores, média de curtidas e total de curtidas acumuladas.

Após o pré-processamento e a análise exploratória, as variáveis independentes selecionadas foram **avg_likes** e **new_post_avg_like**, enquanto a variável dependente escolhida foi **60_day_eng_rate**, que representa a taxa de engajamento dos últimos 60 dias. O modelo foi avaliado usando métricas como R^2, MAE, MSE e RMSE, além de gráficos de correlação e pairplots para entender melhor as relações entre as variáveis. Os resultados fornecem insights valiosos sobre os fatores que mais impactam o engajamento, destacando a média de curtidas recentes como uma das principais influências.

Sumário

1. Introdução

No marketing de influência, a avaliação do impacto de um influenciador vai além de métricas básicas, como curtidas, compartilhamentos ou até o número total de seguidores. Com isso, a taxa de engajamento destaca-se como uma medida mais robusta de sucesso, pois reflete o grau de interação genuína do público com o conteúdo, indo além do alcance bruto das postagens. Dessa forma, conforme o Relatório de Referência do Setor de Mídia Social de 2024 do Rival IQ (Rival IQ, 2024), a taxa de engajamento por postagem se torna uma métrica ideal, ajustando-se ao volume de postagens e ao tamanho do público do influenciador, permitindo uma análise proporcional e qualitativa do impacto. Assim, entender e prever essa taxa torna-se essencial para campanhas de marketing mais eficazes e estratégias bem direcionadas, fornecendo um indicador real do envolvimento e influência reais que cada post gera.

Sendo assim, define-se taxa de engajamento como um indicador que mede a aceitação de um conteúdo digital pela audiência, avaliando seu impacto por meio das interações realizadas na postagem, como curtidas, comentários e compartilhamentos. Esse cálculo consiste na divisão do total de reações pelo alcance ou impressões, indicando se o conteúdo foi relevante o suficiente para engajar o público e influenciar o algoritmo a entregá-lo a mais usuários (Dantas, 2020). A autora ainda explica que, “O instagram entende engajamento como a somatória de curtidas, comentários, encaminhamentos e salvamentos em relação ao alcance”. Por isso, é fundamental que essa métrica seja calculada com precisão, evitando erros de inferência que possam distorcer a interpretação do impacto real do conteúdo.

Para o propósito de analisar e inferir a taxa de engajamento dos influenciadores, a técnica de regressão linear é bem recomendada. Suas propriedades bem conhecidas e o processo de treinamento rápido tornam a regressão linear uma ferramenta consolidada, com aplicação prática em problemas de inferência e predição (IBM, 2024). No contexto deste projeto, a regressão linear se destaca pela capacidade de modelar relações proporcionais entre variáveis e gerar insights sobre os fatores que influenciam a taxa de engajamento.

Nesse contexto, este projeto tem o objetivo de desenvolver um modelo preditivo usando o algoritmo de Regressão Linear para resolver um problema de inferência sobre taxa de engajamento dos principais influenciadores do instagram listados no conjunto de dados "*Top Instagram Influencers Data*". Este contém 10 atributos organizados com base no ranking dos influenciadores, determinado pelo número de seguidores. Abaixo está a descrição de cada atributo:

- **rank**: Posição do influenciador, ordenada com base na quantidade de seguidores.
- **channel_info**: Nome de usuário do influenciador no Instagram.
- **influence_score**: Pontuação de influência do usuário, calculada com base em menções, importância e popularidade.
- **posts**: Quantidade de postagens realizadas até o momento.
- **followers**: Número total de seguidores do influenciador.
- **avg_likes**: Média de curtidas nas postagens (total de curtidas dividido pelo total de postagens).
- **60_day_eng_rate**: Taxa de engajamento dos últimos 60 dias, representando a fração de interações nas postagens recentes.
- **new_post_avg_like**: Média de curtidas nas postagens mais recentes.
- **total_likes**: Total de curtidas acumuladas nas postagens do usuário (em bilhões).
- **country**: País ou região de origem do influenciador.

O conjunto de dados possui 200 registros e 10 colunas, conforme descrito a seguir na Tabela 1:

Tabela 1 – Descrição do conjunto de dados

Coluna	Tipo de Dados	Contagem de Valores Não-Nulos	Descrição
rank	int64	200	Ranking do influenciador com base no número de seguidores.
channel_info	object	200	Nome de usuário do influenciador no Instagram.
influence_score	int64	200	Pontuação de influência, baseada em menções, importância e popularidade.
posts	object	200	Quantidade de postagens realizadas.

followers	object	200	Número total de seguidores do influenciador.
avg_likes	object	200	Média de curtidas nas postagens.
60_day_eng_rate	object	200	Taxa de engajamento dos últimos 60 dias.
new_post_avg_like	object	200	Média de curtidas nas postagens mais recentes.
total_likes	object	200	Total de curtidas acumuladas em todas as postagens.
country	object	138	País ou região de origem do influenciador.

Os dados contêm uma mistura de variáveis numéricas e categóricas, totalizando aproximadamente 15,8 KB em memória. Observa-se que apenas a coluna `country` possui valores ausentes, com 138 registros preenchidos.

2. Metodologia

2.1 Preparação dos dados

Como pôde ser observado na Tabela 1 acima, há campos numéricos no dataset que tem seu tipo object, sendo necessária a conversão para float ou int, para a realização de operações numéricas.

Os campos `posts`, `followers`, `avg_likes`, `new_post_avg_like` e `total_likes` foram convertidos para inteiro (int), enquanto que `60_day_eng_rate`, como expressa uma porcentagem, foi convertido para ponto flutuante (float). Durante este processo, foi observado que apenas um valor de `60_day_eng_rate` foi inserido como NaN (Not a Number), sendo assim, esse valor foi dado como 0.

Além disso, como a coluna `country` possui valores ausentes e isso dificulta uma análise por se tratar de uma variável categórica cuja predição se torna imprecisa devido a possibilidade de prever o país em que os influenciadores moram. Como o objetivo é prever a taxa de engajamento das contas, a variável `country` não se mostra útil para a previsão, pois não tem uma correlação direta com o

comportamento de engajamento nas redes sociais, o que torna irrelevante para a modelagem.

Portanto, essa variável pode ser descartada do modelo para evitar complicações e melhorar a qualidade da análise. Em vez disso, seria mais benéfico focar em variáveis `influence_score`, `posts`, `followers`, `avg_likes`, `new_post_avg_like` e `total_likes`, que impactam diretamente a taxa de engajamento, pois elas têm uma correlação mais forte com o objetivo de prever o `60_day_eng_rate`. Realizando esse pré-processamento adequado, podemos garantir que o modelo use apenas as variáveis mais relevantes, resultando em previsões mais precisas.

2.2 Análise exploratória

Para obter-se um entendimento inicial sobre o conjunto de dados, buscamos identificar padrões, anomalias, relações entre variáveis e características gerais dos dados através de técnicas estatísticas e visualizações.

Desse modo, foi realizado um resumo estatístico com o método `df.describe()`, da biblioteca Pandas. Ele calcula e exibe automaticamente estatísticas resumidas para cada coluna numérica. Essas estatísticas incluem:

- **count**: A quantidade de valores não nulos na coluna.
- **mean**: A média dos valores na coluna.
- **std**: O desvio padrão, que mede a dispersão ou variação dos dados.
- **min**: O menor valor na coluna.
- **25%**: O 25º percentil (primeiro quartil), que é o valor abaixo do qual 25% dos dados se encontram.
- **50%**: 50º percentil (mediana), que é o valor abaixo do qual 50% dos dados se encontram.
- **75%**: O 75º percentil (terceiro quartil), que é o valor abaixo do qual 75% dos dados se encontram.
- **max**: O maior valor na coluna.

A fim de visualizar os dados, foram plotados histogramas, boxplots, gráficos de dispersão e matriz de correlação com o auxílio das bibliotecas `Seaborn` e `Matplotlib` do Python que podem ser vistos nos resultados.

2.3 Modelo de regressão linear

A regressão linear é uma das técnicas mais amplamente utilizadas para análise preditiva, sendo especialmente útil para modelar relações entre variáveis. Esse método baseia-se na estimativa de coeficientes que definem uma equação linear, utilizando uma ou mais variáveis independentes para prever, com maior precisão, o valor da variável dependente. Ao ajustar uma linha reta ou superfície, a regressão linear minimiza as diferenças entre os valores observados e os previstos, tornando-se uma ferramenta eficiente para análise quantitativa (IBM, 2024).

Equação da regressão linear:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Onde:

- y é a variável dependente;
- β_0 é o termo de viés;
- $\beta_1, \beta_2, \ldots, \beta_n$ são os coeficientes dos modelos de parâmetros;
- $x_1, x_2, \ldots, x_n$ são as variáveis independentes;
- ϵ é o erro aleatório.

Assim, sua simplicidade e precisão a tornam ideal para investigar a influência de fatores específicos sobre um fenômeno, como a taxa de engajamento no contexto deste projeto. Em Python, a implementação do modelo de regressão linear é comumente realizada utilizando a classe `LinearRegression`, disponível na biblioteca `scikit-learn`.

Conforme a análise exploratória realizada, definiu-se que a variável dependente para o modelo seria `60_day_eng_rate` e as variáveis independentes seriam duas: `avg_likes` e `new_post_avg_like`, por serem consideradas

relevantes para prever a taxa de engajamento dos influenciadores. Em seguida, os dados foram divididos em dois conjuntos: 80% dos dados foram destinados ao treinamento do modelo, e os 20% restantes, para o teste.

O procedimento incluiu os seguintes passos:

1. **Treinamento do modelo:**

O modelo foi ajustado (método `fit`) ao conjunto de treinamento, onde os coeficientes (β_1 e β_2) e o viés (β_0) foram calculados com base nos dados fornecidos.

2. **Predição dos valores:**

Após o treinamento, as previsões da variável dependente (`60_day_eng_rate`) foram geradas para o conjunto de teste utilizando o método `predict`.

2.4 Validação do modelo

Para avaliar a eficácia do modelo, métricas como o erro médio absoluto (MAE), o erro quadrático médio (MSE) e o coeficiente de determinação (R^2) foram calculadas, permitindo medir a proximidade entre os valores previstos e os valores reais. A performance do modelo foi avaliada utilizando métricas como:

- R² (Coeficiente de Determinação): Mede a proporção de variabilidade explicada pelo modelo.

$$R^2 = 1 - \frac{\sum_{i=1}^{n} |Y_i - \hat{Y}_i|^2}{\sum_{i=1}^{n} |Y_i - \overline{Y}|^2}$$

- MAE (Erro Absoluto Médio): Avalia a magnitude média do erro.

$$MAE = \frac{1}{n} \sum_{i=1}^{n} |Y_i - \hat{Y}_i|$$

- MSE (Mean Squared Médio):

$$MSE = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i|^2$$

- RMSE (Root Mean Squared Error):

$$RMSE = \sqrt{\sum_{i=1}^{n} |Y_i - \hat{Y}_i|^2}$$

2.4.1 Validação cruzada

A validação cruzada é uma técnica utilizada para avaliar a performance de modelos de aprendizado de máquina, especialmente para evitar problemas como o overfitting. Nessa abordagem, o conjunto de dados é dividido em partes menores chamadas de *folds*. O modelo é treinado em uma parte dos dados (conjunto de treino) e validado em outra parte (conjunto de teste), repetindo o processo várias vezes para obter uma avaliação mais robusta.

Para aplicá-lo, foi utilizado o método `K-Fold Cross-Validation`, configurado com os seguintes parâmetros:

- Número de Folds (`n_splits`): 2, dividindo o conjunto de dados em duas partes iguais.
- Embaralhamento dos Dados (`shuffle`)**:** Ativado para garantir que os dados sejam aleatoriamente embaralhados antes de serem divididos em folds, promovendo maior generalização.
- Semente Aleatória (`random_state`): Fixada em 42 para garantir que o embaralhamento seja consistente entre as execuções, promovendo reprodutibilidade nos resultados.

A validação foi realizada com a função `cross_val_score`, que avalia o modelo em cada fold e retorna o coeficiente de determinação (R^2) para cada rodada.

3. Resultados

3.1 Análise exploratória

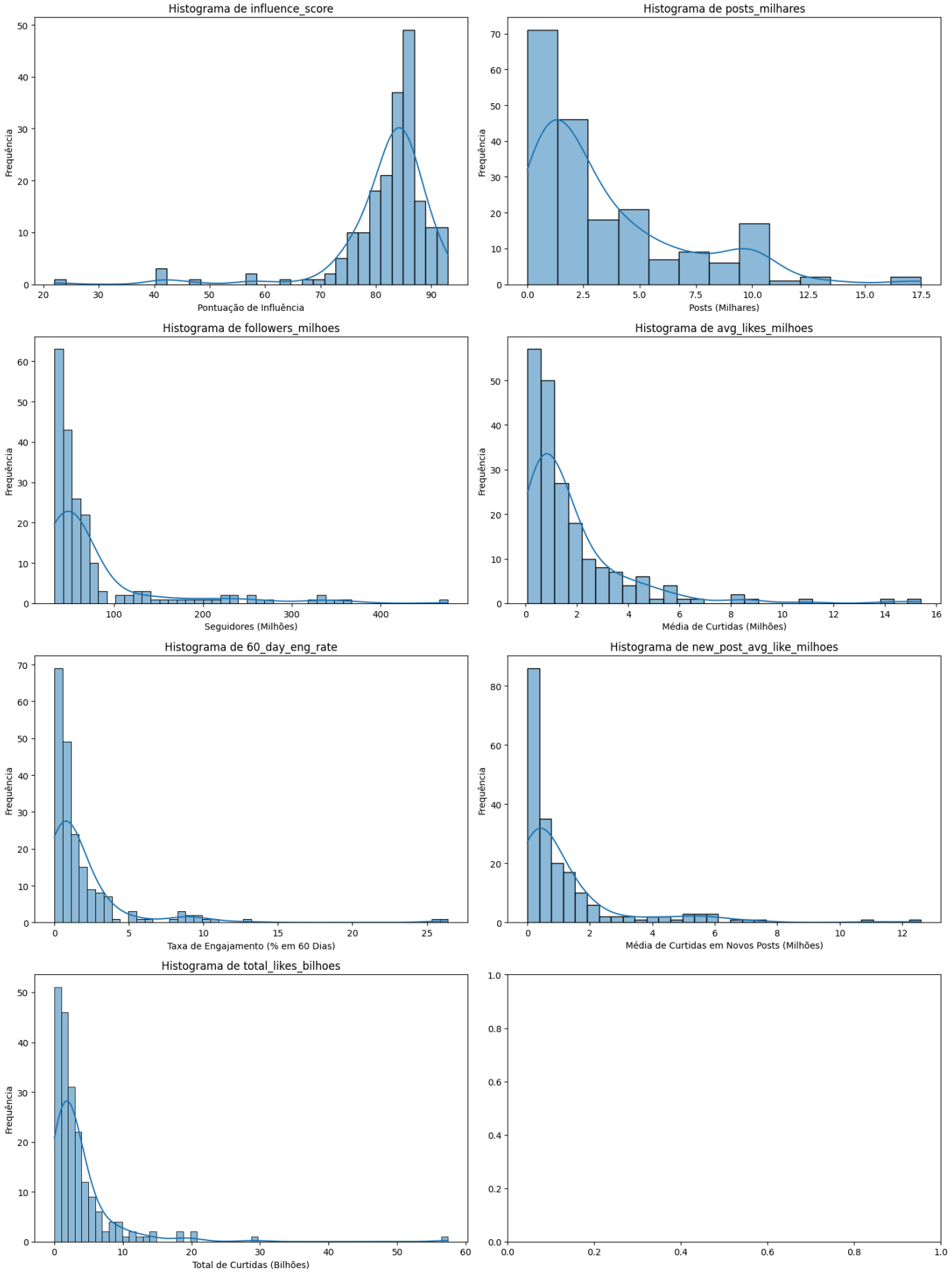
Ao aplicar o `df.describe()`, obteve-se os resultados de acordo com a Tabela 2:

Tabela 2 — Resumo estatístico dos campos numéricos do conjunto de dados

	influence_score	posts (milhares)	followers (milhões)	avg_likes (milhões)	60_day_eng_rate (%)	new_post_avg_like (milhões)	total_likes (bilhões)
count	200.00	200.00	200.00	200.00	200.00	200.00	200.00
mean	81.82	3.50	77.41	1.79	1.89	1.21	3.66
std	8.88	3.48	73.69	2.19	3.32	1.86	5.56
min	22.00	0.01	32.80	0.07	0.00	0.00	0.02
25%	80.00	0.95	40.00	0.50	0.41	0.20	1.00
50%	84.00	2.10	50.05	1.10	0.86	0.53	2.00
75%	86.00	5.03	68.90	2.10	2.03	1.32	3.90
max	93.00	17.50	475.80	15.40	26.41	12.60	57.40

Em seguida foram visualizadas a distribuição dos valores dos campos nos gráficos representados pelas figuras a seguir.

Figura 1 — Histogramas dos campos numéricos



Pontuação de Influência (`influence_score`)

- A distribuição apresenta assimetria negativa (leve concentração na ponta superior da escala).
- A maioria dos influenciadores apresenta pontuações acima de 75, indicando que são figuras com grande impacto.

Quantidade de Posts em Milhares (`posts_milhares`)

- A maioria dos influenciadores tem menos de 5 mil posts, com concentração significativa em menos de 2 mil.
- Há outliers com até 17 mil posts, mas são casos raros.

Seguidores em Milhões (`followers_milhoes`)

- Os valores estão altamente concentrados entre 0 e 100 milhões.
- Existem influenciadores com até 400 milhões de seguidores, mas são raríssimos e podem distorcer análises.

Média de Curtidas em Milhões (`avg_likes_milhoes`)

- A maioria dos influenciadores recebe menos de 2 milhões de curtidas, com concentração em valores muito baixos (<1 milhão).
- Apesar disso, existem influenciadores com uma média de curtidas superior a 12 milhões.

Taxa de Engajamento em 60 Dias (`60_day_eng_rate`)

- A distribuição é altamente assimétrica, com a maioria dos influenciadores tendo engajamento abaixo de 5%.
- Valores acima de 15% são menos frequentes e considerados elevados.

Média de Curtidas em Novos Posts (`new_post_avg_like_milhoes`)

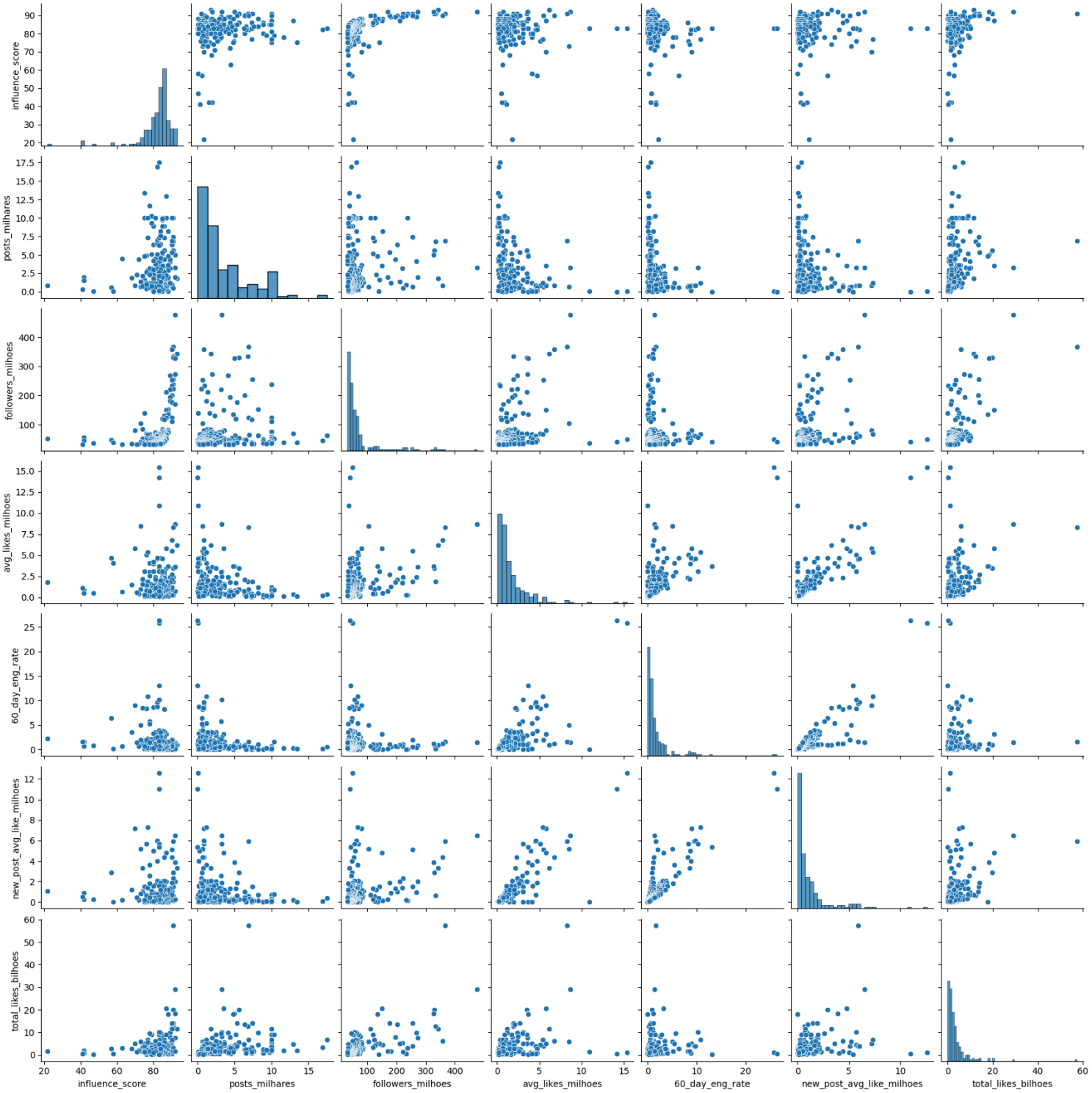
- A distribuição segue padrão semelhante ao de curtidas médias: altamente concentrada abaixo de 2 milhões.
- Influenciadores que atingem mais de 8 milhões de curtidas são muito raros.

Total de Curtidas em Bilhões (`total_likes_bilhoes`)

- O total de curtidas apresenta concentração abaixo de 20 bilhões, com poucos casos acima de 40 bilhões.
- Valores extremos são observados, o que sugere a necessidade de tratamento para evitar impactos no modelo.

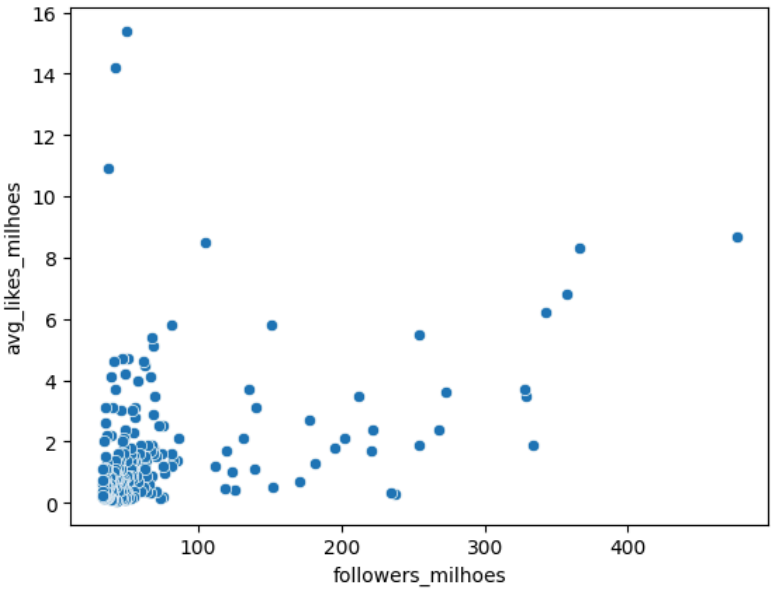
Para explorar as relações entre as variáveis do dataset, foi gerado um gráfico de pares (pairplot) que permite visualizar a dispersão entre pares de variáveis e observar padrões, correlações e possíveis outliers (valores fora do esperado).

Figura 2 — Pairplot do conjunto de dados



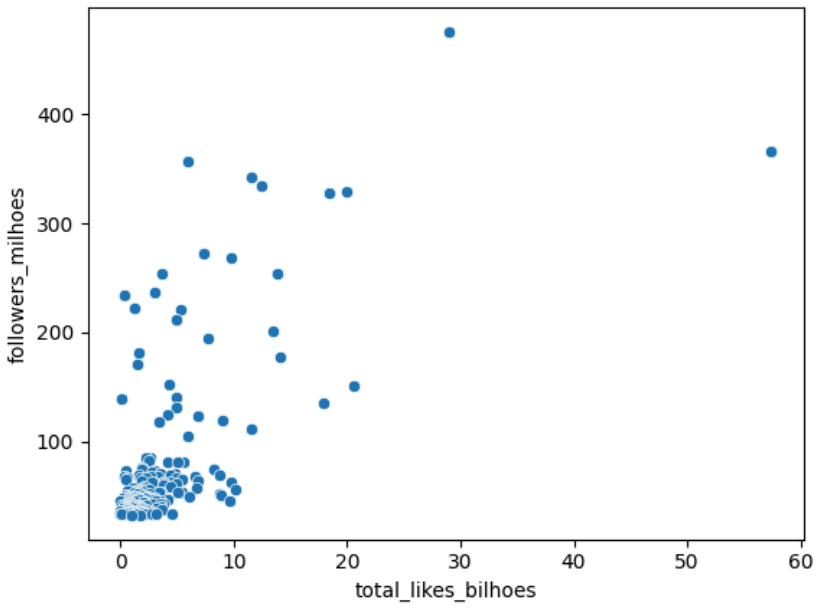
Correlação entre `followers_milhoes` e `avg_likes_milhoes`

Observa-se uma tendência positiva: influenciadores com mais seguidores tendem a ter maior média de curtidas. Contudo, a dispersão é significativa, indicando que o número de seguidores sozinho não é suficiente para explicar a média de curtidas.



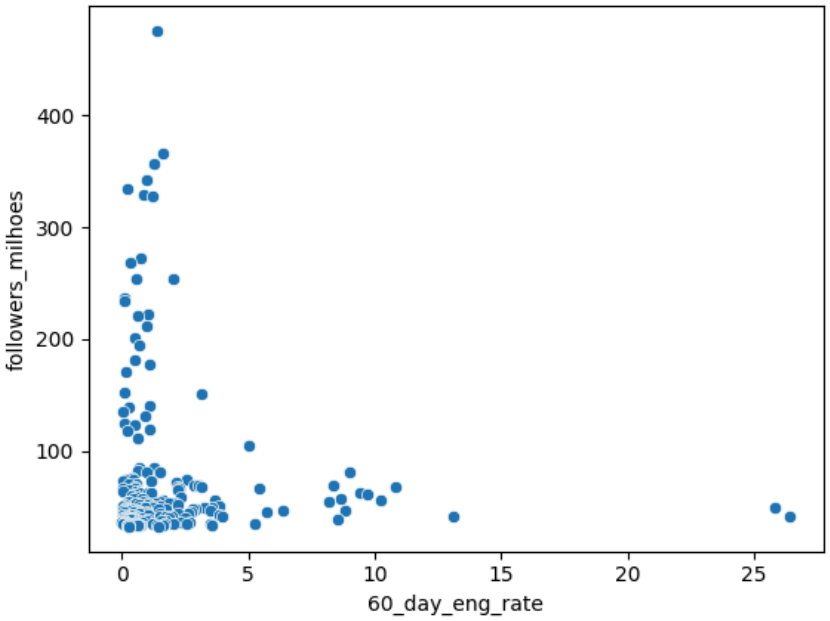
Correlação entre `total_likes_bilhoes` e `followers_milhoes`

Existe uma relação positiva clara: influenciadores com mais seguidores geralmente acumulam maior número total de curtidas.



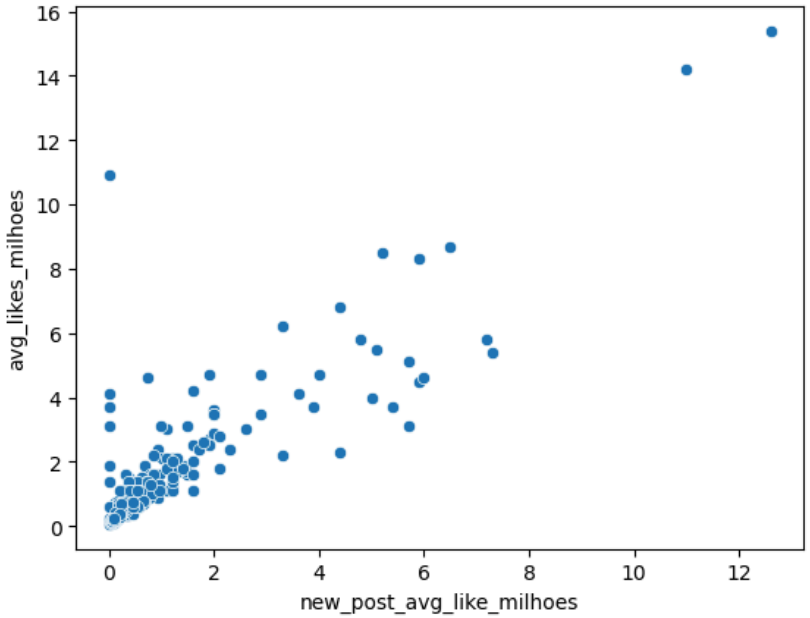
Relação entre `60_day_eng_rate` e `followers_milhoes`

A taxa de engajamento parece ser inversamente proporcional ao número de seguidores. Influenciadores com muitos seguidores tendem a ter taxas de engajamento mais baixas, o que é consistente com a literatura sobre redes sociais.



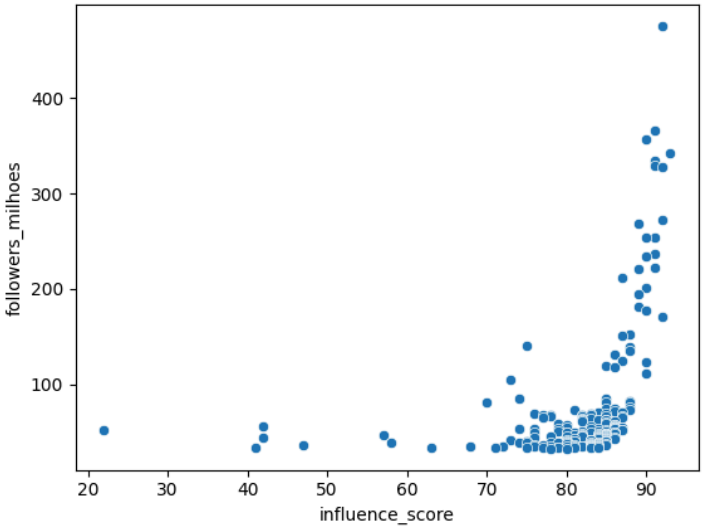
Correlação entre `new_post_avg_like_milhoes` e `avg_likes_milhoes`

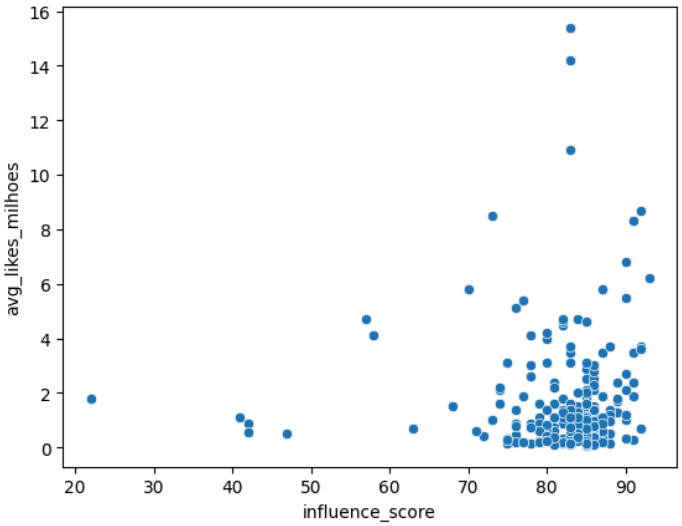
Há uma correlação linear forte entre a média de curtidas em novos posts e a média geral de curtidas, indicando consistência no engajamento ao longo do tempo.



Distribuição de `influence_score` em relação às demais variáveis

A pontuação de influência tem uma relação complexa com as variáveis. Embora haja uma leve tendência de aumento com `followers_milhoes` e `avg_likes_milhoes`, os pontos estão dispersos, sugerindo que a pontuação leva em consideração múltiplos fatores.



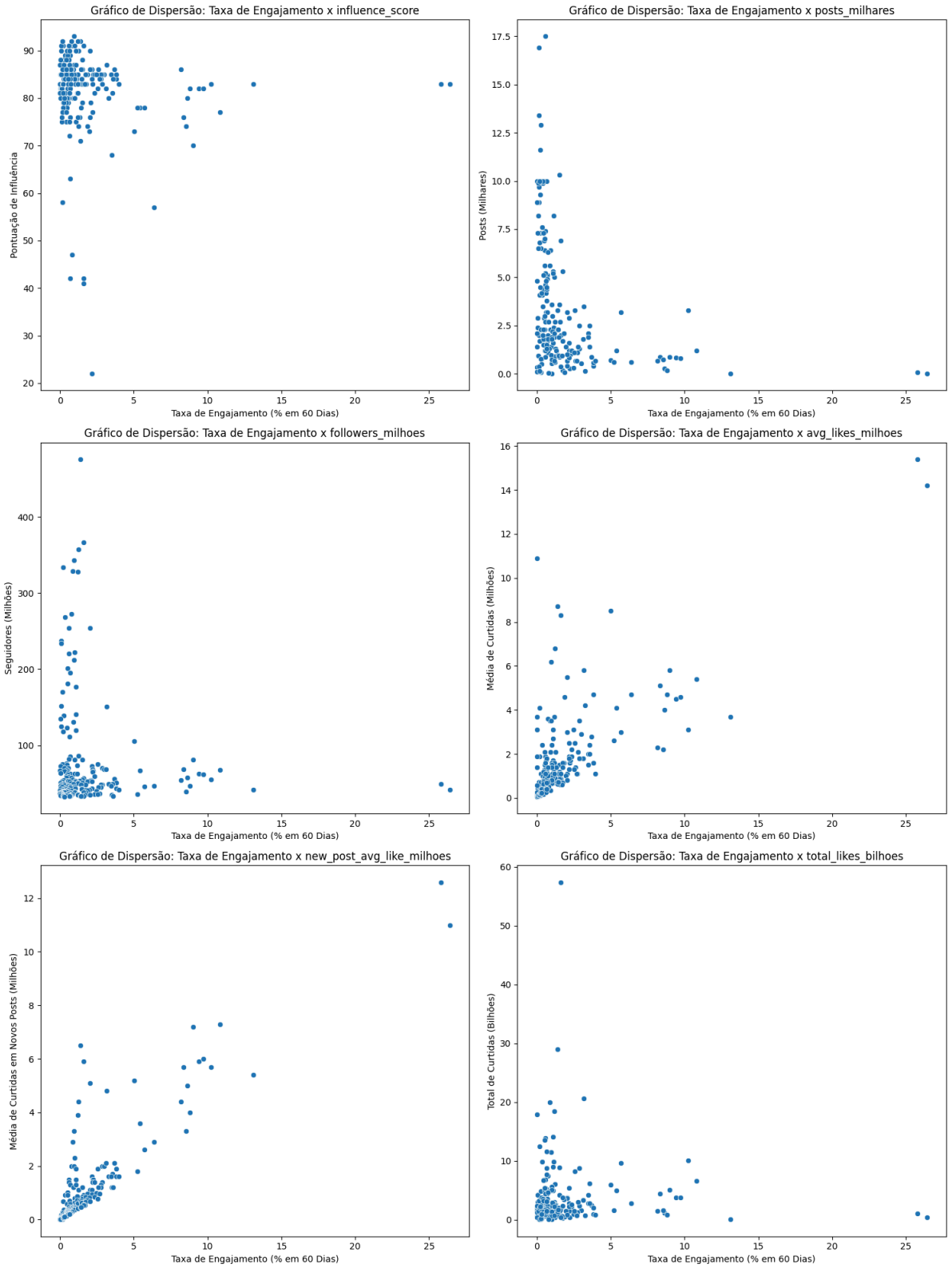


Outliers

Alguns pontos indicam valores extremos, como influenciadores com muitos seguidores, mas baixa média de curtidas ou engajamento. Esses casos devem ser investigados, pois podem representar estratégias de compra de seguidores ou comportamento atípico.

A fim de detalhar as relações entre a variável de taxa de engajamento (`60_day_eng_rate`) com outras variáveis foram feitos os gráficos de dispersão a seguir:

Figura 3 — Gráficos de dispersão entre 60_day_eng_rate e as outras variáveis



`60_day_eng_rate` x `influence_score`

Parece não haver uma relação linear clara entre essas variáveis, indicando que a taxa de engajamento pode não ser diretamente proporcional ao score de influência.

`60_day_eng_rate` x `posts_milhares`

No gráfico podemos perceber que a maior parte dos influenciadores apresenta menos de 5 mil posts, independentemente da taxa de engajamento. Sendo assim, não há um padrão óbvio entre a quantidade de posts e a taxa de engajamento. Influenciadores com muitos posts não necessariamente têm maior ou menor engajamento.

`60_day_eng_rate` x `followers_milhoes`

Há uma concentração alta de influenciadores com menos de 100 milhões de seguidores e taxas de engajamento menores que 10%.

`60_day_eng_rate` x `avg_likes_milhoes`

Uma relação positiva é evidente, uma vez que influenciadores com maior taxa de engajamento tendem a ter uma média mais alta de curtidas.

`60_day_eng_rate` x `new_post_avg_like_milhoes`

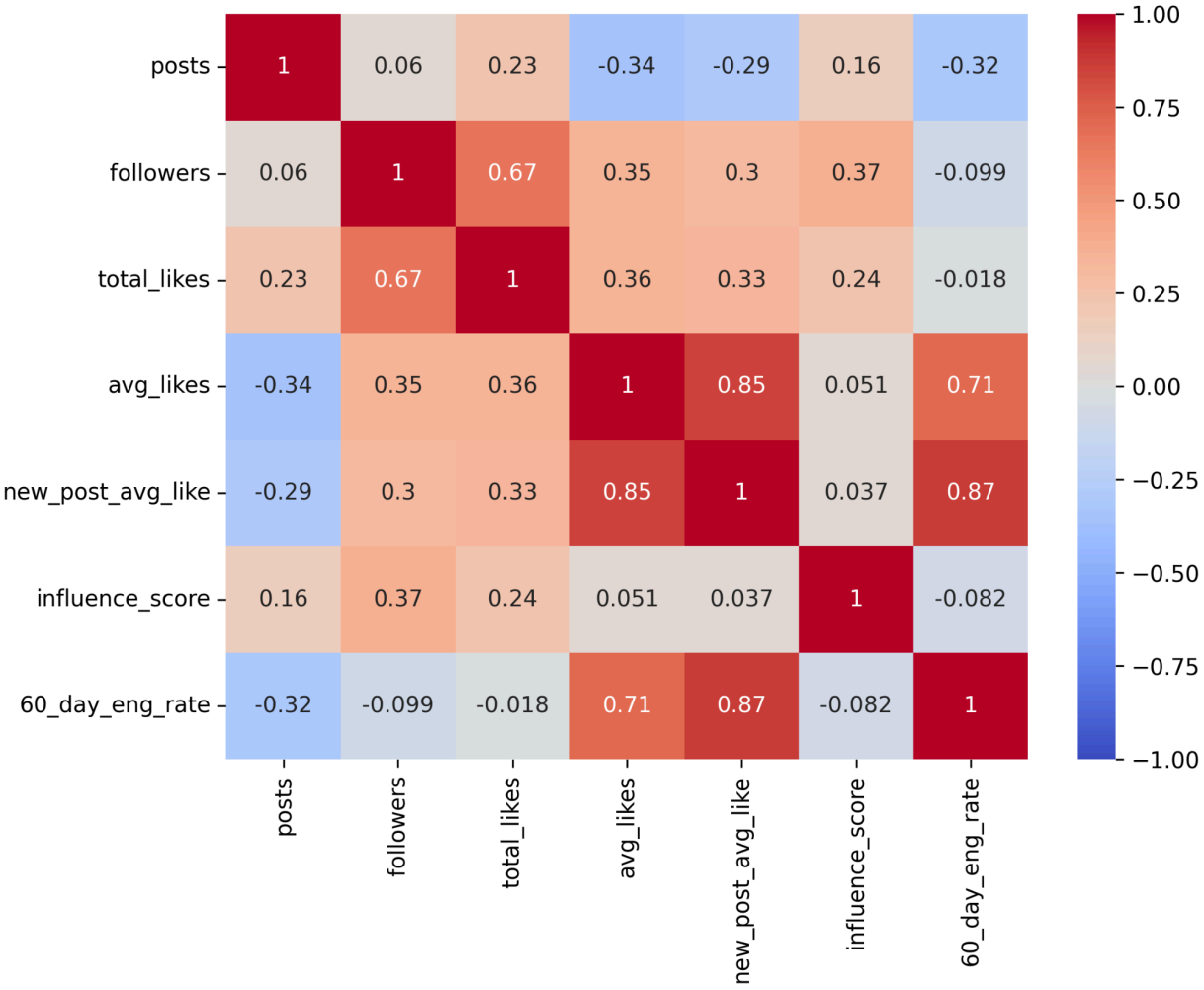
Similar ao gráfico anterior, há uma relação positiva entre a taxa de engajamento e a média de curtidas em novos posts.

`60_day_eng_rate` x `total_likes_bilhoes`

Embora haja uma tendência de influenciadores com mais curtidas totais apresentarem maior engajamento, a dispersão é considerável. Desse modo, o total de curtidas parece menos confiável como preditor direto da taxa de engajamento, possivelmente por incluir curtidas acumuladas ao longo do tempo, que não refletem o engajamento recente.

A Figura 4 mostra a matriz de correlação das variáveis usadas no modelo, ajudando a identificar o grau de associação linear entre elas. Os valores variam de -1 a 1: valores próximos de 1 indicam uma forte correlação positiva, valores próximos de -1 indicam uma forte correlação negativa, e valores próximos de 0 indicam pouca ou nenhuma correlação.

Figura 4 – Matriz de Correlação das variáveis



A variável alvo, `60_day_eng_rate`, está fortemente correlacionada com:

- `avg_likes` (0.71): Isso indica que o engajamento nos últimos 60 dias está diretamente relacionado à média de curtidas nos posts.
- `new_post_avg_like` (0.87): Sugerindo que o engajamento recente é altamente influenciado pela média de curtidas nos posts mais novos.

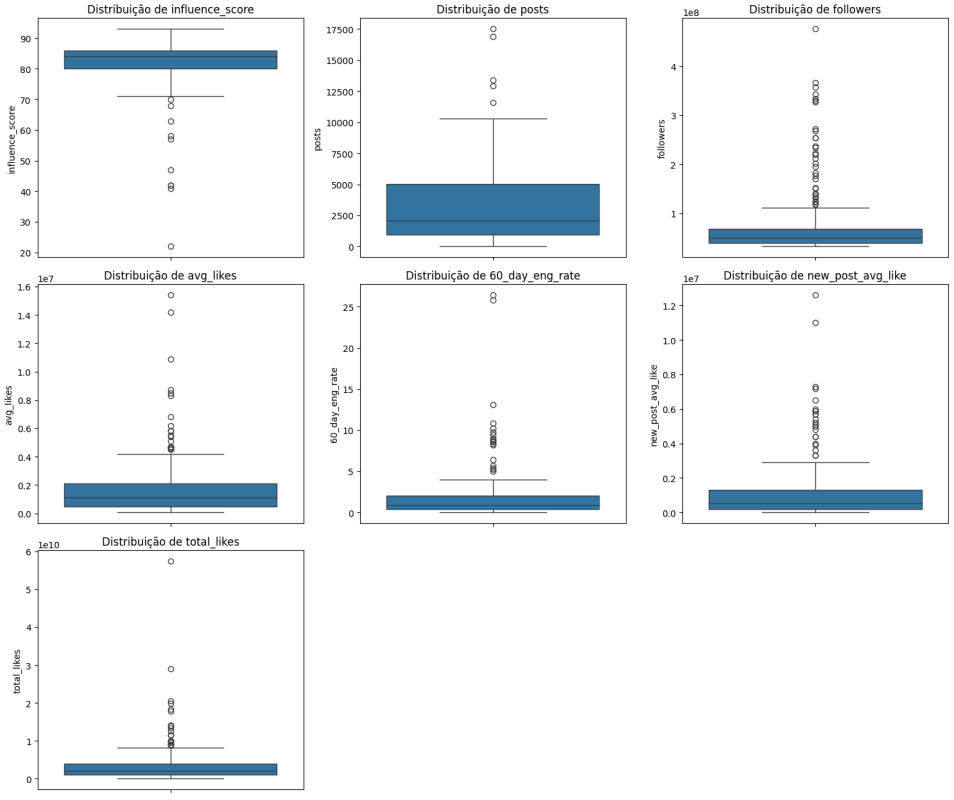
Outras correlações importantes observadas são:

- `followers` tem uma correlação fraca negativa com `60_day_eng_rate` (-0.099), indicando que o número de seguidores pode não ter uma influência direta no engajamento recente.
- `posts` apresenta uma correlação negativa moderada (-0.32), sugerindo que um aumento no número de postagens pode reduzir o engajamento recente.

Analisar a matriz de correlação é fundamental para selecionar as variáveis preditoras mais relevantes para o modelo de Regressão Linear, garantindo que os fatores com maior impacto na variável alvo sejam priorizados. Diante deste cenário, optou-se por utilizar as variáveis que possue*m correlações que realmente podem contribuir para a análise preditiva:* `avg_likes` *e* `new_post_avg_like`*.*

Realizando a análise dos outliers (Figura 5), vimos que, como temos apenas 200 registros e considerando que os dados representam as interações dos seguidores nas redes sociais, os outliers podem conter informações valiosas e refletir variações no comportamento de interação dos usuários. Exemplos disso são posts que se tornam virais ou influenciadores que experimentam picos de engajamento. Portanto, os outliers serão mantidos.

Figura 5 – Boxplots das variáveis



3.2 Regressão linear

As variáveis independentes foram definidas como:

$$x_1 = avg\_likes$$

$$x_2 = new\_post\_avg\_like$$

Conforme a equação de regressão linear apresentada anteriormente, β_0 é o termo de viés, ou intercepto, que representa o valor previsto de y quando todas as variáveis independentes x_1, x_2, …, x_n são iguais a zero. Este foi obtido ao executar-se o comando `model.intercept_`, sendo igual a:

$$\beta_0 = 0.05875218910937319$$

Já os coeficientes das variáveis independentes foram obtidos através do comando `model.coef_`:

$$\beta_1 = -1.31489707 * 10^{-7}$$
$$\beta_2 = 1.67806974 * 10^{-6}$$

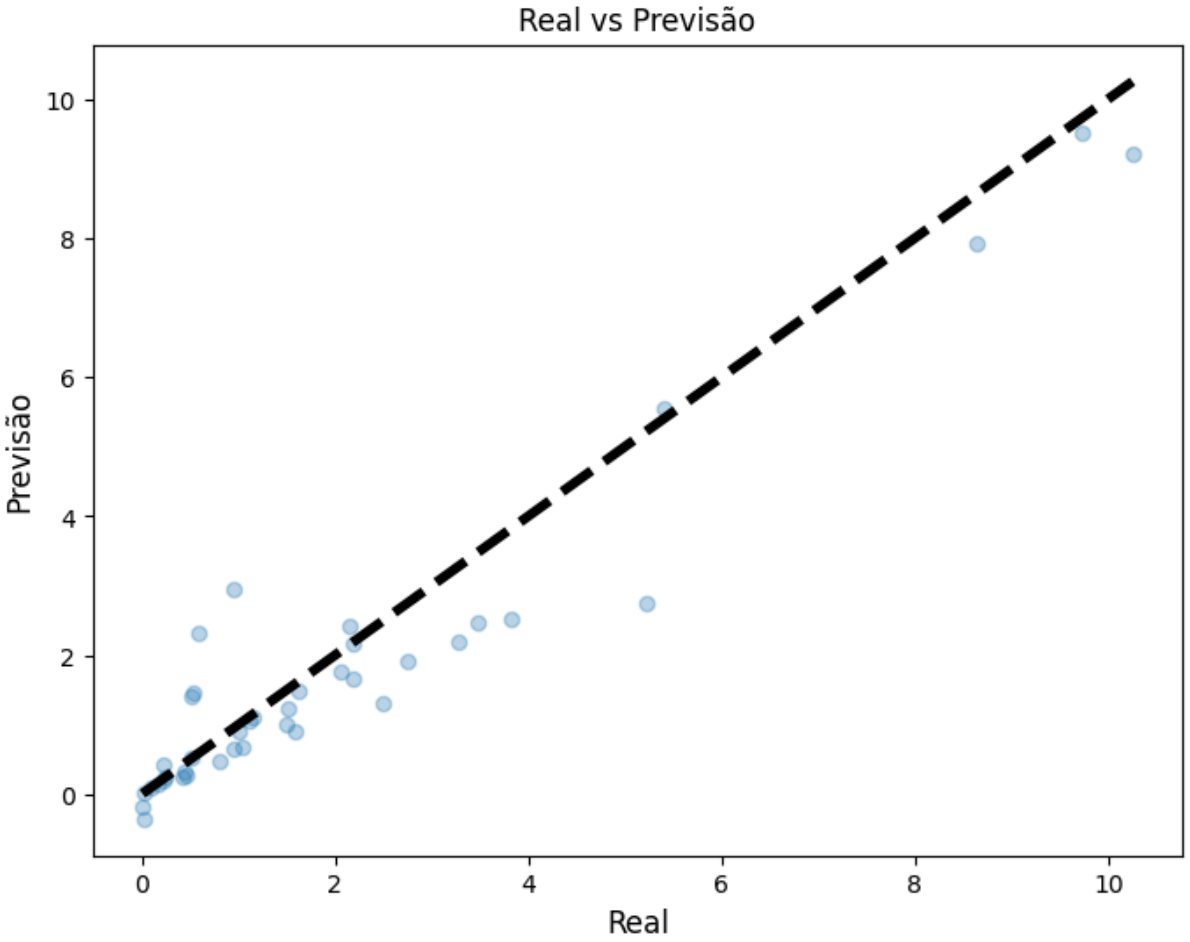
Portanto, a equação da regressão linear para esse conjunto de dados é:

$$y = 1.8569375 - 1.31489707 * 10^{-7} * avg\_likes + 1.67806974 * 10^{-6} * new\_post\_avg\_like$$

Ao rodar o modelo, os seguintes valores foram previstos, sendo demonstrados na Figura 6:

Real	Previsão	Real	Previsão	Real	Previsão	Real	Previsão
1.52	1.244412	1.01	0.910143	0.44	0.337332	0.59	2.326026
0.96	2.954678	0.17	0.137968	0.53	1.460694	0.51	1.420900
0.22	0.427264	0.02	0.025078	2.49	1.319807	0.01	-0.191078
0.43	0.248836	10.25	9.216132	2.18	1.667947	8.63	7.923142
0.96	0.649387	0.02	-0.348866	1.12	1.069404	0.22	0.189860
3.47	2.480684	5.23	2.737404	2.19	2.158219	3.28	2.191407
9.72	9.522318	5.40	5.560695	2.06	1.759990	1.05	0.676645
0.10	0.095011	0.46	0.268570	1.15	1.120310	1.49	1.009878
0.80	0.470388	0.25	0.260516	2.76	1.927797	1.58	0.913460
2.16	2.414940	0.52	0.533616	3.82	2.533280	1.62	1.485217

Figura 6 – Correlação entre Valores Reais e Previstos



3.3 Validação do modelo

A validação do modelo de Regressão Linear foi realizada utilizando as métricas R^2, MAE, MSE e RMSE, que avaliam a capacidade preditiva e a precisão do modelo. Os resultados obtidos foram os seguintes:

- Coeficiente de Determinação (R^2): 0,9042. Esse valor indica que aproximadamente 90,42% da variabilidade da variável dependente **60_day_eng_rate** é explicada pelas variáveis independentes **avg_likes** e **new_post_avg_like**. Esse resultado mostra que o modelo capturou bem a relação entre as variáveis analisadas.
- Erro Absoluto Médio (MAE): 0,5182. O MAE representa o erro médio absoluto entre as previsões e os valores reais. Esse valor mostra que, em média, as

previsões do modelo diferem dos valores reais em 0,5182 unidades, o que é aceitável no contexto da análise.

- Erro Quadrático Médio (MSE): 0,6066. O MSE mede a magnitude média dos erros ao quadrado, penalizando desvios maiores. Um valor relativamente baixo indica que os erros do modelo estão contidos.
- Raiz do Erro Quadrático Médio (RMSE): 0,7788. Esse valor é a raiz quadrada do MSE e fornece uma métrica mais intuitiva para interpretar o erro médio. O RMSE confirma a boa precisão do modelo, com valores de erro alinhados à escala dos dados.

Tabela 6 – Métricas de Desempenho do Modelo

Métrica	Valor
R^2	0.9042
MAE	0.5182
MSE	0.6066
RMSE	0.7788

Os resultados da validação apontam para um desempenho satisfatório do modelo, com alto coeficiente de determinação e erros contidos. Isso sugere que as variáveis escolhidas são adequadas para prever a taxa de engajamento **60_day_eng_rate**.

3.4 Validação cruzada

Ao realizar a validação cruzada para os dois folds, foram obtidos os respectivos valores de R^2:

Fold 1: 0.76156693

Fold 2: 0.67659147

Para medir a performance do modelo, foram calculados a média e o desvio padrão destes valores pelo coeficiente de determinação médio, expresso em porcentagem.

Coeficiente de Determinação Médio: 71.91%

Este valor indica que o modelo é capaz de explicar uma parcela significativa da variância da variável dependente. Deve-se levar em consideração também que a diferença entre os R^2 dos folds (7.85%) é relativamente pequena, indicando que o modelo é consistente ao lidar com diferentes divisões dos dados.

4. Discussão

Existe uma tendência de engajamento inversamente proporcional ao número de seguidores. Influenciadores com muitos seguidores apresentam taxas de engajamento menores, o que é comum no contexto de redes sociais. No entanto, a análise dos dados permite identificar que, mesmo com essa tendência, variáveis como o número médio de curtidas em postagens recentes **avg_likes** e a média de curtidas em novas postagens **new_post_avg_like** têm uma forte relação com a taxa de engajamento **60_day_eng_rate**.

O modelo de Regressão Linear utilizado conseguiu capturar essa relação, demonstrando que a combinação dessas variáveis pode ser eficaz na previsão do engajamento futuro. A forte correlação evidenciada pelo R^2 de 0,9042 indica que o modelo é capaz de representar com alta precisão a variabilidade do engajamento a partir das variáveis preditoras selecionadas.

Ainda assim, é importante considerar que a relação inversa entre engajamento e número de seguidores sugere a necessidade de personalização na análise. Para influenciadores de grande porte, estratégias como segmentação do público, criação de conteúdos direcionados e estímulo à interação ativa podem ser exploradas para reduzir a queda no engajamento.

A análise dos resíduos e a visualização da dispersão dos dados sugerem que o modelo pode apresentar limitações em capturar outliers ou padrões não lineares, indicando que há espaço para explorar abordagens mais alternativas. Modelos não lineares poderiam complementar a análise e fornecer insights mais profundos, especialmente em cenários onde há grande variação no comportamento dos dados.

5. Conclusão e Trabalhos Futuros

Este estudo demonstrou que é viável usar a Regressão Linear para modelar o engajamento em redes sociais, destacando os fatores mais relevantes para a taxa de engajamento de influenciadores. Apesar das limitações observadas, como a precisão relativamente baixa e ajustes inadequados, os resultados oferecem um ponto de partida valioso para futuras análises.

Trabalhos Futuros

- Explorar Modelos Preditivos Mais Complexos: Considerar o uso de regressão polinomial ou algoritmos baseados em aprendizado de máquina para melhorar a precisão das previsões.
- Ampliar a Base de Dados: Incluir influenciadores de diferentes plataformas, o que aumentaria a generalização dos resultados e a aplicabilidade do modelo.
- Incorporar Variáveis Contextuais: Adicionar variáveis como tipo de conteúdo ou sazonalidade para refinar ainda mais as previsões.

Essa abordagem contínua permitirá um entendimento mais profundo do comportamento de engajamento, contribuindo para estratégias mais eficazes no marketing digital.

6. Referências

DANTAS, Isadora. **Taxa de engajamento**: Por que está ruim? Isa Dantas, 2020. Disponível em: https://isadoradantasm.com.br/taxa-de-engajamento-por-que-esta-ruim/. Acesso em: 09 nov. 2024.

INTERNATIONAL BUSINESS MACHINES CORPORATION (IBM). **O que é regressão linear?** IBM, 2024. Disponível em: https://www.ibm.com/br-pt/topics/linear-regression. Acesso em: 09 nov. 2024.

RIVAL IQ. **2024 Rival IQ Social Media Industry Benchmark Report**: Industry benchmarks across the most important social media metrics. 1ª ed. Santa Clara: Rival IQ, 2024.