Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Resumo

O projeto visa implementar um modelo preditivo para estimar a taxa de engajamento de influenciadores em diversas regiões do mundo. A previsão será realizada com base em variáveis independentes que estão fortemente correlacionadas com a variável dependente, ou em casos onde as variáveis não apresentam correlação entre si. O objetivo principal é criar um sistema eficiente de previsão, usando técnicas de regressão e análise de dados.

Metodologia

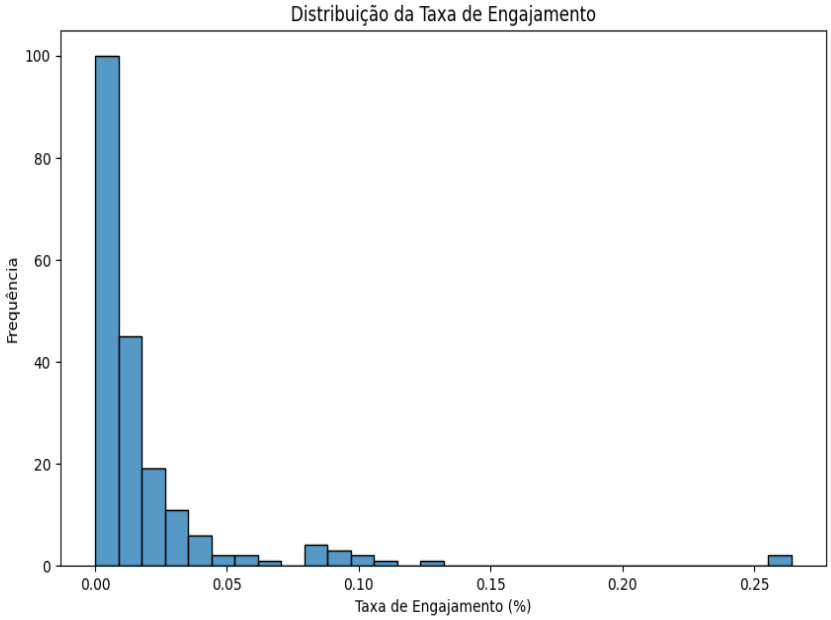
O projeto utilizou três variações de modelos de regressão linear para prever a taxa de engajamento dos influenciadores: regressão linear simples, Ridge e Lasso. O processo de modelagem incluiu a análise e o pré-processamento dos dados, onde foram realizadas etapas de remoção de outliers para garantir a qualidade dos dados e minimizar a influência de valores extremos no modelo. Além disso, as variáveis independentes foram normalizadas para garantir que todas as features tivessem uma escala similar, o que favorece o desempenho dos modelos de regressão. A seleção dos modelos foi feita com base na eficiência de cada um para lidar com dados altamente correlacionados e na capacidade de regularização de Ridge e Lasso para evitar overfitting.

Análise Exploratória

Conhecendo os Dados

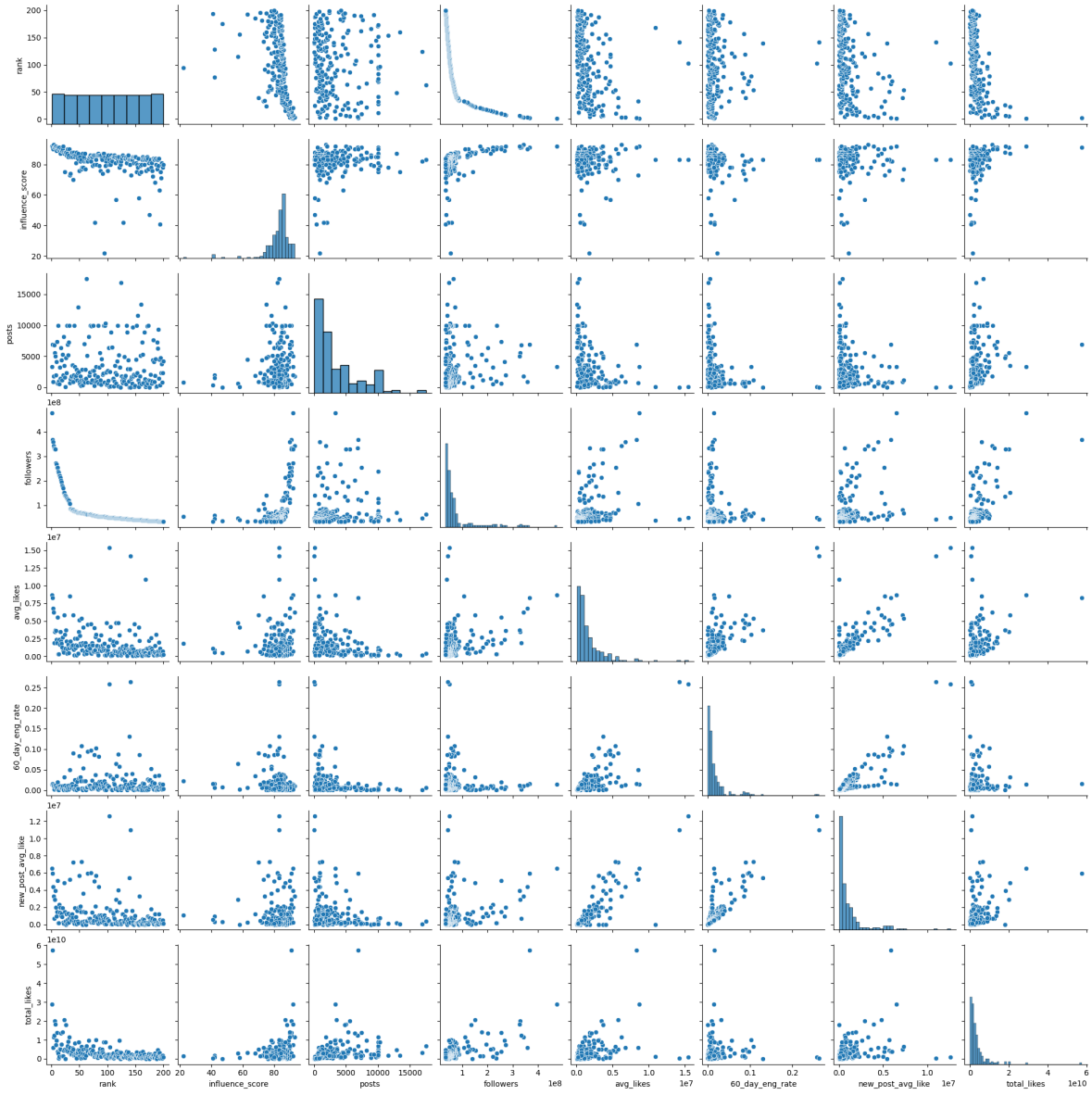
1° Analisamos a distribuição da nossa variável dependente

Percebe-mos alguns outliers com taxa de engajamento acima de 25%, no qual é muito alto comparado ao resto dos dados. Devemos tratar isso.



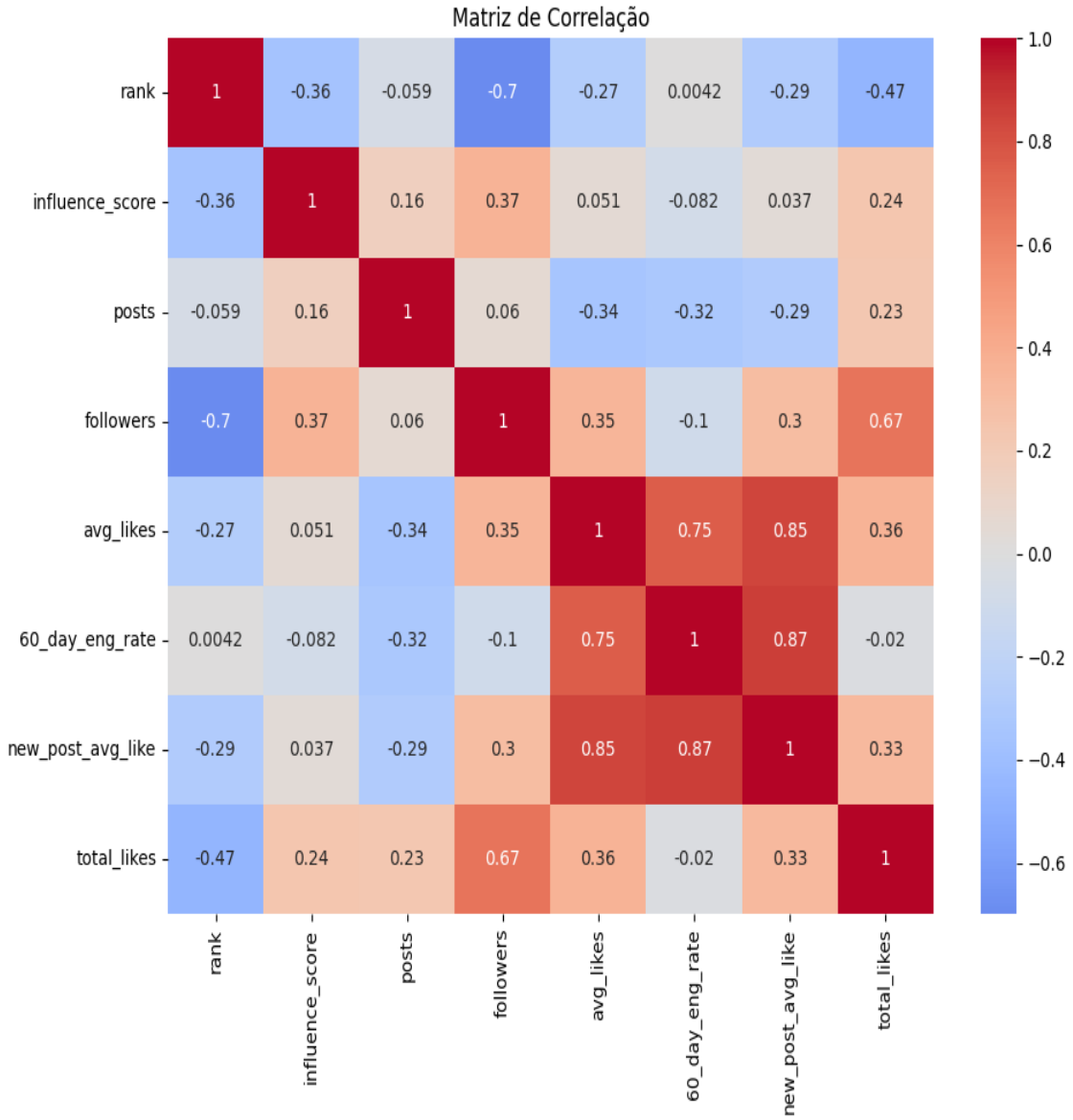
2° Analisamos a distribuição e relação de cada uma das variáveis

Percebe-se que a variável independente rank segue uma distribuição qualse constante, e tem uma relação não linear com a váriavel dependente. A variável independente new_post_avg_like, segue uma distribuição bastante semelhante ao da nossa variável dependente, e de uma forma não tão clara, parece se relacionar com a variável dependente. de forma linear. E também, a variável new_post_avg_like parece se relacionar de forma qualse linear com avg_likes. Outra observação, é praticamente todas as variáveis exceto rank, possuem outliers que devem ser tratados para melhorar o desempenho do nosso modelo.



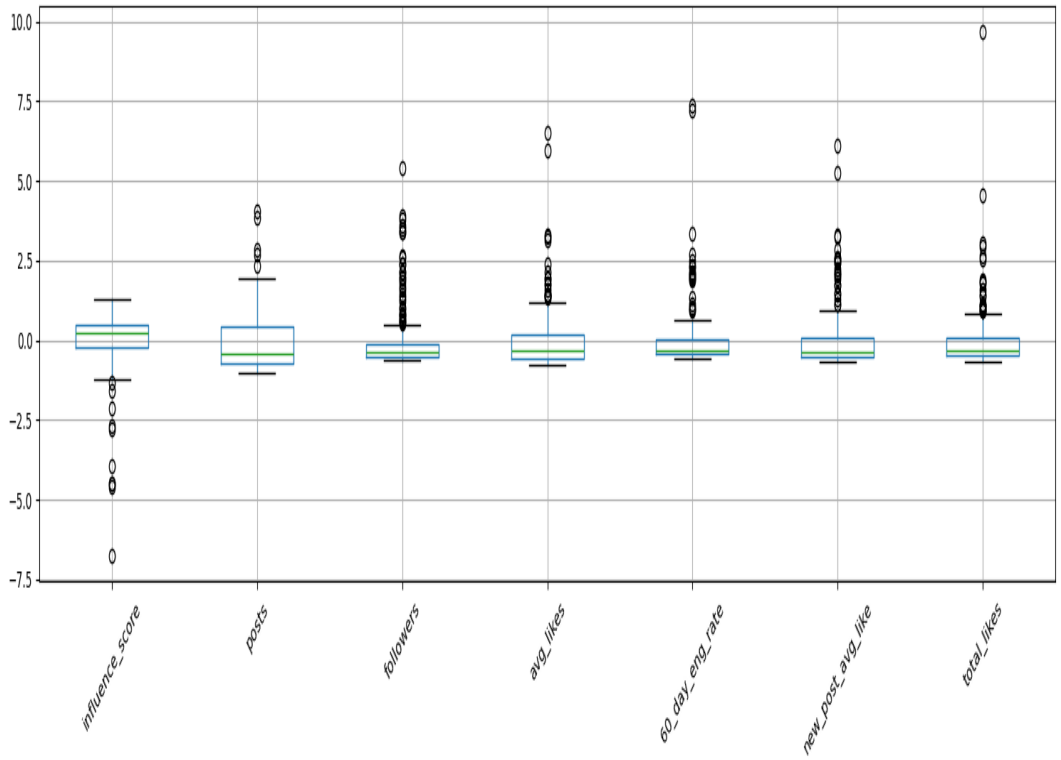
3° Analisamos a matriz de correlação

Como esperado a variável rank possui correlação qualse nula com a variável dependente oque nos permite remover tranquilamente essa variável, da mesma forma as variáveis independentes channel_info e country foram removidas pelo mesmo motivo em testes anteriores. E percebemos uma forte correlação entre a variável dependente 60_day_eng_rate, e duas variáveis independentes new_post_avg_like e avg_likes. A correlação entre as variáveis independentes new_post_avg_like e avg_likes também está alta, devemos reduzi-la no preprocessamento. A variável followers também está um pouco correlacionada com total_likes, mas nada muito alarmante.



4° Distribuição das Features antes tratamento de Outliers

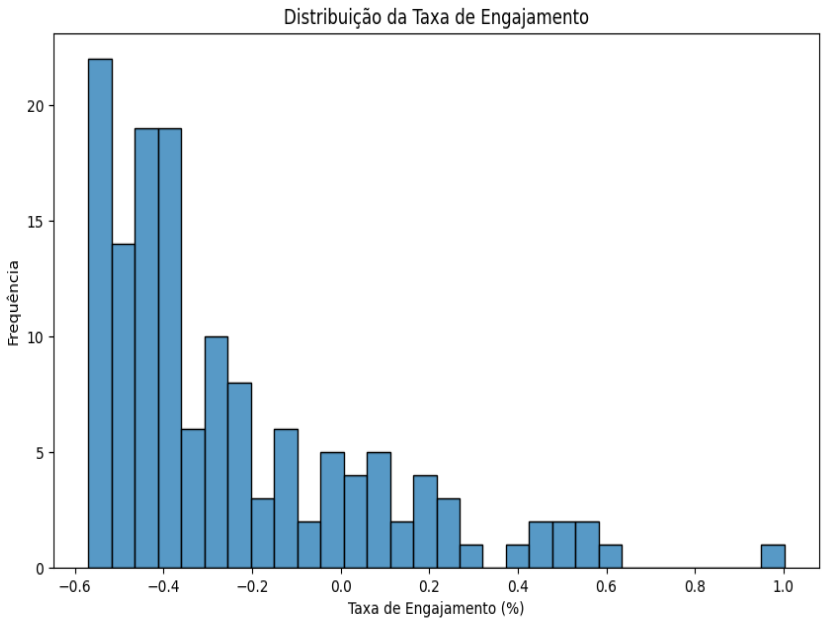
Percebe-se uma vasta quantidade de outliers em métricas como followers, avg_likes, new_post_avg_like e na variável dependente 60_day_eng_rate. Esses outliers devem ser tratados da melhor forma possível para que o modelo seja menos enviesado pelo overfitting.



Dados Após Pré-processamento

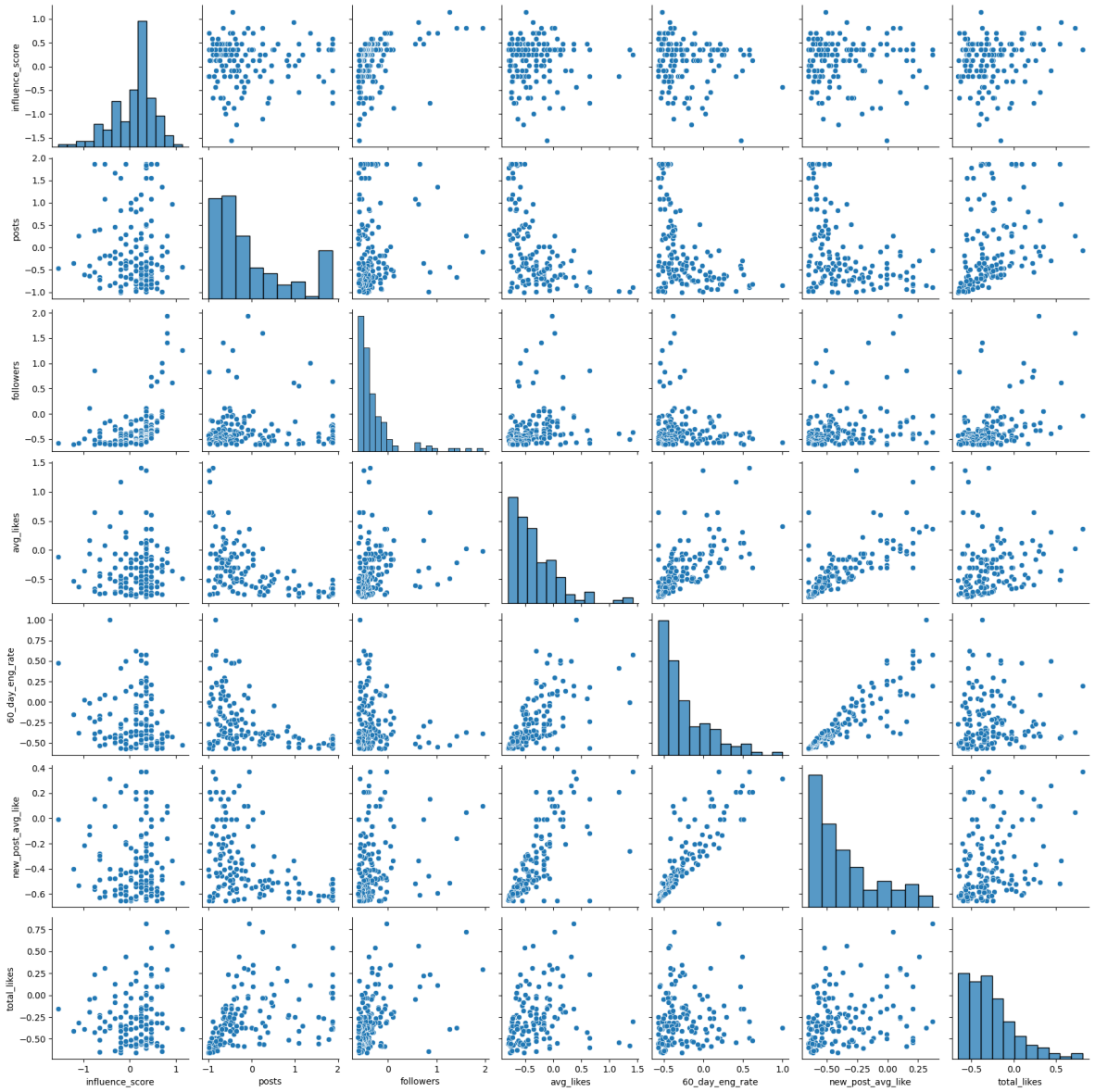
1° Nova distribuição da variável dependente

Após a remoção de outliers e a normalização dos dados a nova distribuição ficou mais clara. O pico inicial de dados foi reduzido, e a distribuição ficou mais suave nos evidenciando que temos muito mais dados com taxa de engajamento de porcentagem baixa doque de porcentagem alta.



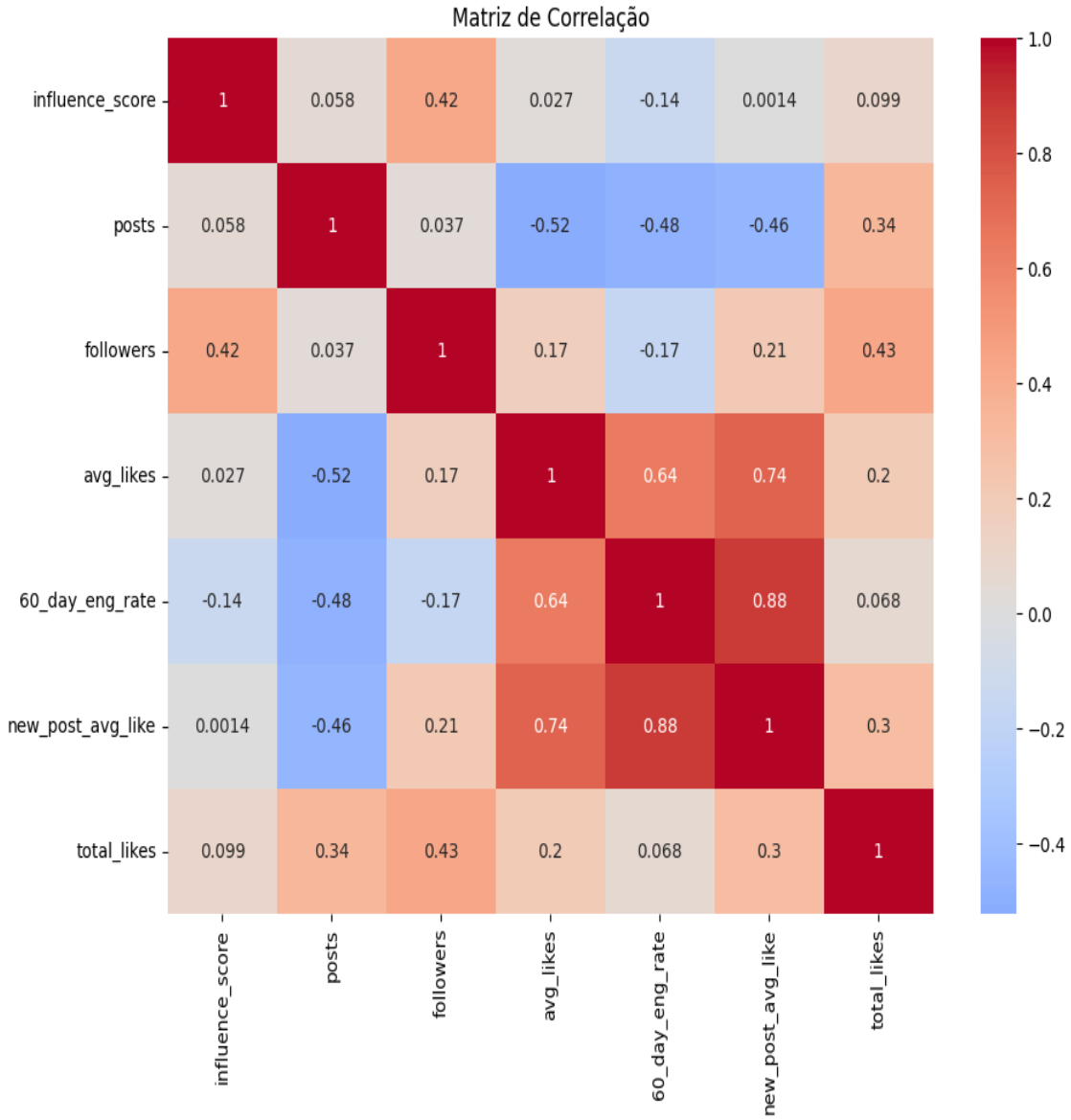
2° Nova distribuição e relação entre variáveis

Conseguimos evidenciar mais ainda a relação linear entre a variável independente new_post_avg_like e a variável dependente 60_day_eng_rate, oque é ótimo para precisão do nosso modelo. Percebe-se também, que qualse todas as distribuições de variáveis seguem o mesmo padrão com muitos valores baixos e menos valores altos, exceto pela variável independente influence_score, que segue uma distribuição mais central. As variáveis rank, country e channel_info foram removidas por terem relação qualse nula com a variável dependente.



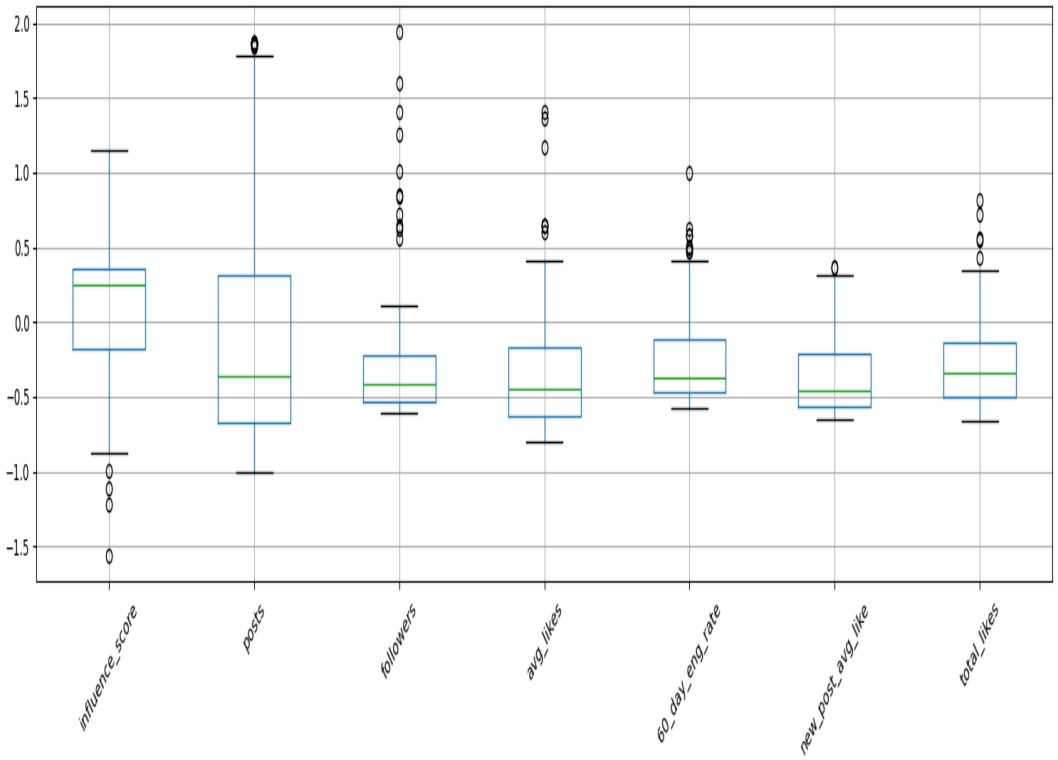
3° Nova Matriz de Correlação

Percebe-se que a relação entre a variável dependente 60_day_eng_rate e as variáveis independentes new_post_avg_like e avg_likes permanecem da mesma forma. Entretanto, conseguimos reduzir a correlação entre new_post_avg_like e avg_likes para um nível aceitável, abaixo de 0.75. Decidimos por não usar tecnicas de redução de dimensionalidade em variáveis independetes com correlação abaixo de 0.8, e também optamos por não fundir as duas variáveis porque avg_likes também está com uma correlação considerável com a variável dependente. A variável independente total_likes que possuia uma correlação considerável com followers não possui mais, a correlação entre as duas foi bastante reduzida. Dessa forma, os dados ficaram melhores para treinar o modelo.



4° Distribuição das Features após tratamento de Outliers

Aqui ja temos uma redução considerável na quantidade de outliers em todas as variáveis independentes citadas anteriormente, principalmente na new_post_avg_like que apenas sobrou um outlier. Dessa forma, nosso modelo não será tão afetado pelos outliers no data set.



Implementação do Algoritmo

O algoritmo foi implementado usando 3 modelos diferentes da biblioteca sklearn.linear_model sendo eles o modelo regular de regressão linear, o ~de Lasso e o de Ridge. Eles foram treinados com os mesmos dados para comparação posteriormente. Depois de alguns testes decidimos que 0.75% é o valor ideal para treinamento do nosso modelo. E claro, todos os modelos foram treinados com dados preprocessados.

Validação e Ajuste de Hiperparâmetros

Escolha das variaveis

As variáveis independentes foram escolhidas por dois principais motivos, 1° forte correlação com a variável dependente, 2° baixa correlação entre as outras variáveis independentes. Dessa forma temos, que as variáveis independentes como influence_score, followers e total_likes, atendem o primeiro critério, enquanto as avg_likes e new_post_arg_like atendem o segundo e por fim a variável independente post que é um meio termo entre o primeiro e segundo. Variáveis como rank, country e channel_info possuiam uma correlação com a variável dependente muito baixa, por isso foram removidas e como elas não eram correlacionadas também, tecnicas de redução de dimensionalidade não iriam ajudar.

Validação dos Modelos

Para validar os modelos foram usadas métricas de validação comuns como R^2, MSE e MAE. E para evitar o maximo de overfitting usamos validação cruzada com o R^2, e calculamos a média e o desvio padrão, da validação cruzada, para análise.

Otimização dos Modelos

Para otimizar todos os modelos, além de ajustar os hiperparametros nós deveriamos otimizar a base de dados, e assim fizemos. A base de dados foi melhorada com os seguintes passos:

- Garantimos que todos os dados estão normalizados.

- Evitamos a multicolinearidade entre variáveis independentes muito correlacionadas.

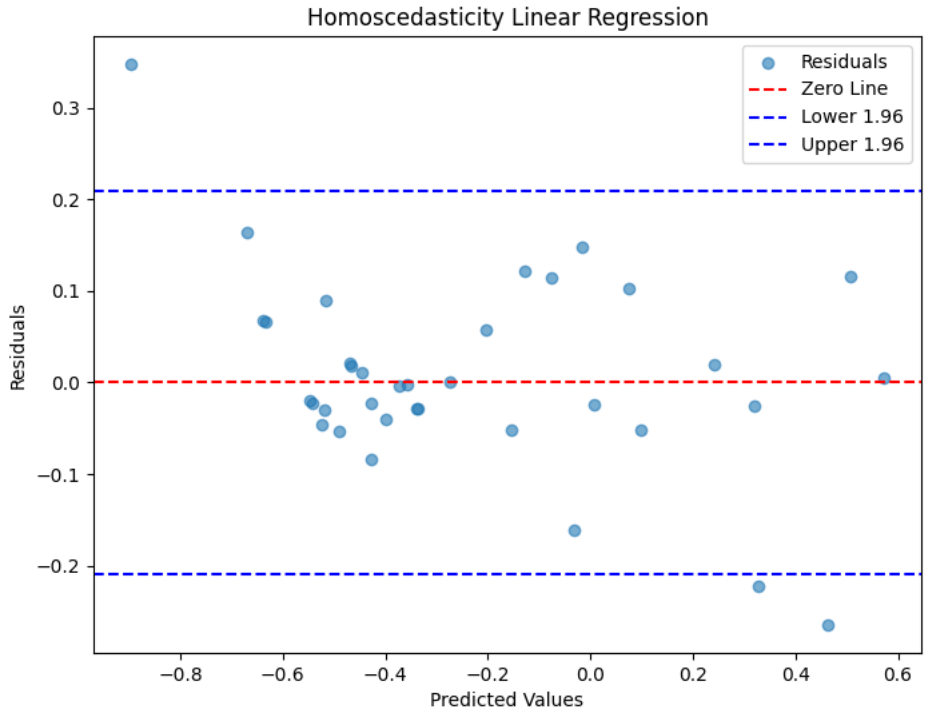
Entretando, permitimos que avg_like e new_post_avg_like pois ambas compartilhavam uma correlação de alto valor com a variável dependente.

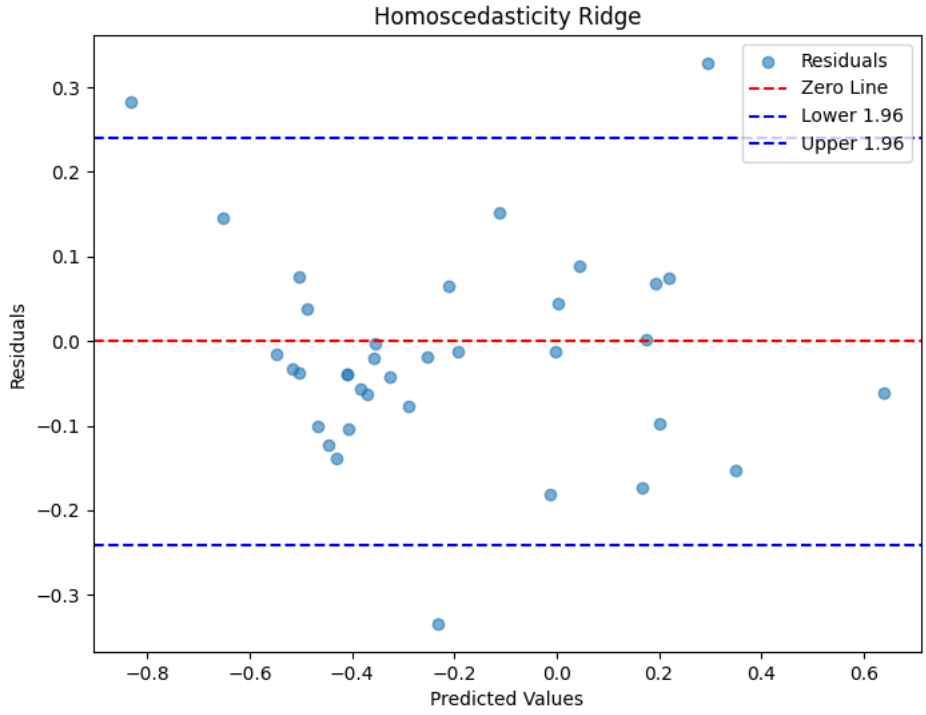
- Ajustamos o hiperparamêtro alpha de Ridge e Lasso.

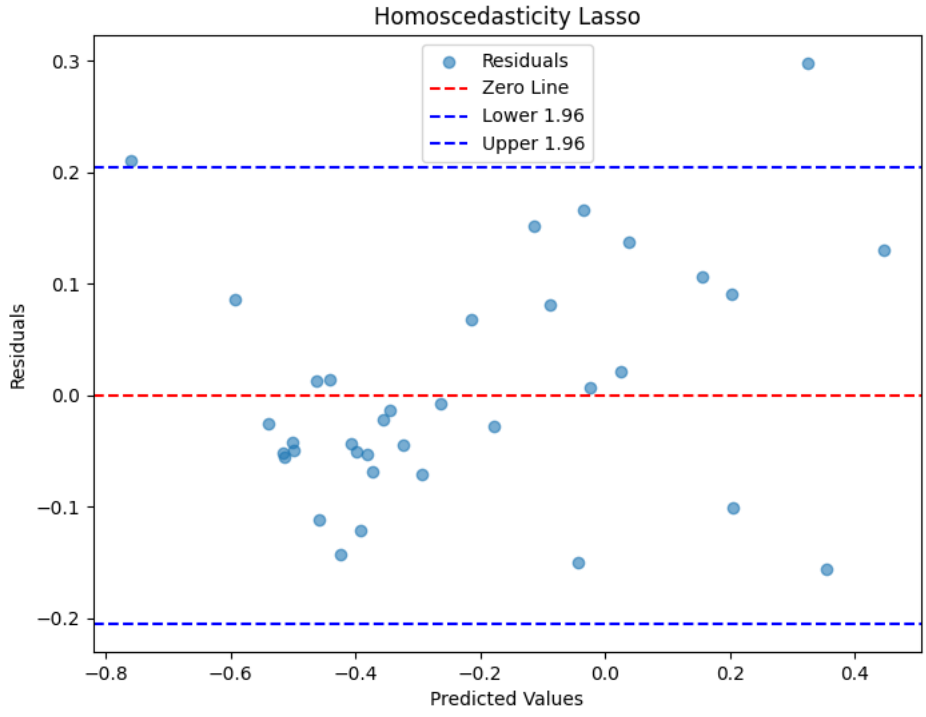
Para o modelo Ridge um hiperparamêtro alpha = 1 consideramos ideal, penalizando moderadamente os coeficientes do modelo. E para o modelo de Lasso, alpha = 0.01 ja é o suficiente para zerar alguns coeficientes e e fornecer boas métricas de desvio padrão e média do R^2 da validação cruzada.

- Evitamos que os dados sofram de heteroscedasticidade.

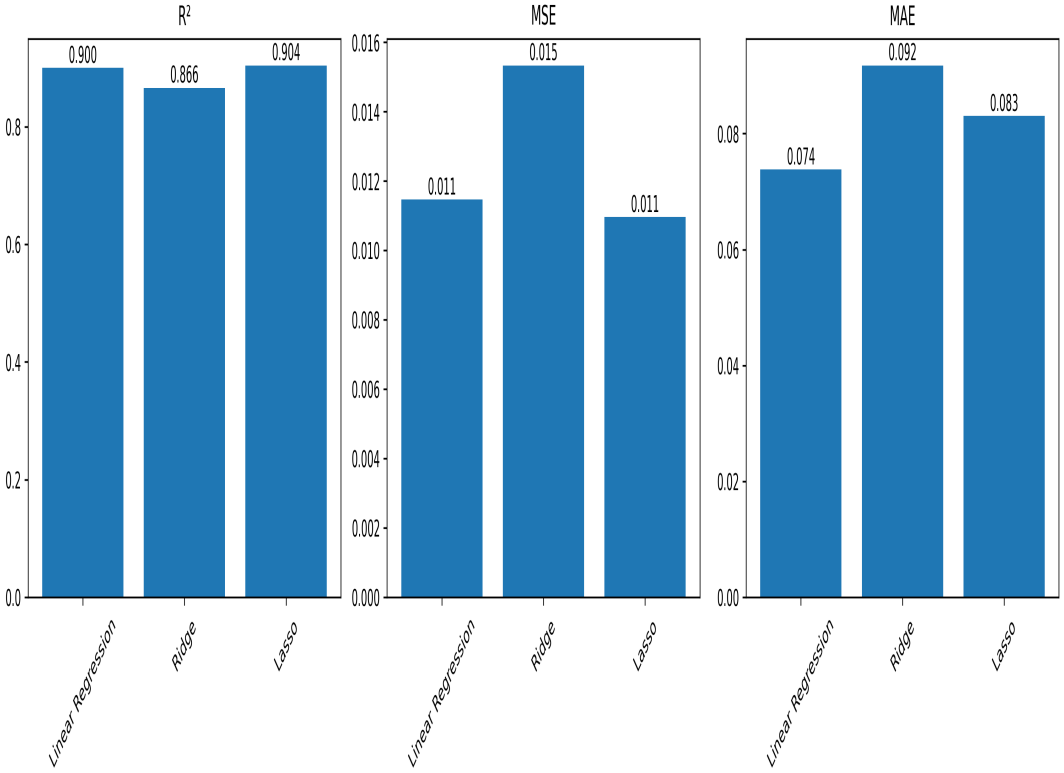
Percebe-se que temos pouquissimos dados que saem do intervalo de confiança dos modelos, e ainda assim os valores que saem do intervalo de confiança não são extrapolantes, dessa forma eles podem ser aceitos.




Homoscedasticity Ridge
Residuals
Zero Line
Lower 1.96
Upper 1.96
Residuals
Predicted Values
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.3
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6


Homoscedasticity Lasso
Residuals
Zero Line
Lower 1.96
Upper 1.96
Residuals
Predicted Values
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.8
−0.6
−0.4
−0.2
0.0
0.2
0.4

Resultados



RESULTADOS DOS MODELOS ========== Linear Regression: ---------- R²: 0.8996 MSE: 0.0115 MAE: 0.0738 CV Score (média): 0.8326 CV Score (desvio): 0.1446 Coeficientes: influence_score: 0.0072 posts: -0.0061 followers: -0.2722 avg_likes: -0.0483 new_post_avg_like: 1.2016 total_likes: -0.0582 ========== Ridge: ------ R²: 0.8657 MSE: 0.0153 MAE: 0.0918 CV Score (média): 0.8187 CV Score (desvio): 0.0829 Coeficientes: influence_score: -0.0038 posts: -0.0236 followers: -0.2445 avg_likes: 0.1458 new_post_avg_like: 0.8073 total_likes: -0.0303 ========== Lasso: ------ R²: 0.9040 MSE: 0.0110 MAE: 0.0831 CV Score (média): 0.8556 CV Score (desvio): 0.0726 Coeficientes: influence_score: -0.0000 posts: -0.0310 followers: -0.1999 avg_likes: 0.0000 new_post_avg_like: 0.8960 total_likes: -0.0000 ==========

Conclusão

Com base nos resultados obtidos, o modelo de Regressão Linear demonstrou melhor performance para a predição de taxas de engajamento, apresentando um equilíbrio adequado entre complexidade e acurácia.