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.

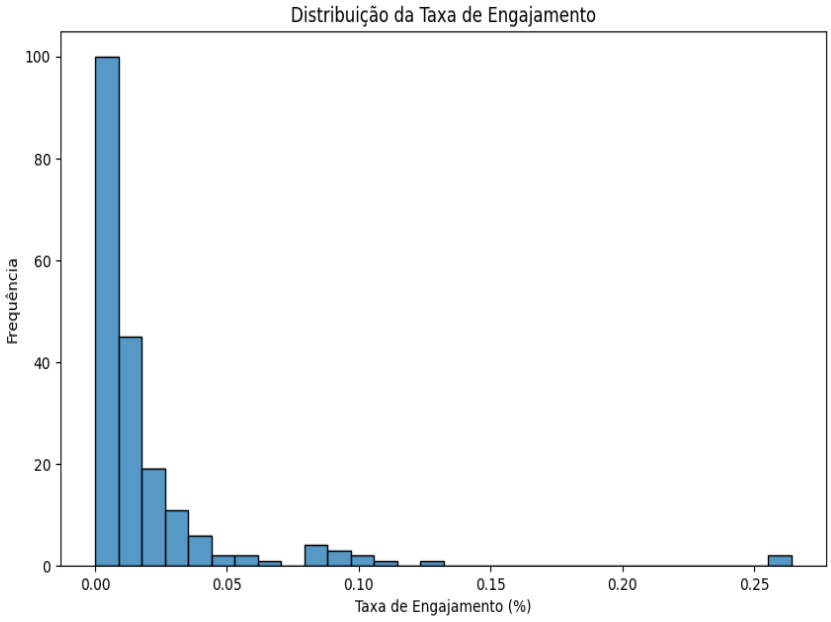
Metodologia

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.

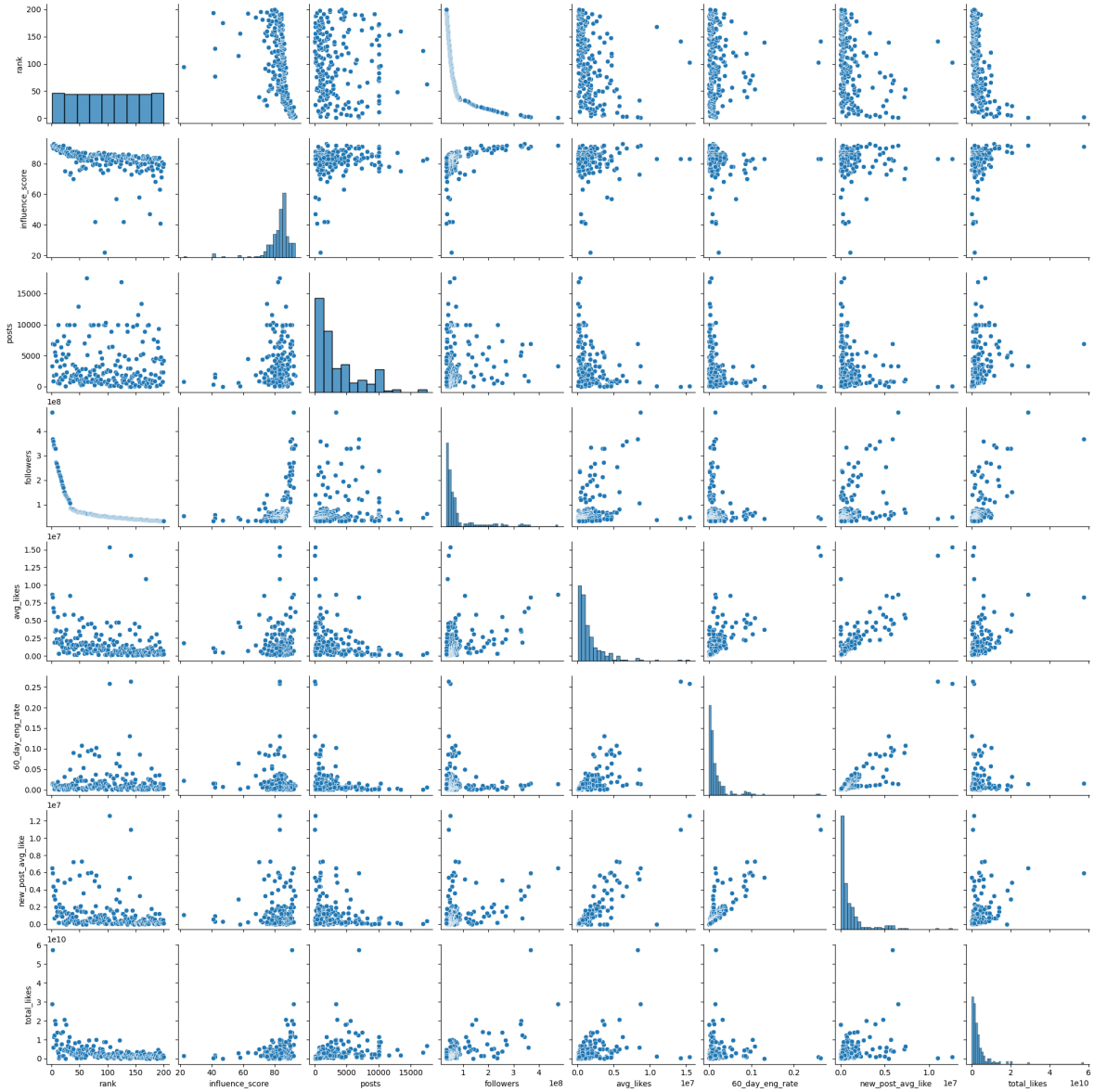
Análise Exploratória

Conhecendo os Dados

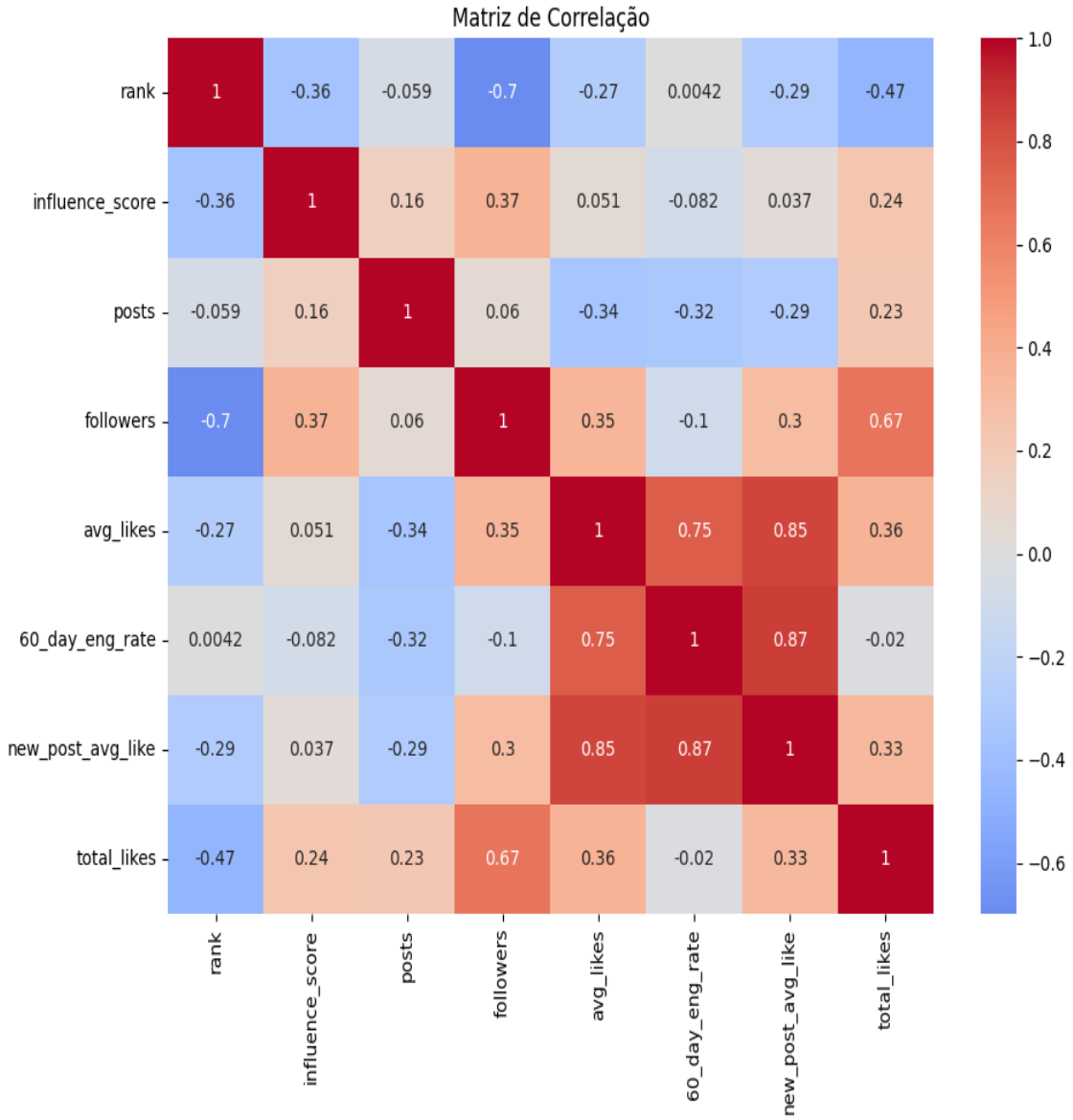
1° Analisamos a distribuição da nossa variável dependente, e percebe-mos alguns outliers com taxa de engajamento acima de 25%, no qual é muito alto comparado ao resto dos dados. Devemos tratar isso.



2° Analisamos a distribuição de cada uma das variáveis independentes e dependente, e a relação entre cada uma delas. 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.

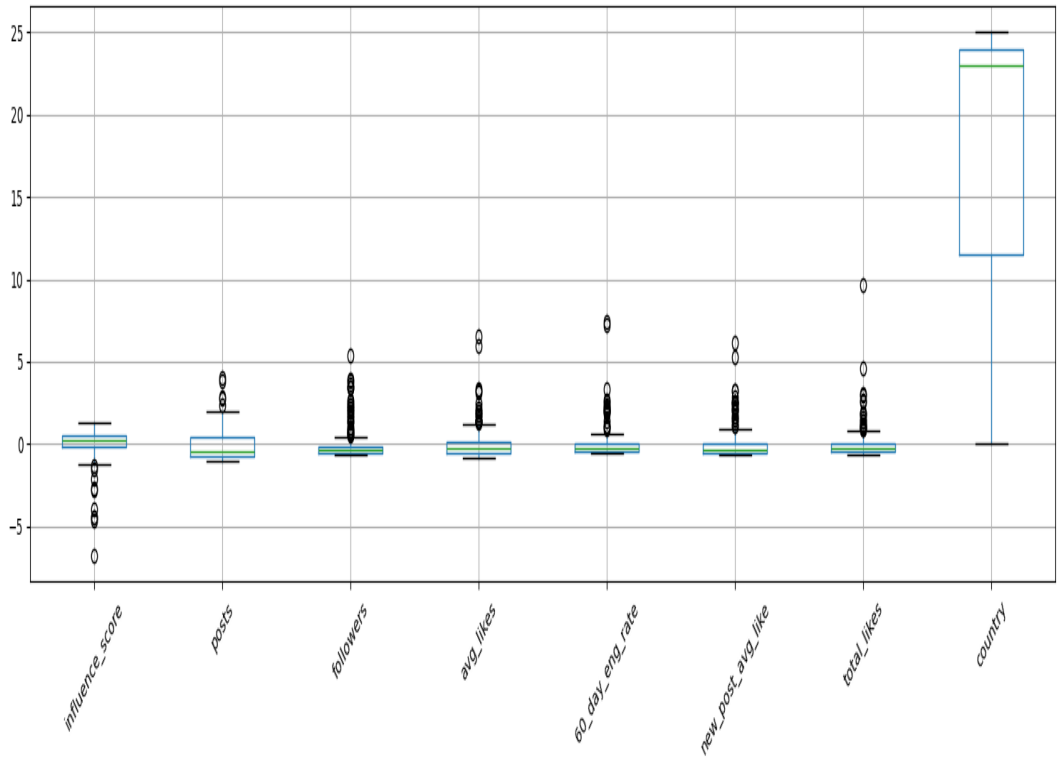


3° Analisamos a matriz de correlação, como esperado a variável rank possui correlação qualse nula com a variável dependente oque nos permite remover tranquilamente essa variável, da mesma forma as variáveis independentes channel_info e country foram removidas pelo mesmo motivo em testes anteriores. E 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.



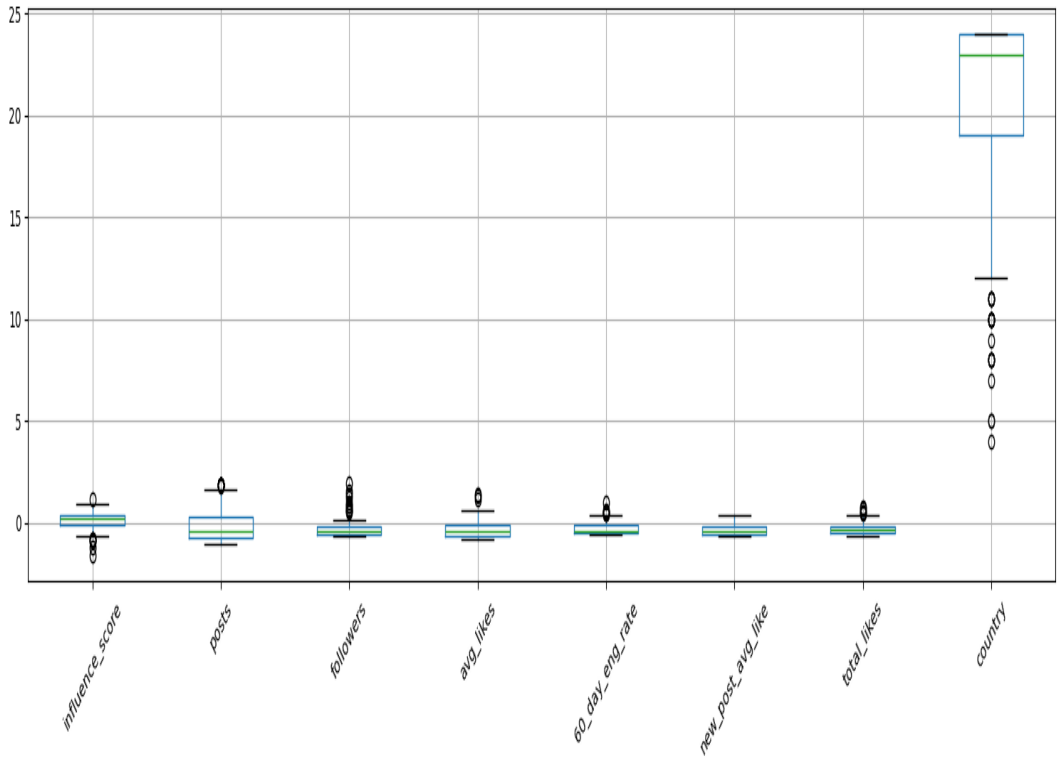
Distribuição das Features antes tratamento de Outliers

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.

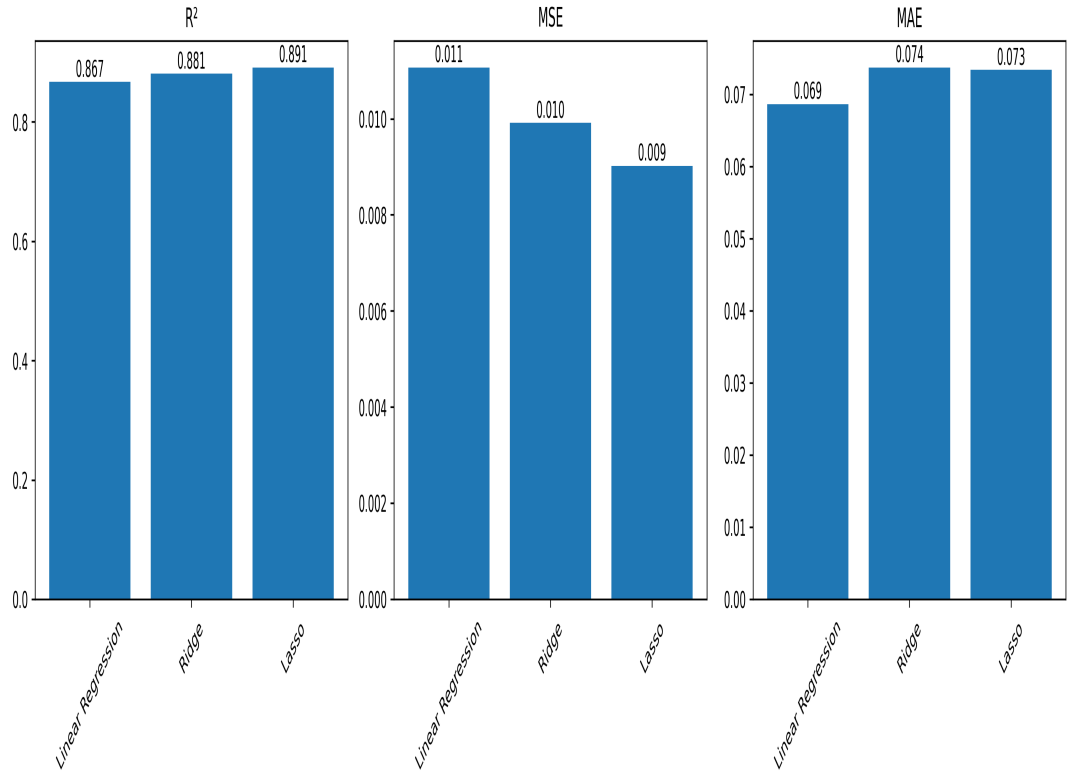


Distribuição das Features após tratamento de Outliers

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.



Resultados



RESULTADOS DOS MODELOS ========== Linear Regression: ---------- R²: 0.8668 MSE: 0.0111 MAE: 0.0687 CV Score (média): 0.8785 CV Score (desvio): 0.0484 Coeficientes: influence_score: 0.0075 posts: -0.0044 followers: -0.2666 avg_likes: -0.0253 new_post_avg_like: 1.1849 total_likes: -0.0787 country: -0.0002 ========== Ridge: ------ R²: 0.8808 MSE: 0.0099 MAE: 0.0738 CV Score (média): 0.8562 CV Score (desvio): 0.0910 Coeficientes: influence_score: -0.0113 posts: -0.0330 followers: -0.2449 avg_likes: 0.0595 new_post_avg_like: 0.8827 total_likes: -0.0075 country: -0.0000 ========== Lasso: ------ R²: 0.8915 MSE: 0.0090 MAE: 0.0734 CV Score (média): 0.8735 CV Score (desvio): 0.0644 Coeficientes: influence_score: -0.0000 posts: -0.0308 followers: -0.2093 avg_likes: 0.0000 new_post_avg_like: 0.9362 total_likes: -0.0000 country: -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.