Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Residentes: Saulo Alves e Thailane Carmo

Data de entrega: 17/11/2024

1.Resumo

O objetivo deste projeto foi criar um modelo preditivo com o uso do algoritmo de Regressão Linear para abordar uma questão sobre a taxa de engajamento dos principais influenciadores no Instagram. O trabalho abrangeu desde a exploração inicial dos dados até a otimização e validação do modelo. Para atingir esse fim, foram aplicadas técnicas de Estatísticas Descritivas, identificação de outliers, análise de correlação e transformações de dados. Além disso, para a validação, foram utilizadas abordagens como o IQR, validação cruzada e análise dos parâmetros do modelo. O projeto incluiu implementações que melhoraram os parâmetros dos modelos, tanto por meio do tratamento dos dados de entrada quanto do ajuste dos hiperparâmetros empregados na regressão.

.

2.Introdução

Diversos fatores influenciam o sucesso dos influenciadores em suas plataformas, sendo o algoritmo da mídia em que operam o principal responsável pelas métricas observadas. Embora a lógica desse algoritmo possa ser obscurecida, as métricas estão acessíveis. Com base nisso, foi aplicada uma técnica de Machine Learning (ML) de regressão para realizar inferências nos dados numéricos. A escolha da ML se justifica pela natureza da variável dependente, considerando que as variáveis independentes também são numéricas.

O conjunto de dados compila informações sobre os principais influenciadores do Instagram, abrangendo uma diversidade de 200 tipos diferentes de influenciadores. Ele inclui perfis que possuem um grande número de seguidores, elevado engajamento e um amplo alcance em suas publicações.

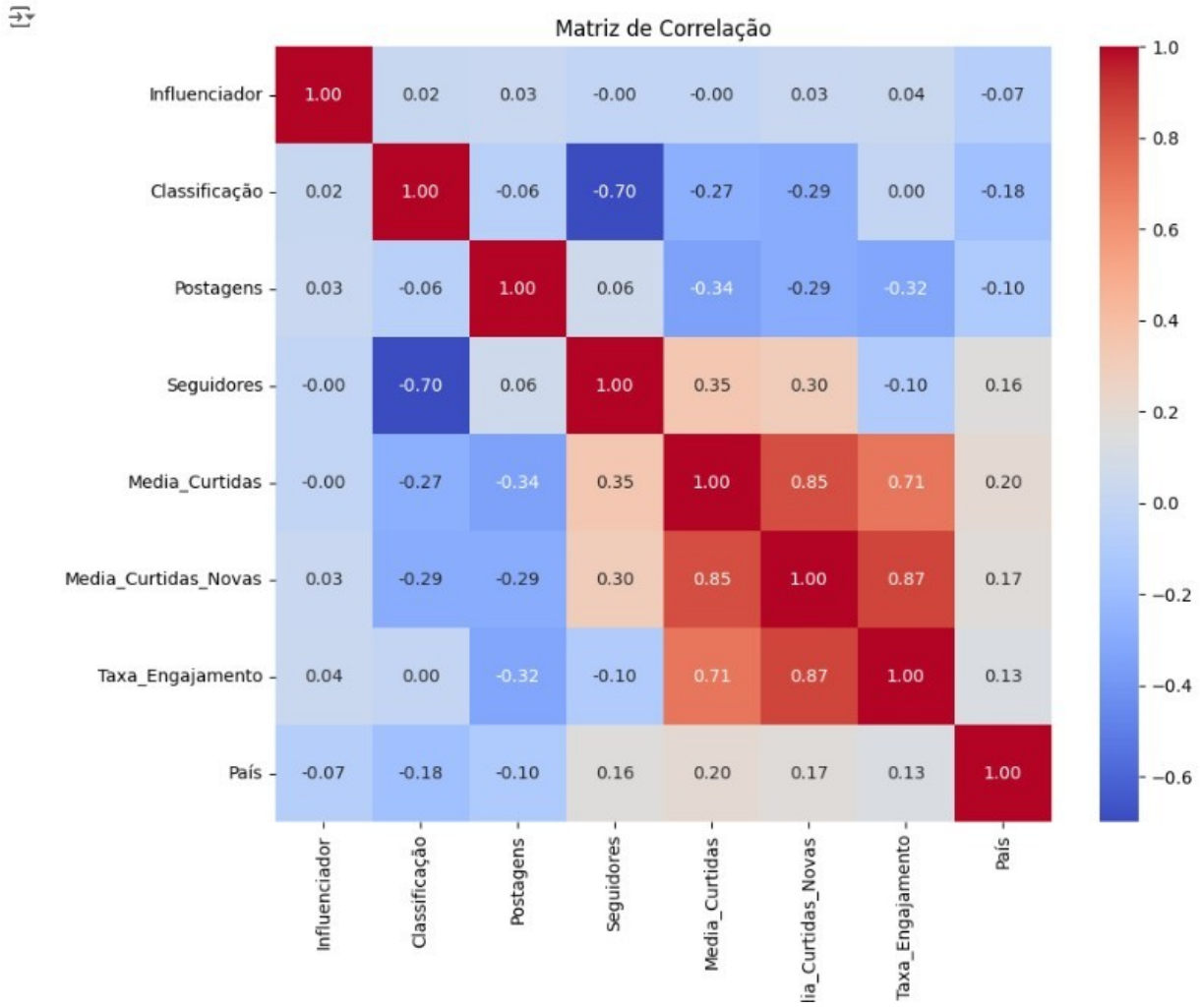
3.Metodologia

A análise exploratória teve início com uma visão geral da estatística descritiva do conjunto de dados, com a finalidade de avaliar a magnitude dos valores e identificar padrões superficiais. Também foi verificada a qualidade dos dados, em busca de valores ausentes ou nulos. Como apenas uma célula apresentava valor nulo, optou-se por preenchê-la com base na mediana da coluna. Transformações foram realizadas nos dados para examinar como a variável alvo responde às alterações nas variáveis preditoras.

Foi empregado o LabelEncoder, uma técnica de pré-processamento de dados que converte valores categóricos em numéricos, atribuindo um número exclusivo a cada categoria e estabelecendo uma

correspondência entre a categoria original e um número inteiro. As variáveis que fariam parte do modelo foram analisadas por meio da matriz de correlação e classificadas de acordo com sua relevância.

Figura 1 – Matriz de Correlação



Fonte: Autores, 2024

A multicolinearidade foi verificada com o Variance Inflation Factor (VIF).

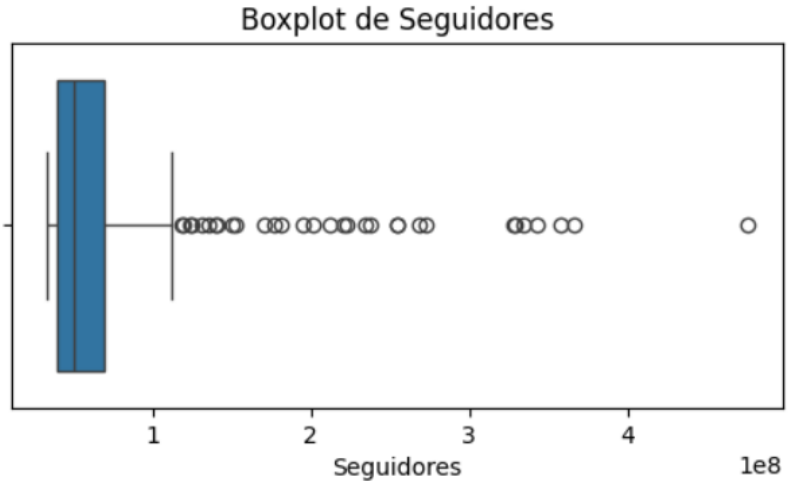
Figura 2 – Grau de Correlação

	Variável	VIF
0	const	27.142000
1	Influenciador	1.010406
2	Classificação	2.019417
3	Postagens	1.191250
4	Seguidores	2.111342
5	Media_Curtidas	3.900003
6	Media_Curtidas_Novas	3.603287
7	País	1.070629

Fonte: Autores, 2024

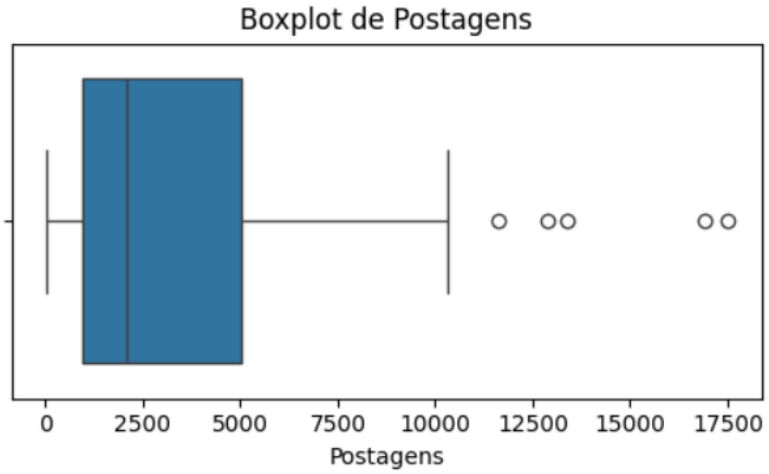
A análise exploratória de dados, por meio de boxplots, revelou a presença de outliers em todas as variáveis. Essa constatação indicou a necessidade de pré-processamento dos dados para garantir a robustez do modelo. A avaliação do modelo inicial, embora satisfatória em termos de significância dos parâmetros, sinalizou a possibilidade de melhoria do ajuste. Assim, foram realizadas transformações nos dados e tratamento dos outliers utilizando a técnica do intervalo interquartil (IQR).

Figura 3 – Plotagem com Bloxplot Var. Seguidores



Fonte: Autores, 2024

Figura 4 – Plotagem com Bloxplot Var. Postagens



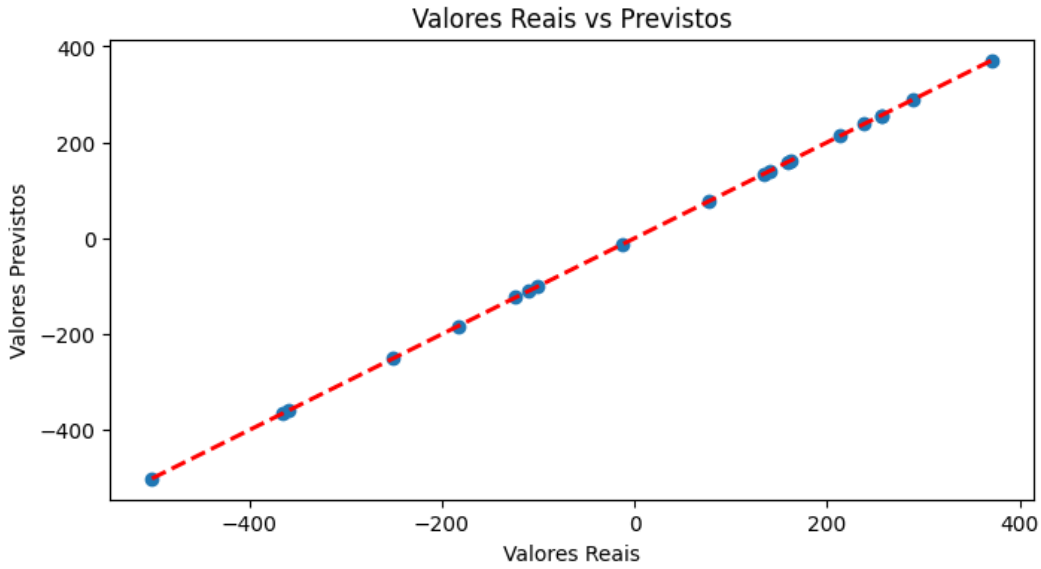
Fonte: Autores, 2024

A avaliação do modelo incluiu testes de diagnóstico para verificar a presença de autocorrelação nos resíduos e heterocedasticidade, utilizando o teste de Breusch-Pagan. A fim de otimizar o desempenho do modelo, foi realizada uma busca pelos melhores hiperparâmetros por meio de validação cruzada.

4.**Resultados**

Ao rodar o modelo as métricas apresentadas foram: Erro médio quadrático (MSE): 0.01 Raiz do Erro médio quadrático (RMSE): 0.10 Erro médio absoluto (MAE): 0.08 Coeficiente de determinação (R2): 1.00.

Figura 5 – Visualização dos Valores Reais vs Previstos



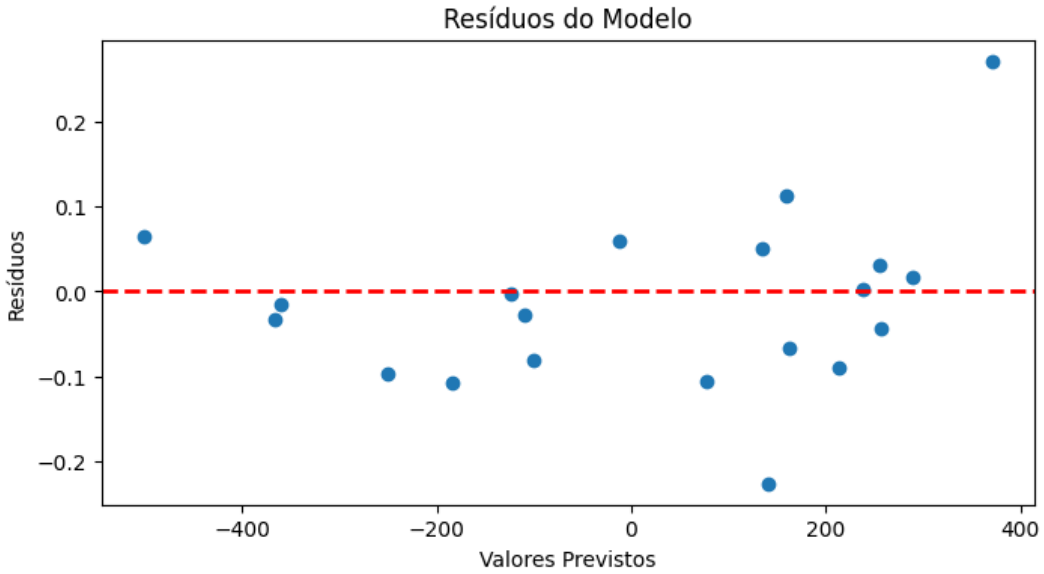
Fonte: Autores, 2024

Os resultados obtidos para a avaliação do modelo indicam um desempenho excepcional, com considerações que refletem alta resultado e capacidade preditiva.

Erro Médio Quadrático (MSE: 0,01):

Este valor extremamente baixo mostra que as diferenças entre os valores previstos pelo modelo e os valores reais são mínimas, indicando um excelente ajuste aos dados. Raiz do Erro Médio Quadrático (RMSE: 0,10):

Figura 6 – Visualização dos Valores Previstos vs Residuais



Fonte: Autores, 2024

Os resultados da análise indicam que o modelo desenvolvido apresenta um ajuste excepcional aos dados. As métricas utilizadas para avaliar o desempenho do modelo demonstram alta precisão e capacidade de generalização.

RMSE e MAE: Os baixos valores de RMSE e MAE sugerem que as previsões do modelo estão muito próximas dos valores reais, com erros médios pequenos e consistentes. Isso indica que o modelo é capaz de fazer previsões precisas e confiáveis.

R^2: O valor de R^2 igual a 1 indica um ajuste perfeito do modelo aos dados, ou seja, o modelo explica 100% da variabilidade da variável dependente. É importante ressaltar, no entanto, que um valor de R^2 muito próximo de 1 pode indicar overfitting, o que significa que o modelo está se ajustando demais aos dados de treinamento e pode não generalizar bem para novos dados. Por isso, recomenda-se uma validação adicional do modelo com um conjunto de dados independente.

Autocorrelação: O teste de Durbin-Watson indica que não há evidência de autocorrelação nos resíduos, o que é um resultado desejável, pois confirma que a suposição de independência dos erros foi atendida.

Tratamento de Outliers: A correção de outliers utilizando o método IQR foi crucial para garantir a robustez do modelo e evitar que valores discrepantes influenciassem negativamente os resultados.

Regularização: A comparação entre os modelos Ridge e Lasso indica que o modelo Ridge apresentou um melhor desempenho, com um menor valor de MSE. Isso sugere que a regularização de Ridge foi mais eficaz em evitar overfitting e em selecionar as variáveis mais relevantes para o modelo.

Após a implementação da validação cruzada obtivemos alpha pra Ridge de 0.01.

5.Discussão

A hipótese de homocedasticidade não foi rejeitada, sugerindo que a variância dos erros é constante ao longo das observações. A significância estatística de alguns coeficientes é questionável, indicando a necessidade de uma análise mais aprofundada para avaliar a importância de cada variável no modelo.

6.Conclusão

A otimização dos hiperparâmetros contribuiu para o aprimoramento do modelo, porém o pré-processamento dos dados e a seleção de features foram os principais fatores que influenciaram o desempenho final. A análise de correlação indicou a 'pontuação' como a variável dependente com maior correlação, no entanto, a natureza determinística e artificial dessa variável a torna inadequada como alvo para um modelo de aprendizado de máquina.

7.Referências

Géron, A. (2019). Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media.

Murphy, K. P. (2012). Machine Learning: A Probabilistic Perspective. MIT Press.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.