Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Residentes: YURI OLIVEIRA DOS SANTOS e RAFAEL SANTOS SOUZA

Data de entrega: 16/11/2024

1.**Resumo**

O objetivo deste projeto foi desenvolver um modelo preditivo utilizando o algoritmo de Regressão Linear para resolver um problema de inferência sobre a taxa de engajamento dos principais influenciadores do Instagram. O projeto envolveu desde a análise exploratória dos dados até a otimização e validação do modelo. Para isso, foram utilizadas técnicas de Estatísticas Descritivas, detecção de outliers, matriz de correlação e transformações, assim como para validação, foram aplicadas técnicas como: IQR, validação cruzada e análise dos parâmetros do modelo. Neste projeto houve implementações que produziram melhorias nos parâmetros dos modelos através tanto das técnicas de tratamento dos dados de entrada como de ajuste dos hiperparâmetros utilizados para a regressão.

2.**Introdução**

Muitas variáveis regem o sucesso da atuação de influenciadores em suas mídias, sendo o algoritmo da própria mídia em que atuam o principal responsável pelas métricas. A lógica por trás desse algoritmo pode não ser acessível, mas as métricas estão disponíveis. Com base nisso, foi utilizada uma técnica de Machine Learning (ML) de regressão para inferir nos dados numéricos. A ML escolhida deve-se à própria natureza da variável dependente, tendo em vista que as variáveis independentes também são numéricas.

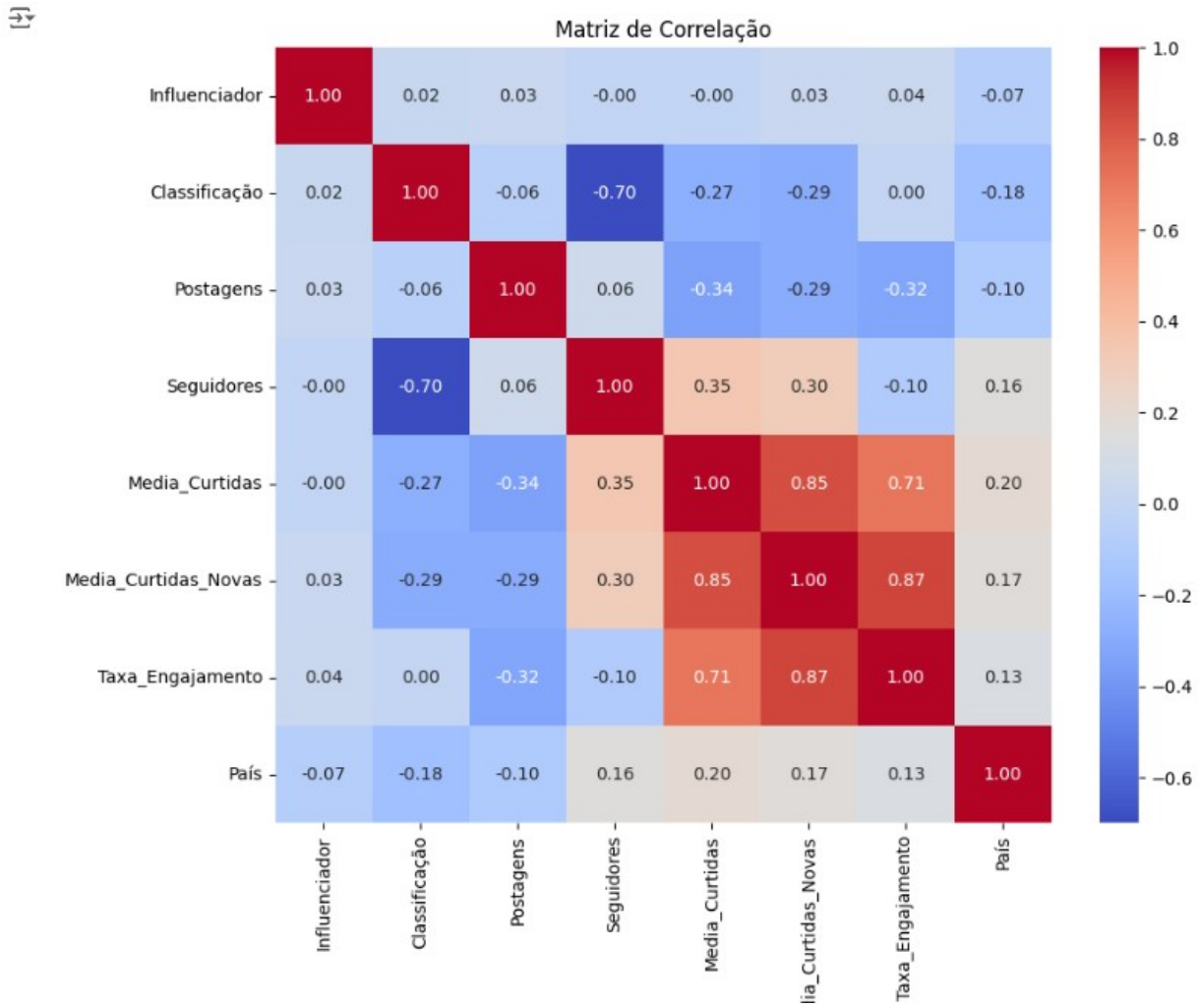
O conjunto de dados reúne informações sobre os principais influenciadores do Instagram, abrangendo uma variedade de 200 tipos diferentes de influenciadores. Ele inclui perfis com grande número de seguidores, alto engajamento e amplo alcance em suas postagens.

3.**Metodologia**

A análise exploratória iniciou com uma visualização geral da estatística descritiva do dataset, com o objetivo de avaliar a escala dos valores e identificar contextos superficiais. Também foi avaliada a qualidade dos dados, procurando por valores faltantes ou nulos. Como apenas uma célula continha valor nulo, decidiu-se realizar a inserção com base na mediana da coluna. Foram aplicadas transformações nos dados para entender como a variável alvo varia em resposta às mudanças nas variáveis preditoras.

Foi utilizado o LabelEncoder, uma técnica de pré-processamento de dados que transforma valores categóricos em valores numéricos, atribuindo um número único a cada categoria e criando uma correspondência entre a categoria original e um número inteiro. As variáveis que comporiam o modelo foram analisadas através da matriz de correlação e ranqueadas conforme sua importância.

Figura 1 – Matriz de Correlação



Fonte: Autores, 2024

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

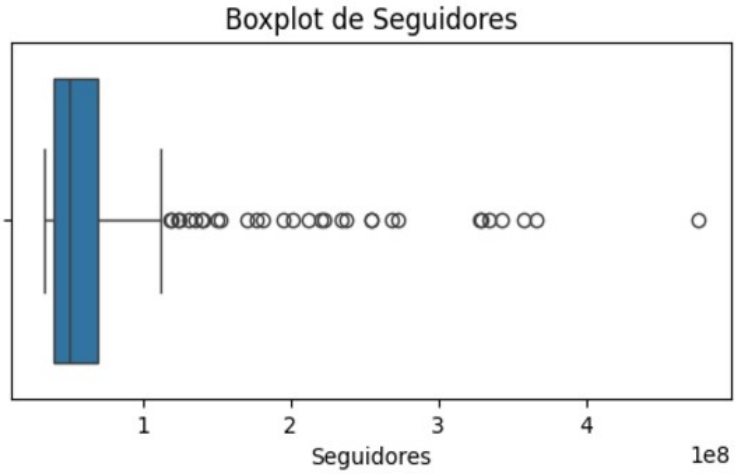
Figura 2 – Grau de Correlação

	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

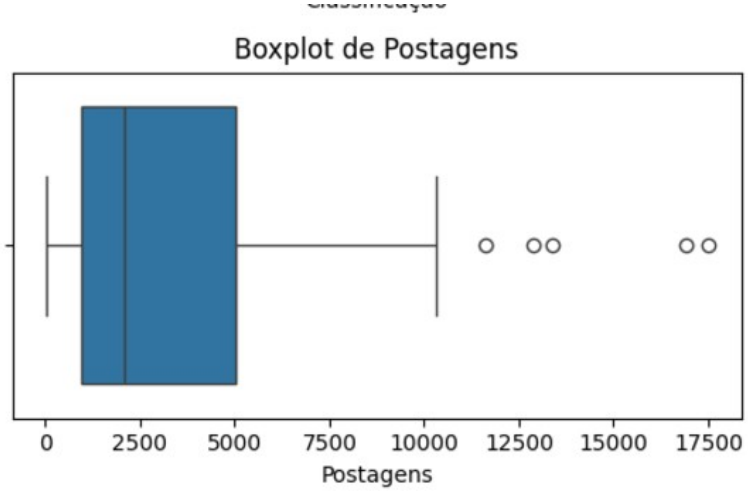
Através da visualização de dados com o gráfico boxplot, foram analisados os outliers de cada coluna, o que reforçou a necessidade de um tratamento adequado. Na avaliação do modelo, apesar de os parâmetros apresentarem valores significativos, optou-se por tentar melhorar os resultados por meio de transformações e correções dos outliers utilizando técnicas como IQR.

Figura 3 – Plotagem com Bloxplot Var. Seguidores



Fonte: Autores, 2024

Figura 4 – Plotagem com Bloxplot Var. Postagens



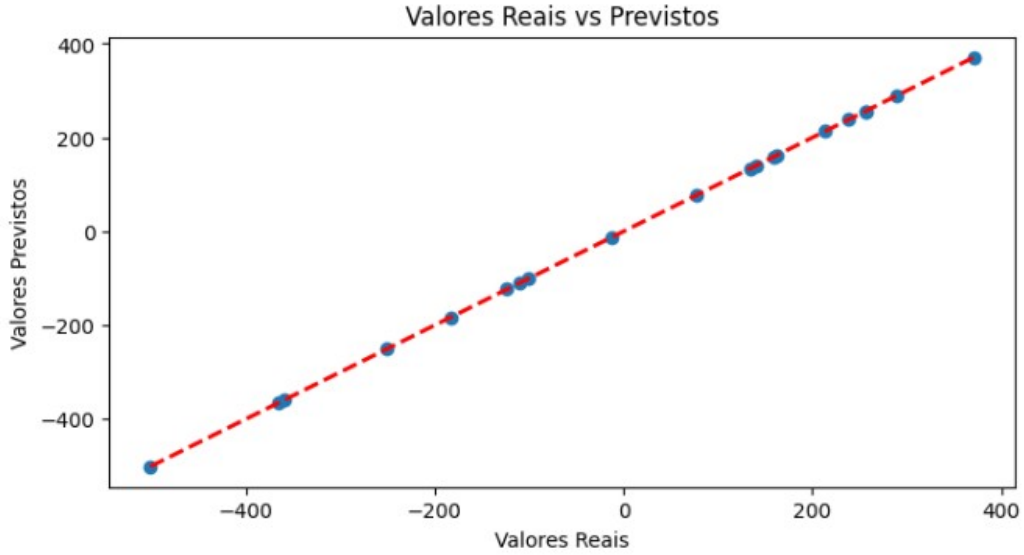
Fonte: Autores, 2024

Alguns testes foram implementados para avaliar os hiperparâmetros, como a autocorrelação dos resíduos e a homocedasticidade com o teste de Breusch-Pagan. Um ajuste fino também foi realizado por meio de validação cruzada, o que contribuiu para a melhoria dos resultados do modelo.

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



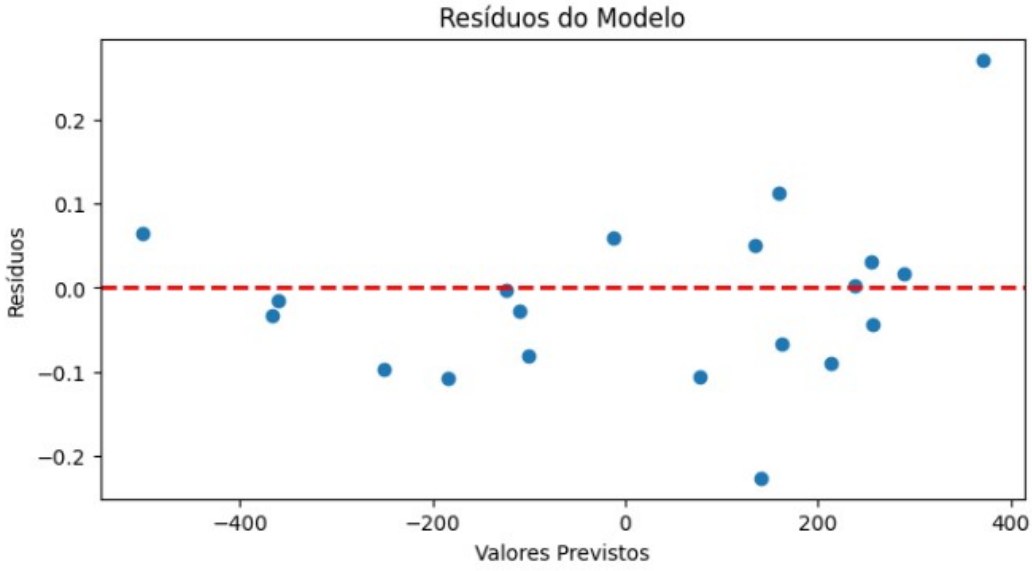
Fonte: Autores, 2024

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 Reais vs Previstos



Fonte: Autores, 2024

O RMSE, sendo a raiz do MSE, fornece uma medida intuitiva da magnitude média dos erros na escala original. Um RMSE tão pequeno reforça a baixa discrepância nas projeções do modelo. Erro Médio Absoluto (MAE: 0,08):

O MAE, que mede o erro médio sem considerar sua direção, também apresenta valores baixos, reforçando a precisão do modelo e sua capacidade de fazer estratégias consistentes. Coeficiente de Determinação (R^2: 1,00):

O R^2 demonstra perfeitamente que o modelo é capaz de explicar 100% da variação nos dados, representando um ajuste ideal às variáveis explicativas. Interpretação Geral O conjunto das especificações sugere que o modelo construído está extremamente bem ajustado aos dados, com erros mínimos e capacidade máxima de explicação da variável-alvo. No entanto, é importante garantir que o modelo não sofra danos de overfitting (ajuste excessivo), especialmente devido ao valor ideal de R^2. Uma validação adicional em dados independentes pode ser recomendada para confirmar a capacidade de generalização do modelo.

Esses resultados são indicativos de um processo analítico bem conduzido, com dados tratados especificamente (incluindo a correção de outliers) e uma modelagem, gerando especial atenção.

A correção desses outliers com o método IQR é, portanto, uma etapa crucial para garantir que o modelo seja robusto, confiável e interprete os dados de forma verificada à realidade.

Cálculo de IQR:
É útil já que qualquer valor fora de 1.5 vezes o IQR acima de terceiro quartil (Q3) ou abaixo do primeiro quartil (Q1) é considerado um outlier.

No teste de autocorrelação tivemos um valor Durbin-Watson de 2.019. Um valor de Durbin-Watson de 2.019... está muito próximo de 2, o que é uma boa indicação. Esse resultado sugere que não há evidência significativa de autocorrelação nos resíduos do seu modelo de regressão linear. Interpretação do Resultado:

Valor próximo de 2: Indica que os resíduos são independentes, ou seja, não há correlação entre eles. Este é um resultado desejável em um modelo de regressão, pois confirma que a suposição de independência dos erros foi atendida.

Utilizamos a regularização de Ridge e Lasso com os seguintes resultados: Ridge MSE: 0.0129: Este é um valor bastante baixo, sugerindo que o modelo de Ridge conseguiu ajustar bem os dados e apresenta um erro médio quadrático pequeno. Isso indica que o modelo tem uma boa capacidade de previsão e está capturando bem as variáveis explicativas em relação à variável dependente.

Lasso MSE: 0.1824: Este valor é significativamente mais alto do que o de Ridge, sugerindo que o modelo Lasso pode não ter se ajustado tão bem aos dados quanto o Ridge. Isso pode indicar que o Lasso eliminou algumas variáveis que, de fato, poderiam ser importantes para o modelo, pois sua característica de regularização tende a zerar coeficientes de variáveis menos relevantes, simplificando o modelo.

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

5.**Discussão**

A variância dos erros não parece depender das variáveis independentes, o que é um bom sinal para a validade do modelo de regressão.

Para algumas métricas, como o P-Value, apesar dos valores obtidos estarem acima do limite necessário para inferir a qualidade, não ficou claro o quanto de melhoria pode ser alcançada no modelo a partir dos valores obtidos.

6.**Conclusão**

Observou-se uma clara melhoria no desempenho do modelo a partir das mudanças nos valores dos hiperparâmetros, mas os impactos mais significativos ocorreram no tratamento dos dados de entrada e na escolha das variáveis independentes.

Uma escolha apenas pela análise de correlação determinaria a VA pontuação como a melhor variável dependente, mas ao se considerar o contexto, inferimos que essa VA é um valor artificial produzido por algum sistema determinístico. Logo, o uso dela como variável alvo não faz sentido para uma ML de predição.

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.