PROGRAMA RESIDÊNCIA EM TECNOLOGIA DA INFORMAÇÃO E COMUNICAÇÃO - TIC 36

JÉSSICA PEREIRA DA SILVA
REBECCA SANTANA SANTOS

RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO DE REGRESSÃO LINEAR

ILHÉUS - BAHIA
2024

RESUMO

Com o crescimento do Instagram, a análise de fatores que influenciam o engajamento digital é estratégica. Este estudo utilizou Regressão Linear para identificar variáveis relevantes no engajamento de influenciadores, como seguidores, publicações e curtidas médias. Técnicas como regularizações e validação cruzada foram aplicadas para evitar overfitting. Os resultados indicaram padrões não lineares, sugerindo limitações do modelo linear em capturar complexidades. Concluiu-se que variáveis como seguidores apresentam saturação, destacando a necessidade de métodos não lineares e inclusão de variáveis qualitativas em estudos futuros. O trabalho contribui para estratégias otimizadas no marketing digital e de influência.

1 INTRODUÇÃO

Com a crescente popularidade das redes sociais, entender as dinâmicas que influenciam o engajamento nas plataformas digitais tornou-se uma necessidade estratégica para empresas e influenciadores. Nesse cenário, o Instagram, como uma das plataformas mais relevantes, desempenha um papel crucial na conexão entre marcas e consumidores, oferecendo interações diretas e personalizadas. Este fenômeno é destacado por Kotler (2017) no conceito de "marketing 4.0", que enfatiza o uso de canais digitais para fomentar conexões significativas.

No contexto das redes sociais, influenciadores digitais têm se estabelecido como mediadores dessas interações. O sucesso de suas estratégias é frequentemente avaliado por métricas como taxas de engajamento, alcance e consistência nas interações com o público. Essas métricas são determinantes para compreender o impacto das ações realizadas, ajudando tanto na definição de estratégias para influenciadores quanto no planejamento de campanhas por parte das empresas.

Dada a importância dessas métricas, este projeto busca investigar os fatores que impactam diretamente as taxas de engajamento dos influenciadores no Instagram. A análise é realizada por meio da aplicação de um modelo de Regressão Linear, amplamente utilizado para prever variáveis contínuas e identificar relações entre fatores. Para garantir um processo robusto e confiável, etapas fundamentais como preparação, limpeza dos dados e seleção de recursos são implementadas. A seleção das variáveis mais relevantes é realizada por meio de métodos estatísticos, como a análise de correlação e o uso da ferramenta SelectKBest, que asseguram o foco nas métricas com maior impacto.

Além disso, o projeto aborda a otimização do modelo por meio de regularizações, como Ridge (penalização L2) e Lasso (penalização L1), bem como a busca de hiperparâmetros ideais utilizando validação cruzada com GridSearchCV. Essas técnicas ajudam a evitar o overfitting e garantem um melhor desempenho em dados não vistos, destacando a importância da modelagem preditiva para a cocriação de valor no marketing de influência, conforme observado por Marques (2023).

Assim, a análise realizada neste estudo alinha-se à crescente relevância do marketing digital e de influência, trazendo uma abordagem preditiva baseada em

dados. Os resultados obtidos oferecem percepções valiosas para influenciadores e empresas interessadas em otimizar sua presença no ecossistema digital, reforçando a importância das métricas de engajamento para o sucesso no ambiente online.

2 METODOLOGIA

A metodologia empregada neste projeto foi estruturada em quatro etapas principais, desde a preparação dos dados até a análise dos resultados, com o objetivo de explorar os fatores que impactam o engajamento de influenciadores no Instagram por meio de um modelo de Regressão Linear.

> Hoje em dia, num mundo cada vez mais dependente da informação, a Estatística tornou-se uma ferramenta imprescindível na tomada de decisões, em áreas tão diversas como a Agricultura, a Medicina, a Engenharia ou o Marketing, entre muitas outras (Santos, 2007, p.15).

Santos (2007) destaca a importância da Estatística em diversas áreas, sublinhando seu papel fundamental na tomada de decisões informadas. No âmbito deste projeto, a aplicação de métodos estatísticos, como a Regressão Linear, se torna essencial para compreender e prever os fatores que influenciam o engajamento de influenciadores no Instagram. Ao utilizar essas técnicas, é possível extrair informações valiosas a partir de grandes volumes de dados, permitindo uma análise precisa das variáveis que impactam a interação dos usuários nas redes sociais. Dessa forma, a Estatística não só fundamenta a construção do modelo preditivo, mas também proporciona um suporte estratégico para decisões no contexto digital, cada vez mais dependente de dados (Pereira, 2015).

A primeira etapa consistiu na definição e preparação do problema, utilizando um conjunto de dados armazenado no arquivo data/top_insta_influencers_data.csv. Esse dataset incluía informações relevantes como o número de seguidores (followers), número de publicações (posts), média de curtidas por publicação (avg_likes) e a taxa de engajamento (influence_score). Inicialmente, foi realizada uma análise exploratória dos dados, incluindo a construção de uma matriz de correlação para identificar relações entre as variáveis, histogramas para observar a distribuição das mesmas e gráficos de dispersão para explorar tendências entre as variáveis independentes e a variável alvo (influence_score).

Na segunda etapa, procedeu-se à implementação do algoritmo de Regressão Linear utilizando a biblioteca Scikit-Learn. Para garantir que apenas as variáveis mais relevantes fossem consideradas no modelo, foram aplicadas técnicas de seleção de recursos, como análise de correlação e o método SelectKBest. Além disso, todas as variáveis independentes foram normalizadas, permitindo que seus

valores fossem escalonados uniformemente, o que contribuiu para melhorar a performance do modelo.

A terceira etapa envolveu a otimização e ajustes do modelo por meio de técnicas de regularização, utilizando os métodos Ridge (L2) e Lasso (L1). Enquanto o Ridge introduziu uma penalidade baseada na soma dos quadrados dos coeficientes, o Lasso aplicou penalidades baseadas na soma dos valores absolutos dos coeficientes, ajudando a reduzir possíveis problemas de overfitting. O ajuste dos hiperparâmetros, como o valor de alpha, foi realizado por meio de validação cruzada com o uso de GridSearchCV, garantindo maior robustez ao modelo.

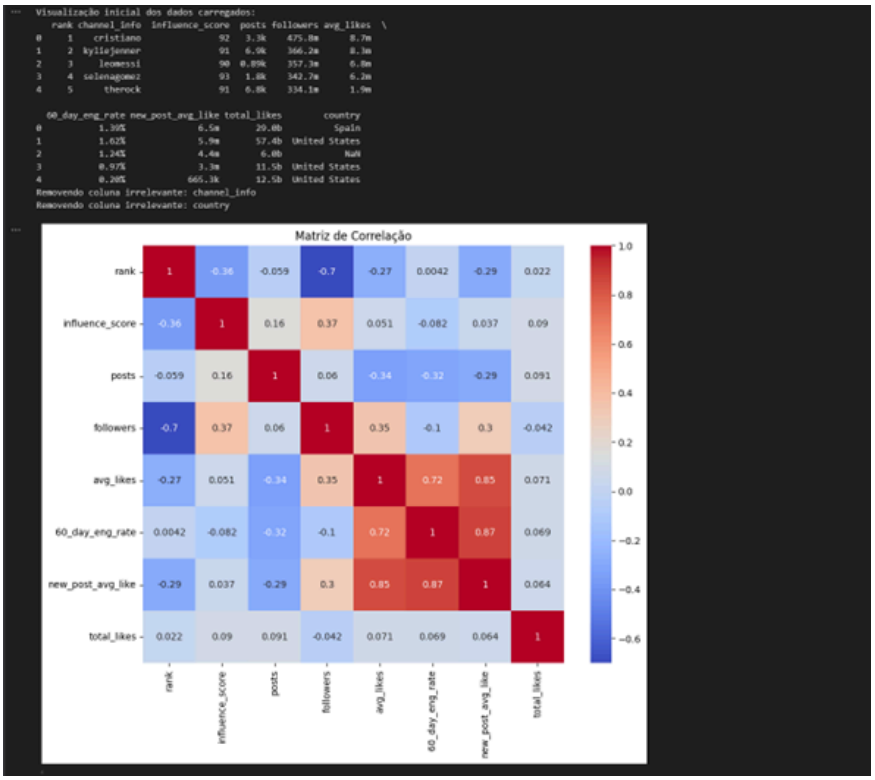
Por fim, a análise e visualização dos resultados focou na avaliação do desempenho do modelo com base em métricas como Erro Médio Quadrático (MSE), Raiz do Erro Médio Quadrático (RMSE) e Erro Absoluto Médio (MAE). Os resultados foram apresentados graficamente, comparando os valores reais e preditos pelos modelos Linear, Ridge e Lasso, permitindo uma interpretação clara sobre a precisão e eficácia das predições realizadas.

Essa abordagem metodológica garantiu uma análise consistente dos dados e contribuiu para a obtenção de percepções relevantes sobre os fatores que influenciam o engajamento dos influenciadores no Instagram.

3 RESULTADOS

Os resultados obtidos neste estudo demonstram o impacto de diferentes variáveis nas taxas de engajamento dos influenciadores no Instagram. A análise inicial destacou as relações entre as variáveis independentes e a variável influence_score, conforme ilustrado na Figura 1, que apresenta a matriz de correlação. Essa matriz revelou conexões significativas entre variáveis como rank, followers e posts.

Figura 1 - Matriz de correlação entre as variáveis independentes e a taxa de engajamento (influence_score).



Gráficos de dispersão, como os apresentados nas Figuras 2 a 6, detalham as relações entre as variáveis selecionadas e a taxa de engajamento. Por exemplo, a Figura 3 demonstra que a variável followers possui um padrão de saturação, no qual altos números de seguidores não correspondem necessariamente a engajamentos proporcionais. Outros gráficos, como a Figura 4 (avg_likes) e a Figura 5 (60_day_eng_rate), indicam tendências não lineares, reforçando a complexidade das relações.

Figura 2 - Distribuição das variáveis no conjunto de dados analisados.

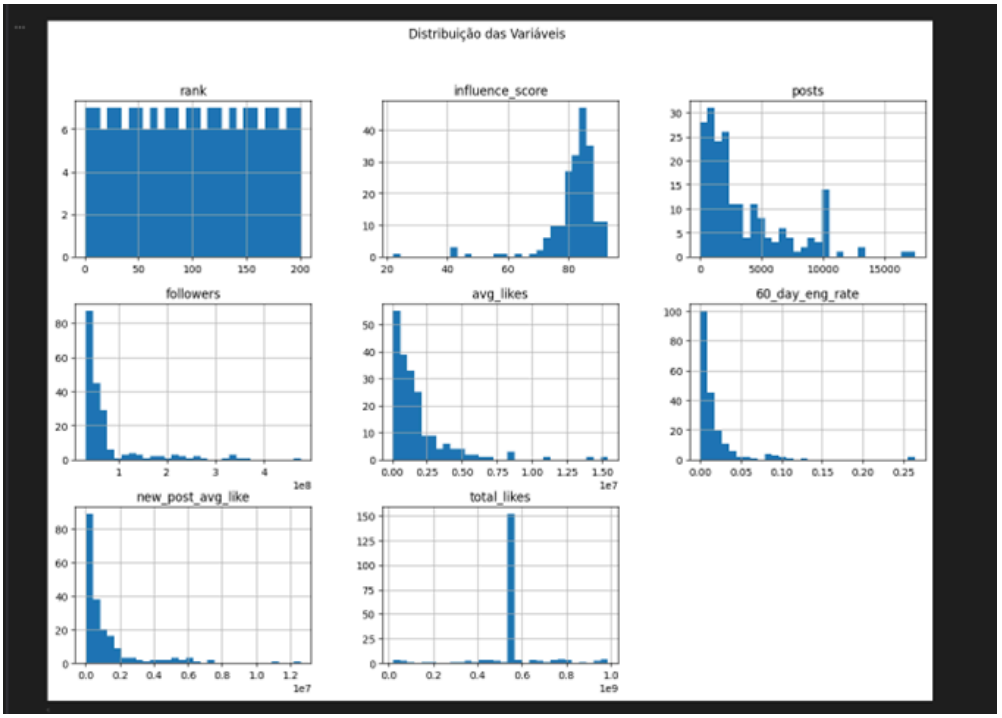


Figura 3 - Relação entre a variável followers e o índice de engajamento (influence_score).

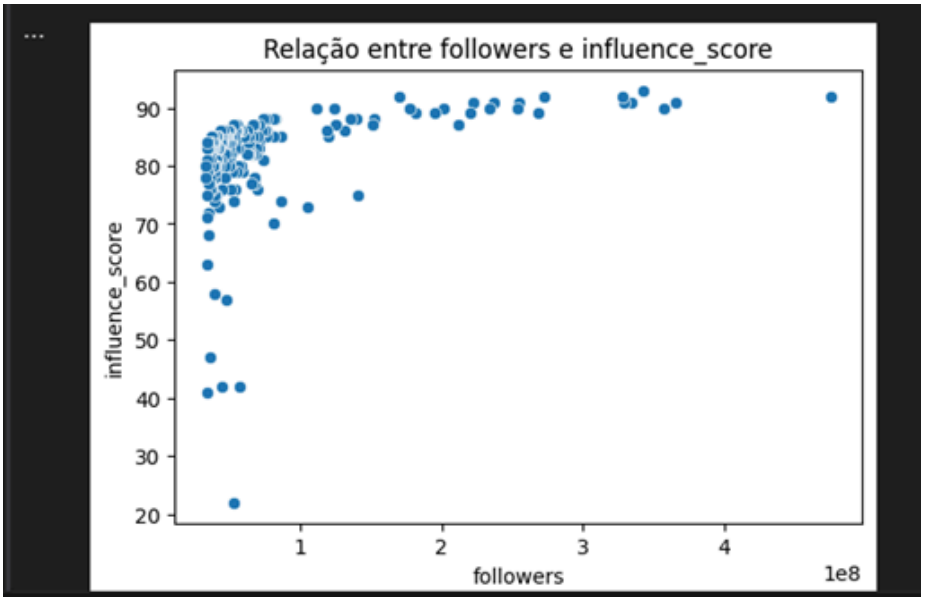


Figura 4 - Relação entre a variável avg_likes e o índice de engajamento (influence_score).

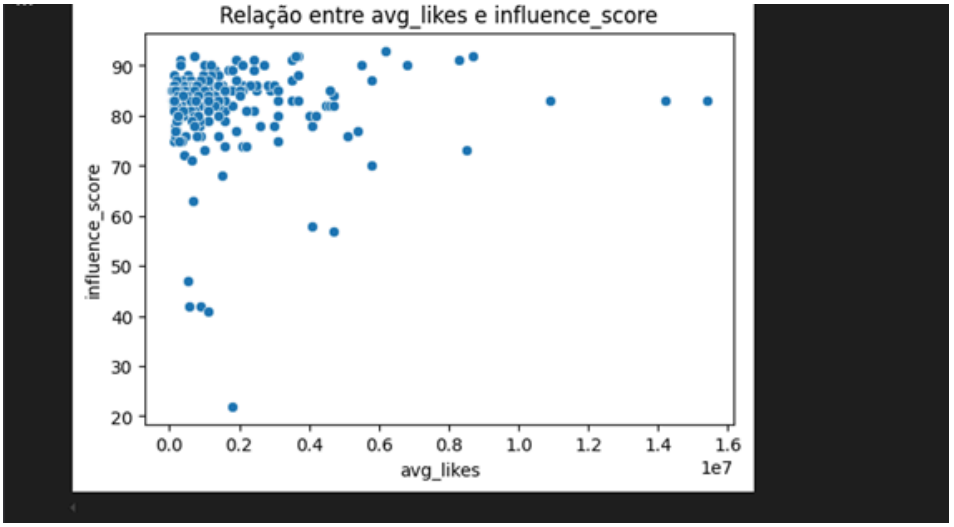
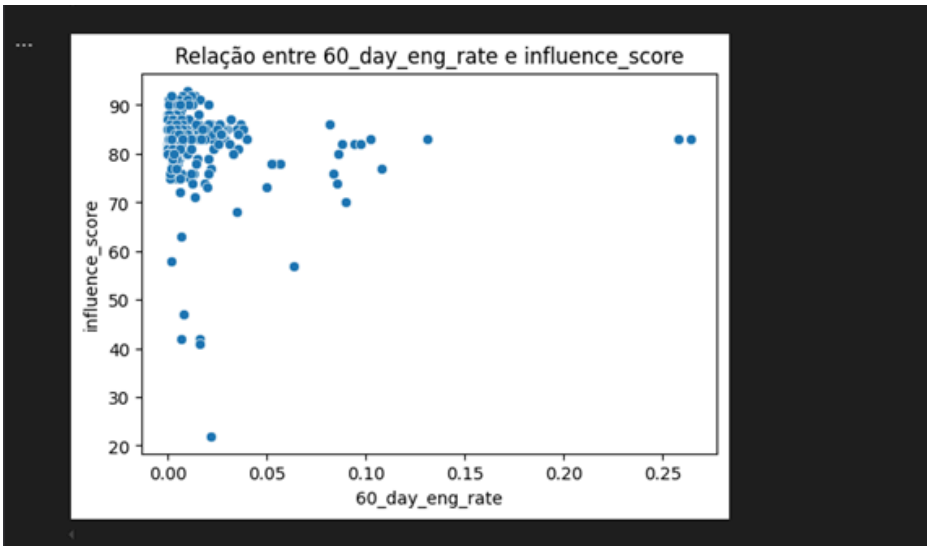


Figura 5 - Relação entre a variável 60_day_eng_rate e o índice de engajamento (influence_score).



No modelo de Regressão Linear, o RMSE foi calculado como 12.74, conforme demonstrado na Figura 7. Modelos com regularização, como Ridge e Lasso, apresentaram RMSEs de 12.74 e 12.75, respectivamente, sugerindo que a adição de regularização não trouxe melhorias significativas para a generalização. Esse comportamento é corroborado pela proximidade entre os valores preditos e os valores reais, evidenciado também na Figura 7, que mostra a comparação entre esses valores.

Figura 6 - Distribuição da variável total_likes, indicando assimetria nos dados.

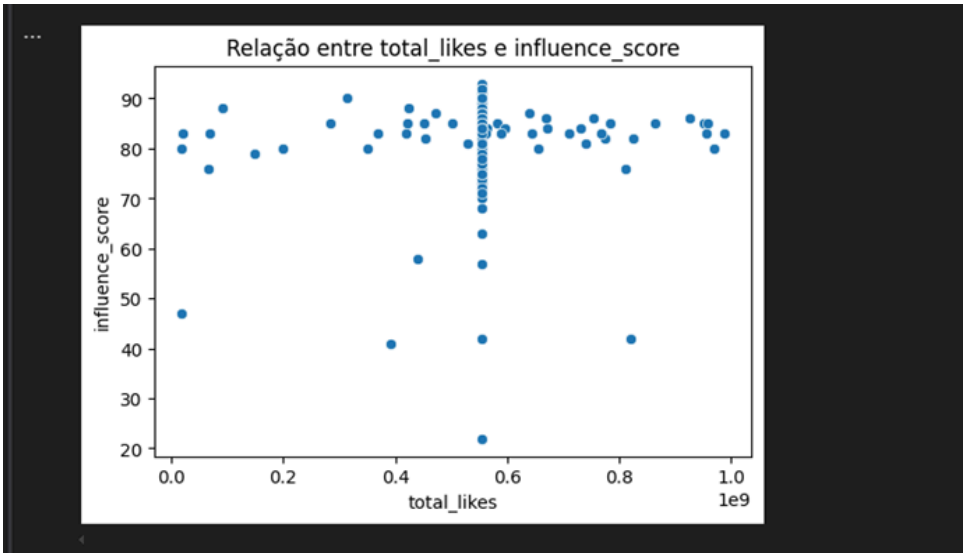


Figura 7 - Comparação entre valores reais e preditos pelos modelos Regressão Linear, Ridge e Lasso.

4 DISCUSSÃO

Os resultados apontam que as variáveis rank, followers e posts possuem o maior impacto direto na previsão do engajamento. No entanto, os padrões de saturação e não linearidade observados em variáveis como followers e avg_likes indicam que o engajamento digital é multifatorial e depende de variáveis qualitativas não capturadas neste modelo linear.

A análise exploratória revelou distribuições assimétricas em variáveis como total_likes (Figura 6), destacando a influência de outliers no conjunto de dados. Essa observação sugere a necessidade de métodos mais robustos para lidar com valores extremos, como o uso de técnicas de normalização avançadas.

Embora o modelo linear apresenta boa capacidade de previsão, os gráficos reforçam a necessidade de explorar métodos não lineares, como Random Forest ou Gradient Boosting, para capturar padrões mais complexos. Variáveis como avg_likes e 60_day_eng_rate merecem atenção especial em estudos futuros devido à sua relação não linear com o influence_score.

5 CONCLUSÃO E TRABALHOS FUTUROS

Conclui-se que o modelo de Regressão Linear é eficaz na análise de fatores que impactam o engajamento de influenciadores no Instagram, mas apresenta limitações ao capturar relações mais complexas. As variáveis rank, followers e posts foram as mais relevantes no contexto do estudo, sendo consistentes entre a análise exploratória e o modelo preditivo.

Como próximos passos, recomenda-se:

5.1 Explorar modelos mais sofisticados para capturar relações não lineares.

5.2 Realizar uma análise mais detalhada de outliers para evitar impactos negativos nas métricas de desempenho.

5.3 Incorporar variáveis qualitativas que possam complementar os fatores já analisados.

6 REFERÊNCIAS

KOTLER, Philip. Marketing 4.0: dal tradizionale al digitale. 2017.

LUCCHESI, Reinaldo Nascimento. **Marketing: dá origem à sociedade de consumo.** *Revista Hórus*, v. 5, n. 1, p. 79-101, 2010.

MARQUES, Alzira**. Marketing Relacional**. 3. ed. Lisboa: Edições Sílabo, 2023.

Pereira, J. L. (2015). Análise Preditiva em Sistemas de Informação no contexto do Big Data.

SANTOS, Carla. Estatística descritiva. **Manual de auto-aprendizagem**, v. 2, 2007.