RELATÓRIO TÉCNICO

Implementação e Análise do Algoritmo de Regressão Linear

Marya de Souza Fernandes Matos
Daiane Gomes Meira
17/nov/2024

Marya de Souza Fernandes Matos

Daiane Gomes Meira

Resumo

O Instagram é uma das plataformas de redes sociais mais populares da atualidade e serve como uma grande oportunidade de marcas, empresas e indivíduos se conectarem com a sua audiência da forma mais assertiva possível. Nesse cenário, a famosa métrica “taxa de engajamento” serve como um importante indicador de desempenho na avaliação da eficácia das estratégias adotadas por essas contas. O objetivo principal desse projeto foi desenvolver um modelo preditivo robusto que consiga prever a taxa de engajamento com base em outras métricas do Instagram, e, consequentemente, permitir uma otimização das estratégias de conteúdo.

A metodologia envolveu a aplicação de padronização dos dados; técnicas de transformação das variáveis para estabilizar a variância e melhorar a normalidade das distribuições; a exploração de modelos de regressão linear, e também de diversos modelos de regularização, que buscavam mitigar problemas de multicolinearidade e melhorar a interpretabilidade das variáveis; validação cruzada para garantir que os resultados fosse robustos e também para buscar pelos melhores hiperparâmetros nas técnicas de regularização.

Os resultados obtidos indicaram que o modelo linear foi capaz de explicar 99% da variância na taxa de engajamento, com diferenças muito pequenas entre os erros de treino e teste, com um erro quadrático médio extremamente baixo, sugerindo uma alta precisão preditiva. Indicaram também a ausência de sinais de overfitting ou underfitting e de heterocedasticidade, reforçando a robustez do modelo.

Introdução

Inicialmente, o Instagram surgiu como uma rede social onde as pessoas pudessem postar fotos e vídeos em suas contas pessoais e, assim, interagir com outras pessoas do mundo todo diariamente. Atualmente, com o intuito de estreitar laços com seus consumidores e de marcar presença nos lugares onde eles passam a maior parte do tempo, as marcas, empresas e influenciadores usam o Instagram não

só para uma questão de visibilidade, mas também como uma peça-chave na estratégia de marketing digital. Nesse contexto, a métrica de “taxa de engajamento” se tornou um importante indicador que reflete o nível de interação dos usuários com os conteúdos publicados, conseguindo fornecer insights valiosos sobre a eficácia de campanhas e a repercussão do conteúdo com o seu público alvo.

Dada a importância estratégica da taxa de engajamento, o desenvolvimento de modelos preditivos para estimar essa métrica pode oferecer vantagens competitivas significativas. Algoritmos de regressão, como os utilizados neste projeto, são particularmente adequados para este fim, já que permitem entender como diferentes fatores contribuem para o engajamento e prever essa taxa com base em outras métricas mensuráveis do Instagram, como o número de seguidores e a média de curtidas.

O conjunto de dados utilizado neste estudo foi obtido a partir do Kaggle e é composto por métricas coletadas de várias contas do Instagram. O dataset apresenta 200 observações e 10 variáveis abrangendo dados quantitativos e qualitativos, incluindo o número de seguidores, de postagens, média de curtidas por postagem, total de likes e outras variáveis relevantes. Cada uma dessas métricas foi pré-processada e transformada conforme necessário, buscando facilitar o entendimento e o manuseio desses dados, garantindo a qualidade deles e a robustez das previsões.

Metodologia

A etapa inicial do projeto constituiu na análise exploratória dos dados, com o objetivo de entender a estrutura e as características do conjunto de dados, compreender a distribuição das variáveis, identificar padrões relevantes e detectar possíveis problemas, como valores ausentes, valores duplicados, outliers e correlação entre as variáveis.

Foram gerados histogramas para verificar a distribuição de cada variável e identificar a presença de outliers, que foram tratados pelo IQR (Método de Distânia Interquartil). Como os histogramas das variáveis não apresentaram uma distribuição normal que facilita o uso de modelos lineares, transformações logarítmica e boxcox

foram aplicadas nessas variáveis. Um teste de Shapiro-Wilk foi executado pra avaliar os valores-p e as estatísticas W e confirmar a normalidade das distribuições. Os resultados obtidos apontaram pra um sucesso maior associado à transformação boxcox.

Uma das variáveis do dataset é categórica e diz respeito ao país de origem da conta relacionada. Então uma análise de variância (ANOVA) foi realizada para verificar se ela contribui significativamente pro modelo e, de acordo com as métricas obtidas, pôde-se perceber que não valia a pena fazer nenhuma codificação dessa variável já que ela não era relevante para o modelo.

A regressão linear foi escolhida como o algoritmo a ser trabalhado por ser adequado a problemas de previsão contínua e por sua interpretabilidade. Os dados começaram sendo normalizados com o *MinMaxScaler*, depois a variável-alvo foi separada das variáveis independentes e, por último, os dados foram divididos em um conjunto de treino (80%) e um conjunto e teste (20%), o que garantia que o modelo fosse avaliado em dados não vistos. O modelo de regressão linear foi implementado usando a biblioteca *scikit-learn*, que permite uma integração eficiente com diversas outras ferramentas de análise. Feito isso, foram realizadas previsões no conjunto de dados de teste e seguiu-se para a avaliação do modelo utilizando métricas como o R^2, o RMSE, o MSE e o MAE tanto no conjunto de dados de treino quanto nos de teste. Foram gerados, também, curvas de aprendizado com base no R^2 e no MSE e um gráfico de dispersão relacionando os valores reais e os valores previstos pelo modelo. Ambos sugeriram uma necessidade de ajustes no modelo e uma possível presença de overfitting, que acontece quando o modelo se ajusta demais aos dados de treino, mas não performa tão bem nos dados de teste.

Tentando entender melhor a relação entre as variáveis, foi realizada uma análise dos coeficientes do modelo, que oferece, de forma geral, um panorama sobre quaia variáveis são significativas ou não para o modelo. Para garantir a robustez do modelo, foi utilizada a validação cruzada com a técnica de *k-folds cross-validation*. Nesse caso, o conjunto de dados foi dividido em 10 subconjuntos que foram utilizados para teste e a cada iteração realizaram a avaliação do modelo, fornecendo métricas para cada subconjunto.

Foram gerados histogramas, gráficos Q-Q e gráficos de dispersão dos resíduos para verificar possíveis padrões neles, o que poderia sugerir um problema de heterocedasticidade. Como, através dos gráficos, não foram encontrados indícios claros desse problema, um teste de Breusch-Pagan também foi realizado pra confirmar esse diagnóstico. De fato, o p-valor encontrado mostrou que não houveram evidências significativas de heterocedasticidade.

Uma matriz de correlação foi construída para identificar relações entre as variáveis independentes e a variável-alvo, além de detectar possíveis multicolinearidades entre elas. Com base nos resultados dessa matriz, foi realizada uma análise de multicolinearidade (VIF), que sugeriu uma correlação moderada, mas não preocupante, entre duas variáveis do modelo. A inclusão de interação entre variáveis foi efetuada na tentativa de melhorar essa questão da multicolinearidade, mas acabou piorando.

Embora a regressão linear tradicional não possua hiperparâmetros complexos, um ajuste de hiperparâmetros foi realizado através de uma busca em grade, utilizando o *GridSearchCV*, com o objetivo de testar uma série de valores para o parâmetro de regularização *alpha*, nos modelos Ridge, Lasso e no Elastic Net, que é uma combinação dos dois ao mesmo tempo. Esse processo permitiu identificar o parâmetro que minimizava o MSE e maximizava o R^2. Esses parâmetros foram aplicados em um loop que identificava o tipo de regularização ideal pro modelo, que, nesse caso, foi o Ridge. Aplicando esse modelo, as métricas obtidas foram muito boas (R^2 de 0.98 e MSE de 3.38e-06) e sem sinal de overfitting, mas a análise dos coeficientes do modelo sugeriam um problema de multicolinearidade.

Por último, foi testado um modelo só com as variáveis que a análise dos coeficientes classificou como significantes pro modelo e o R^2 aumentou um pouco (0.99) e o MSE continuou bem baixo, também sem sinal de overfitting, sem evidências de heterocedasticidade e sem problemas de multicolinearidade entre as variáveis.

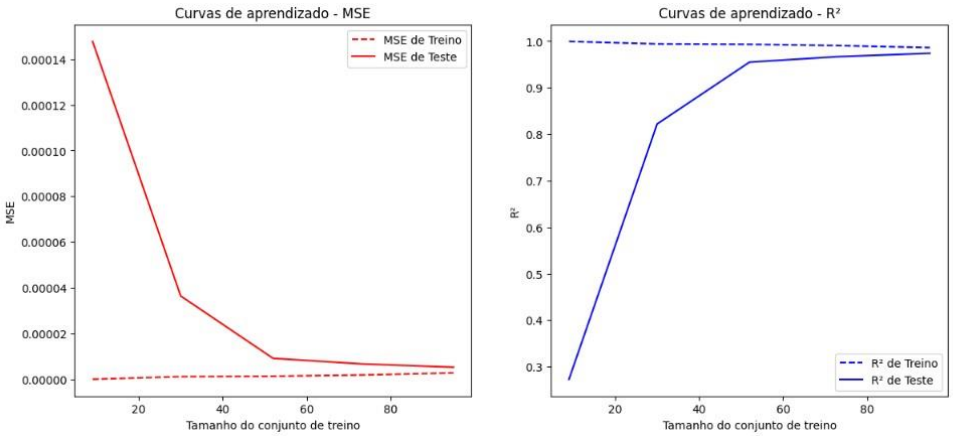
A escolha do melhor modelo foi determinada com base nas melhores métricas e no melhor conjunto de características observadas.

Resultados

Durante o processo do pojeto, alguns modelos foram testados para identificar a solução que apresentava o melhor equilíbrio entre desempenho e simplicidade. De uma forma geral, todos eles performaram bem, mas sugeriam uma margem para melhorias.

O primeiro modelo testado apresentou boas métricas, desempenhos semelhantes nos conjuntos de treino e teste, nenhuma evidência significativa de heterocedasticidade, mas apresentou sintomas de leves problemas relacionados à overfitting, já que nas curvas de aprendizado, antes de convergir, há uma diferença entre os desempenho de treino e teste, como mostra a Figura 1.

Figura 1 – Curvas de aprendizado do modelo 1 testado.



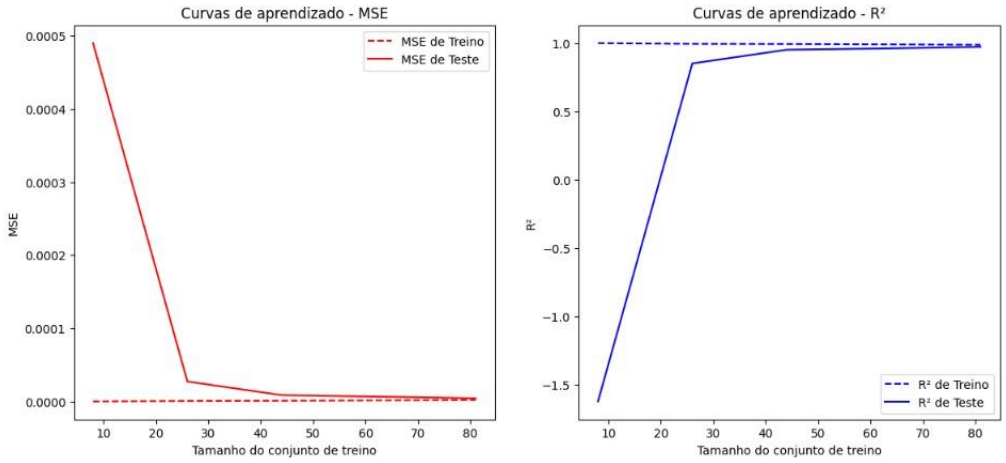
Fonte: Próprio Autor (2024).

Além disso, os resultados obtidos na análise dos coeficientes do modelo, sugeriram a presença de uma forte multicolinearidade entre variáveis. Com uma matriz de correlação e uma verificação de VIF, foi percebido uma correlação preocupante, mesmo não sendo tão problemática. Com essas informações em mãos, novas variáveis foram criadas a partir da interação entre variáveis pré-existentes, tentando buscar um impacto positivo no modelo.

Esse segundo modelo, que contava com novas variáveis, apresentou métricas levemente superiores ao primeiro modelo e mostrou que o desempenho de teste

estabiliza mais rapidamente, se aproximando do desempenho de treino, mostrando uma possível melhora na questão referente ao overfitting, como mostra a Figura 2.

Figura 2 – Curvas de aprendizado do modelo 2 testado.



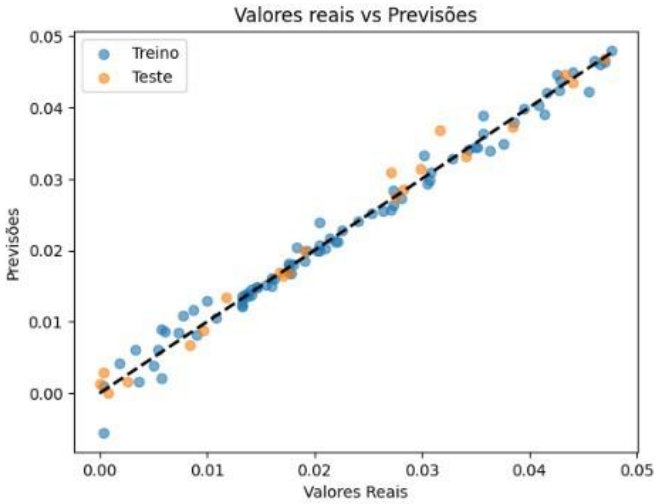
Fonte: Próprio Autor (2024).

Apesar do problema de overfitting ter sido reduzido, uma nova análise dos coeficientes do modelo, seguidos de uma nova verificação de VIF, mostraram que o problema com a multicolinearidade continuou acontecendo e, inclusive, apresentou uma piora.

A análise dos coeficientes do modelo fornece, além de informações sobre colinearidade, métricas que avaliam a significância estatística de cada variável na previsão da variável-alvo. Nos dois modelos, essas métricas foram consistentes e apontaram que as variáveis significativas na previsão eram “media de curtidas novas” e “seguidores”. Essa informação serviu como direcionamento para o treinamento do próximo modelo.

Esse modelo apresentou métricas muito boas, com um R^2 de 0.986, indicando que 98,6 % da variância da variável dependente foi explicada pelas variáveis independentes selecionadas, e um MSE de 3.12e-06, que, por ser um valor bem próximo de zero, indica que as previsões estão muito próximas dos valores reais, como mostra a Figura 3.

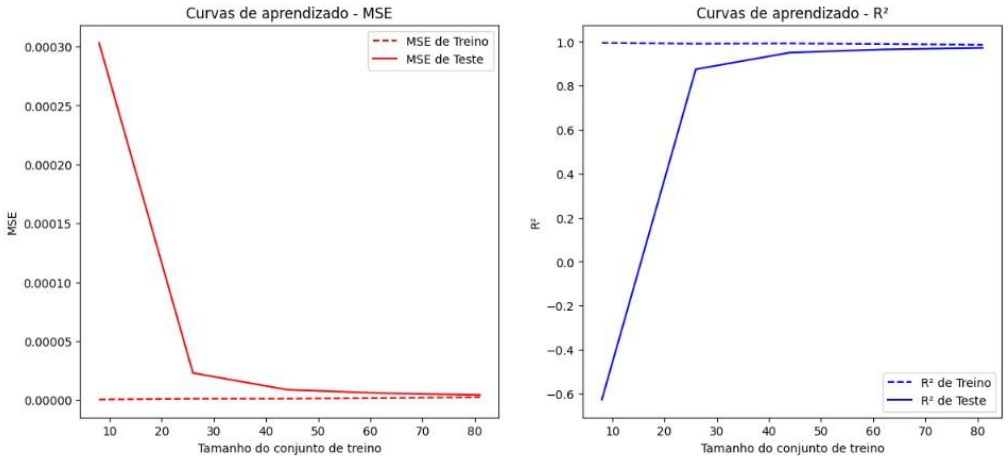
Figura 3 – Gráfico de Dispersão de Valores Reais vs Previsões.



Fonte: Próprio Autor (2024).

Os resultados do modelo foram consistentes nos conjuntos de treino e teste, com um R^2 de 0.9851 no treino e 0.9856 no teste, indicando boa capacidade de generalização. As curvas de aprendizado mostram como o desempenho de teste estabiliza rapidamente, se aproximando do de treino, como mostra a Figura 4.

Figura 4 – Curvas de aprendizado do modelo 3 testado.

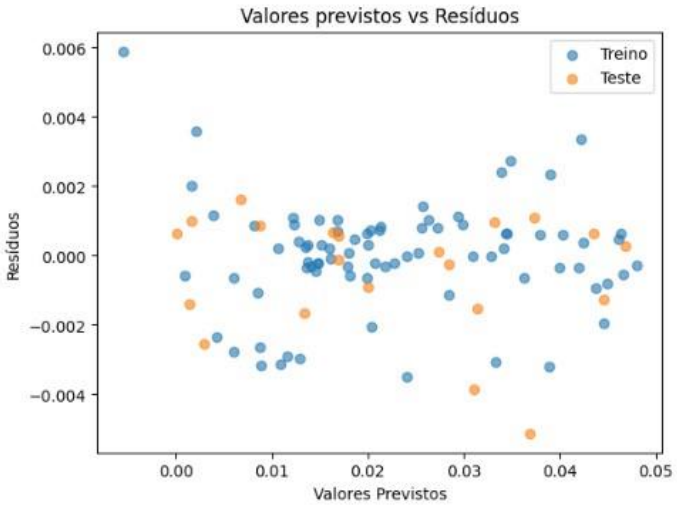


Fonte: Próprio Autor (2024).

Uma validação cruzada do modelo mostrou valores constantes de R^2 nos 10 subconjuntos de dados analisados e um desvio padrão bem baixo, mostrando que o modelo generaliza bem. A análise dos coeficientes do modelo e a verificação do VIF mostrou que o problema na multicolinearidade foi resolvido. Além disso, uma análise

do gráfico de dispersão dos resíduos, associada ao teste de Breusch-Pagan, mostraram a falta de indícios de heterocedasticidade. A dispersão dos resíduos não seguem nenhum padrão visível ao longo dos valores previstos, espalhados ao redor de zero, como mostra a Figura 5.

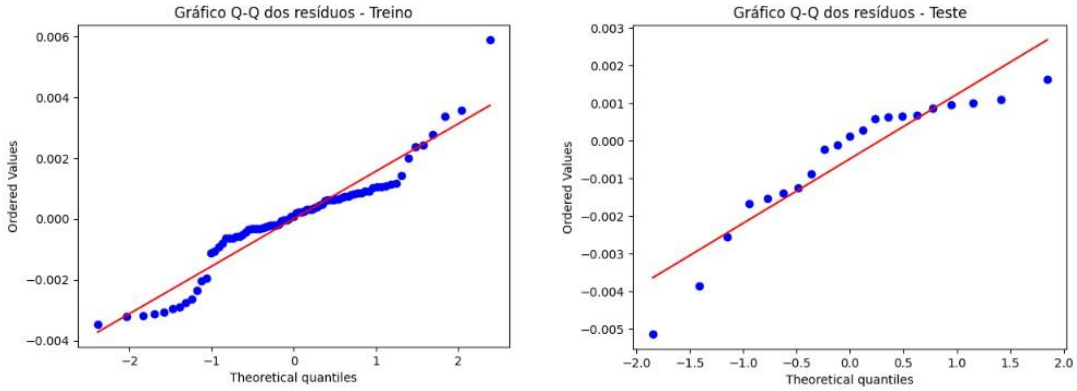
Figura 5 – Gráfico de dispersão dos resíduos.



Fonte: Próprio Autor (2024).

Um gráfico Q-Q dos resíduos sugeriu que eles seguem uma distribuição aproximadamente normal, com alguns indícios de desvios da normalidade em alguns pontos, como mostra a Figura 6.

Figura 6 – Gráfico Q-Q dos resíduos.

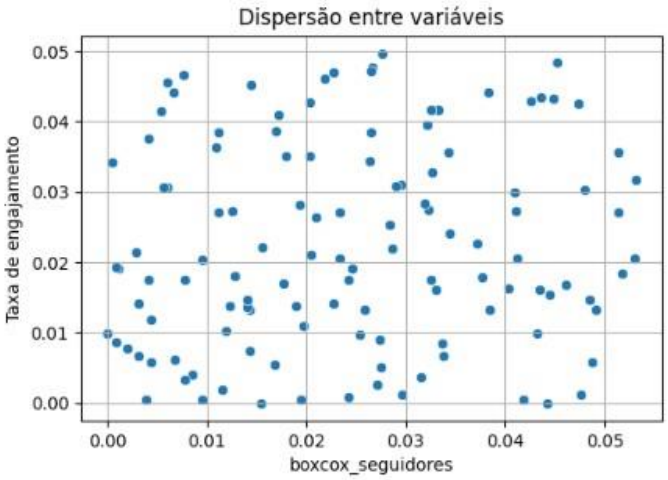


Fonte: Próprio Autor (2024).

Essa situação pode ter sido causada pela não linearidade na relaçao entre a variável “seguidores” e a variável-alvo (“taxa de engajamento”). O modelo de

regressão linear assume linearidade entre as variáveis e, como a relação entre “seguidores” e a “taxa de engajamento” é não linear, como mostra a figura 7, o modelo pode ter tido dificuldade em capturar padrões nos dados, gerando resíduos sistematicamente grandes.

Figura 7 – Relação não-linear entre “seguidores” e “taxa de engajamento”.



Fonte: Próprio Autor (2024).

Outras diferentes configurações com aplicação de regularização, como Elastic Net, Ridge e Lasso foram testadas, mas a regressão linear simples se destacou por ser suficientemente precisa para o contexto do problema e por facilitar a interpretação dos coeficientes das variáveis independentes.

Discussão

Os resultados do modelo de regressão linear foram satisfatórios, explicando cerca de 99% da variância na taxa de engajamento no conjunto de treino e teste, com valores consistentes, sugerindo que o modelo foi capaz de generalizar bem para dados nã vistos. O MSE demonstrou um erro absoluto médio baixo, confirmando a boa qualidade das previsões.

Os gráficos de curva de aprendizado mostraram que o modelo está generalizando bem, com diferenças muito pequenas entre os erros de treino e teste. O R^2 de teste cresce rapidamente à medida que o tamanho do conjunto de treino

aumenta e converge próximo ao R^2 de treino. Acontece o mesmo com o MSE de teste que também baixo e próximo ao de treino. O comportamento de ambos sugere que o modelo generaliza bem para o conjunto de teste e que não há sinais claros de overfitting ou de underfitting.

Apesar de um bom desempenho, no geral, o modelo apresentou algumas limitações. A regressão linear assume uma relação linear entre as variáveis e, embora transformações tenham sido aplicadas para aproximar relações não lineares, algumas variáveis não se ajustaram perfeitamente ao pressuposto de linearidade, o que pode ter introduzido alguma tendência no modelo. Além disso, apesar de ter sido feito um bom tratamento dos outliers, ainda é possível que alguns dados extremos tenham permanecido e isso pode ter influenciado os resultados. De uma forma geral, outras abordagens, como modelos mais complexos, poderiam ser explorados para capturar relações mais complexas.

As escolhas metodológicas tomadas ao longo do projeto tiveram impacto direto no desempenho do modelo. Por exemplo:

- as transformações boxcox foram cruciais para estabilizar a variância e normalizar a distribuição dos dados, permitiu que o modelo fizesse previsões mais precisas, mesmo que não tenha resolvido todos os casos de não linearidade;

- a escolha de variáveis com base na matriz de correlação, na análise de colinearidade e na análise de coeficientes do modelo contribuiu para evitar redundâncias e melhorar a interpretabilidade e a precisão do modelo, assegurando que apenas as variáveis mais relevantes fossem incluídas.

Conclusão e Trabalhos Futuros

Esse projeto proporcionou uma boa compreensão sobre a aplicação de modelos de regressão linear para prever a taxa de engajamento no Instagram. A análise inicial e o pré-processamento dos dados foi essencial para identificar padrões de distriuição dos dados, a presença de outliers, correlações entre as variáveis e colinearidade. Essa etapa guiou decisões críticas, como o tratamento dos outliers, as

transformações para ajustar relações e a escolha de variáveis significativas, por exemplo.

A aplicação de transformações foi essencial para lidar com problemas de variância e normalidade, mesmo que não tenha resolvido a questão da não-linearidade de algumas variáveis. Essa é mais uma prova da importância de uma boa preparação dos dados.

Todas as decisões tomadas durante o projeto impactam diretamente no desempenho do modelo. Isso reforça a necessidade de fundamentar e basear todas as escolhas em evidências e não fazer escolhas arbitrárias.

O modelo desenvolvido apresentou resultados excelentes em termos de desempenho geral e estabilidade, mesmo com algumas limitações. Ele oferece uma ferramenta valiosa para prever a taxa de engajamento no Instagram.

Para melhorar ainda mais os resultados e superar as limitações identificadas, algumas sugestões poderiam ser exploradas em trabalhos futuros, como a inclusão de termos polinomiais para capturar relações mais complexas, a incorporação de modelos mais complexos que poderiam capturar relações não lineares de forma mais eficaz, ou ainda trabalhar com um conjunto de dados maior e mais diversificado, tentando a inclusão de outras métricas, como a média de comentários, buscando melhorar a capacidade de generalização do modelo.

Referências

MORAIS, N. S. D. .; BRITO, M. L. de A. . Marketing digital através da ferramenta Instagram. **E-Acadêmica**, *[S. l.]*, v. 1, n. 1, p. e5, 2020. Disponível em: https://www.eacademica.org/eacademica/article/view/5. Acesso em: 17 nov. 2024.