CEPEDI- RESTIC

Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Abigail Cruz dos Santos

Vinícius Amorim Santos Rozario

Feira de Santana - BA
17/11/2024

RESUMO

Este relatório apresenta a implementação e análise do algoritmo de Regressão Linear, com foco em sua aplicação prática e teórica no tema influenciadores do Instagram. O principal objetivo foi explorar como esse método pode ser utilizado para modelar relações entre variáveis em um conjunto de dados sobre métricas digitais.

A metodologia abrangeu a coleta, limpeza e análise preliminar do conjunto de dados, que inclui variáveis como número de posts, seguidores, e curtidas, entre outras. O algoritmo foi aplicado para identificar relações entre essas métricas e avaliar seu impacto em variáveis dependentes. A performance do modelo foi avaliada por métricas como R^2 e erro médio absoluto, permitindo verificar sua eficácia.

Os resultados indicaram que a **Regressão Linear** foi capaz de capturar algumas relações significativas. Por exemplo, para a variável "posts", o R^2 foi de 0,1226, indicando que cerca de 12,26% da variação na variável dependente foi explicada por essa métrica. Apesar de sua simplicidade, o método demonstrou potencial para prever comportamentos e fornecer insights úteis para a análise de influenciadores digitais.

INTRODUÇÃO

A técnica estatística da Regressão Linear Múltipla é usada para estudar a relação entre uma variável dependente e várias variáveis independentes explicativas; ela também é bastante aplicada pela comunidade científica (BAPTISTELLA, Marisa).

A análise preditiva tem se tornado uma ferramenta essencial em diversas áreas do conhecimento, permitindo modelar e prever comportamentos com base em dados históricos. No contexto deste relatório, a Regressão Linear foi escolhida por sua simplicidade e eficiência na modelagem de relações lineares entre variáveis.

A análise de regressão com a inclusão de covariadas nos permite considerar diferenças observáveis entre os dois grupos, como, por exemplo, a escolaridade dos pais, o fato de se trabalhar ou não, a idade, entre outras características que afetam

a decisão de estudar, além da própria restrição de crédito, que estaria sendo realizada pela política pública. (Chein, Flávia)

O conjunto de dados utilizado neste estudo foi obtido de uma fonte pública e contém informações sobre **influenciadores digitais no Instagram**. As variáveis analisadas incluem métricas como o **número de posts**, **seguidores**, **curtidas** e a **origem geográfica (país)** dos influenciadores. Essas métricas foram selecionadas por sua relevância no estudo do comportamento digital e pela facilidade de acesso a esses dados.

Este relatório justifica-se pela relevância de explorar métodos estatísticos para a análise de métricas digitais e pelo crescente interesse em entender o impacto de ações como publicações e engajamento sobre a performance de influenciadores. A escolha do conjunto de dados foi fundamentada em sua **pertinência ao estudo da influência digital**, proporcionando uma base sólida para a aplicação prática da Regressão Linear em problemas reais.

METODOLOGIA

A análise exploratória foi a etapa inicial deste estudo, cujo objetivo principal foi compreender as características do conjunto de dados e identificar possíveis padrões ou relações entre as variáveis. A inspeção preliminar do dataset incluiu a verificação de valores ausentes, tipos de variáveis e distribuições estatísticas. Para isso, utilizamos ferramentas como pandas para manipulação de dados e seaborn para visualizações gráficas.

O cálculo do coeficiente de correlação de Pearson foi realizado para mensurar as relações lineares entre variáveis. Essa etapa foi essencial para selecionar as variáveis mais relevantes e entender a dinâmica do conjunto de dados.

Com base nos insights obtidos na análise exploratória, foi implementado o algoritmo de Regressão Linear utilizando a biblioteca sklearn. As variáveis independentes escolhidas incluíram aquelas que apresentaram maior relevância na análise exploratória. Antes do treinamento do modelo, os dados foram divididos em dois subconjuntos: 80% para treino e 20% para teste, utilizando a função train_test_split.

Essa divisão garantiu a independência entre os dados utilizados para ajuste do modelo e aqueles reservados para validação. Para evitar possíveis vieses

causados por diferenças de escala entre variáveis, aplicou-se a normalização com StandardScaler, assegurando que todas as variáveis estivessem na mesma unidade de medida.

A validação do modelo foi realizada para garantir sua capacidade de generalização e identificar as configurações ideais. O método de validação cruzada K-Fold (k = 5) foi utilizado, permitindo avaliar o desempenho do modelo em diferentes divisões do conjunto de dados.

Além disso, para selecionar as variáveis independentes mais significativas, aplicamos o método SelectKBest, que utilizou o teste f-regression para ordenar as variáveis com base em sua relevância estatística.

A otimização do modelo foi conduzida por meio da inclusão de técnicas de regularização, como Ridge e Lasso. No caso do Ridge, o parâmetro de regularização α foi ajustado por meio de uma busca em grade (Grid Search), enquanto no Lasso, as variáveis menos relevantes foram eliminadas automaticamente, reduzindo a complexidade do modelo.

A comparação entre o modelo padrão e os modelos regularizados foi feita utilizando métricas como R^2, erro médio absoluto (MAE) e erro médio quadrático (MSE), selecionando-se o modelo com melhor equilíbrio entre precisão e simplicidade.

RESULTADOS

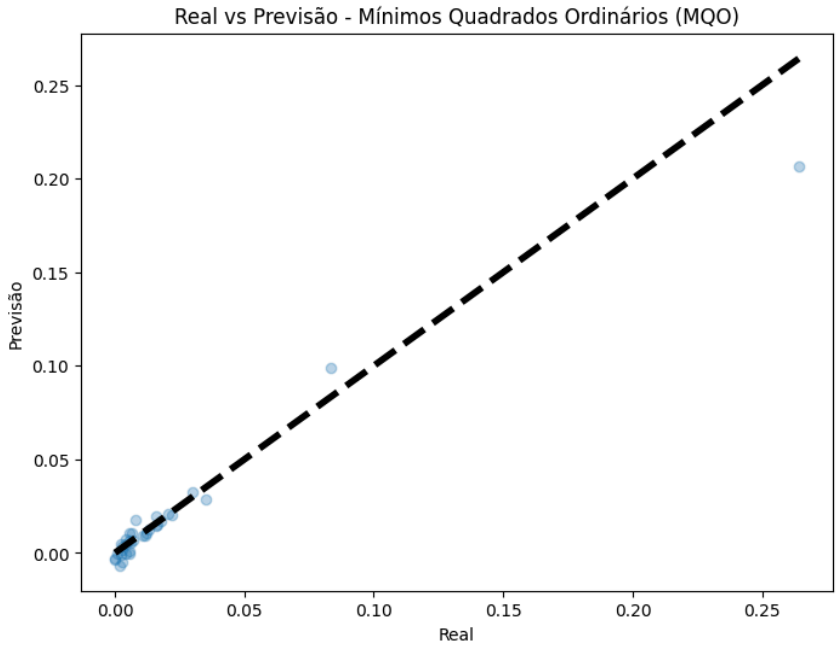
O desempenho do modelo foi avaliado utilizando o coeficiente de determinação (R^2), o erro médio absoluto (MAE) e o erro médio quadrático (MSE). Os resultados foram comparados para os modelos de Regressão Linear por Mínimos Quadrados Ordinários (MQO), Ridge e Lasso, considerando tanto os dados de treinamento quanto os dados de teste.

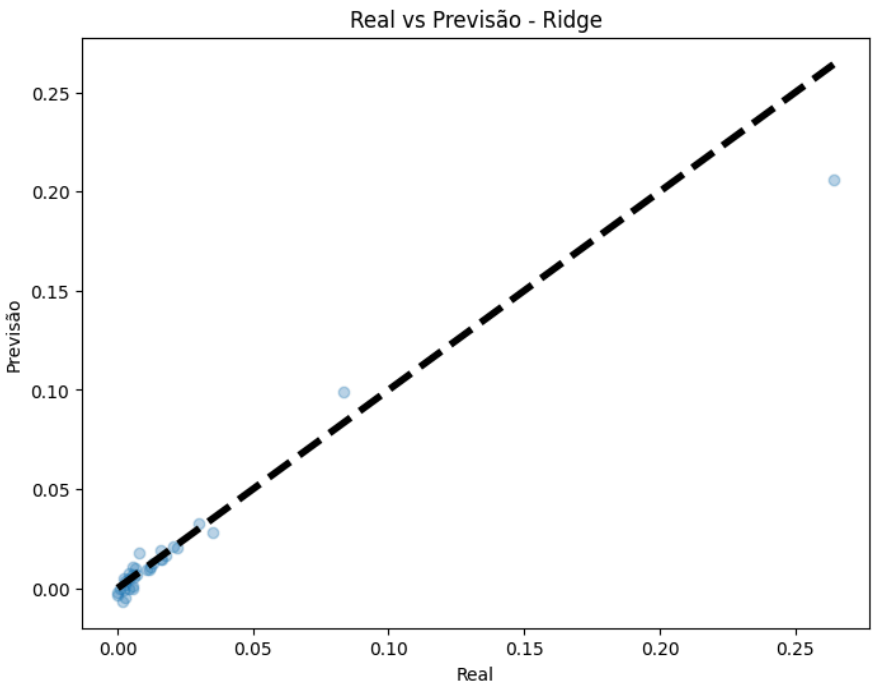
- **Mínimos Quadrados Ordinários (MQO):** O modelo padrão apresentou um R^2 de **0,9568** no conjunto de treinamento e **0,9424** no conjunto de teste, indicando uma alta capacidade de explicação da variância dos dados. O MSE foi extremamente baixo, com valores de **0,0000** no treinamento e **0,0001** no teste, sugerindo uma boa precisão. O MAE foi de **0,0042** no treinamento e **0,0049** no teste, refletindo um erro absoluto médio pequeno nas previsões.

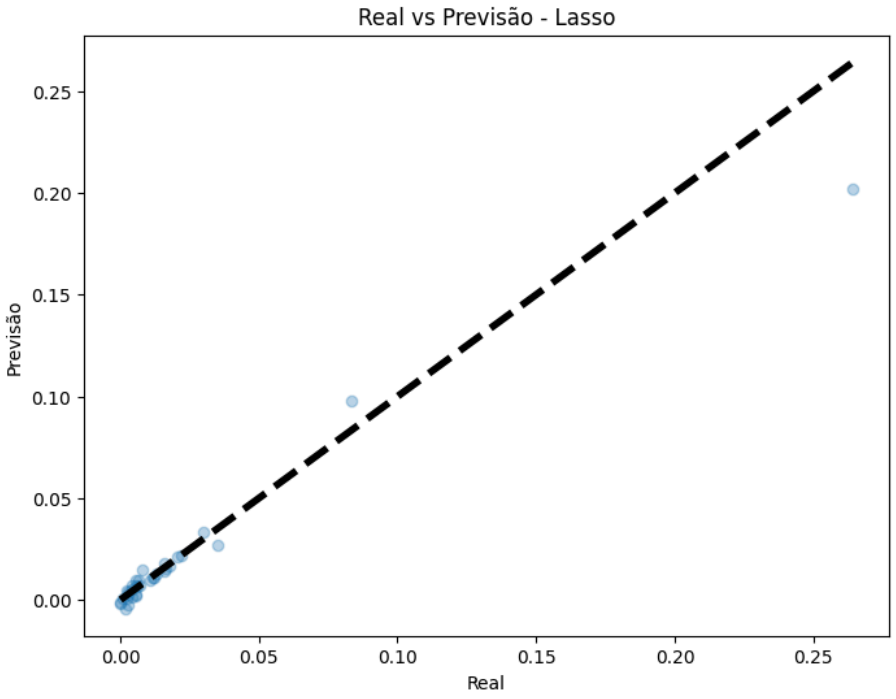
- **Ridge:** Para o modelo Ridge, o melhor valor de α foi encontrado como **0,6158** após a busca em grade. Este modelo teve um desempenho similar ao MQO, com R^2 de **0,9567** no treinamento e **0,9421** no teste. O MSE e MAE permaneceram praticamente idênticos aos do modelo padrão, com **0,0000** e **0,0042** no treinamento, e **0,0001** e **0,0049** no teste, respectivamente. A inclusão da regularização Ridge, portanto, não comprometeu a performance, mas ajudou a controlar a complexidade do modelo.
- **Lasso:** O modelo Lasso apresentou uma ligeira redução no desempenho, mas ofereceu um maior nível de simplificação ao modelo. Com o melhor valor de α identificado como **0,0007**, o R^2 foi de **0,9551** no treinamento e **0,9376** no teste. O MSE foi de **0,0000** no treinamento e **0,0001** no teste, enquanto o MAE teve valores de **0,0039** e **0,0043**, respectivamente. Embora o desempenho tenha sido marginalmente inferior, o modelo Lasso eliminou algumas variáveis de menor relevância, promovendo maior interpretabilidade.

Para complementar a análise quantitativa, foram geradas visualizações que ilustram o desempenho dos modelos:

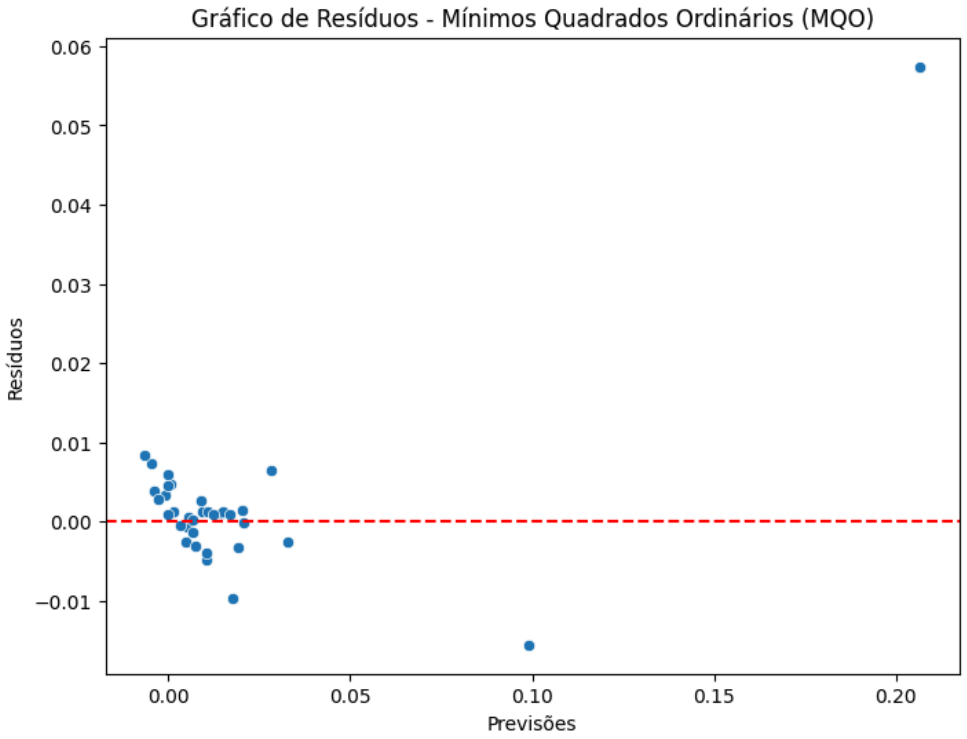
- **Gráfico de Dispersão:** Apresentou os valores reais em comparação com os valores previstos para os três modelos (MQO, Ridge e Lasso). Em todos os casos, os pontos se alinharam próximo à diagonal, indicando que as previsões foram consistentes com os valores observados.

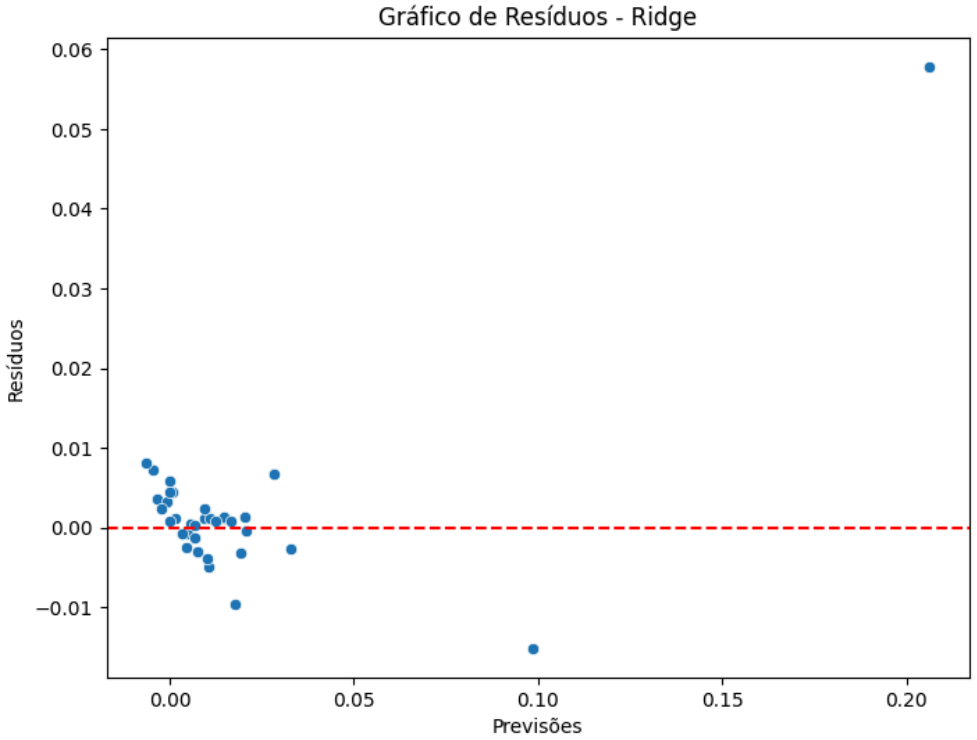


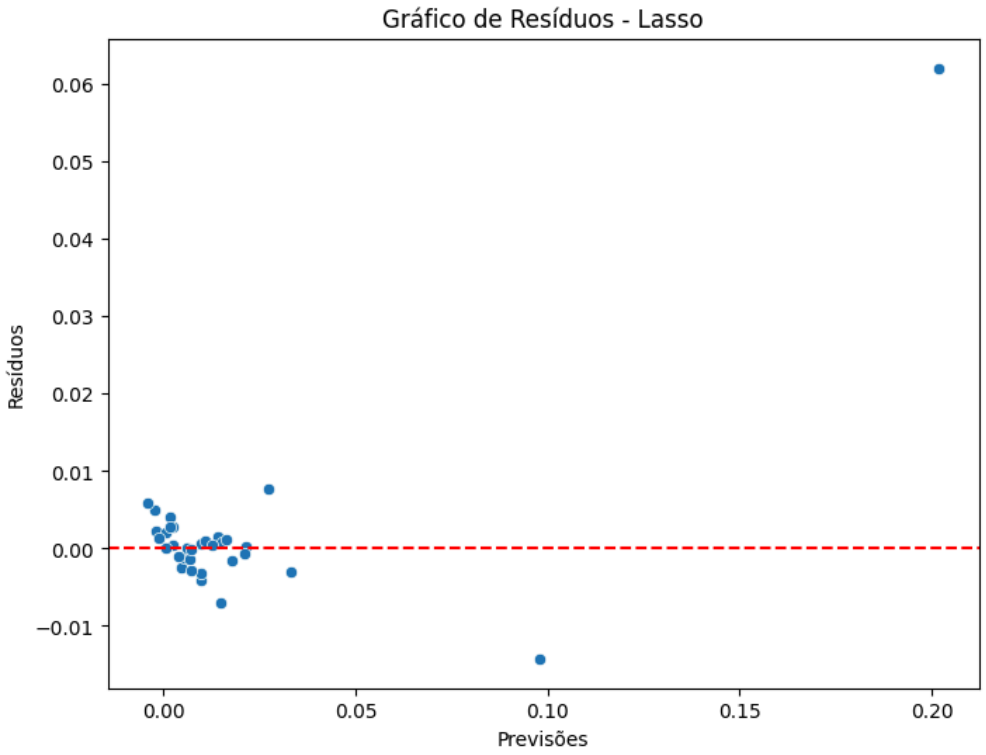

Real vs Previsão - Ridge
Previsão
Real
0.00
0.05
0.10
0.15
0.20
0.25


Real vs Previsão - Lasso
Previsão
Real
0.00
0.05
0.10
0.15
0.20
0.25

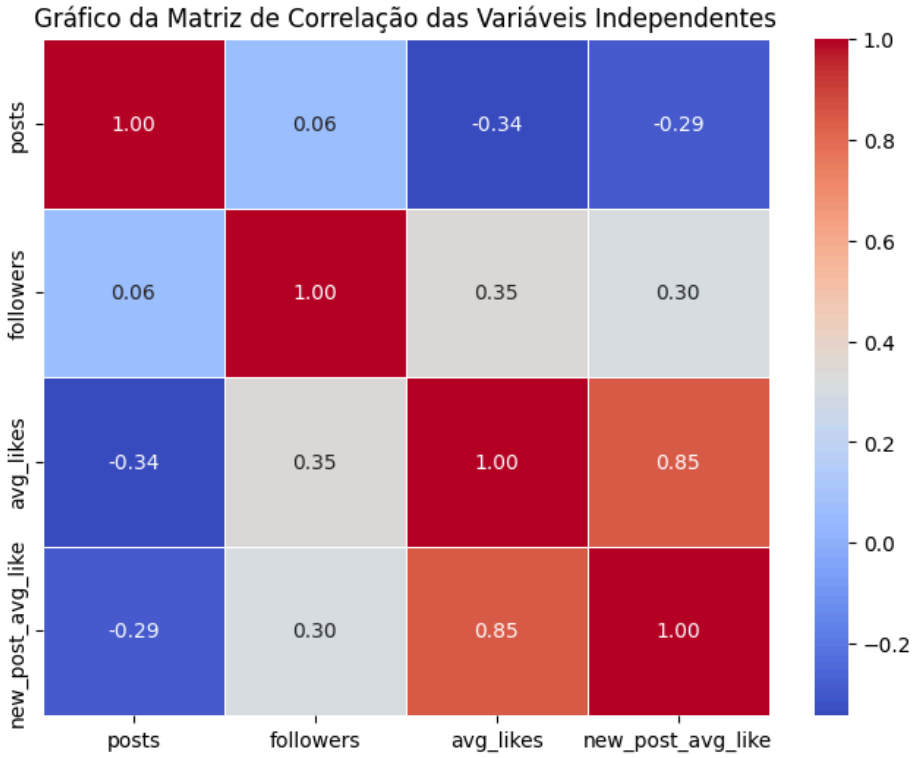
- **Gráfico de Erros Residuais:** Representou a distribuição dos resíduos para cada modelo. A análise revelou que os erros se distribuem de forma homogênea ao longo das previsões, com alguns outliers mais pronunciados, especialmente no modelo Lasso, o que reflete sua menor capacidade de ajustar totalmente a variância dos dados.







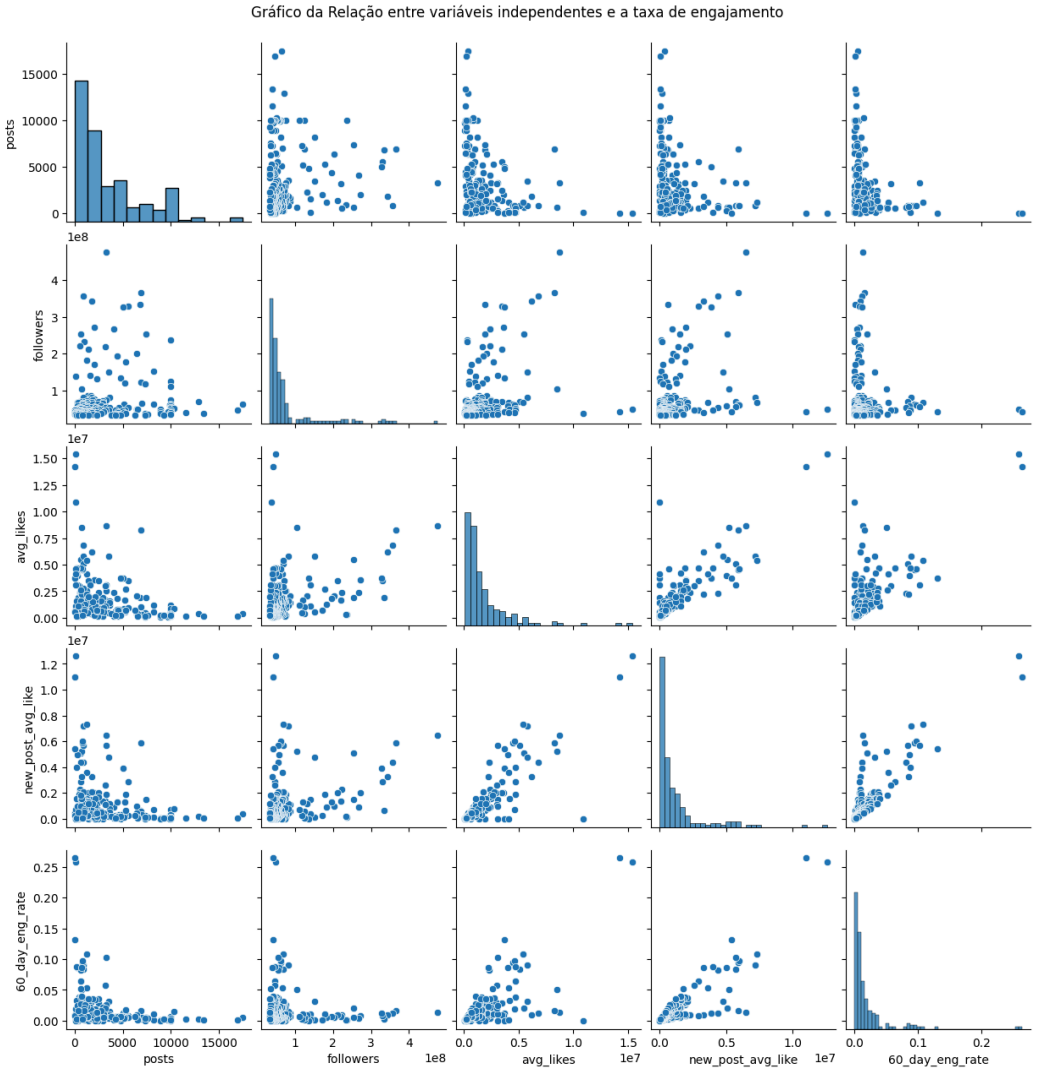
- **Gráfico de Correlação:** O gráfico de correlação foi utilizado para avaliar as relações lineares entre as variáveis presentes no conjunto de dados. Por meio de uma matriz de correlação visualizada com um heatmap (mapa de calor), foi possível identificar os coeficientes de Pearson que quantificam a força e a direção dessas relações.



Os resultados indicaram que variáveis como **número de seguidores (followers)** e **curtidas (likes)** possuem correlações positivas moderadas com a taxa de engajamento, enquanto a variável **número de posts (posts)** apresentou uma correlação negativa e fraca. Isso sugere que, em geral, influenciadores com mais posts tendem a ter taxas de engajamento ligeiramente menores, o que pode ser explicado por possíveis saturações no conteúdo ou baixa qualidade em publicações excessivas.

A análise da matriz de correlação foi fundamental para selecionar variáveis relevantes para o modelo e para entender como cada métrica contribui para prever o engajamento.

- **Gráfico da Relação entre Variáveis Independentes e a Taxa de Engajamento:** O gráfico que explorou as relações entre as variáveis independentes e a taxa de engajamento utilizou diagramas de dispersão para ilustrar tendências.



DISCUSSÃO

A análise dos resultados revelou que todos os modelos – MQO, Ridge e Lasso – tiveram um desempenho satisfatório, com altos valores de R^2 tanto nos dados de treinamento quanto nos de teste. Isso indica que o modelo foi eficaz em explicar a variância na taxa de engajamento dos influenciadores digitais.

Apesar dos bons resultados, algumas limitações foram observadas. A Regressão Linear assume uma relação linear entre as variáveis, mas, na prática, o engajamento de influenciadores pode depender de interações não lineares, que o modelo pode não ter capturado. Comportamentos como a saturação de conteúdo e padrões complexos de engajamento podem não ser bem representados, o que limita a aplicabilidade dos resultados para certos cenários.

Além disso, a amostra de dados utilizada pode não representar adequadamente a diversidade dos influenciadores no Instagram, considerando que a base de dados pode ter influência de fatores como origem geográfica e nicho de atuação. Esse viés pode restringir a generalização dos resultados para influenciadores com diferentes perfis ou para outros contextos de redes sociais. Por fim, é possível que o conjunto de dados inclua influenciadores com taxas de engajamento muito diferentes, resultando em outliers que afetam a precisão dos modelos, como observado com os resíduos mais pronunciados no Lasso.

A regularização com Ridge e Lasso foi fundamental para equilibrar a precisão do modelo com sua simplicidade. Ridge ajudou a controlar a complexidade ao evitar pesos extremos nas variáveis, enquanto Lasso foi útil para simplificar, zerando os coeficientes de variáveis menos importantes. No entanto, o uso de Lasso também resultou em uma pequena redução de desempenho, refletindo a troca entre precisão e interpretabilidade.

A divisão dos dados em 80% para treino e 20% para teste, junto com a validação cruzada K-Fold, proporcionou um modelo bem ajustado e independente do conjunto de teste. Contudo, a variação dos dados em cada "fold" da validação cruzada pode ter influenciado o resultado final, dependendo de como os influenciadores foram distribuídos nesses subconjuntos.

CONCLUSÃO E TRABALHOS FUTUROS

Este estudo demonstrou a eficácia da Regressão Linear na previsão do engajamento de influenciadores digitais no Instagram. O uso de técnicas de regularização trouxe simplificação e melhor interpretabilidade ao modelo.

Contudo, a suposição de linearidade e a limitação dos dados podem ter comprometido a captura de comportamentos não lineares e a representatividade da amostra.

Para melhorar, sugerem-se o uso de modelos não lineares, como Árvores de Decisão ou Redes Neurais, e a inclusão de variáveis contextuais, como o tipo de conteúdo postado e o horário das postagens.

Trabalhos futuros podem incluir a ampliação da base de dados para abranger diferentes nichos e a implementação de modelos mais avançados, para melhorar a precisão das previsões.

A análise de dados em tempo real e a personalização das recomendações de engajamento também seriam abordagens valiosas para aprimorar o modelo e torná-lo mais dinâmico e aplicável a influenciadores de diversos perfis.

REFERÊNCIAS

BAPTISTELLA, Marisa; STEINER, Maria Teresinha Arns; NETO, Anselmo Chaves. O uso de redes neurais e regressão linear múltipla na engenharia de avaliações: Determinação dos valores venais de imóveis urbanos. **Diss., Universidade Federal do Paraná**, 2005. Disponível em: http://www.din.uem.br/~ademir/sbpo/sbpo2006/pdf/arq0172.pdf. Acesso em: 16 nov. 2024.

Chein, Flávia. Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas / Flávia Chein. -- Brasília: **Enap**, 2019. Disponível em: https://repositorio.enap.gov.br/bitstream/1/4788/1/Livro_Regress%C3%A3o%20Linear.pdf. Acesso em: 17 nov. 2024