RESIDÊNCIA
EM SOFTWARE
BAHIA + TECNOLOGIA + EMPREENDEDORISMO

Projeto Regressão Linear

Grupo 47: Isis Caroline Lima Viana e Marco Antonio Oliveira Santos

Sumário

Resumo

Este relatório apresenta a implementação de um modelo de Regressão Linear utilizando um conjunto de dados de influenciadores do Instagram. O objetivo foi prever a métrica *60-day-engagement-rate*(o engajamento futuro de um perfil com base nos dados do influenciador) com base em variáveis como seguidores, número de postagens, e média de curtidas por novas postagens. A metodologia incluiu a análise exploratória, tratamento de dados, validação cruzada e ajuste de hiperparâmetros. Os resultados destacam a eficácia do modelo Ridge,com R^2 médio de 0.93 nos testes e evidências de correlações significativas entre as variáveis independentes e a dependente.

Introdução

Nos últimos anos, plataformas como o Instagram cresceram muito e ressignificaram o funcionamento das redes sociais. Nesse novo cenário, ser um influenciador se tornou mais do que apenas ser popular na internet, se tornou um trabalho."Em 2022, 72% dos profissionais de marketing usaram o Instagram para campanhas de influenciadores, consolidando a rede como a principal plataforma de marketing nesse segmento"(Sprout Social, 2024). Os *Influencers* atuais criam diariamente grandes parcerias com empresas a fim de promover produtos, campanhas e marcas para seus seguidores. Por causa disso, muitos novos desafios relacionados ao engajamento desses influenciadores digitais foram surgindo. Afinal é preciso sempre criar mais conteúdo, monitorar a reação do público a esse conteúdo e buscar crescer mais e mais na plataforma. Este estudo tem o intuito de criar um modelo capaz de prever a taxa de engajamento de um perfil nos próximos 60 dias (*60-day engagement-rate*), uma métrica crucial para empresas que buscam maximizar o impacto de suas campanhas digitais.

O conjunto de dados contém informações como número de seguidores, postagens, e métricas relacionadas a curtidas e engajamento. A escolha pela Regressão Linear teve como objetivo aproveitar a sua simplicidade, interpretabilidade e adequação ao escopo inicial do problema.

Metodologia

Análise Exploratória

Os dados foram explorados para identificar relações entre as variáveis. A matriz de correlação revelou forte correlação entre o engajamento e variáveis como curtidas e seguidores, enquanto gráficos de dispersão destacaram tendências não-lineares em algumas relações.

Implementação do Algoritmo

1. **Seleção de Variáveis:** As variáveis independentes escolhidas foram:
    - As variáveis independentes escolhidas foram:
        i. Número de seguidores (*followers*).

   ii. Número total de postagens (*posts*).
   iii. Média de curtidas em novas postagens (*new_post_avg_like*).
   iv. Total de curtidas (*total_likes*).
   v. Média de curtidas (*avg_likes*).
   - A variável dependente foi: o *60-day-eng-rate*.
2. **Pré-processamento:**
   - Conversão de unidades (*k*, *m*, %) para valores numéricos.
   - Tratamento de valores ausentes com imputação pela mediana.
   - Padronização dos dados usando *StandardScaler*.
3. **Treinamento do Modelo:**
   Modelos Ridge e Regressão Linear simples foram testados. O modelo Ridge, com regularização (α=1.0\alpha=1.0α=1.0), foi implementado para evitar o overfitting.
4. **Validação e Ajuste de Hiperparâmetros:**

   A validação cruzada foi realizada com 5 *folds* para comparar os modelos. A métrica R2R^2R2 foi usada como critério de desempenho.

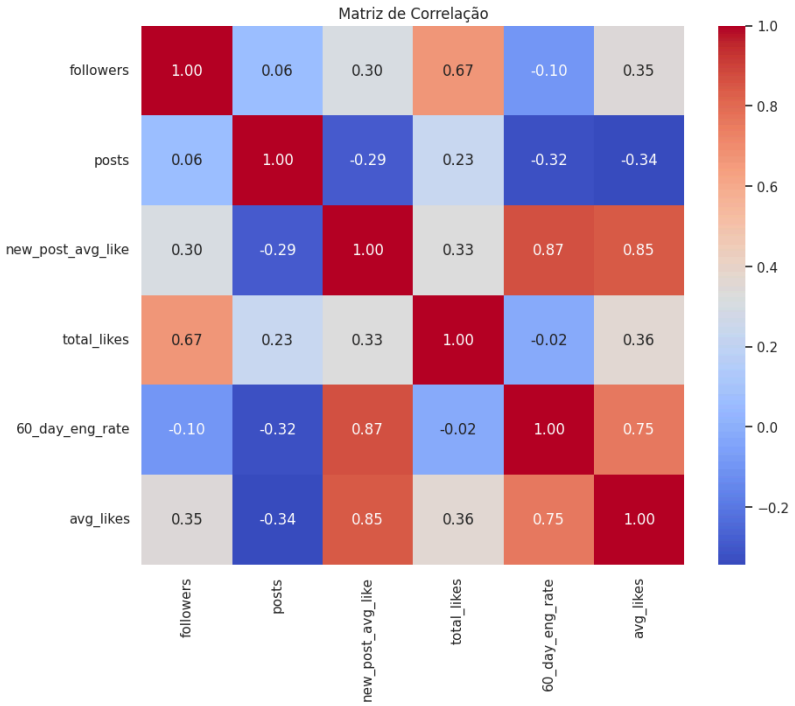
Resultados

Métricas de Avaliação

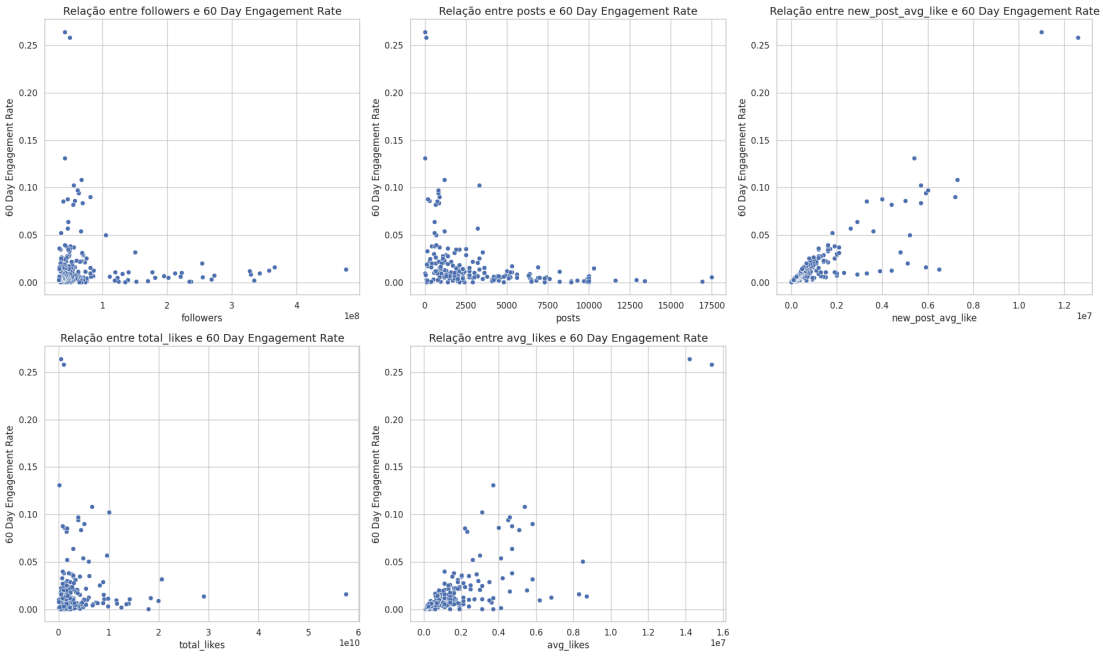
- **Erro Quadrático Médio (RMSE):** 0.01
- **Coeficiente de Determinação (R^2):** 0.93
- **Validação Cruzada:**
  - Ridge: R2R^2R2 médio de 0.70
  - Desvio Padrão dos Scores de R^2: 0.26

Visualizações

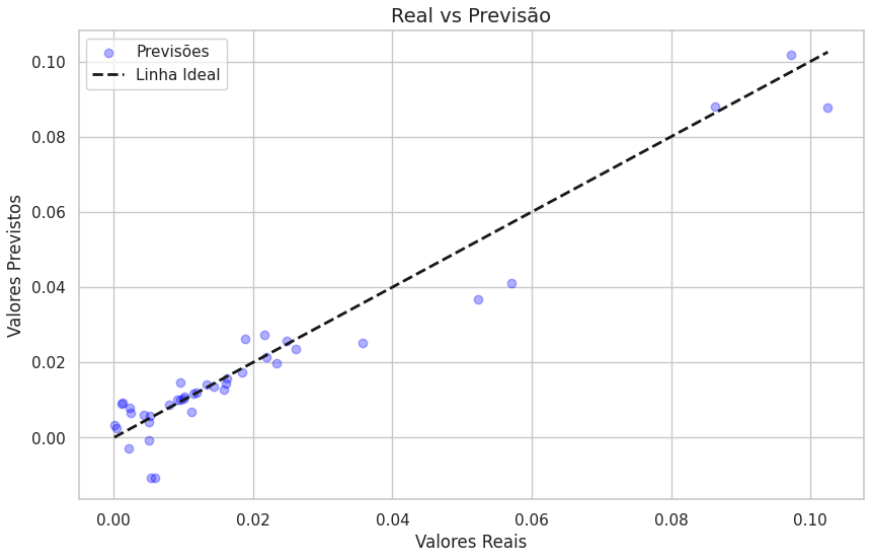
1. **Matriz de Correlação:** Identificou relações positivas significativas entre variáveis-chave e a taxa de engajamento.



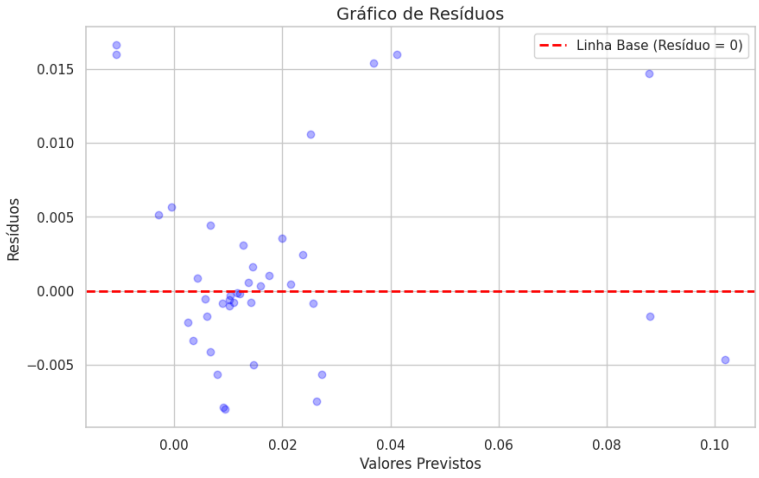
2. **Gráficos de Dispersão:**



3. **Gráficos de Previsão vs Valores Reais:** O gráfico mostrou alinhamento consistente entre valores previstos e reais, com dispersão baixa.



4. **Gráfico de Resíduos:** Os resíduos apresentaram distribuição homogênea em torno de zero, indicando boa adequação do modelo. Mas apesar disso, ainda pode-se observar muitos outliers indicando uma não-linearidade dos dados



Discussão

Os resultados sugerem que o modelo Ridge, com regularização, supera a Regressão Linear simples ao lidar com multicolinearidade e ruído nos dados. Também sendo beneficiado pela regularização, evitando o overfitting.

No entanto, o desempenho do modelo foi prejudicado pela incapacidade de captar possíveis relações não-lineares presentes nos dados. Além disso, a qualidade das previsões foi influenciada pela presença de valores ausentes e pela falta de variáveis que poderiam fornecer informações mais completas para o modelo.

Conclusão e Trabalhos Futuros

Em conclusão, o projeto demonstrou uma boa aptidão para prever taxas de engajamento usando dados captados de perfis do instagram. Contudo, o desempenho foi limitado por possíveis relações não-lineares não capturadas, isto é, a quantidade elevada de outliers. Além disso, pode-se pensar que a precisão do modelo foi impactada pela análise das métricas erradas, afinal outros tópicos como frequência das postagens, quantidade de comentários por post, entre muitos outros seriam muito relevantes para essa análise. Portanto, apesar de ter-se obtido um resultado satisfatório com o modelo, na vida real ele não é tão aplicável e exato.

Por isso, sugere-se que utilize-se um novo banco de dados que possua mais métricas a serem analisadas, pois isso proporcionará um modelo mais preciso e eficaz. Também é válido optar pelo uso de algoritmos não-lineares, já que os dados não parecem seguir um padrão linear. Para tal, indica-se algoritmos como Regressão com Kernel ou Redes Neurais, pois são modelos que podem melhorar a acurácia.

9

Referências

Sprout Social. *22 Influencer Marketing Statistics for Your Brand's Strategy in 2024.* Disponível em: https://sproutsocial.com. Acesso em: 17 nov. 2024.

Fontes de dados: Dataset de influenciadores do Instagram. Disponível em: https://drive.google.com/uc?id=1_F_oUDjtdqZ9s3g3il9i2j6XPfuS4uoT