Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Emanuelle de Araujo da Hora
Victor Augusto Silva de Jesus

Novembro/2024
Feira de Santana/Ba

SUMÁRIO

1. RESUMO

Este relatório descreve a criação de um modelo preditivo com o objetivo de estimar a taxa de engajamento de influenciadores no Instagram em um período de 60 dias (`60_day_eng_rate`). A abordagem incluiu processamento de dados e análise exploratória, conversão de valores baseados em texto em formatos numéricos, padronização e a aplicação de regressão linear para explorar várias combinações de variáveis preditoras. A avaliação do modelo utilizou métricas como R^2 (Coeficiente de Determinação), MAE (Erro Absoluto Médio) e MSE (Erro Quadrático Médio), que ajudaram a identificar as principais variáveis que influenciam o engajamento.

2. METODOLOGIA

A metodologia adotada neste projeto para prever a taxa de engajamento de 60 dias (60_day_eng_rate) dos influenciadores no Instagram foi composta pelas seguintes etapas principais:

2.1. Preparação dos Dados

O conjunto de dados foi carregado a partir do arquivo top_insta_influencers_data.csv, contendo informações sobre influenciadores do Instagram. As principais variáveis analisadas foram: número de postagens, seguidores, curtidas médias por postagem, taxa de engajamento de 60 dias (variável alvo), curtidas médias em postagens recentes e total de curtidas.

Para garantir a qualidade dos dados e torná-los apropriados para análise, foi realizada a conversão de valores categóricos em formatos numéricos. Valores como k (milhares), m (milhões) e % (percentuais) foram convertidos para números correspondentes, aplicando-se multiplicações apropriadas (ex: k → 1.000, m → 1.000.000).

Além disso, foi realizado o tratamento de dados ausentes, onde as linhas com valores nulos na variável dependente foram removidas, assegurando a integridade dos dados durante o treinamento do modelo.

2.2. Definição das Variáveis

O modelo preditivo foi desenvolvido com base nas seguintes variáveis:

- Variáveis preditoras:
  - followers: Quantidade de seguidores do influenciador.
  - avg_likes: Média de curtidas por postagem.
  - new_post_avg_like: Média de curtidas em postagens recentes.
  - total_likes: Total de curtidas recebidas nas postagens.
- Variável alvo:
  - 60_day_eng_rate: Taxa de engajamento média nos últimos 60 dias.

Essas variáveis foram selecionadas com base na sua relevância para estimar o comportamento de engajamento dos influenciadores.

2.3. Divisão dos Dados

O conjunto de dados foi dividido em dois subconjuntos para treinamento e teste do modelo:

- Treinamento (80%): Utilizado para ajustar o modelo, permitindo que ele aprenda os padrões presentes nos dados.
- Teste (20%): Utilizado para avaliar a performance do modelo em dados não vistos.

A divisão foi realizada utilizando o método train_test_split do scikit-learn, com a configuração do parâmetro random_state=42 para garantir a reprodutibilidade dos resultados.

2.4. Padronização dos Dados

Devido à variação nas escalas das variáveis (por exemplo, o número de seguidores pode estar na ordem de milhões, enquanto as curtidas médias podem ser centenas), foi aplicada a técnica de padronização utilizando o StandardScaler do scikit-learn. Isso transformou os dados de forma que cada variável tivesse média 0 e desvio padrão 1, evitando que variáveis com escalas maiores influenciassem desproporcionalmente o modelo.

2.5. Seleção de Variáveis

Para identificar as variáveis mais relevantes, foi utilizado o método SelectKBest com a métrica de avaliação f_regression, que mede a correlação entre as variáveis preditoras e a variável alvo. Diferentes combinações de variáveis foram testadas para avaliar a importância de cada uma delas na previsão da taxa de engajamento.

2.6. Construção e Treinamento do Modelo

O modelo de Regressão Linear foi escolhido devido à sua simplicidade e capacidade de modelar relações lineares entre as variáveis. O treinamento foi realizado de forma incremental, utilizando de 1 até todas as variáveis preditoras, para observar como o desempenho do modelo variava com diferentes combinações de variáveis.

2.7. Avaliação do Modelo

O desempenho do modelo foi avaliado por meio das seguintes métricas:

- R^2 (Coeficiente de Determinação): Mede a proporção da variância da variável alvo explicada pelo modelo.
- MSE (Erro Quadrático Médio): Mede a magnitude dos erros ao quadrado, penalizando grandes erros.
- MAE (Erro Absoluto Médio): Mede a média das diferenças absolutas entre os valores previstos e observados.

Essas métricas permitiram identificar o melhor conjunto de variáveis e a eficácia do modelo na previsão da taxa de engajamento.

Essa metodologia proporcionou um processo estruturado e eficiente para a construção de um modelo preditivo robusto, visando uma previsão precisa da taxa de engajamento de influenciadores no Instagram.

3. RESULTADO

Ao aplicar o modelo preditivo por meio de regressão linear, as principais conclusões obtidas foram as seguintes:

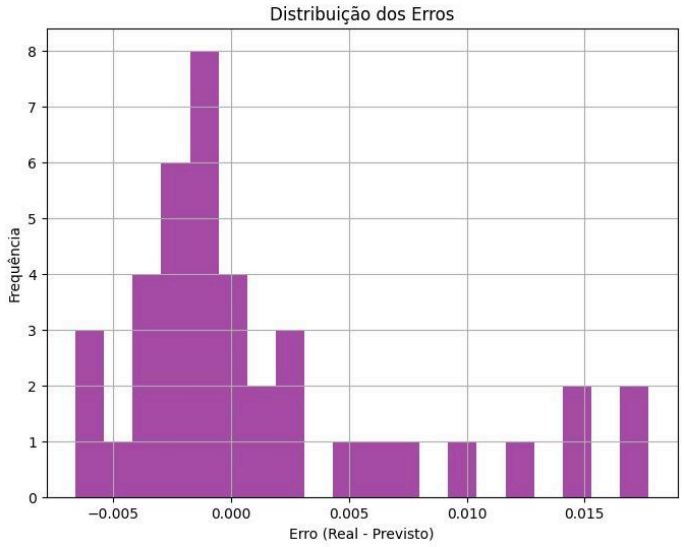
3.1 Desempenho do modelo

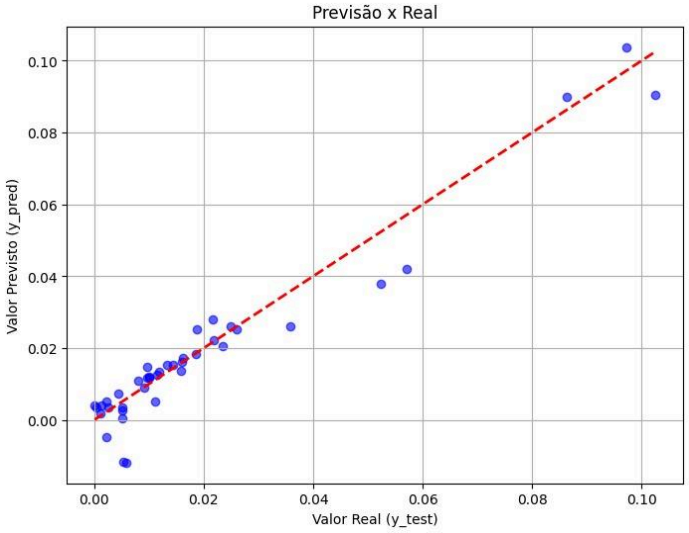
A avaliação do modelo utilizou métricas R^2, MSE e MAE em várias combinações de variáveis preditoras. Os resultados indicaram um R^2 substancial, demonstrando que o modelo foi responsável por uma parcela considerável da variabilidade nas taxas de engajamento dos influenciadores. Além disso, o MSE e o MAE produziram valores aceitáveis, o que implica que o modelo é proficiente em fazer previsões, embora ainda haja potencial para aprimoramento, principalmente por meio da adição de mais variáveis ou refinamento adicional do modelo.

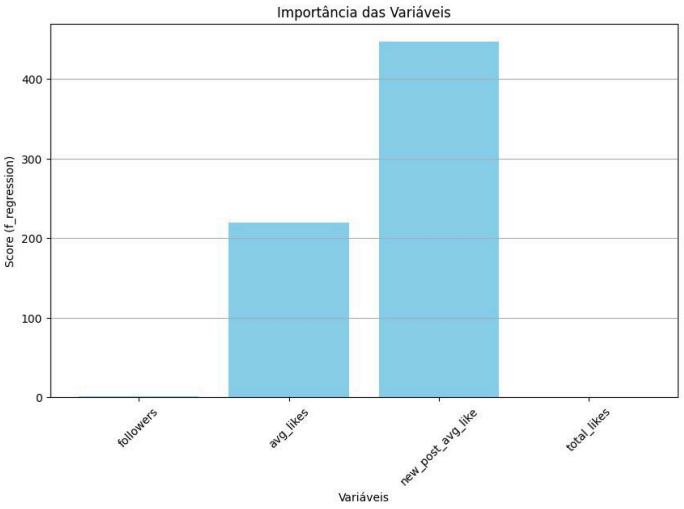
3.2 Variáveis mais relevantes

Ao analisar a significância das variáveis com o método SelectKBest, foi determinado que seguidores, avg_likes e new_post_avg_like foram os mais influentes na previsão da taxa de engajamento, enquanto outras variáveis contribuíram menos para a eficácia do modelo.

3.3. Gráficos que ilustram o desempenho do modelo







4. DISCUSSÃO

4.1 Interpretação dos Resultados

Os resultados do modelo indicam que a regressão linear, embora simples, é uma ferramenta eficaz para prever a taxa de engajamento dos influenciadores no Instagram com base em variáveis como o número de seguidores, a média de curtidas por postagem e a interação em postagens recentes. O valor de R^2 obtido sugere que o modelo é capaz de capturar uma boa parte da variabilidade dos dados. No entanto, o MSE e o MAE ainda indicam que o modelo não é perfeito, com erros que podem ser atribuídos a fatores não considerados ou complexidades adicionais do comportamento dos influenciadores.

4.2 Limitações do Modelo

Uma das principais limitações do modelo está na escolha da regressão linear, que assume uma relação linear entre as variáveis preditoras e a variável alvo. No entanto, o comportamento dos influenciadores pode não ser totalmente linear, e outros fatores, como o tipo de conteúdo postado ou as interações específicas com os seguidores, podem ter um impacto significativo na taxa de engajamento, mas não foram considerados nesta análise. A inclusão de variáveis qualitativas, como o engajamento por tipo de conteúdo ou as interações com marcas, poderia enriquecer o modelo.

5. CONCLUSÃO E TRABALHOS FUTUROS

5.1. Conclusão

Este projeto teve como objetivo construir um modelo preditivo capaz de estimar a taxa de engajamento de influenciadores no Instagram com base em métricas quantitativas, como número de seguidores, curtidas por postagem e engajamento recente. Através da aplicação de uma regressão linear, foi possível obter uma boa aproximação dos dados, com um desempenho satisfatório nas métricas de avaliação, como R^2, MSE e MAE. Apesar das limitações do modelo, como a simplicidade da regressão linear e a ausência de variáveis qualitativas, os resultados oferecem perspectivas valiosos para a seleção de influenciadores em

campanhas de marketing. Para futuras melhorias, a adoção de modelos mais complexos e a inclusão de variáveis adicionais podem aprimorar a precisão das previsões e a adaptação do modelo a diferentes contextos.

5.2. Ajustes e melhorias

Embora o modelo tenha mostrado um desempenho forte, melhorias futuras — como incorporar variáveis adicionais ou empregar modelos mais sofisticados, como regressão não linear ou redes neurais — podem aumentar ainda mais a precisão preditiva.

Esses resultados estabelecem uma base sólida para o desenvolvimento de modelos preditivos mais sofisticados e a utilização dessas previsões dentro dos domínios do marketing digital e da análise de influenciadores.

5.3. Direções Futuras

Futuros estudos podem expandir este modelo, considerando outras variáveis, como o comportamento dos seguidores ao longo do tempo, o impacto de campanhas publicitárias específicas, e fatores externos, como sazonalidade e eventos globais. O desenvolvimento de modelos mais dinâmicos, que integrem dados em tempo real, também poderia melhorar ainda mais a precisão das previsões e a capacidade de adaptação do modelo ao comportamento dos influenciadores.

6. REFERÊNCIAS

https://towardsdatascience.com/8-plots-for-explaining-linear-regression-to-a-layman-489b753da696

https://paginas.fe.up.pt/~mam/regressao.pdf