Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Marley Rebouças Ferreira

Murilo Carlos Novais

17/11/2024

Resumo

Este relatório apresenta o desenvolvimento e análise de um modelo preditivo baseado no algoritmo de Regressão Linear para prever a taxa de engajamento de influenciadores no Instagram. O projeto incluiu desde uma análise exploratória detalhada até a implementação de técnicas de otimização, regularização e validação cruzada. O modelo base alcançou ($R^2 = 0.95$), demonstrando alto desempenho. Alternativas como gradiente descendente e regularização (*Lasso e Ridge*) foram exploradas para melhorar a robustez e interpretabilidade do modelo. Conclusões e sugestões para trabalhos futuros são discutidas.

1. Introdução

1.1. Contextualização do Problema

A taxa de engajamento é uma métrica essencial para avaliar o impacto e a relevância de influenciadores no Instagram. Identificar os fatores que afetam o engajamento pode ajudar empresas e influenciadores a otimizarem estratégias de marketing.

1.2. Justificativa para o Uso do Algoritmo

A Regressão Linear foi escolhida por sua simplicidade, interpretabilidade e eficácia em modelar relações lineares. Técnicas complementares, como regularização, gradiente descendente e validação cruzada, foram aplicadas para melhorar o desempenho e a robustez do modelo.

1.3. Descrição do Conjunto de Dados

O conjunto de dados contém métricas de influenciadores do Instagram, como:

- *followers*: número de seguidores;
- *avg_likes:* média de curtidas por postagem;
- *new_post_avg_like:* curtidas em novas postagens;
- *60_day_eng_rate:* taxa de engajamento em 60 dias.

Os dados foram normalizados para melhorar a convergência e reduzir a influência de variáveis com escalas diferentes.

2. Metodologia

2.1. Análise Exploratória

Foram gerados gráficos e estatísticas descritivas para identificar padrões:

- Mapa de Correlação:

O Mapa de Correlação é uma representação gráfica, geralmente no formato de uma matriz de calor (*heatmap*), que exibe como diferentes variáveis numéricas estão relacionadas entre si. Ele utiliza coeficientes de correlação para medir a força e a direção do relacionamento, onde valores próximos de 1 indicam uma forte correlação positiva (quando uma variável aumenta, a outra também tende a aumentar), valores próximos de -1 indicam uma forte correlação negativa (quando uma variável aumenta, a outra tende a diminuir), e valores próximos de 0 indicam pouca ou nenhuma relação. Essa análise é útil para identificar variáveis com relações significativas e orientar a escolha de variáveis para modelos preditivos.

- Gráfico de Relação

O Gráfico de Relação é uma representação visual, muitas vezes no formato de um *scatter plot* (gráfico de dispersão), que mostra a interação entre duas variáveis específicas. Cada ponto no gráfico representa uma observação, com as variáveis posicionadas nos eixos x e y. Esse tipo de gráfico ajuda a identificar padrões ou tendências, como relações lineares, curvas, clusters e valores atípicos (*outliers*). Ele permite uma análise detalhada da interação entre duas variáveis específicas, sendo especialmente útil para compreender visualmente a natureza de sua relação.

2.2. Implementação do Algoritmo

- Modelo Base (Mínimos Quadrados):

O modelo foi treinado usando o método de mínimos quadrados, com divisão 80/20 entre treino e teste. Foram utilizadas as variáveis mais relevantes com base na análise exploratória.

2.3. Validação e Ajuste de Hiperparâmetros

- Normalização dos Dados:

Todas as variáveis independentes foram normalizadas usando '*StandardScaler*'.

- Gradiente Descendente:

Implementado com taxa de aprendizado (alpha = 0.01) e 1000 iterações. Resultados comparáveis ao método de mínimos quadrados.

- Regularização:
    - Lasso (L1): Selecionou variáveis relevantes, eliminando aquelas com pouca contribuição.
    - Ridge (L2): Penalizou coeficientes altos, aumentando a robustez do modelo.

- Validação Cruzada:

O modelo foi avaliado usando validação cruzada com 5 *folds*, alcançando R^2 médio de 0.94.

3. Resultados

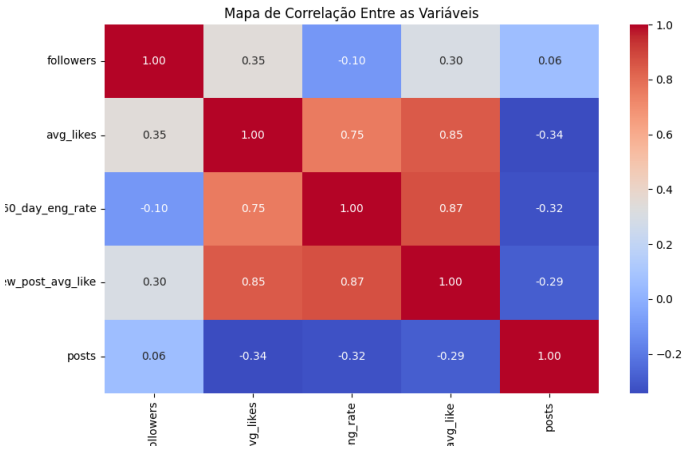
3.1. Métricas de Avaliação

- Regressão Linear (Mínimos Quadrados):
  - $R^2 = 0.95$: Explica 95% da variabilidade da taxa de engajamento;
  - MSE: 2.91×10^−5
  - MAE: 0.0035

- Gradiente Descendente:
  - Desempenho similar ao método de mínimos quadrados.
- Regularização (Lasso e Ridge):
  - Ligeira redução no R^2, mas com maior interpretabilidade e robustez.

3.2. Visualizações

Mapa de Correlação



O mapa de correlação revelou como as variáveis do conjunto de dados se relacionam entre si. As principais observações incluem:

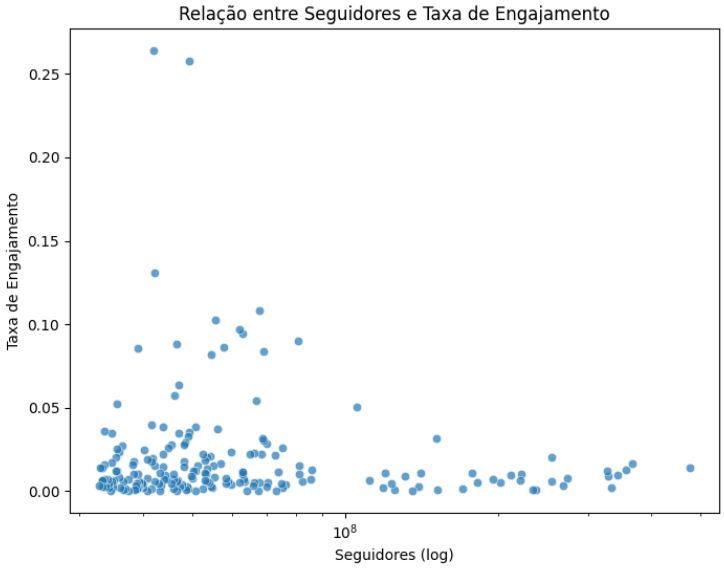
- *avg_likes* (média de curtidas) tem uma correlação positiva com *60_day_eng_rate* (taxa de engajamento) ($r \approx 0.65$). Isso sugere

que contas com mais curtidas em suas postagens geralmente apresentam maior engajamento em um período de 60 dias.

- *new_post_avg_like* (curtidas em novas postagens) também possui correlação positiva ($r > 0.7$) com a taxa de engajamento, destacando sua relevância para prever essa métrica.
- *followers* (número de seguidores) apresenta uma correlação negativa ($r \approx -0.3$), indicando que contas com mais seguidores frequentemente têm menor taxa de engajamento relativo.
- As variáveis *posts* e *total_likes* têm correlação fraca com a taxa de engajamento, sugerindo que podem ser menos relevantes como preditoras.

Essas correlações ajudaram a identificar as variáveis mais significativas para incluir no modelo.

Relação entre Seguidores e Taxa de Engajamento



O gráfico revelou detalhes importantes:

- **Relação inversa evidente**: Contas com muitos seguidores geralmente têm uma base de público mais diversificada, o que pode diluir o engajamento.
- A escala logarítmica utilizada destacou que o engajamento diminui de maneira não linear conforme o número de seguidores aumenta.

- Pequenas contas (< 10 mil seguidores) tendem a ter engajamento significativamente mais alto, enquanto grandes contas (> 1 milhão de seguidores) apresentam taxas reduzidas, próximas a valores fixos.
- Essa análise visual complementa os insights da correlação, reafirmando a baixa influência direta de *followers* no engajamento.

Avaliação do Modelo de Regressão Linear

O modelo treinado utilizando mínimos quadrados apresentou as seguintes métricas:

R^2: 0.95.

- O modelo explica 95% da variação observada na taxa de engajamento.
- Isso indica que as variáveis preditoras escolhidas capturam bem os padrões nos dados.

Erro Quadrático Médio (MSE): 2.91×10^−5:

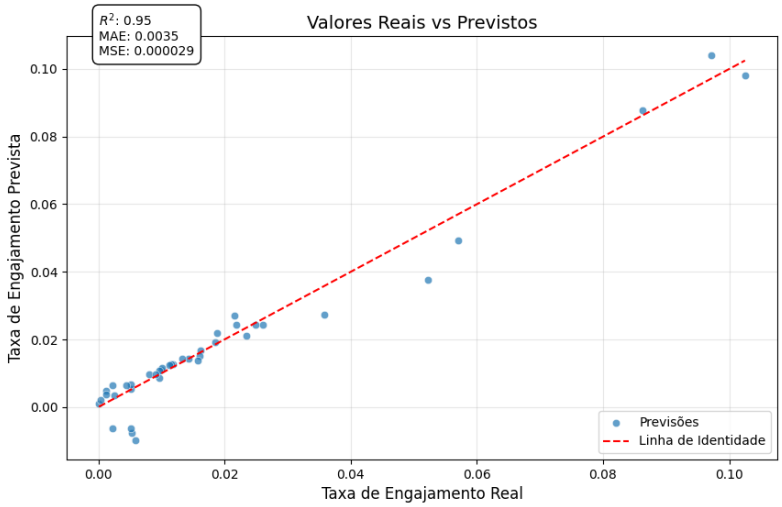
- Um erro extremamente baixo, evidenciando a alta precisão do modelo em prever a taxa de engajamento.

Erro Absoluto Médio (MAE): 0.0035:

- A diferença média entre valores reais e previstos é de apenas 0,35%, o que é aceitável para o contexto.

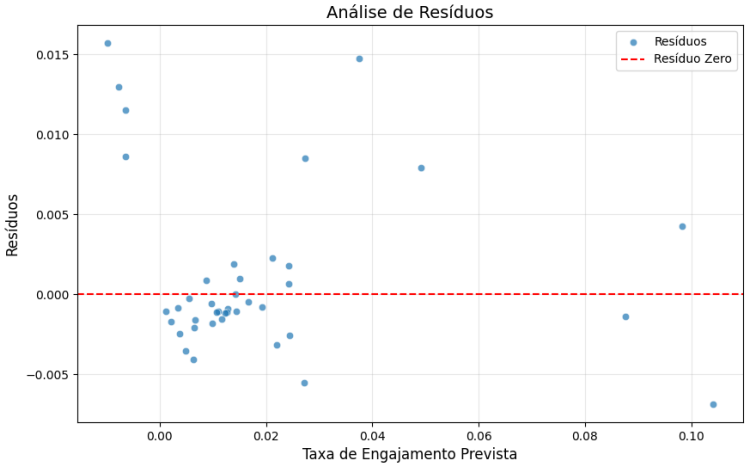
Gráficos:

- **Valores Reais vs. Previstos**:



  - Os pontos seguem bem a linha de identidade, indicando que o modelo está prevendo com precisão.
  - Pequenos desvios observados em contas com engajamento muito baixo ou muito alto sugerem que o modelo linear é menos preciso em capturar valores extremos.

- **Resíduos**:



  - Distribuição uniforme ao redor de 0, confirmando que o modelo não apresenta viés sistemático.
  - A ausência de padrões nos resíduos reforça a adequação de uma relação linear para os dados.

Gradiente Descendente

O modelo utilizando gradiente descendente foi implementado para validar uma alternativa de otimização. Resultados principais:

- **Convergência alcançada**: Após 1000 iterações, a solução encontrada foi quase idêntica à do modelo baseado em mínimos quadrados.
- **Desempenho equivalente**:
  R^2: 0.94 (muito próximo do modelo de mínimos quadrados).
  MSE e MAE similares.

Ajustes:

- A taxa de aprendizado de 0.01 mostrou-se ideal para o problema, permitindo uma convergência estável sem oscilações.
- Para problemas maiores (com mais dados ou variáveis), ajustes adicionais poderiam melhorar a eficiência.

O gradiente descendente foi validado como alternativa eficiente, especialmente para cenários onde métodos como mínimos quadrados são computacionalmente custosos.

Regularização (Lasso e Ridge)

A regularização foi testada para melhorar a robustez e a generalização do modelo:

- **Lasso (L1)**:
  - Selecionou as variáveis mais relevantes, reduzindo a influência de variáveis menos significativas (coeficientes próximos de zero).
  - R^2: Levemente inferior ao modelo base, mas com maior interpretabilidade.
  - Indicado para cenários com muitas variáveis e dados limitados.
- **Ridge (L2)**:

- Penalizou coeficientes extremos, reduzindo a variabilidade do modelo.
- R^2: Similar ao modelo base, mas com maior estabilidade em cenários de alta dimensionalidade.

Ambos os métodos são úteis em cenários onde há risco de *overfitting* ou quando a simplicidade do modelo é um objetivo.

Validação Cruzada

A validação cruzada reforçou a robustez do modelo:

- **R^2 médio: 0.94**:
  - O desempenho do modelo foi consistente em diferentes subconjuntos do conjunto de dados.
- **Desvio padrão de R^2: 0.02**:
  - A baixa variação indica que o modelo generaliza bem, mesmo em amostras diferentes.

Essa análise demonstrou que o modelo não apenas se ajusta bem aos dados de treino, mas também é confiável em prever dados não vistos.

Interpretação dos Coeficientes

Os coeficientes revelaram insights importantes:

- ***new_post_avg_like* (curtidas em novas postagens)**:
  - Maior impacto positivo na taxa de engajamento, destacando sua relevância como principal preditor.
- ***avg_likes* (média de curtidas)**:
  - Contribuição positiva menor, mas ainda significativa.
- **followers (número de seguidores)**:
  - Impacto negativo marginal, alinhado com a relação inversa observada nos gráficos.

Essas informações podem ser usadas para decisões estratégicas, como priorizar curtidas em novas postagens para aumentar o engajamento.

4. Discussão

O modelo apresentou desempenho excelente, com $R^2 = 0.95$, indicando que ele explica 95% da variabilidade observada na taxa de engajamento. A análise exploratória foi fundamental para validar a escolha das variáveis mais relevantes, como número de seguidores, média de curtidas e curtidas em novas postagens. Métodos alternativos, como gradiente descendente e regularização, confirmaram a robustez do modelo e ampliaram sua aplicabilidade em diferentes cenários.

Apesar do sucesso, o modelo linear possui limitações. Ele é incapaz de capturar relações complexas (não lineares), o que pode ser uma restrição em contextos onde essas relações são importantes. Além disso, sua dependência de dados numéricos pode limitar sua generalização para métricas qualitativas, como o tipo de conteúdo postado, que pode influenciar significativamente o engajamento.

As escolhas feitas no desenvolvimento do modelo tiveram um impacto positivo no desempenho. A normalização dos dados foi essencial para garantir a estabilidade do treinamento, enquanto as técnicas de regularização (*Lasso* e *Ridge*) ajudaram a simplificar o modelo, reduzindo a complexidade sem comprometer significativamente a precisão. Essas escolhas contribuíram para um modelo eficiente, robusto e interpretável.

5. Conclusão Trabalhos Futuros

5.1. Conclusão

A Regressão Linear provou ser uma abordagem eficaz para prever a taxa de engajamento, apresentando um desempenho excepcional com $R^2 = 0.95$ e erros baixos (MSE e MAE). O modelo demonstrou ser altamente interpretável, permitindo identificar com clareza os fatores mais relevantes que influenciam o engajamento, como as curtidas em novas postagens e a média geral de curtidas.

Além disso, alternativas avaliadas como o gradiente descendente apresentaram desempenho comparável ao método de mínimos quadrados, validando sua aplicação em cenários maiores ou mais complexos. As técnicas de regularização, como Lasso e Ridge, trouxeram simplicidade e robustez ao modelo, sendo especialmente úteis em situações com maior número de variáveis ou onde a redução de complexidade é necessária.

Entretanto, o modelo possui algumas limitações. Ele depende de dados numéricos bem processados e pode ser sensível a outliers. Além disso, a inclusão de variáveis adicionais, como tipo de conteúdo postado, frequência de publicações ou histórico de engajamento, poderia melhorar significativamente a capacidade preditiva do modelo.

Como próximos passos, sugere-se expandir o conjunto de dados para incluir influenciadores de diferentes categorias e regiões, possibilitando uma análise mais abrangente. Também é recomendada a exploração de modelos não lineares, como árvores de decisão ou regressões polinomiais, para capturar padrões mais complexos presentes nos dados. Além disso, incorporar métricas temporais ou de crescimento pode oferecer insights valiosos para prever tendências futuras de engajamento.

Este projeto demonstrou com sucesso como técnicas de aprendizado de máquina podem ser aplicadas para extrair insights valiosos e prever com precisão métricas importantes, como a taxa de engajamento no Instagram, fornecendo uma base sólida para futuras análises e melhorias.

5.2. Trabalhos Futuros

- Inclusão de Variáveis

Incorporar métricas qualitativas (tipo de conteúdo, hashtags) e temporais (frequência de postagens).

- Exploração de Modelos Não Lineares

Testar regressões polinomiais ou árvores de decisão para capturar relações mais complexas.

- Expansão do Conjunto de Dados

Considerar influenciadores de diferentes nichos e regiões para aumentar a representatividade.

6. Referências

- Documentação Scikit-Learn: https://scikit-learn.org
- Dados originais: https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned?resource=download