Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Álisson Natan dos Anjos Pinheiro
Larissa Ribeiro Firminio

16/11/2024

Resumo

Este projeto visa desenvolver um modelo de Regressão Linear para prever a taxa de engajamento de influenciadores no Instagram, analisando um conjunto de dados real de métricas de influência. A principal meta é criar um modelo preditivo eficaz, avaliar seu desempenho e documentar o processo para disponibilização no GitHub.

Metodologia

1. Durante a análise exploratória de dados, as principais variáveis do conjunto de dados foram examinadas, incluindo seguidores, média de curtidas e posts, identificando padrões e relações com a taxa de engajamento.
2. Já no pré-processamento, valores em formato de texto, como aqueles com sufixos de "k" e "m", foram convertidos para números. Ademais, linhas com valores ausentes, na variável de saída, foram removidas para evitar erros.
3. Para o desenvolvimento do modelo de Regressão Linear, foi feito uso da biblioteca “scikit-learn” para realizar tanto a implementação quanto o treinamento. Além disso, variáveis relevantes foram normalizadas para otimizar o treinamento.
4. Para a realização da avaliação do modelo, métricas como R^2 e Erro Médio Absoluto foram utilizadas. Já para a sua validação, foi utilizado o método de validação cruzada para garantir a generalização.

Principais Resultados

O modelo conseguiu capturar tendências na taxa de engajamento com base nas variáveis escolhidas, apresentando um desempenho razoável. A análise dos coeficientes revelou a influência relativa de cada variável independente, e as métricas de avaliação indicaram que o modelo pode ser ajustado ou aprimorado em projetos futuros para melhor acurácia.

Conclusão

Este estudo fornece uma base para avaliar taxas de engajamento, com possibilidade de aprimoramentos em variáveis adicionais ou em algoritmos alternativos. O código e a documentação estão disponíveis em repositório GitHub, junto com um relatório técnico detalhado.

Introdução

Com o crescimento das redes sociais, plataformas como o Instagram se tornaram fundamentais para influenciadores digitais e empresas que desejam aumentar sua visibilidade e engajamento com o público. A taxa de engajamento – uma métrica que mede a interação dos seguidores com o conteúdo publicado – é um dos principais indicadores de sucesso nessas plataformas.

Este projeto busca entender e prever a taxa de engajamento de influenciadores, o que é essencial para estratégias de marketing digital, permitindo que empresas e criadores de conteúdo avaliem o impacto de suas campanhas e otimizem suas abordagens.

Para atingir esse objetivo, utilizamos um modelo de Regressão Linear, uma técnica estatística simples e eficiente para modelar relações entre variáveis. A Regressão Linear é particularmente útil neste contexto porque permite quantificar o impacto de diferentes variáveis sobre a taxa de engajamento. Essa abordagem fornece uma base interpretável e transparente, adequada para uma análise preditiva que ajude a identificar padrões e tendências importantes.

Descrição do Conjunto de Dados

O conjunto de dados utilizado neste projeto contém informações sobre influenciadores de destaque no Instagram, com variáveis relacionadas ao desempenho e alcance de suas publicações. Cada linha representa um influenciador, e as principais variáveis incluem:

- Número de seguidores (followers): Quantidade de seguidores do influenciador.
- Número de posts (posts): Total de publicações feitas no perfil.
- Média de curtidas (avg_likes) e média de curtidas em novos posts (new_post_avg_like): A média de curtidas nos posts, representando o engajamento médio por publicação.
- Taxa de engajamento a cada 60 dias (60_day_eng_rate): A variável alvo, que indica a taxa média de interação dos seguidores nos posts recentes.

Essas variáveis fornecem uma visão abrangente das características de popularidade e interação de cada influenciador, tornando o conjunto de dados uma base valiosa para desenvolver e avaliar um modelo de Regressão Linear que capture fatores relacionados ao engajamento no Instagram.

Metodologia

Análise Exploratória de Dados

A primeira etapa do projeto foi a análise exploratória dos dados, essencial para entender a distribuição, padrões e correlações entre as variáveis disponíveis.

As colunas com informações em formatos não numéricos foram convertidas para valores numéricos, facilitando o uso direto no modelo. Em seguida, foi realizada uma inspeção visual das variáveis, utilizando gráficos de dispersão e uma matriz de correlação para observar relações entre variáveis.

Verificou-se uma correlação positiva entre seguidores e média de curtidas, embora a taxa de engajamento não acompanhasse necessariamente esse crescimento, indicando a relevância de incluir variáveis que capturem não apenas o alcance, mas também a qualidade da interação.

Implementação do Algoritmo

Para a modelagem, optou-se pelo uso de Regressão Linear, implementada com a biblioteca scikit-learn. A regressão linear foi escolhida por ser uma abordagem transparente e de fácil interpretação, permitindo entender como cada variável independente contribui para a taxa de engajamento dos influenciadores.

Após a preparação dos dados, o modelo foi treinado com followers, posts, avg_likes, e new_post_avg_like como variáveis independentes, e 60_day_eng_rate como variável dependente.

Durante a implementação, a normalização das variáveis independentes foi realizada para melhorar a convergência e estabilidade do modelo. O uso de uma escala padronizada permitiu que o modelo tratasse todas as variáveis de forma balanceada, evitando o domínio de variáveis com valores maiores.

Em seguida, o modelo foi treinado com o conjunto de dados de treino, aplicando o método dos mínimos quadrados para minimizar a função de custo e encontrar os coeficientes mais adequados para cada variável.

Validação e Ajuste de Hiperparâmetros

Para garantir a robustez e a capacidade de generalização do modelo, aplicamos validação cruzada, dividindo o conjunto de dados em múltiplas partições e avaliando o modelo em cada uma delas.

A validação cruzada ajudou a obter uma medida média de desempenho que representasse melhor a generalização. Além disso, vários valores de hiperparâmetros foram testados, como a taxa de aprendizado e o número de épocas no caso do uso de gradiente descendente.

Em relação à escolha de variáveis, foram realizadas análises de correlação e seleção de recursos, mantendo apenas as variáveis que demonstraram impacto relevante na taxa de engajamento, visando melhorar a precisão do modelo e reduzir a variância.

Por fim, as métricas de avaliação foram calculadas para cada iteração da validação cruzada, utilizando R^2, Erro Médio Quadrático (MSE), e Erro Absoluto Médio (MAE) para avaliar o desempenho final do modelo em dados de teste.

Resultados

Métricas de Avaliação

Após o treinamento e a validação do modelo de Regressão Linear, seu desempenho foi avaliado usando métricas adequadas para análise de regressão. As métricas calculadas incluem:

- Coeficiente de Determinação (R^2): Esta métrica indica a proporção da variação na taxa de engajamento explicada pelas variáveis independentes escolhidas. Um valor de R^2 próximo de 1 indicaria que o modelo explica bem a variabilidade dos dados.
- Erro Médio Quadrático (MSE): O MSE fornece uma ideia da magnitude média do erro ao quadrado entre as previsões do modelo e os valores reais, destacando penalizações maiores para erros mais significativos. Um MSE mais baixo é preferível, pois indica um modelo mais preciso.
- Erro Absoluto Médio (MAE): O MAE mede a média dos erros absolutos entre as previsões e os valores reais, o que ajuda a entender a precisão média sem amplificar o efeito de outliers, como ocorre no MSE.

Avaliando os resultados obtidos em um conjunto de dados de teste, o modelo mostrou um desempenho razoável com um valor de R^2 que indica uma explicação moderada da variância dos dados. O MAE e MSE demonstraram que o modelo tem uma margem de erro considerável, o que pode ser um indicativo de que há outros fatores não considerados ou uma possível complexidade nas relações entre as variáveis que não é totalmente capturada pela Regressão Linear.

Visualizações

Para melhor compreensão e comunicação dos resultados, foram gerados gráficos e visualizações que ilustram o desempenho do modelo:

- Gráfico de Dispersão dos Valores Preditos vs. Reais: Este gráfico compara as previsões do modelo com os valores reais de taxa de engajamento no conjunto de dados de teste. A linha ideal é uma linha reta onde valores reais

e previstos coincidem. Neste caso, observamos uma dispersão ao redor da linha ideal, o que indica uma margem de erro nas previsões.

- Histograma dos Resíduos: O histograma dos resíduos (diferença entre valores reais e previstos) ajuda a verificar a distribuição dos erros. Uma distribuição aproximadamente normal sugere que o modelo está ajustado de forma adequada, enquanto distribuições enviesadas podem indicar que há variáveis não consideradas ou outros padrões não capturados pelo modelo.
- Heatmap de Correlação Final: Após a seleção de variáveis, um heatmap foi usado para visualizar as correlações finais entre as variáveis independentes e a taxa de engajamento. Isso ajuda a identificar quais variáveis têm maior impacto e como elas interagem entre si, ajudando na interpretação dos coeficientes do modelo.

Essas visualizações permitiram uma análise crítica do desempenho do modelo, sugerindo que, embora o modelo capture alguns dos padrões de engajamento, ajustes adicionais ou a inclusão de outras variáveis poderiam melhorar a precisão e a capacidade preditiva para contextos mais variados no Instagram.

Discussão

A análise dos resultados obtidos no modelo de Regressão Linear aplicada à inferência de taxa de engajamento de influenciadores no Instagram revela pontos relevantes sobre o comportamento dos dados e a eficácia do modelo em capturar padrões de engajamento.

Discussão Crítica dos Resultados

Os resultados mostram que o modelo consegue captar parcialmente a relação entre o número de seguidores, média de curtidas, média de comentários e a taxa de engajamento. Contudo, as métricas de avaliação indicam que o modelo não é totalmente preciso, sugerindo que há outras variáveis ou relações complexas que não foram capturadas de forma completa.

O valor moderado de R^2 indica que apenas uma parte da variabilidade na taxa de engajamento pode ser explicada pelas variáveis independentes escolhidas. Isso reflete a complexidade do fenômeno de engajamento, que pode depender de fatores adicionais como qualidade do conteúdo, horário das postagens, tipo de público, e outras características comportamentais dos usuários.

Limitações Encontradas

Um dos principais desafios enfrentados foi a qualidade e a complexidade dos dados. A presença de variáveis categóricas e a necessidade de conversão de unidades impactaram diretamente no pré-processamento. Embora a normalização e seleção de variáveis tenham ajudado a melhorar o desempenho do modelo, pode ter havido perda de informações valiosas.

Outro ponto importante é que o modelo de Regressão Linear, por ser uma abordagem linear, pode não capturar todas as interações entre as variáveis, especialmente em dados onde as relações não são lineares. Isso limita sua capacidade de generalizar em cenários onde o engajamento não se comporta de maneira proporcional e simples em relação às variáveis independentes

Impacto das Escolhas no Desempenho do Modelo

As escolhas de pré-processamento, como a normalização e a aplicação de técnicas de regularização, foram fundamentais para melhorar a estabilidade e reduzir o risco de overfitting. No entanto, o uso de um modelo linear impôs limitações à análise, já que esse tipo de modelo assume uma relação direta entre as variáveis independentes e a taxa de engajamento, o que pode não refletir a realidade de redes sociais dinâmicas como o Instagram.

Por fim, a escolha de variáveis independentes baseada em correlações iniciais foi eficaz para reduzir o número de variáveis irrelevantes, mas essa análise não capturou possíveis interações não-lineares que poderiam melhorar a previsão. Técnicas mais sofisticadas, como modelos de regressão não-linear ou de machine learning, poderiam fornecer uma análise mais detalhada e acurada para capturar essas interações complexas.

Conclusão e Trabalhos Futuros

Este projeto demonstrou a aplicação do modelo de Regressão Linear para prever a taxa de engajamento de influenciadores do Instagram com base em variáveis como número de seguidores, média de curtidas e comentários.

A abordagem possibilitou uma visão inicial sobre o impacto dessas variáveis na taxa de engajamento, mostrando que, embora seja possível capturar parte das tendências gerais, o modelo linear apresenta limitações quando aplicado a dados com relações complexas, típicos de redes sociais.

Principais Aprendizados

Durante o desenvolvimento do modelo, foi possível observar a importância do pré-processamento de dados, especialmente ao lidar com variáveis em diferentes escalas e com unidades inconsistentes. A aplicação de técnicas de normalização, bem como a seleção cuidadosa de variáveis, mostrou-se fundamental para a melhoria do desempenho do modelo.

Outro ponto importante foi a necessidade de avaliar criticamente os resultados e identificar limitações inerentes ao modelo linear, reforçando a complexidade de prever métricas de engajamento que dependem de múltiplos fatores.

Referências

1. Documentação do Scikit-Learn: Disponível em: https://scikit-learn.org/stable/documentation.html
2. Biblioteca Pandas: Disponível em: https://pandas.pydata.org/
3. Biblioteca Seaborn: Disponível em: https://seaborn.pydata.org/
4. Gradiente Descendente e Métodos de Otimização: Disponível em: https://arxiv.org/abs/1609.04747