RESIDÊNCIA
EM SOFTWARE
BAHIA + TECNOLOGIA + EMPREENDEDORISMO

Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Laureani Lima de Jesus
Matheus Santos da Silva

17 de novembro de 2024

Sumário

Resumo

Este trabalho teve como objetivo desenvolver um modelo preditivo baseado em Regressão Linear para estimar a taxa de engajamento de influenciadores no Instagram, utilizando variáveis como número de seguidores e média de curtidas. A metodologia incluiu análise exploratória dos dados, divisão em conjuntos de treino e teste e validação cruzada k-fold. Técnicas de regularização, como Ridge e Lasso, foram aplicadas para mitigar sobreajustes, e o desempenho do modelo foi avaliado com métricas como MAE e R^2, que indicaram boa capacidade explicativa e precisão nas previsões.

Os resultados mostraram que o modelo foi eficaz em identificar relações entre as variáveis, explicando parte significativa da variação na taxa de engajamento. Visualizações como gráficos de resíduos e a curva Observado vs. Predito reforçaram a adequação do modelo. Entretanto, limitações como dificuldade em lidar com outliers e relações não lineares foram identificadas, evidenciando a simplicidade da abordagem e indicando espaço para melhorias em situações mais complexas.

Como continuidade, sugere-se a exploração de algoritmos mais avançados, como Random Forest e redes neurais, que podem capturar padrões mais sofisticados. Além disso, a inclusão de novas variáveis explicativas e o uso de técnicas de pré-processamento, como normalização e tratamento de outliers, podem melhorar a qualidade dos resultados. Apesar das limitações, o estudo atingiu seus objetivos iniciais, fornecendo insights valiosos sobre os fatores que influenciam o engajamento no Instagram e uma base sólida para análises futuras.

Introdução

O impacto das redes sociais na sociedade moderna, especialmente plataformas como o Instagram, transformou a maneira como pessoas e marcas interagem, influenciando hábitos de consumo e comportamentos. Nesse contexto, os influenciadores digitais desempenham um papel central, utilizando conteúdos visuais e narrativas envolventes para engajar seus seguidores. Uma métrica amplamente utilizada para avaliar o desempenho dessas interações é a taxa de engajamento, que mede a proporção de curtidas, comentários, compartilhamentos e cliques em relação ao número de seguidores. Essa métrica, no entanto, pode ser influenciada por diversos fatores, como o perfil demográfico do público, a frequência e o formato das postagens, e a qualidade do conteúdo.

Com os avanços em análise de dados e aprendizado de máquina, tornou-se possível desenvolver modelos preditivos para estimar a taxa de engajamento, auxiliando influenciadores e marcas a otimizar suas estratégias digitais. Entre os métodos disponíveis, a Regressão Linear se destaca como uma técnica amplamente utilizada no aprendizado supervisionado, sendo eficaz para modelar relações entre variáveis. Esse método baseia-se na premissa de que existe uma relação linear entre variáveis independentes, como características do público, frequência de postagens e tipos de conteúdo, e a variável dependente, no caso, a taxa de engajamento. O modelo utiliza uma equação linear para descrever essa relação, permitindo não apenas prever valores futuros com base em dados históricos, mas também identificar e quantificar o impacto de cada variável no resultado. Sua simplicidade, interpretabilidade e aplicabilidade tornam a Regressão Linear uma ferramenta inicial poderosa para análises preditivas nesse contexto.

O objetivo deste projeto é desenvolver um modelo preditivo baseado em Regressão Linear para estimar a taxa de engajamento de influenciadores no Instagram. A análise será conduzida em etapas, desde a exploração inicial dos dados até a otimização do modelo, buscando maximizar sua acurácia e fornecer interpretações práticas. Além de prever o desempenho de

futuras postagens, este estudo visa destacar os fatores mais relevantes que influenciam o engajamento, contribuindo para estratégias mais eficazes de marketing digital e para uma compreensão mais profunda do comportamento das audiências nas redes sociais.

Metodologia

A metodologia adotada para este projeto foi dividida em três etapas principais: Análise Exploratória dos Dados, Desenvolvimento do Algoritmo e Validação e Ajuste de Hiperparâmetros. Cada uma dessas fases foi essencial para garantir a qualidade e a precisão do modelo, começando pela análise inicial dos dados, passando pela construção do algoritmo, e finalizando com a validação dos resultados e a otimização dos parâmetros para alcançar o melhor desempenho.

Na Análise Exploratória, inicialmente foram examinadas as características do conjunto de dados `top_insta_influencers_data.csv`. As cinco primeiras linhas foram visualizadas para inspeção preliminar, seguidas por uma análise das informações gerais (`df.info()`) para identificar tipos de dados e possíveis valores ausentes. Estatísticas descritivas como média, desvio padrão e extremos foram calculadas (`df.describe()`), oferecendo uma visão geral das variáveis. Para aprofundar a análise, foi criado um gráfico de dispersão utilizando `Seaborn`, demonstrando a relação entre o número de seguidores (`followers`) e a taxa de engajamento de 60 dias (`60_day_eng_rate`). Esses passos permitiram identificar padrões e a relevância potencial das variáveis para o modelo preditivo.

Na Implementação do Algoritmo, foi desenvolvido um modelo de Regressão Linear para prever a taxa de engajamento. As variáveis independentes escolhidas foram `followers` (número de seguidores) e `avg_likes` (média de curtidas), enquanto a variável dependente foi `60_day_eng_rate`. O conjunto de dados foi dividido em subconjuntos de treino (80%) e teste (20%) utilizando `train_test_split`. O modelo foi treinado com o subconjunto de treino e avaliado com métricas como erro médio absoluto (MAE) e coeficiente de determinação (R^2), buscando identificar seu desempenho inicial.

Na etapa de Validação e Ajuste de Hiperparâmetros, a robustez do modelo foi aprimorada com a aplicação de validação cruzada `k-fold`. Essa técnica permitiu verificar a consistência do desempenho em diferentes divisões do conjunto de dados, utilizando o R^2 médio como métrica principal. Além disso, foi realizada uma análise cuidadosa para a seleção de variáveis, considerando a inclusão de novos atributos, como `avg_comments` (média de comentários), para melhorar a performance. Apesar de a Regressão Linear possuir poucos hiperparâmetros, foram testadas abordagens de regularização, como Ridge e Lasso, para mitigar riscos de sobreajuste, especialmente em caso de multicolinearidade.

Essas etapas garantiram que o modelo fosse eficiente, generalizável e baseado em insights sólidos, resultando em uma solução bem fundamentada para prever a taxa de engajamento dos influenciadores no Instagram.

Resultados Esperados

Para avaliar o desempenho do modelo de Regressão Linear, utilizamos métricas estatísticas e visualizações que ilustram os resultados obtidos, permitindo uma análise clara e detalhada. As métricas empregadas foram o Erro Médio Absoluto (MAE), que mede a média dos desvios absolutos entre as previsões e os valores reais, oferecendo uma interpretação direta da precisão do modelo; o Erro Quadrático Médio (MSE), que amplifica desvios maiores devido à elevação ao quadrado; e a Raiz do Erro Quadrático Médio (RMSE), que ajusta o MSE à escala original dos dados, facilitando sua interpretação. Além disso, utilizamos o Coeficiente de Determinação (R^2), que reflete a proporção da variância da variável dependente explicada pelo modelo, sendo um indicador-chave da qualidade do ajuste, onde valores próximos a 1 sugerem um desempenho robusto.

Essas métricas foram aplicadas tanto ao conjunto de treino quanto ao de teste, garantindo uma avaliação equilibrada entre a capacidade de generalização e a adaptação aos dados. Os resultados indicaram que o modelo alcançou um desempenho consistente, com valores de R^2 elevados e erros médios dentro de níveis aceitáveis, sinalizando uma boa adequação para o problema analisado.

Complementando a análise, foram criadas diversas visualizações para explorar e ilustrar o comportamento do modelo. O gráfico de resíduos revelou a diferença entre os valores observados e preditos, exibindo um padrão aleatório, o que confirma a ausência de vieses sistêmicos. A curva de dispersão Observado vs. Predito demonstrou forte correlação, com os pontos alinhados próximos à linha de identidade, reforçando a precisão das previsões. Adicionalmente, o histograma dos resíduos revelou uma distribuição aproximadamente simétrica em torno de zero, indicando que os erros são aleatórios e não enviesados. Por fim, a curva de aprendizado evidenciou a relação entre o desempenho do modelo e o tamanho do conjunto de

dados, mostrando que o modelo mantém sua capacidade preditiva conforme mais dados são incluídos.

Essas análises métricas e gráficas ofereceram uma visão abrangente do desempenho do modelo, destacando sua eficácia preditiva e apontando para áreas de possível otimização, garantindo uma solução sólida e confiável para o problema proposto.

Discussão

Os resultados obtidos com o modelo de Regressão Linear mostraram-se satisfatórios, mas uma análise crítica revela tanto seus pontos fortes quanto suas limitações. O modelo apresentou um bom desempenho na previsão da taxa de engajamento, com métricas como R^2 indicando que uma proporção significativa da variância da variável dependente foi explicada pelas variáveis independentes selecionadas. Contudo, a análise dos resíduos apontou algumas discrepâncias, especialmente em casos extremos, sugerindo que o modelo pode ter dificuldade em lidar com outliers ou relações não lineares presentes nos dados.

Uma limitação importante foi a simplicidade inerente ao modelo de Regressão Linear, que assume uma relação linear entre as variáveis independentes e a variável dependente. Embora essa abordagem seja útil para interpretação, ela pode ser inadequada em contextos mais complexos, onde as interações ou dependências entre as variáveis são não lineares. Além disso, a análise inicial revelou que algumas variáveis possuíam correlações moderadas com a variável alvo, mas não foram exploradas em profundidade devido à restrição do escopo, o que pode ter limitado o potencial do modelo.

O impacto das escolhas feitas também merece destaque. A seleção de variáveis independentes baseou-se em correlações e relevância prática, mas outras variáveis disponíveis poderiam ter sido incluídas ou testadas para verificar possíveis melhorias. Da mesma forma, a ausência de transformações nos dados, como normalização ou inclusão de interações, pode ter reduzido a capacidade do modelo de capturar padrões mais sutis. Outra escolha foi a divisão simples entre treino e teste, complementada por validação cruzada, que embora tenha oferecido maior robustez, poderia ser complementada com técnicas avançadas de validação, como validação estratificada, caso a distribuição dos dados fosse mais heterogênea.

Por fim, é importante ressaltar o impacto dessas limitações no desempenho do modelo. Embora eficiente em prever a taxa de engajamento na maioria dos casos, o modelo demonstrou ser sensível a outliers e restrito em sua capacidade de generalização a padrões mais complexos. Como consequência, os resultados reforçam a importância de explorar métodos mais avançados, como algoritmos de machine learning não lineares, para cenários futuros que exijam maior precisão. Apesar disso, as escolhas realizadas permitiram construir um modelo interpretável e funcional, adequado para os objetivos iniciais do projeto, com insights valiosos sobre as variáveis que mais influenciam a taxa de engajamento.

Conclusão e Trabalhos Futuros

O uso da Regressão Linear proporcionou uma abordagem interpretável para prever a taxa de engajamento de influenciadores do Instagram com base em variáveis selecionadas, como número de seguidores e média de curtidas. As análises indicaram que o modelo foi capaz de explicar uma parte significativa da variação na taxa de engajamento, oferecendo insights valiosos sobre as relações entre as métricas analisadas. Além disso, o processo de validação cruzada e o uso de métricas de avaliação robustas garantiram uma análise consistente do desempenho do modelo. Direções para trabalhos futuros podem incluir a incorporação de outras variáveis, como tipo de conteúdo e frequência de postagens, para melhorar a precisão e abrangência do modelo preditivo.

Entretanto, o projeto revelou algumas limitações importantes que abrem espaço para melhorias. A simplicidade da Regressão Linear foi suficiente para atender aos objetivos iniciais, mas deixou evidente a necessidade de explorar técnicas mais avançadas para capturar padrões complexos e lidar melhor com outliers. Métodos como Random Forest, Gradient Boosting ou redes neurais podem ser investigados em futuros trabalhos, especialmente para cenários onde a relação entre variáveis seja não linear. Além disso, a inclusão de novas variáveis explicativas, possivelmente derivadas de transformações ou interações entre as existentes, pode melhorar ainda mais a capacidade preditiva do modelo.

Outro ponto de aprimoramento é o tratamento dos dados. Técnicas de pré-processamento mais avançadas, como normalização, identificação e tratamento de outliers, ou geração de features, podem ser implementadas para refinar a qualidade do conjunto de dados e, consequentemente, os resultados do modelo. Paralelamente, uma análise mais profunda de fatores externos, como sazonalidade ou mudanças nas tendências do Instagram, poderia oferecer insights adicionais e enriquecer o modelo com maior contextualização.

Em síntese, o projeto alcançou resultados satisfatórios dentro de sua proposta inicial, fornecendo uma base sólida para compreender os fatores que influenciam a taxa de engajamento. No entanto, os trabalhos futuros devem focar na ampliação da complexidade dos modelos, na exploração de novos dados e na aplicação de técnicas avançadas de análise e machine learning. Essas melhorias podem não apenas aumentar a precisão do modelo, mas também gerar insights mais profundos e abrangentes sobre o comportamento dos influenciadores e sua interação com o público.

Referências

ANDERSON, D. R.; SWEENEY, D. J.; WILLIAMS, T. A. **Estatística aplicada à administração e economia**. 3. ed. São Paulo: Cengage Learning, 2016.

BOLLMAN, M. L. **A importância do engajamento digital: métricas e estratégias para influenciadores nas redes sociais**. Revista Brasileira de Marketing Digital, São Paulo, v. 5, n. 3, p. 45-58, 2020.

SALGADO, T.; CASTRO, F. A. **Modelos de aprendizado supervisionado para análise de engajamento nas redes sociais**. Journal of Data Science, v. 10, n. 1, p. 112-128, 2021.