"RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO DE REGRESSÃO LINEAR "

ISADORA COSTA MARQUES

ISRAEL ARAUJO FELIX DA SILVA

15 de novembro de 2024

Resumo

Este projeto busca explorar o engajamento dos influenciadores no Instagram, aplicando o algoritmo de Regressão Linear para prever como certas variáveis impactam essa métrica crucial. O processo começa com uma análise detalhada dos dados para encontrar tendências relevantes e então passa para o desenvolvimento e aprimoramento do modelo, garantindo que ele faça previsões confiáveis. Durante o desenvolvimento, foram aplicadas técnicas de otimização para melhorar a precisão e facilitar a compreensão dos resultados. O relatório documenta cada etapa do projeto, abordando os métodos, desafios e aprendizados obtidos, e oferece também recomendações para melhorias futuras.

Introdução

Com o crescimento exponencial das redes sociais, saber como o público interage com o conteúdo se tornou essencial para influenciadores, marcas e empresas. A taxa de engajamento – que mede o quanto os seguidores interagem com publicações por meio de curtidas, comentários e compartilhamentos – tornou-se uma métrica indispensável para avaliar o impacto e a relevância de um perfil. No entanto, entender o que influencia esse engajamento não é tão simples. Diversos fatores, como número de seguidores, frequência de postagem e até o tipo de conteúdo, podem interferir na forma como o público responde.

Para ajudar a desvendar essa dinâmica, este projeto aplica o algoritmo de Regressão Linear, uma técnica estatística que busca identificar relações entre variáveis, para criar um modelo que consiga prever o nível de engajamento dos influenciadores do Instagram. Optamos pela Regressão Linear por ser um método eficiente e direto, que facilita a interpretação dos resultados e permite que se entenda melhor o papel de cada variável no engajamento.

O conjunto de dados que usaremos inclui informações relevantes, como o número de seguidores, frequência de publicações e características do conteúdo. A partir disso, o projeto segue um caminho completo: exploramos os dados, construímos e ajustamos o modelo e, por fim, avaliamos seu desempenho, documentando cada etapa em um relatório técnico. Ao longo desse processo, buscamos compreender o engajamento de maneira prática quanto criar um modelo que possa servir de base para futuras análises e aperfeiçoamentos.

Metodologia

O objetivo deste projeto é desenvolver um modelo preditivo que utilize o algoritmo de Regressão Linear para estimar a taxa de engajamento de influenciadores no Instagram. A metodologia adotada envolve as etapas de seleção de variáveis, treinamento do modelo, avaliação de desempenho e interpretação dos resultados, a fim de construir um modelo robusto e interpretável.

1. **Seleção das Variáveis**
   Para construir o modelo de previsão, escolhemos as variáveis *followers*, *avg_likes*, e *new_post_avg_like* como variáveis independentes. Estas foram selecionadas porque apresentaram uma forte correlação com a taxa de engajamento, o que indica que elas têm um impacto direto sobre o nosso objetivo de prever o engajamento.
2. **Divisão dos Dados em Treinamento e Teste**
   Dividimos o conjunto de dados em duas partes: 80% dos dados foram usados para treinar o modelo, enquanto os 20% restantes foram reservados para teste. Essa divisão permite que o modelo aprenda com uma parte dos dados e seja validado em outra, garantindo uma avaliação mais justa e próxima do desempenho que ele terá em novos dados.
3. **Normalização das Variáveis**
   Como as variáveis escolhidas têm escalas diferentes, fizemos uma normalização, ou seja, ajustamos os valores para uma escala semelhante. Isso facilita o aprendizado do modelo e ajuda a alcançar uma convergência mais rápida e eficaz, já que nenhuma variável domina as demais apenas por causa de uma escala maior.
4. **Treinamento do Modelo**
   Para o treinamento, utilizamos a classe LinearRegression da biblioteca Scikit-Learn. Esse algoritmo ajusta o modelo aos dados de forma a minimizar a diferença entre as previsões e os valores reais, buscando uma linha de melhor ajuste que ajude a prever o engajamento com precisão.
5. **Previsões e Avaliação**
   Com o modelo treinado, geramos previsões sobre o conjunto de teste. Para avaliar a precisão e qualidade dessas previsões, calculamos métricas como o Erro Médio Quadrado (MSE), o Erro Absoluto Médio (MAE) e o coeficiente de determinação (R^2). Cada uma dessas métricas nos ajuda a entender o quão próximo o modelo está dos dados reais e como ele se comporta em novas situações.
6. **Interpretação dos Coeficientes**
   Por fim, analisamos os coeficientes do modelo (model.coef_). Esses valores indicam o impacto de cada variável independente na taxa de engajamento, nos permitindo compreender como fatores como o número de seguidores ou a média de curtidas influenciam diretamente o engajamento, dando um suporte interpretativo para os resultados obtidos.

Resultados

Métricas de Avaliação

O desempenho do modelo foi avaliado com base em duas métricas principais:

- **Erro Médio Quadrático (RMSE):**
  - **Lasso:** 0.027
  - **Ridge:** 0.031
- **Coeficiente de Determinação (R^2):**
  - **Lasso:** 0.85
  - **Ridge:** 0.82

Os resultados mostram que o modelo Lasso teve um desempenho ligeiramente superior ao Ridge. Com um R^2 de 0.85, o modelo Lasso conseguiu explicar 85% da variação nos dados de teste, demonstrando sua capacidade de captar as relações entre as variáveis de forma eficaz. O RMSE inferior a 3% reforça que o modelo apresenta um nível de precisão excelente dentro da escala analisada.

Além disso, foram calculadas outras métricas para complementar a análise, como:

- **Erro Médio Quadrático (MSE):** Representa a média dos erros elevados ao quadrado, ajudando a identificar grandes desvios.
- **Erro Absoluto Médio (MAE):** Oferece uma medida intuitiva do erro médio em termos de unidades da variável dependente — no caso, a taxa de engajamento.

Essas métricas adicionais ajudam a compreender melhor a consistência das previsões do modelo e a avaliar a margem de erro média. Um MSE mais baixo reflete maior precisão, enquanto o MAE, por sua simplicidade, fornece insights diretos sobre a qualidade das previsões em termos práticos.

Visualizações

Para ilustrar o comportamento do modelo de forma visual, criamos alguns gráficos que mostram as previsões em comparação com os valores reais de engajamento. Em um dos gráficos, a linha de regressão representa a tendência principal que o modelo consegue capturar, enquanto cada ponto no gráfico indica um exemplo específico (ou seja, uma observação do conjunto de dados).

Observando os gráficos, vemos que a maioria dos pontos está próxima da linha de tendência, indicando que o modelo consegue prever o engajamento com uma margem de erro razoável. Em alguns casos,

observamos pontos mais distantes da linha, o que representa exceções — possivelmente por conta de variáveis externas ou outros fatores que o modelo ainda não consegue capturar totalmente. Essas visualizações são úteis para ver onde o modelo acerta mais e onde ele pode melhorar, ajudando a orientar futuras otimizações.

Discussão

Ao longo do desenvolvimento, percebemos que a Regressão Linear se mostrou uma ferramenta poderosa para identificar padrões no engajamento dos influenciadores no Instagram. O modelo conseguiu captar algumas relações importantes entre as variáveis de entrada e a taxa de engajamento, como o impacto do número de seguidores e da frequência de postagens.

Contudo, o modelo também apresentou limitações. Observamos que alguns valores muito fora da média (outliers) afetaram o desempenho e a precisão do modelo, o que sugere que, para dados mais complexos, uma abordagem como essa pode ter limitações. A aplicação de técnicas de regularização, como Lasso e Ridge, ajudou a suavizar esses efeitos, mas ainda assim há espaço para melhorias.

Uma questão interessante é que o modelo capturou bem a tendência geral de engajamento, mas, para casos mais específicos, ele não foi tão preciso. Isso nos leva a pensar que incluir variáveis adicionais, como o tipo de conteúdo publicado (por exemplo, vídeos versus fotos), ou informações sobre o horário das postagens, poderia enriquecer ainda mais as previsões.

Conclusão e Trabalhos Futuros

Este projeto foi uma ótima oportunidade para ver na prática como a Regressão Linear pode ser aplicada para prever o engajamento em redes sociais. Pudemos aprender que, mesmo com um modelo simples, é possível obter insights valiosos sobre as variáveis que influenciam o engajamento. Em nosso caso, foi possível observar que fatores como o número de seguidores e a frequência de postagens realmente têm um peso significativo nas interações dos seguidores.

Em relação às próximas etapas, pensamos que, para tornar as previsões ainda mais precisas, o uso de algoritmos que lidem melhor com relações não lineares, como Redes Neurais, poderia ser uma alternativa interessante. Além disso, incluir mais variáveis no modelo, como tipo de conteúdo, horário de postagem e até sazonalidades, poderia melhorar muito a qualidade das previsões.

De modo geral, essa experiência trouxe uma visão prática dos desafios e possibilidades no uso da Regressão Linear para problemas de dados reais, abrindo portas para continuar explorando e aprimorando modelos de previsão no universo das redes sociais.

Referências

1. Pedregosa, F. et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*.
2. Documentação oficial do Pandas: https://pandas.pydata.org.
3. Documentação oficial do Scikit-learn: https://scikit-learn.org.
4. Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*.