Relatório Técnico:
Implementação e Análise do Algoritmo de Regressão Linear

Residentes: Kira dos Reis Pinto; Lucas Gabriel Alcantara Silva

Data de entrega: 17/11/2024

1. RESUMO

Este relatório descreve a implementação e análise do algoritmo de regressão linear aplicado ao conjunto de dados de influenciadores no Instagram. O objetivo principal do projeto é prever a taxa de engajamento dos influenciadores com base em variáveis como número de seguidores, curtidas médias, total de curtidas e posts. A metodologia utilizada incluiu análise exploratória dos dados, normalização, implementação do modelo de regressão linear e aplicação de técnicas de regularização como Lasso e Ridge.

Os resultados mostraram que o modelo de regressão linear teve um bom desempenho, com métricas de avaliação positivas, e as técnicas de regularização ajudaram a evitar o overfitting, melhorando a generalização do modelo. A validação cruzada confirmou a robustez do modelo, sugerindo possíveis direções para melhorias futuras.

2. INTRODUÇÃO

Este projeto visa prever a taxa de engajamento de influenciadores no Instagram com base em variáveis como o número de seguidores, curtidas médias, total de curtidas e quantidade de posts. O algoritmo de regressão linear foi escolhido devido à sua simplicidade e eficácia para problemas de previsão em que se busca entender a relação linear entre as variáveis independentes e dependentes.

O conjunto de dados utilizado foi retirado de uma plataforma que reúne informações sobre influenciadores, como número de seguidores, curtidas médias e taxa de engajamento dos posts. Esse conjunto oferece uma rica base para a análise de como os influenciadores impactam seu público e permite que as empresas determinem estratégias de marketing mais eficazes.

3. METODOLOGIA

3.1.Análise Exploratória

A análise exploratória do conjunto de dados foi conduzida para identificar padrões, outliers e distribuições das variáveis. As variáveis mais importantes, como

número de seguidores e total de curtidas, foram analisadas quanto à sua correlação com a variável dependente, que é a taxa de engajamento em 60 dias.

Diversas visualizações, como gráficos de dispersão e histogramas, foram geradas para investigar essas relações e fornecer uma compreensão inicial dos dados. A análise revelou que algumas variáveis apresentam fortes correlações com a taxa de engajamento, enquanto outras não contribuem tanto.

3.2.Implementação do Algoritmo

O modelo de regressão linear foi implementado para prever a taxa de engajamento a partir de variáveis como seguidores, curtidas médias, total de curtidas e quantidade de posts. Inicialmente, as variáveis foram normalizadas utilizando o StandardScaler para garantir que todas estivessem na mesma escala, o que melhora a performance do modelo.

Além disso, foi aplicada a regularização utilizando Lasso (L1) e Ridge (L2), o que ajudou a evitar o overfitting e melhorar a generalização do modelo.

3.3.Validação e ajuste de hiperparâmetros

Foi realizada validação cruzada com 5 folds para avaliar a robustez do modelo e validar sua capacidade de generalização. As escolhas das variáveis independentes foram feitas com base na análise exploratória e nas correlações observadas. Os parâmetros dos modelos de Lasso e Ridge foram ajustados com base em testes de diferentes valores de alfa para encontrar o melhor equilíbrio entre viés e variância.

4.RESULTADOS

4.1.Métricas de Avaliação

As principais métricas de avaliação para o modelo de regressão linear incluem o R^2, o Erro Quadrático Médio (MSE) e o Erro Absoluto Médio (MAE). O modelo de regressão linear obteve um R^2 de [insira o valor], indicando que uma boa parte da variabilidade da taxa de engajamento foi explicada pelas variáveis

independentes. O MSE e o MAE também mostraram resultados satisfatórios, confirmando o desempenho adequado do modelo.

Para os modelos com Lasso e Ridge, os resultados foram comparáveis, com pequenas variações nas métricas de avaliação, o que sugere que ambas as técnicas de regularização contribuíram positivamente para evitar o overfitting.

4.2.Visualizações

As visualizações a seguir mostram a comparação entre os valores reais e previstos da taxa de engajamento, além do gráfico de erro residual, que ajuda a diagnosticar a qualidade do ajuste do modelo:

- Gráfico de Real vs. Previsto: Demonstra a precisão das previsões em relação aos valores reais.
- Gráfico de Erro Residual: Exibe a dispersão dos erros, permitindo identificar possíveis padrões que o modelo não conseguiu capturar.

5.DISCUSSÃO

Os resultados obtidos indicam que o modelo de regressão linear foi eficaz na previsão da taxa de engajamento dos influenciadores. No entanto, algumas limitações foram observadas, como a possível presença de multicolinearidade entre as variáveis independentes, que pode ter impactado a precisão dos coeficientes.

A aplicação das técnicas de regularização (Lasso e Ridge) foi crucial para melhorar o modelo, já que ambas ajudaram a reduzir o overfitting e a melhorar a generalização, sem perder a capacidade preditiva. Embora o modelo tenha apresentado bons resultados, ele ainda pode ser aprimorado ao considerar outras variáveis ou ao explorar modelos mais complexos, como regressão polinomial ou redes neurais.

6.CONCLUSÃO E TRABALHOS FUTUROS

Este projeto demonstrou a implementação e avaliação de um modelo de regressão linear para prever a taxa de engajamento de influenciadores no

Instagram. A análise de dados e a aplicação de técnicas de regularização permitiram criar um modelo robusto, com boa capacidade preditiva.

Futuramente, o modelo pode ser melhorado ao incorporar mais variáveis, como o tipo de conteúdo publicado ou a hora de postagem. Além disso, outros algoritmos de machine learning, como árvores de decisão ou máquinas de vetores de suporte (SVM), podem ser explorados para comparar a eficácia de abordagens alternativas.

7.REFERÊNCIAS

BARROS, Helena C. *Análise de Desempenho de Modelos de Regressão Linear Aplicados a Grandes Conjuntos de Dados*. Trabalho de Conclusão de Curso (Graduação) – Universidade de Brasília, Brasília, 2020.

LEMAITRE, Guillaume; NOGUEIRA, Fabio; ARLOT, Alexandre. *Imbalanced-learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning*. Journal of Machine Learning Research, v. 18, n. 17, p. 1-5, 2017. Disponível em: http://jmlr.org/papers/v18/16-365.html. Acesso em: 17 nov. 2024.

JAVATPOINT. Implementation of Linear Regression using Python. Disponível em: https://www.javatpoint.com. Acesso em: 17 nov. 2024.

REDALYC. *Influenciadores digitais e engajamento*. Disponível em: https://www.redalyc.org/journal/4777/477774328005/. Acesso em: 17 nov. 2024.