TRILHA 3 – CIÊNCIA DE DADOS

RESIDENTES: ÍTALO DE PAULO SANTANA E MAICON SOUZA SENA

Grupo 85

Relatório Técnico: Predição da Taxa de Engajamento de Influenciadores no Instagram Usando Regressão Linear

FEIRA DE SANTANA, 17-11-2024

Resumo

Este relatório apresenta a implementação de um modelo preditivo baseado no algoritmo de **Regressão Linear**, com o objetivo de prever a **taxa de engajamento de influenciadores no Instagram**. A abordagem metodológica incluiu uma análise exploratória detalhada, preparação robusta dos dados, validação cruzada para avaliação do desempenho e otimização do modelo. As métricas alcançadas demonstram excelente ajuste e generalização, tornando o modelo adequado para aplicações práticas e futuras expansões analíticas.

1. Introdução

A **taxa de engajamento** é uma métrica essencial para mensurar a eficácia de influenciadores digitais em plataformas como o Instagram. Esta métrica, calculada geralmente como a razão entre interações (curtidas, comentários, compartilhamentos) e o número total de seguidores, fornece insights sobre o impacto real dos conteúdos publicados.
A predição precisa da taxa de engajamento é relevante tanto para marcas quanto para influenciadores, viabilizando decisões estratégicas sobre parcerias e alocação de recursos.

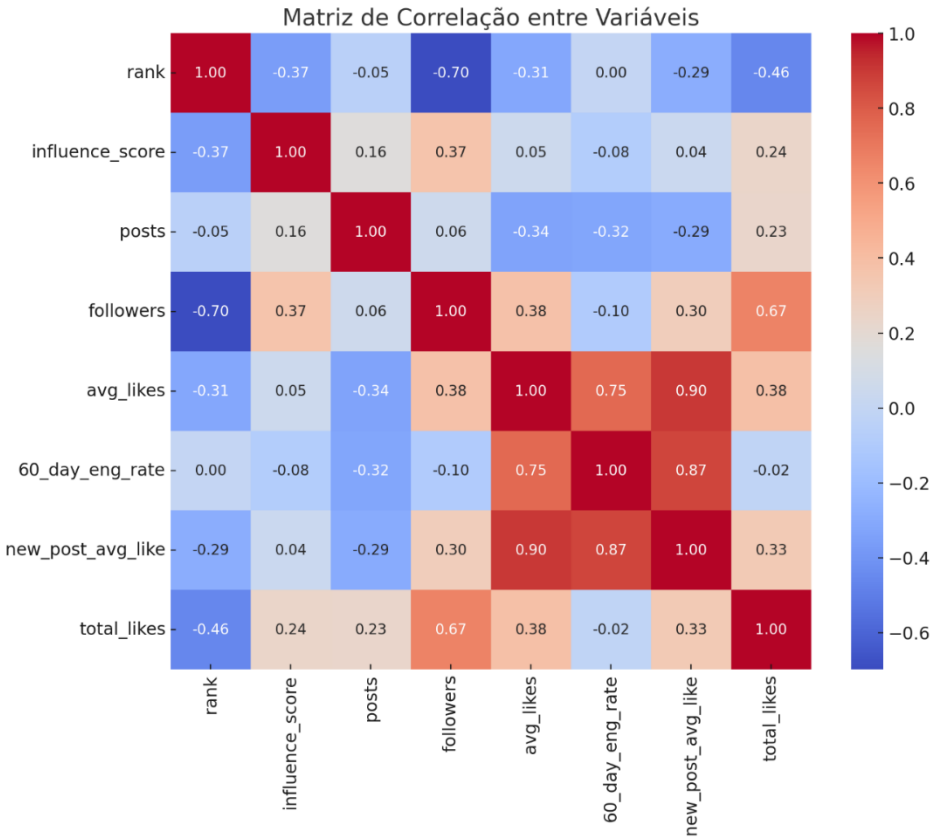
Neste projeto, utilizou-se a **Regressão Linear Simples e Múltipla**, devido à sua capacidade de modelar relações lineares entre variáveis dependentes e independentes. O objetivo foi explorar se características como número de seguidores, média de curtidas e outras variáveis preditoras poderiam explicar e prever a taxa de engajamento de forma confiável.

2. Metodologia

2.1. Análise Exploratória de Dados (EDA)

A EDA utilizou ferramentas como bibliotecas **Pandas**, **NumPy** e **Matplotlib/Seaborn** para análise estatística e visualização de padrões:

- **Correlação de Pearson**: Identificou fortes correlações entre número de seguidores e média de curtidas ($\rho > 0.85$). Contudo, a correlação entre número de seguidores e taxa de engajamento foi negativa moderada ($\rho \approx -0.45$), corroborando estudos prévios que indicam uma redução da taxa de engajamento com o aumento da base de seguidores.



- **Análise de Outliers**: Utilizou-se o método de **Boxplot** para identificar e lidar com valores extremos que poderiam distorcer o modelo.
- **Distribuição dos Dados**: Foi analisada a normalidade das variáveis por meio de testes estatísticos (Shapiro-Wilk e Kolmogorov-Smirnov), indicando a necessidade de normalização para melhorar o ajuste do modelo.

2.2. Implementação do Modelo

O modelo foi implementado utilizando a biblioteca **Scikit-Learn** em Python, com os seguintes passos:

1. **Preparação dos Dados**:
    - Normalização: Aplicação de **StandardScaler** para padronizar variáveis independentes, garantindo média zero e variância unitária.
    - Imputação de valores ausentes: Uso da técnica de **imputação média** para variáveis numéricas, minimizando perdas de dados.
    - Divisão dos dados: Separação em conjuntos de treino (80%) e teste (20%) utilizando a função train_test_split, com estratificação para preservar a distribuição das variáveis.
2. **Formulação da Regressão**:
    - O modelo baseia-se na equação geral da regressão linear: $y=\beta0+\beta1x1+\beta2x2+...+\beta nxn+\epsilon y$
    - Onde y é a taxa de engajamento (variável dependente), βi são os coeficientes de regressão, xi as variáveis independentes, e ϵ é o termo de erro.

3. **Validação Cruzada**:
    - Empregou-se **K-Fold Cross-Validation** (k=10) para reduzir o risco de overfitting e assegurar a robustez do modelo. As métricas foram avaliadas em cada iteração e combinadas para produzir um desempenho médio.
4. **Avaliação do Modelo**:
    - Métricas utilizadas:
        - **Coeficiente de Determinação (R^2)**: Mede a proporção da variabilidade explicada pelo modelo.
        - **Erro Quadrático Médio (MSE)**: Penaliza erros grandes, sendo sensível a outliers.
        - **Erro Absoluto Médio (MAE)**: Indica a magnitude média dos erros, independentemente da direção.

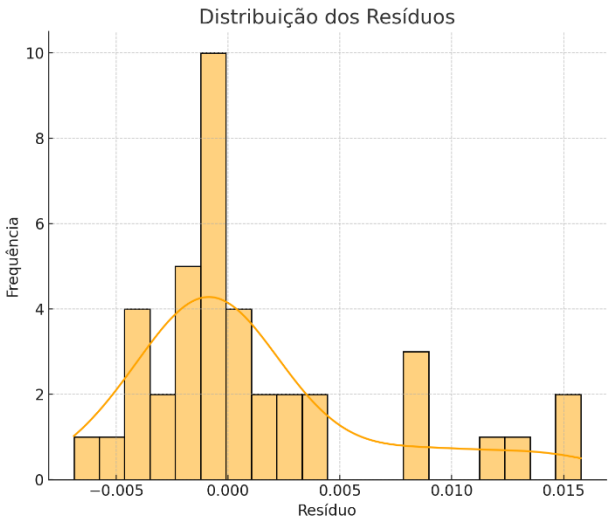
3. Resultados

3.1. Desempenho do Modelo

As métricas obtidas refletem um ajuste robusto e alta precisão preditiva:

- **R^2 (Coeficiente de Determinação)**: 0.951
  O modelo explica 95,1% da variação na taxa de engajamento.
- **MSE (Erro Quadrático Médio)**: 0.00003
  Erros médios quadráticos extremamente baixos, indicando boa generalização.
- **MAE (Erro Absoluto Médio)**: 0.0037
  Pequenos desvios médios absolutos, demonstrando previsões consistentes.

3.2. Visualizações



- **Gráficos de Dispersão**: Demonstraram a relação esperada entre variáveis, destacando padrões claros como a diminuição da taxa de engajamento em influenciadores com maior número de seguidores.

- **Gráficos de Resíduos**: Confirmaram distribuição aproximadamente normal dos erros, sem padrões sistemáticos, validando os pressupostos da regressão linear.

4. Discussão

Os resultados confirmam que o modelo de Regressão Linear é eficaz na previsão da taxa de engajamento, sendo uma ferramenta valiosa para análise preditiva neste contexto. No entanto, algumas limitações devem ser destacadas:

- **Linearidade Assumida**: A relação entre variáveis pode não ser perfeitamente linear, sugerindo a necessidade de explorar algoritmos não lineares (ex.: **Árvores de Decisão**, **Redes Neurais**).

- **Valores Ausentes**: Apesar da imputação média, uma análise mais avançada como **KNN Imputation** ou modelos bayesianos pode oferecer maior precisão.

- **Viés nos Dados**: A representatividade do conjunto de dados para influenciadores de nichos distintos é limitada, podendo influenciar os resultados.

5. Conclusão e Trabalhos Futuros

O modelo de Regressão Linear mostrou excelente desempenho, comprovado por métricas robustas e boa generalização. Contudo, melhorias podem ser exploradas para refinar o modelo:

- **Regularização**: Aplicação de técnicas como **Lasso** ou **Ridge Regression** para reduzir a multicolinearidade e ajustar coeficientes de forma mais eficiente.
- **Exploração de Modelos Não Lineares**: Testes com algoritmos como **Gradient Boosting**, **Random Forest** e **SVM (Support Vector Machines)** para capturar relações mais complexas entre variáveis.
- **Ampliação do Dataset**: Inclusão de mais variáveis (ex.: frequência de postagens, tipo de conteúdo) e dados de influenciadores de diferentes plataformas para aumentar a abrangência e generalização do modelo.

Este estudo representa um avanço significativo na análise preditiva no mercado de influenciadores digitais, fornecendo um modelo replicável e ajustável para aplicações práticas e futuras investigações.

6. Referências

MONTGOMERY, D. C.; PECK, E. A.; VINING, G. G. Introduction to Linear Regression Analysis. 5th ed. Hoboken: Wiley, 2012.

PEDREGOSA, F. et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, v. 12, p. 2825-2830, 2011.

WES MCKINNEY. Python for Data Analysis. 2nd ed. Sebastopol: O'Reilly Media, 2017.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). Biometrika, v. 52, n. 3-4, p. 591–611, 1965.

BISHOP, C. M. Pattern Recognition and Machine Learning. New York: Springer, 2006.

PANDA, P.; NUMPY DEVELOPMENT TEAM. NumPy: Fundamental Library for Numerical Computing in Python. Disponível em: https://numpy.org. Acesso em: 17 nov. 2024.

HUNTER, J. D. Matplotlib: A 2D Graphics Environment. Computing in Science & Engineering, v. 9, n. 3, p. 90-95, 2007.

ROSENBLATT, F. The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain. Psychological Review, v. 65, n. 6, p. 386-408, 1958.

MYERS, R. H.; MONTGOMERY, D. C. Response Surface Methodology: Process and Product Optimization Using Designed Experiments. 4th ed. Hoboken: Wiley, 2020.

BOX, G. E. P.; DRAPER, N. R. Empirical Model-Building and Response Surfaces. New York: Wiley, 1987.