POLO FEIRA DE SANTANA

CEPEDI

RESTIC36

TRILHA CIÊNCIA DE DADOS

Adriel Henrique Oliveira Nunes

Gabriel Dias Santana

Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Relatório Técnico

Feira de Santana

2024

1 RESUMO

Este relatório apresenta a implementação de um modelo de Regressão Linear para prever a taxa média de engajamento de 60 dias com base em variáveis como número de seguidores, curtidas médias e score de influência. Foram realizadas análise exploratória de dados, validação do modelo e ajustes utilizando Ridge e Lasso. O modelo inicial apresentou um coeficiente de determinação R^2 de 0.5359186590632095 e foi ajustado para melhorar a robustez dos resultados.

2 INTRODUÇÃO

2.1 Contextualização do Problema

Com o crescimento do marketing digital, a análise de engajamento de influenciadores tornou-se essencial para campanhas estratégicas. Este trabalho utiliza um modelo de Regressão Linear para analisar e prever a taxa de engajamento em um conjunto de dados de influenciadores do Instagram.

2.2 Descrição do Conjunto de Dados

O dataset contém informações sobre influenciadores, incluindo número de seguidores, curtidas médias, taxa de engajamento e score de influência. Os dados passaram por pré-processamento para normalizar colunas contendo sufixos como “k” e “m” e lidar com valores faltantes.

3 METODOLOGIA

3.1 Análise Exploratória

- Conversão de valores em colunas com sufixos como “k”, “m” e “%”.
- Inspeção de tipos de dados e estatísticas descritivas para identificar padrões e outliers.
- Visualização da relação entre variáveis por meio de gráficos de dispersão e da matriz de correlação.

3.2 Implementação do Algoritmo

- Divisão dos dados em conjuntos de treino (80%) e teste (20%).
- Implementação da Regressão Linear para modelar a taxa de engajamento.
- Avaliação inicial usando métricas como Mean Squared Error (MSE), Mean Absolute Error (MAE) e R^2.

3.3 Validação e Ajustes

- Normalização dos dados usando StandardScaler para evitar problemas de escala.
- Ajustes com Ridge e Lasso para regularização e prevenção de overfitting.
- Comparação das métricas de desempenho entre os três modelos (Linear, Ridge e Lasso).

4 RESULTADOS

4.1 Métricas de Avaliação

- **Regressão Linear**:
  - MSE: 0.0002862337661557805
  - MAE: 0.0108624655082814
  - R^2: 0.5359186590632095
- **Ridge Regression**:
  - MSE: 0.000284565785440814
  - R^2: 0.5386230175924434
- **Lasso Regression**:
  - MSE: 0.0006178603580257902
  - R^2: -0.0017597410510941103

4.2 Visualizações

1) **Dispersão de Valores Preditos vs Reais**
   O gráfico mostra a relação entre os valores preditos pelo modelo e os reais, indicando o desempenho do modelo.
2) **Matriz de Correlação**
   A matriz revelou forte correlação entre as variáveis followers e avg_likes, justificando sua inclusão no modelo.

4.3 Coeficientes do Modelo

Os coeficientes do modelo linear indicam o impacto de cada variável independente na taxa de engajamento:

Tabela 1 – Legenda

Variável	Coeficiente
Followers	-2.226924e-10
Avg Likes	1.480789e-08
Influence Score	2.911285e-04

5 DISCUSSÃO

Os resultados sugerem que **as curtidas médias possuem uma influência positiva na taxa de engajamento (coeficiente: 1.48e-08), enquanto o número de seguidores apresentou um impacto negativo marginal (-2.22e-10)**. O score de influência também contribuiu positivamente para o modelo, mas com menor magnitude (2.91e-04). A performance inicial do modelo, medida pelo coeficiente de determinação R2Rˆ2R2, foi de **0.536**, indicando que aproximadamente 53,6% da variação na taxa de engajamento em 60 dias foi explicada pelas variáveis independentes selecionadas. A regressão Ridge apresentou uma ligeira melhoria no R2Rˆ2R2 (0.539), enquanto a Lasso teve um desempenho inferior, com R2Rˆ2R2 negativo (-0.002), sugerindo uma subajustagem ao modelo.

Limitações incluem a ausência de variáveis como o tipo de conteúdo publicado e a frequência de postagem, que podem ser fatores significativos para o engajamento. A regularização com Ridge contribuiu para maior estabilidade do modelo, enquanto a Lasso não trouxe melhorias significativas devido à possível inadequação dos hiperparâmetros para este conjunto de dados.

6 CONCLUSÃO E TRABALHOS FUTUROS

Este estudo demonstrou a eficácia da Regressão Linear e da regularização para análise de engajamento de influenciadores. Para trabalhos futuros, sugere-se:

- Inclusão de mais variáveis explicativas, como frequência de postagem.
- Uso de algoritmos mais avançados, como regressão polinomial ou redes neurais.

7 REFERÊNCIAS

- Scikit-learn Documentation: https://scikit-learn.org
- Seaborn Documentation: https://seaborn.pydata.org