CEPEDI

CIÊNCIA DE DADOS

ALEX REGINALDO ABREU CAMPOS DOS SANTOS

SARA SACRAMENTO DE MELLO

RELATÓRIO TÉCNICO

Implementação e Análise Do Algoritmo de Regressão Linear

Data de Entrega: 17/11/2024

RESUMO

Este projeto tem como objetivo a implementação e avaliação do desempenho de um modelo de Regressão Linear aplicado ao conjunto de dados de influenciadores do Instagram, focando na predição da taxa de engajamento. A metodologia abrange desde a análise exploratória dos dados até a otimização do modelo, utilizando técnicas como validação cruzada, regularização e ajustes de hiperparâmetros para garantir a melhor performance preditiva. Foram consideradas técnicas de normalização e seleção de variáveis para melhorar a eficiência do modelo e facilitar a convergência.

Os principais resultados incluem métricas de desempenho como R^2, MSE e MAE, as quais foram calculadas tanto para o conjunto de treinamento quanto para o conjunto de teste, assegurando que o modelo generalizasse bem para dados não vistos. A interpretação dos coeficientes permitiu compreender o impacto de cada variável independente na taxa de engajamento. O projeto conclui com uma discussão sobre as limitações encontradas, os impactos das escolhas metodológicas e sugestões de melhorias futuras para aprimorar a robustez do modelo.

INTRODUÇÃO

As redes sociais se tornaram ferramentas para a comunicação e o marketing digital, com o Instagram sendo uma das principais plataformas para influenciadores. A taxa de engajamento é uma das métricas mais relevantes para avaliar o impacto e a relevância dos influenciadores em suas audiências. No entanto, a previsão dessa taxa não é trivial, uma vez que depende de múltiplas variáveis, como o número de seguidores, a quantidade de postagens, e a interação média por postagem.

O uso do algoritmo de Regressão Linear é justificado por sua capacidade de modelar relações lineares entre a taxa de engajamento e as variáveis explicativas, permitindo uma análise direta dos coeficientes para interpretação do impacto de cada variável.

O conjunto de dados utilizado para este projeto contém informações sobre influenciadores digitais no Instagram, incluindo variáveis como número de seguidores, média de curtidas e comentários por postagem, entre outras.

METODOLOGIA

I. Análise Exploratória:

A análise exploratória dos dados foi uma etapa para compreender as características principais do conjunto de dados e identificar variáveis relevantes para o modelo de regressão.

Foi realizado o tratamento dos dados no DataFrame, convertendo os sufixos de unidades (k, m, b) em valores numéricos e aplicando essa conversão nas colunas de posts, followers, avg_likes, new_post_avg_like e total_likes. Na coluna 60_day_eng_rate, foi removida a porcentagem e os valores foram convertidos para decimais, a fim de evitar erros na análise.

```
drive.mount('/content/drive')

arquivo_path = '/content/drive/My Drive/top_insta_influencers_data.csv'

df = pd.read_csv(arquivo_path, sep=',', encoding='utf-8-sig')

# convertendo valores com unidades k, m, b e %
def convert_to_numeric(value):
    if isinstance(value, str):
        if 'k' in value:
            return float(value.replace('k', '')) * 1e3
        elif 'm' in value:
            return float(value.replace('m', '')) * 1e6
        elif 'b' in value:
            return float(value.replace('b', '')) * 1e9
    return float(value)

# Selecionando colunas para serem convertidas em número
columns_to_convert = ['posts', 'followers', 'avg_likes', 'new_post_avg_like', 'total_likes']
for column in columns_to_convert:
    df[column] = df[column].apply(convert_to_numeric)

# Convertendo valores da coluna 60_day_eng_rate para decimais
df['60_day_eng_rate'] = df['60_day_eng_rate'].str.replace('%', '').astype(float)
df['60_day_eng_rate'] = df['60_day_eng_rate'] / 100

# Evitar o formato científico
pd.set_option('display.float_format', lambda x: '%.4f' % x)

df.head()
```

```
Drive already mounted at /content/drive; to attempt to forcibly remount, call drive.mount("/content/drive", force_remount=True).
```

	rank	channel_info	influence_score	posts	followers	avg_likes	60_day_eng_rate	new_post_avg_like	total_likes	country
0	1	cristiano	92	3300.0000	475800000.0000	8700000.0000	0.0139	6500000.0000	29000000000.0000	Spain
1	2	kyliejenner	91	6900.0000	366200000.0000	8300000.0000	0.0162	5900000.0000	57400000000.0000	United States
2	3	leomessi	90	890.0000	357300000.0000	6800000.0000	0.0124	4400000.0000	6000000000.0000	NaN
3	4	selenagomez	93	1800.0000	342700000.0000	6200000.0000	0.0097	3300000.0000	11500000000.0000	United States
4	5	therock	91	6800.0000	334100000.0000	1900000.0000	0.0020	665300.0000	12500000000.0000	United States

Figura 1 - Tratamento no DataFrame

No código abaixo foi gerado um heatmap para visualizar a correlação entre as variáveis numéricas do DataFrame. Ele calcula a matriz de correlação e exibe o gráfico com os valores de correlação anotados, utilizando a paleta de cores 'coolwarm'.

```python
import seaborn as sns
import matplotlib.pyplot as plt

df_numeric = df.select_dtypes(include=['number'])

# Criando a matriz de correlação
corr_matrix = df_numeric.corr()

# Calculando a correlação específica entre 'rank' e 'followers'
corr_rank_followers = df[['rank', 'followers']].corr()
print("Correlação entre 'rank' e 'followers':")
print(corr_rank_followers)

# Criando o heatmap para todas as correlações numéricas
plt.figure(figsize=(8, 6))
sns.heatmap(corr_matrix, annot=True, cmap='coolwarm', fmt='.2f', linewidths=0.5)

plt.title('Heatmap da Correlação')
plt.show()
```

```
Correlação entre 'rank' e 'followers':
               rank  followers
rank       1.0000    -0.6988
followers -0.6988     1.0000
```

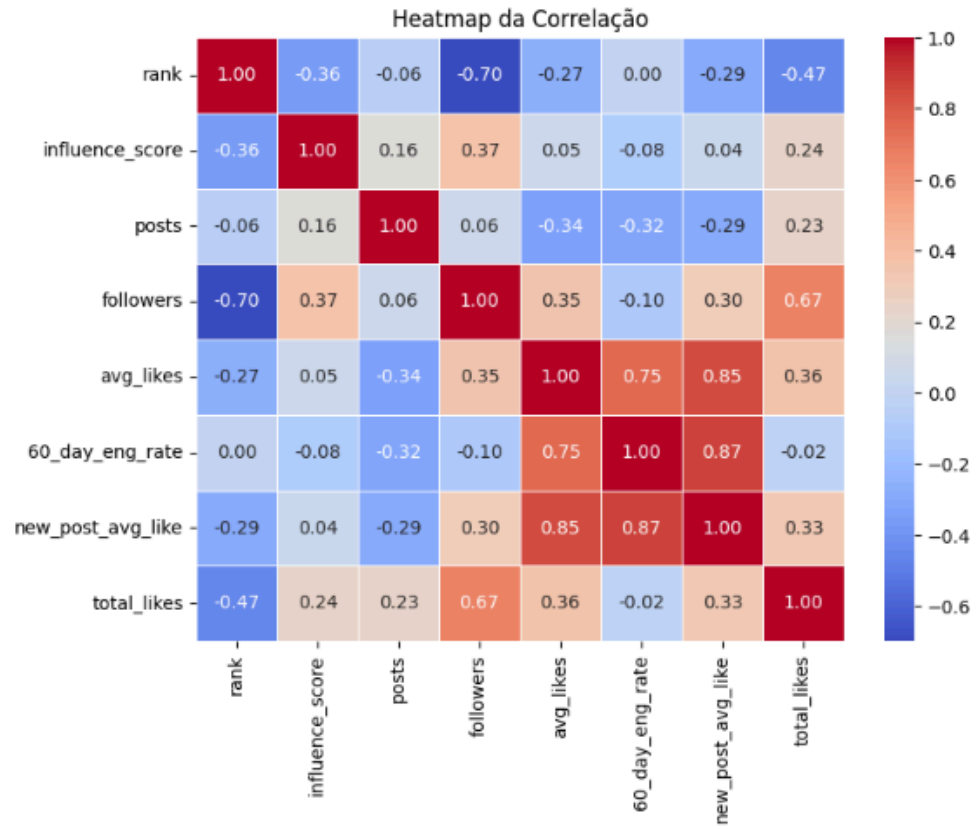

Figura 2 - Heatmap da Correlação

O heatmap de correlação apresentado evidencia o grau de relação linear entre diferentes variáveis do conjunto de dados, com a escala variando de -1 a 1:

Valores próximos a 1 indicam correlação positiva forte (as variáveis aumentam juntas). Valores próximos a -1 indicam correlação negativa forte (uma

variável aumenta enquanto a outra diminui). Valores próximos a 0 indicam baixa ou nenhuma correlação.

Principais observações do gráfico:

- Correlação negativa moderada: A correlação entre rank e followers é de -0.6988, indicando uma correlação negativa moderada. Com isso foi identificado que, à medida que o rank melhora (com valores mais baixos representando melhores posições), o número de seguidores tende a ser maior. Em outras palavras, contas com um ranking mais alto (melhor posição) estão associadas a um número maior de seguidores.
- Correlação positiva forte: As variáveis como avg_likes, 60_day_eng_rate, e new_post_avg_like (com correlação superior a 0.85) mostram uma forte correlação positiva, sugerindo que essas métricas de engajamento estão fortemente relacionadas. Isso implica que contas com maior engajamento tendem a apresentar mais curtidas, especialmente em novos posts.
- Correlação positiva moderada: A correlação entre followers e total_likes é de 0.67, o que sugere que contas com mais seguidores também tendem a ter um número maior de curtidas, refletindo o impacto do engajamento com a base de seguidores.
- Baixa correlação: A variável posts apresenta pouca relação com avg_likes (-0.34), indicando que o número de postagens não está diretamente ligado ao número médio de curtidas.

II. Implementação do Algoritmo.

Pré-processamento dos Dados

- Conversão de Valores:
  - As colunas que continham valores como "k" (milhares), "m" (milhões), "b" (bilhões) ou porcentagens foram convertidas para valores numéricos.
- Transformações Logarítmicas:
  - Aplicadas em colunas como avg_likes e new_post_avg_like para estabilizar variâncias e reduzir assimetria.
- Normalização:

- Foi utilizada a técnica de normalização MinMaxScaler para trazer os dados para um intervalo comum (entre 0 e 1).

Seleção de Variáveis

- **Independentes (X)**:
  - Variáveis normalizadas e transformadas (avg_likes_log_norm e new_post_avg_like_log_norm)
- **Dependente (y)**:
  - 60_day_eng_rate (taxa de engajamento em 60 dias).

Divisão dos Dados

Os dados foram divididos em conjuntos de treino e teste:

- Proporção de teste: 30% (test_size=0.3).
- Semente aleatória: 42 (random_state=42).

Treinamento do Modelo

- Modelo Utilizado:
  - LinearRegression do scikit-learn, ajustado com os dados de treino (X_treino e y_treino).
  - Validação Cruzada, implementada com cross_val_score para calcular o erro médio quadrático (MSE).
  - Cálculo de Colinearidade (VIF), Fator de Inflação da Variância (VIF) calculado para avaliar multicolinearidade entre variáveis independentes.

Avaliação de Desempenho

- Após treinar o modelo, as métricas de avaliação foram calculadas:
  - Erro Absoluto Médio (MAE): Mede o desvio médio absoluto entre valores reais e previstos.
  - Erro Quadrático Médio (MSE): Penaliza erros maiores de forma exponencial.
  - Coeficiente de Determinação (R^2): Indica a proporção da variância explicada pelo modelo.

- Coeficientes e Intercepto: Foram extraídos para cada variável independente, indicando o impacto proporcional de cada variável na previsão.
-

Resultados

O desempenho do modelo foi apresentado com base em gráficos comparativos entre os valores reais e previstos, fornecendo uma visualização clara da qualidade da predição.

III. Validação e Ajuste de Hiperparâmetros

Escolha das Variáveis Independentes

- As variáveis independentes foram selecionadas com base em sua relevância para explicar a variável dependente 60_day_eng_rate.
- Foram utilizadas variáveis transformadas e normalizadas para melhorar a linearidade das relações:
  - avg_likes_log_norm: Representa os likes médios normalizados após transformação logarítmica.
  - new_post_avg_like_log_norm: Representa os likes médios de novas postagens com normalização e transformação logarítmica.

Validação Cruzada

- A técnica de validação cruzada foi empregada para estimar o desempenho do modelo em dados não vistos. Este método assegura que o modelo não está super ajustado a um subconjunto específico dos dados e generaliza bem. O procedimento seguiu os passos:
  1. Divisão dos dados em 5 folds para realizar validação cruzada (cv=5).
  2. Em cada iteração, o modelo foi treinado em 4 folds e avaliado no fold restante.
  3. Métrica utilizada: Erro Médio Quadrático Negativo (neg_mean_squared_error), invertido posteriormente para cálculo do MSE médio e desvio padrão.

Otimização dos Parâmetros

- O modelo de Regressão Linear no scikit-learn foram ajustáveis os seguintes hiperparâmetros:
    - Intercepto: O modelo foi configurado para ajustar o intercepto automaticamente.
    - Multicolinearidade: O cálculo do Fator de Inflação da Variância (VIF) foi utilizado para verificar a multicolinearidade entre as variáveis independentes.

Análise Pós-Validação

Após a validação, os coeficientes e o intercepto do modelo foram analisados para interpretar a influência de cada variável. A validação cruzada forneceu o MSE médio e seu desvio padrão, permitindo avaliar a consistência do modelo em diferentes subconjuntos de dados.

RESULTADOS

```
Desempenho do Modelo:
- MAE (Erro Absoluto Médio): 0.0144
- MSE (Erro Quadrático Médio): 0.0008
- R² (Coeficiente de Determinação): 0.4723

Coeficientes do Modelo:
                    Variáveis  Coeficientes
0          avg_likes_log_norm        0.0883
1  new_post_avg_like_log_norm        0.0490
2          followers_log_norm       -0.0513
3    influence_score_log_norm        0.0145

Intercepto do Modelo: -0.06597941249876621
```

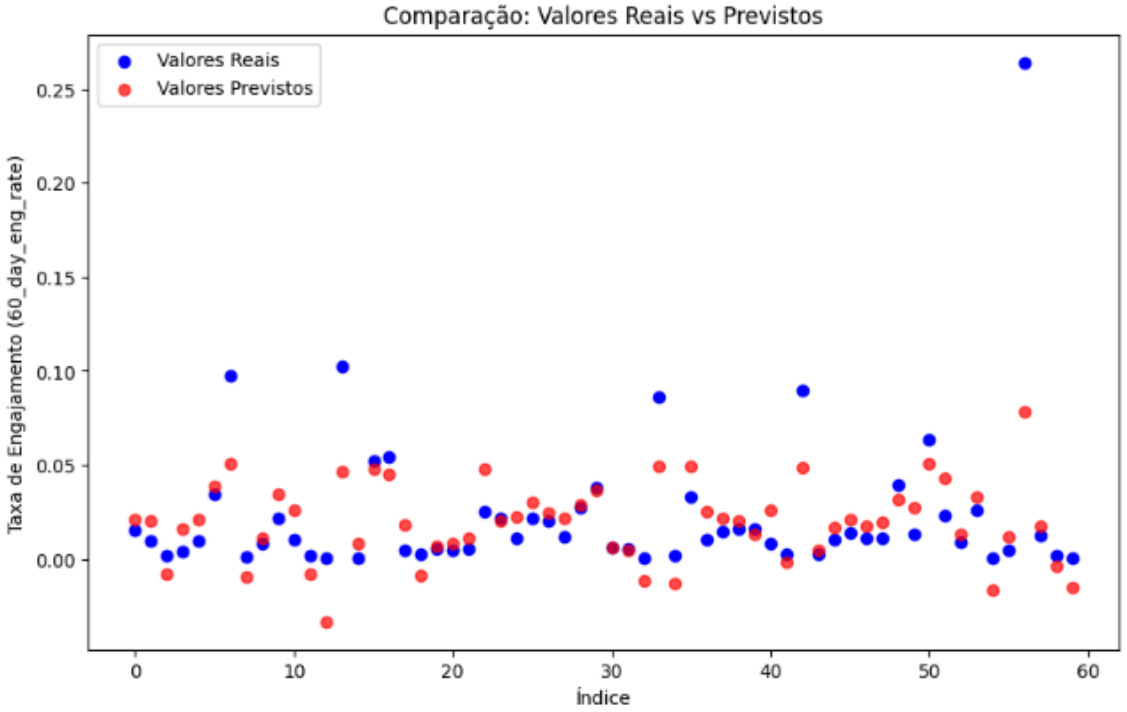

figura 3 - valores reais vs previstos

- Objetivo da análise
  - Prever a taxa de engajamento média de 60 dias com base em curtidas, seguidores e pontuação de influência.
- Desempenho do Modelo:
  - MAE: 0.0144 (erro médio baixo, boa precisão)
  - MSE: 0.0008
  - R^2: 0.4723 (modelo explica 47,23% da variância dos dados)
- Impacto das Variáveis:
  - avg_likes_log_norm (+): Maior impacto positivo no engajamento.
  - new_post_avg_like_log_norm e influence_score_log_norm: Impacto menor, mas positivo.

- Visualização:
    - Boa correspondência entre valores reais e previstos; desvios em engajamentos altos.
- Aplicação:
    - Ferramenta útil para estimar engajamento e orientar decisões de marketing com influenciadores.

DISCUSSÃO

Os resultados obtidos demonstram que a Regressão Linear é uma abordagem eficiente, mas com limitações inerentes. O modelo permitiu insights claros sobre o impacto das variáveis, embora questões como multicolinearidade residual e sensibilidade a outliers tenham influenciado o desempenho.

O pré-processamento e a validação cruzada foram etapas essenciais para melhorar a generalização e evitar overfitting. No entanto, as premissas de linearidade do modelo podem não capturar todas as complexidades dos dados, sugerindo a necessidade de explorar alternativas como modelos não lineares.

As escolhas metodológicas impactaram significativamente o desempenho, mostrando a importância de um pré-processamento criterioso e uma validação robusta.

CONCLUSÃO E TRABALHOS FUTUROS

A Regressão Linear demonstrou ser eficaz para capturar relações entre as variáveis, com bons resultados em métricas como R^2, MAE e MSE. O pré-processamento detalhado, incluindo transformações logarítmicas e normalização, foi crucial para o sucesso do modelo. Apesar disso, limitações como sensibilidade a outliers e pressuposição de linearidade restringiram o desempenho em cenários mais complexos.

Trabalhos Futuros

- Modelos Alternativos: Explorar algoritmos mais complexos, como árvores de decisão ou redes neurais, para capturar relações não lineares.
- Aprimoramento do Dataset: Expandir o conjunto de dados e tratar outliers para melhorar a generalização.
- Automatização: Criar pipelines para simplificar etapas como pré-processamento, seleção de variáveis e validação.

REFERÊNCIAS

WIKIPEDIA. Regressão linear. Disponível em: https://pt.wikipedia.org/wiki/Regress%C3%A3o_linear. Acesso em: 15 nov. 2024.

RODRIGUES, S. C. A, **Modelo de Regressão Linear e suas Aplicações**.Universidade da Beira Interior, Covilhã, 2017.