RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO DE REGRESSÃO LINEAR

JOSÉ EDUARDO MENDES DA SILVA

TAÍS FARIAS RIBEIRO DE SOUZA

17 DE NOVEMBRO DE 2024

RESUMO

O seguinte relatório expõe o desenvolvimento de um modelo preditor baseado no algoritmo de Regressão Linear para prever a taxa de engajamento de influenciadores na rede social Instagram. Foram analisadas as variáveis principais através de técnicas de análise exploratória, seguidas da implementação e otimização do modelo com validação cruzada e ajuste de hiperparâmetros. Os resultados obtidos no final mostram as limitações e os potenciais do modelo.

INTRODUÇÃO

Com o crescimento das redes sociais, prever a taxa de engajamento de influenciadores se tornou crucial para estratégias de marketing digital. Com isso, esse projeto utiliza da Regressão Linear pela sua simplicidade e interpretabilidade para modelar essa relação. Os dados utilizados foram obtidos através do Top Instagram Influencers Data, que possui informações como o número de seguidores, curtidas e taxa de engajamento.

METODOLOGIA

- **Análise Exploratória:**

A análise demonstrou que há uma correlação positiva entre o número de seguidores e a média de likes, enquanto outras variáveis, como o engajamento em 60 dias, mostraram padrões que indicam influência significativa na variável dependente. Visualizações gráficas, incluindo gráficos de dispersão e pairplots, foram geradas para identificar padrões e possíveis outliers.

- **Implementação do Algoritmo:**

O algoritmo de Regressão Linear foi implementado com a biblioteca Scikit-Learn. Para desenvolver o modelo, as variáveis independentes escolhidas foram *followers*, *avg_likes*, *60_day_eng_rate*, *total_likes*, e *country*. A variável dependente foi definida como *influence_score*. Os dados foram divididos em treino e teste, utilizando 80% dos dados para treinamento e 20% para avaliação. As variáveis independentes foram normalizadas utilizando *StandardScaler* para melhorar a estabilidade do treinamento.

Validação e Ajuste de Hiperparâmetros:

Para garantir que o modelo não se ajustasse excessivamente aos dados de treino, foi aplicada a validação cruzada com 5 dobras.

```
cv_scores = cross_val_score(knn, X_train_scaled, y_train, cv=5,
scoring='neg_mean_squared_error')

cv_rmse = np.sqrt(-cv_scores)

print(f'CV RMSE: {cv_rmse.mean()}')
```

1. **Variáveis Independentes**

A seleção das variáveis independentes foi feita com base na análise exploratória inicial do conjunto de dados. Como citado anteriormente, as variáveis escolhidas foram:

- followers: Número de seguidores de um influenciador, uma métrica direta de alcance.
- avg_likes: Média de curtidas por postagem, representando o engajamento médio.
- 60_day_eng_rate: Taxa de engajamento dos últimos 60 dias, algo essencial para medir a interação recente.
- total_likes: Total de curtidas acumuladas, indicando popularidade histórica.
- country: Representado numericamente como valores associados a continentes, essa variável captura o contexto regional, que pode influenciar o comportamento do público.

```
sns.pairplot(data[['followers', 'avg_likes', '60_day_eng_rate',
'total_likes']])

plt.show()

sns.scatterplot(x='followers', y='avg_likes', data=data)

plt.title('Relação entre Followers e Avg Likes')

plt.show()
```

```
X = data[['followers', 'avg_likes', '60_day_eng_rate', 'total_likes',
'country']]

y = data['influence_score']
```

Essas variáveis foram escolhidas pela sua relação lógica com a variável dependente (*influence_score*).

Com o intuito de melhorar o desempenho do modelo, foi utilizada a ferramenta **`GridSearchCV`**, responsável por testar combinações específicas de hiperparâmetros para encontrar a configuração ideal. O processo foi aplicado ao modelo KNN implementado no código, ajustando os seguintes parâmetros:

- **n_neighbors:** Número de vizinhos considerados no modelo.
- **metric:** Métrica utilizada para calcular a distância entre os pontos (euclidiana ou manhattan).

Como mostrado no código abaixo:

```
param_grid = {'n_neighbors': [3, 5, 7, 9, 11], 'metric': ['euclidean', 'manhattan']}

grid_search = GridSearchCV(KNeighborsRegressor(), param_grid, cv=5,
scoring='neg_mean_squared_error')

grid_search.fit(X_train_scaled, y_train)

best_knn = grid_search.best_estimator_

print(f'Best Parameters: {grid_search.best_params_}')

print(f'Best CV Score: {-grid_search.best_score_}')
```

RESULTADO

A análise exploratória inicial do conjunto de dados revelou importantes relações entre as variáveis. No pairplot das variáveis numéricas (`followers`, `avg_likes`, `60_day_eng_rate`, e `total_likes`), foi observado a presença de correlações evidentes entre as métricas de engajamento e o total de curtidas, sugerindo que influenciadores com maiores números de seguidores têm, em geral, médias mais altas de curtidas e engajamento. Entretanto, alguns outliers no gráfico indicam que nem todos os influenciadores com muitos seguidores alcançam altos níveis de engajamento.

Além disso, o scatterplot mostrando a relação entre seguidores (`followers`) e média de curtidas (`avg_likes`) reforça a ideia de que há uma correlação positiva entre essas variáveis. No entanto, a dispersão significativa em certos intervalos sugere a existência de outros fatores influenciando a taxa de engajamento, como o tipo de conteúdo ou a geolocalização dos influenciadores.

No treinamento do modelo, após normalização e uso do algoritmo de Regressão Linear, foi possível ajustar uma função preditiva para estimar a pontuação de influência (`influence_score`). Durante a validação, o modelo apresentou valores de métricas como o Erro Médio Absoluto (MAE) e R-quadrado (R^2), que indicam que ele foi capaz de capturar as relações entre as variáveis de entrada e a saída dependente, mas com certa limitação na generalização, provavelmente devido à variabilidade no conjunto de dados.

CONCLUSÃO E TRABALHOS FUTUROS

Este estudo mostrou como o uso de análise exploratória de dados e modelos preditivos, como a Regressão Linear, pode fornecer insights significativos sobre o comportamento de influenciadores digitais. A relação positiva entre os seguidores e as métricas de engajamento confirma o papel do tamanho da audiência na construção de influência. Contudo, as limitações do modelo, como a dependência de variáveis simples, ressaltam a necessidade de incluir fatores mais qualitativos e não-lineares em análises futuras.

DISCUSSÃO

Os resultados apresentados destacam a importância de explorar e entender a relação entre variáveis antes da aplicação de algoritmos preditivos. A forte correlação entre seguidores e engajamento indica que, em muitos casos, a quantidade de seguidores pode ser um bom indicativo de influência. No entanto, os outliers observados nos gráficos indicam que outros fatores, como o tipo de público ou a consistência na produção de conteúdo, podem afetar o desempenho de engajamento.

O modelo de Regressão Linear demonstrou ser uma abordagem adequada para o problema, mas sua simplicidade pode limitar sua eficácia em cenários mais complexos. Embora o desempenho do modelo seja satisfatório em termos de precisão, ajustes futuros, como a aplicação de técnicas de regularização (Lasso e Ridge), poderiam ajudar a minimizar o impacto de outliers e melhorar a capacidade preditiva.

REFERÊNCIAS

JHA, Suraj. Top Instagram Influencers Data Cleaned. Disponível em: https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned. Acesso em: 17 nov. 2024.