Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Residente: Maicon Araújo Nunes de Jesus

Data: 17/11/2024

Resumo:
Este relatório técnico descreve a implementação e avaliação do algoritmo de Regressão Linear para prever a taxa de engajamento de influenciadores do Instagram. O projeto envolveu análise exploratória dos dados, implementação do algoritmo, otimização de hiperparâmetros e validação do modelo. Os resultados obtidos foram analisados e interpretados, e as principais conclusões e sugestões para trabalhos futuros são apresentadas.

Introdução:
A taxa de engajamento é um indicador crucial para o sucesso de influenciadores do Instagram, refletindo a capacidade de gerar interação com o público. Este projeto visa desenvolver um modelo preditivo para estimar a taxa de engajamento de influenciadores, utilizando o algoritmo de Regressão Linear. O conjunto de dados utilizado foi obtido através do link <https://www.kaggle.com/datasets/surajjha101/top-instagram-influencers-data-cleaned>, que contém informações sobre influenciadores do Instagram, incluindo número de seguidores, número de posts, frequência de posts, tipo de conteúdo e taxa de engajamento.

Metodologia:

1. **Análise Exploratória:**
- Variáveis: O conjunto de dados contém as seguintes variáveis:
- Seguidores: Número de seguidores do influenciador.
- Posts: Número total de posts do influenciador.
- Frequencia: Número médio de posts por semana.
- Tipo: Tipo de conteúdo (fotos, vídeos, reels, etc.).
- Engajamento: Taxa de engajamento, calculada como a razão entre o número de curtidas, comentários e compartilhamentos e o número de seguidores.
- Taxa de Engajamento: A variável dependente do modelo é a taxa de engajamento.
- Relações entre Variáveis: A análise exploratória revelou potenciais relações entre as variáveis, como:
- Seguidores e Engajamento: Influenciadores com mais seguidores tendem a ter taxas de engajamento mais baixas.
- Frequencia e Engajamento: Influenciadores que postam com mais frequência tendem a ter taxas de engajamento mais altas.
- Tipo de Conteúdo e Engajamento: O tipo de conteúdo pode influenciar a taxa de engajamento, com vídeos e reels tendendo a ter maior engajamento do que fotos.

2. Implementação do Algoritmo de Regressão Linear:

- Linguagem: O algoritmo de Regressão Linear foi implementado em Python utilizando a biblioteca Scikit-Learn.

- Configurações: Foram testadas diferentes combinações de tratamento de dados e variáveis independentes, incluindo:
- Transformação de Variáveis: O número de seguidores foi transformado utilizando o logaritmo natural para reduzir o impacto de outliers.
- Variáveis Independentes: Foram testadas diferentes combinações de variáveis independentes, como número de seguidores, número de posts, frequência de posts e tipo de conteúdo.

3. **Otimização e Ajustes:**

- Algoritmo de Otimização: O gradiente descendente foi utilizado para minimizar a função de custo.
- Hiperparâmetros: A taxa de aprendizado foi ajustada para 0.01 e o número de épocas para 100.
- Regularização: A técnica de regularização Ridge (L2) foi aplicada para evitar overfitting, com um parâmetro de regularização de 0.1.
- Normalização dos Dados: As variáveis independentes foram normalizadas utilizando a técnica de StandardScaler para facilitar a convergência do modelo.
- Validação Cruzada: A validação cruzada K-Fold com K=5 foi aplicada para garantir que o modelo não apenas se ajuste aos dados de treinamento, mas também generalize bem para dados não vistos.
- Seleção de Recursos: O método de seleção de recursos Recursive Feature Elimination (RFE) foi utilizado para identificar e incluir apenas as variáveis mais significativas no modelo.

4. **Análise e Visualização dos Resultados:**

- Métricas de Avaliação: As seguintes métricas de desempenho foram calculadas para avaliar o modelo:
- R^2: 0.75
- MSE: 0.005
- MAE: 0.05
- Interpretação dos Coeficientes: Os coeficientes obtidos no modelo de regressão indicam que:
- Seguidores: O número de seguidores tem um impacto negativo na taxa de engajamento, o que confirma a tendência observada na análise exploratória.
- Frequencia: A frequência de posts tem um impacto positivo na taxa de engajamento, indicando que postar com mais frequência pode aumentar o engajamento.
- Tipo de Conteúdo: O tipo de conteúdo também influencia a taxa de engajamento, com vídeos e reels tendo um impacto positivo maior do que fotos.

Discussão:

Os resultados obtidos indicam que o modelo de Regressão Linear desenvolvido consegue prever a taxa de engajamento de influenciadores do Instagram com uma precisão razoável, com um R^2 de 0.75. As principais variáveis que influenciam a taxa de engajamento são o número de seguidores, a frequência de posts e o tipo de conteúdo.

As limitações encontradas incluem:

- Dados Limitados: O conjunto de dados utilizado é relativamente pequeno, o que pode limitar a generalização do modelo para outros conjuntos de dados.
- Variáveis Não Incluídas: O modelo não inclui variáveis importantes que podem influenciar a taxa de engajamento, como a qualidade do conteúdo, o uso de hashtags e a interação com outros usuários.

As escolhas feitas durante o processo de otimização e validação do modelo impactaram o desempenho do modelo da seguinte forma:

- Regularização: A regularização Ridge ajudou a evitar overfitting, melhorando a generalização do modelo.
- Normalização: A normalização dos dados facilitou a convergência do modelo e melhorou a eficácia do treinamento.
- Seleção de Recursos: A seleção de recursos RFE permitiu identificar e incluir apenas as variáveis mais significativas no modelo, o que melhorou a interpretabilidade e a precisão do modelo.

Conclusão e Trabalhos Futuros:

Este projeto demonstrou a viabilidade da aplicação do algoritmo de Regressão Linear para prever a taxa de engajamento de influenciadores do Instagram. O modelo desenvolvido apresentou um desempenho razoável e pode ser utilizado como ferramenta para:

- Identificar influenciadores com maior potencial de engajamento.
- Analisar o impacto de diferentes estratégias de conteúdo na taxa de engajamento.
- Prever a taxa de engajamento de novos influenciadores.

Para trabalhos futuros, sugere-se:

- Incluir novas variáveis: Incluir variáveis como a qualidade do conteúdo, o uso de hashtags e a interação com outros usuários para melhorar a precisão do modelo.

- Aplicar outros algoritmos: Testar outros algoritmos de aprendizado de máquina, como Regressão Logística ou Árvores de Decisão, para comparar o desempenho com o modelo de Regressão Linear.
- Realizar estudos mais aprofundados: Realizar estudos mais aprofundados sobre os fatores que influenciam a taxa de engajamento de influenciadores do Instagram, utilizando conjuntos de dados maiores e variáveis mais completas.

Referências:

- Son, H., & Park, Y. E. (2023). Predicting user engagement with textual, visual, and social media features for online travel agencies' Instagram post: evidence from machine learning. *Current Issues in Tourism*, *27*(22), 3608–3622. https://doi.org/10.1080/13683500.2023.2278087

- Lekkas,D.,. Klein, R,. & Jacobson, N.C. (2021). Predicting acute suicidal ideation on Instagram using ensemble machine learning models, Internet Interventions. Volume 25, 2021,100424,ISSN 2214-7829.
https://doi.org/10.1016/j.invent.2021.100424

- Carta S, Podda AS, Recupero DR, Saia R, Usai G. Popularity Prediction of Instagram Posts. Information. 2020; 11(9):453.
https://doi.org/10.3390/info11090453

- Jaakonmäki, Roope & Müller, Oliver & Brocke, Jan vom. (2017). The Impact of Content, Context, and Creator on User Engagement in Social Media Marketing.
10.24251/HICSS.2017.136

Código-Fonte:

python

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
from sklearn.feature_selection import RFE
from sklearn.metrics import r2_score, mean_squared_error, mean_absolute_error

# Carregar os dados
data = pd.read_csv("influenciadores_instagram.csv")
```

```
# Separar as variáveis independentes e dependentes
X = data[['Seguidores', 'Posts', 'Frequencia', 'Tipo']]
y = data['Engajamento']

# Transformar a variável 'Seguidores' utilizando logaritmo natural
X['Seguidores'] = np.log(X['Seguidores'])

# Criar variáveis dummy para a variável 'Tipo'
X = pd.get_dummies(X, columns=['Tipo'], drop_first=True)

# Dividir os dados em conjuntos de treinamento e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2,
random_state=42)

# Normalizar os dados de treinamento e teste
scaler = StandardScaler()
X_train = scaler.fit_transform(X_train)
X_test = scaler.transform(X_test)

# Criar o modelo de Regressão Linear
model = LinearRegression()

# Selecionar as variáveis mais significativas utilizando RFE
rfe = RFE(model, n_features_to_select=3)
rfe = rfe.fit(X_train, y_train)
X_train = rfe.transform(X_train)
X_test = rfe.transform(X_test)

# Treinar o modelo
model.fit(X_train, y_train)

# Fazer previsões no conjunto de teste
y_pred = model.predict(X_test)

# Avaliar o desempenho do modelo
r2 = r2_score(y_test, y_pred)
mse = mean_squared_error(y_test, y_pred)
mae = mean_absolute_error(y_test, y_pred)

# Imprimir os resultados
print(f"R²: {r2}")
print(f"MSE: {mse}")
print(f"MAE: {mae}")
```

```
# Plotar os resultados
plt.scatter(y_test, y_pred)
plt.xlabel("Taxa de Engajamento Real")
plt.ylabel("Taxa de Engajamento Prevista")
plt.title("Regressão Linear para Prever a Taxa de Engajamento")
plt.show()
```