CEPEDI
ResTIC36
Ciência de Dados

ARTHUR LAGO MARTINS
JOÃO VICTOR OLIVEIRA SANTOS

Análise de Perfis do Instagram Usando Regressão Linear

JEQUIÉ
2024

Resumo

Este projeto visa realizar a análise dos perfis de influenciadores no Instagram utilizando a regressão linear. O objetivo é prever a taxa de engajamento (60-day engagement rate) com base em variáveis como número de seguidores, influência e o número de postagens. A análise inclui uma exploração dos dados, a implementação de modelos de regressão (Ridge) e a validação cruzada para otimização do modelo. Os resultados demonstram que a regressão linear pode ser uma ferramenta útil na previsão do engajamento, com a validação cruzada indicando a robustez do modelo.

Introdução

A avaliação do desempenho de influenciadores no Instagram é fundamental para marcas que buscam realizar campanhas de marketing digital mais eficientes. A taxa de engajamento é um indicador crucial, pois reflete a interação dos seguidores com o conteúdo postado. Dado o grande volume de dados disponíveis, a regressão linear se apresenta como uma abordagem eficiente para prever a taxa de engajamento com base em características do perfil dos influenciadores. A utilização de modelos de regressão linear, como o Ridge, é justificada pela sua simplicidade e capacidade de lidar com regularização, o que pode evitar o overfitting, especialmente quando lidamos com um grande número de variáveis explicativas.

O conjunto de dados utilizado contém informações sobre influenciadores do Instagram, incluindo variáveis como número de seguidores, número de postagens, média de likes por postagem, score de influência e taxa de engajamento nos últimos 60 dias. O dataset foi obtido de fontes públicas e foi previamente processado para converter os dados em formato numérico, com tratamento de valores ausentes e variáveis categóricas.

Metodologia

Análise Exploratória

Na análise exploratória, foi realizada uma inspeção inicial dos dados para entender a distribuição das variáveis, as correlações entre elas e o comportamento das variáveis numéricas. Gráficos como mapas de calor de correlação, histogramas e gráficos de dispersão foram utilizados para identificar padrões, tendências e possíveis outliers.

Passos:

1. Tratamento dos Dados: As colunas com valores em formatos como 'K', 'M' e 'B' foram convertidas para valores numéricos adequados.
2. Identificação de Outliers e Missing Data: Variáveis como "followers", "avg_likes" e "total_likes" foram limpas e tratadas para garantir a qualidade dos dados.

Implementação do Algoritmo

A regressão linear foi implementada usando o modelo de Ridge Regression, uma técnica que aplica regularização para evitar o overfitting, especialmente quando o número de características é elevado. A variável dependente escolhida foi a taxa de engajamento nos últimos 60 dias (60_day_eng_rate), enquanto as variáveis independentes incluem o número de seguidores, o rank do influenciador, o score de influência e o número de postagens. A variável categórica "country" foi transformada em variáveis dummies.

Processo:

1. Pré-processamento: Normalização das variáveis independentes usando o StandardScaler para garantir que todas as variáveis tenham a mesma escala.
2. Divisão dos Dados: O conjunto de dados foi dividido em dados de treino (80%) e teste (20%) usando a função train_test_split.
3. Modelo: O modelo foi treinado com o algoritmo Ridge e avaliado em termos de suas métricas de erro.

Validação e Ajuste de Hiperparâmetros

A validação cruzada foi realizada para avaliar a robustez do modelo. O conjunto de dados foi dividido em 5 partes, e o modelo foi treinado e validado em cada uma delas, permitindo uma avaliação mais confiável da performance do modelo. O hiperparâmetro do modelo foi ajustado com o valor alpha=1.0.

Resultados

As métricas utilizadas para avaliar o desempenho do modelo incluem:

- R^2 (Coeficiente de Determinação): Indica o quanto o modelo consegue explicar a variação da variável dependente. O valor obtido foi 0.12, indicando que o modelo é apenas parcialmente eficaz em explicar a variabilidade da taxa de engajamento.

- MSE (Erro Quadrático Médio): O valor obtido foi 0.0025, o que demonstra um erro relativamente baixo entre os valores previstos e os reais.
- MAE (Erro Absoluto Médio): O valor foi 0.0391, indicando a média do erro absoluto entre as previsões e os valores reais.

Visualizações

- Mapa de Calor das Correlações entre Variáveis: A análise de correlação revelou que variáveis como "followers" e "avg_likes" possuem uma correlação moderada com a taxa de engajamento.
- Distribuição de Seguidores: O histograma mostrou que a maioria dos influenciadores tem menos de 1 milhão de seguidores, com algumas exceções no topo da distribuição.
- Gráfico de Dispersão Seguidores vs. Média de Likes: Esse gráfico revelou que influenciadores com maior número de seguidores tendem a ter maior média de likes, embora o relacionamento não seja perfeitamente linear.

Discussão

O modelo de regressão linear apresentou um desempenho moderado. O R^2 relativamente baixo sugere que, embora algumas variáveis possam influenciar a taxa de engajamento, existem outros fatores não considerados que impactam significativamente os resultados.

As escolhas feitas nas variáveis independentes podem não ter capturado completamente a complexidade dos dados, como a influência de variáveis temporais (ex. sazonalidade de postagens) ou interações mais complexas entre as variáveis.

Limitações

1. Quantidade de Dados: O tamanho do conjunto de dados pode não ser suficiente para capturar todas as variações nas taxas de engajamento.

2. Simples Modelo de Regressão Linear: O uso de modelos mais complexos, como regressão polinomial ou redes neurais, poderia melhorar a capacidade do modelo de capturar relações mais complexas.

Conclusão

Síntese dos Principais Aprendizados

O uso de regressão linear foi útil para uma primeira aproximação, mas as limitações indicam que modelos mais sofisticados podem ser necessários. A validação cruzada foi importante para avaliar a estabilidade do modelo, mas melhorias no pré-processamento e mais dados podem aumentar a precisão.

Sugestões de Melhorias:

- Explorar modelos mais complexos, como Random Forest ou Redes Neurais.
- Incorporar mais variáveis, como o conteúdo do post (imagem, vídeo) e o horário da postagem.
- Utilizar técnicas de feature engineering para melhorar a capacidade preditiva do modelo.