CEPEDI
RESTIC36

Relatório Técnico

Implementação e Análise do Algoritmo de Regressão Linear

Nome do Residente: Gleidson da Silva Nascimento
Data de Entrega: 17/11/2024

Neste relatório, apresentamos o processo de **análise e modelagem** da **taxa de engajamento** de postagens em redes sociais utilizando o algoritmo de **regressão linear** com regularização **Lasso (L1)** e **Ridge (L2)**. A **análise exploratória** dos dados foi seguida por um processo de **seleção de variáveis**, **validação cruzada** e **ajuste de hiperparâmetros**. O desempenho dos modelos foi avaliado utilizando as métricas **R^2**, **MSE** e **MAE**, com ênfase na análise dos coeficientes das variáveis mais impactantes. Este trabalho apresenta conclusões sobre os fatores mais influentes para a previsão de engajamento e sugestões para melhorias futuras.

Introdução:

O estudo teve como objetivo prever a **taxa de engajamento** de postagens em plataformas de redes sociais, utilizando dados como **Ponto de Influência**, **Número de Postagens**, **Média de Curtidas por Postagem**, entre outros. A previsão dessa taxa pode ser útil para **estratégias de marketing digital**, pois permite prever o desempenho das postagens com base em variáveis controláveis.

O algoritmo escolhido para esse estudo foi a **regressão linear**, com aplicação de regularizações **Lasso** e **Ridge**. A regularização ajuda a evitar o **overfitting** e melhora a generalização do modelo.

3.2 Descrição do Conjunto de Dados

O conjunto de dados utilizado contém informações sobre postagens nas redes sociais e inclui as seguintes variáveis:

- **PONTO_DE_INFLUENCIA**: Representa o grau de influência da conta ou do ponto onde a postagem é realizada.
- **POSTAGEM**: Número de postagens realizadas.
- **SEGUNDORES**: Indicador de tempo de engajamento ou resposta.
- **MEDIA_CURTIDAS**: Média de curtidas nas postagens.
- **NOVA_POST_MEDIA_CURTIDAS**: Média de curtidas nas novas postagens.
- **TOTAL_CURTIDAS**: Total de curtidas acumuladas nas postagens.
- **TAXA_ENG_60_DIAS**: A variável alvo, representando a taxa de engajamento medida em 60 dias.

Análise Exploratória dos Dados

Foi realizada uma análise exploratória dos dados para entender o comportamento das variáveis e sua relação com a variável alvo (taxa de engajamento). A **matriz de correlação** revelou que variáveis como **Média de Curtidas por Postagem** e **Ponto de Influência** têm uma correlação significativa com a taxa de engajamento, enquanto outras variáveis mostraram correlações mais fracas.

Gráfico de correlação:

Implementação do Algoritmo

A implementação do algoritmo de regressão linear envolveu os seguintes passos:

1. **Seleção de variáveis**: Com base na análise de correlação, foram selecionadas as variáveis mais influentes para prever a **taxa de engajamento**.
2. **Pré-processamento de dados**: Os dados foram normalizados para garantir que as variáveis estivessem na mesma escala.
3. **Treinamento do modelo**: Foram utilizados dois tipos de regressão linear:
    - **Lasso (L1)**: Aplica penalização para tornar alguns coeficientes exatamente zero.
    - **Ridge (L2)**: Aplica penalização para reduzir os coeficientes e controlar o overfitting.

Validação e Ajuste de Hiperparâmetros

A **validação cruzada** foi aplicada para garantir que o modelo generalizasse bem para novos dados. Os hiperparâmetros de regularização foram ajustados utilizando a técnica de **validação cruzada** e **grid search** para encontrar o melhor valor de penalização.

Métricas de Avaliação

As métricas utilizadas para avaliar o desempenho dos modelos foram:

- **R^2 (Coeficiente de Determinação)**: Mede a proporção da variabilidade explicada pelo modelo.
- **MSE (Erro Quadrático Médio)**: Mede a média das diferenças ao quadrado entre as previsões e os valores reais.
- **MAE (Erro Absoluto Médio)**: Mede a média das diferenças absolutas entre as previsões e os valores reais.

Visualizações

A seguir, apresentamos algumas visualizações que ajudam a ilustrar os resultados obtidos:

1. **Matriz de Correlação**:
   A matriz de correlação mostra a relação entre as variáveis independentes e a variável alvo (taxa de engajamento).

2. **Distribuição dos Erros**:
   Um gráfico de dispersão dos **resíduos** foi gerado para verificar se há padrões não capturados pelo modelo

Análise Crítica dos Resultados

Os resultados indicam que tanto os modelos **Lasso** quanto **Ridge** tiveram bom desempenho na previsão da **taxa de engajamento**. A regularização ajudou a controlar o overfitting, especialmente no modelo **Lasso**, que foi mais eficiente em reduzir o número de variáveis.

A escolha das variáveis foi crucial para o sucesso do modelo. A **Média de Curtidas por Postagem** e o **Ponto de Influência** mostraram-se as mais significativas para a previsão do engajamento.

6.2 Limitações e Melhorias Futuras

Uma limitação do modelo é a dependência das variáveis presentes no conjunto de dados. Variáveis adicionais, como **comentários** e **compartilhamentos**, poderiam melhorar a acurácia do modelo. Além disso, seria interessante testar outros modelos, como **Árvores de Decisão** e **Redes Neurais**, para comparar o desempenho.

Conclusão

Este projeto mostrou que a **regressão linear** com regularização é uma abordagem eficaz para prever a **taxa de engajamento** em postagens nas redes sociais. A análise exploratória e a seleção de variáveis ajudaram a identificar os fatores mais relevantes para o modelo. Embora os resultados sejam satisfatórios, existem oportunidades para melhorar o modelo com mais dados e técnicas de modelagem mais avançadas.

Referências:

https://www.youtube.com/watch?v=UMYuk3GkbFM
Alura | Cursos online de Tecnologia