RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO DE REGRESSÃO LINEAR

JEFFERSON ANTÔNIO TANAJURA SILVA
WANDERSON BRENO REIS DE OLIVEIRA

16/11/2024

Resumo

Neste projeto, implementamos um modelo de regressão linear para prever a taxa de engajamento de influenciadores do Instagram. Foram utilizados métodos de otimização como gradiente descendente e mínimos quadrados, com regularizações Lasso e Ridge. A normalização dos dados e a validação cruzada foram aplicadas para melhorar a performance do modelo. Os resultados mostraram que o modelo possui bom desempenho com uma alta precisão, conforme evidenciado pelas métricas R^2, MAE e RMSE.

Introdução

Contexto e Problema:

As redes sociais se tornaram essenciais para o marketing digital, e medir a taxa de engajamento dos influenciadores é fundamental para determinar seu impacto. A regressão linear foi escolhida por sua simplicidade e interpretabilidade, facilitando a análise de como diferentes fatores afetam o engajamento.

Descrição do Conjunto de Dados:

O conjunto de dados contém informações sobre influenciadores, incluindo o número de seguidores, posts, curtidas médias e taxa de engajamento. As variáveis foram selecionadas com base na sua relevância para prever a taxa de engajamento.

Metodologia

Análise Exploratória

O dataset foi carregado e analisado para verificar a existência de valores nulos e outliers. As colunas que continham informações numéricas como "posts", "followers", "avg_likes" e "total_likes" foram convertidas de strings para valores numéricos. Um resumo estatístico foi gerado para entender a distribuição das variáveis e a correlação entre elas.

Implementação do Algoritmo

Foram testados diferentes métodos de otimização:

- **Gradiente Descendente:** Algoritmo iterativo para minimizar a função de custo.
- **Mínimos Quadrados:** Método padrão para resolver a regressão linear.
- **Regularizações Lasso e Ridge:** Utilizadas para lidar com colinearidade e reduzir o risco de overfitting.

Normalização e Validação

Normalização: Utilizamos o `StandardScaler` para padronizar os dados, facilitando a convergência dos algoritmos.
Validação Cruzada: Aplicamos validação cruzada para verificar a generalização do modelo, usando `cross_val_score` com 5 folds.

Seleção de Recursos

Analisamos a correlação entre as variáveis e eliminamos aquelas com correlação muito baixa ou alta redundância. As variáveis selecionadas para o modelo final foram: "influence_score", "posts", "followers", "avg_likes" e "new_post_avg_like".

Resultados

Modelo	MAE	RMSE	R^2
Gradiente Descendente	2.0377	2.1284	0.2655
Mínimos Quadrados	0.4034	0.5714	0.9471
Lasso (L1)	0.3407	0.5065	0.9584
Ridge (L2)	0.4036	0.5714	0.9471

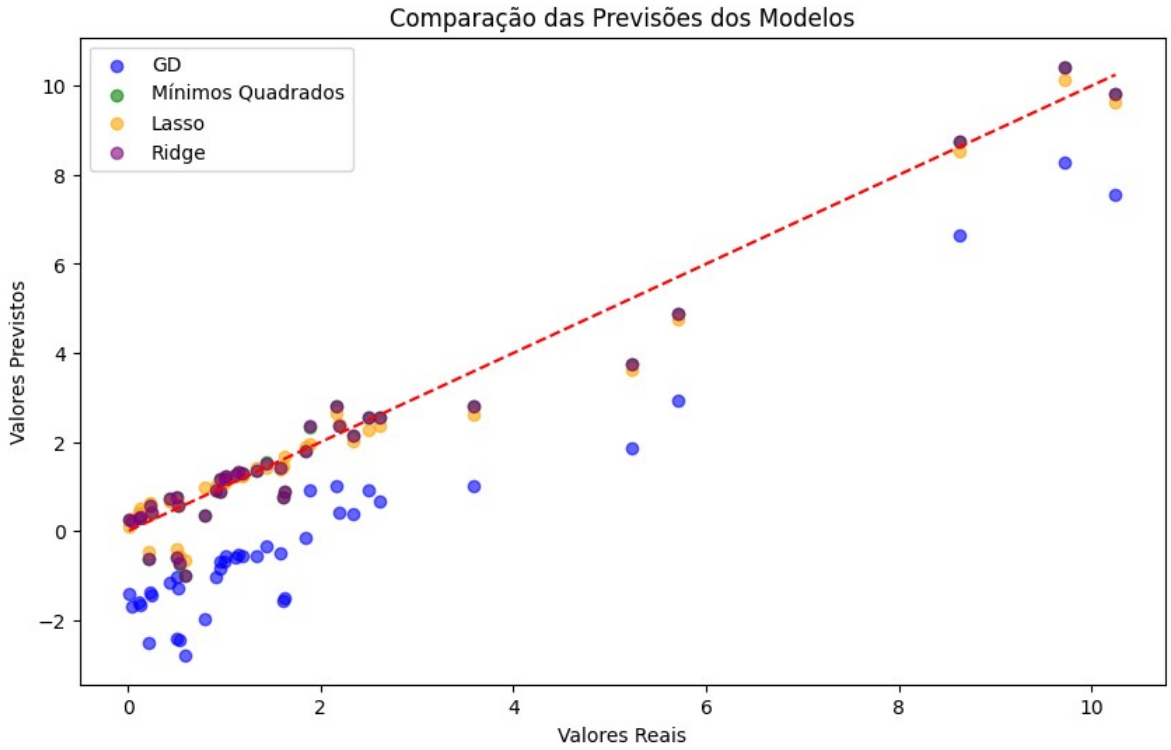

Gráfico de dispersão mostrando a comparação entre os valores reais e previstos para cada modelo.

Discussão

Análise Crítica:

O uso de regularizações Lasso e Ridge ajudou a lidar com multicolinearidade, melhorando a estabilidade do modelo.

A normalização dos dados foi essencial para a convergência eficiente dos algoritmos de otimização, especialmente para o gradiente descendente.

Os modelos Ridge Regression e Lasso tiveram um ótimo desempenho, o que indica que a regularização trouxe benefícios significativos para a construção do modelo. No dataset de influenciadores, várias variáveis, como "número de seguidores" e "curtidas médias", podem ter alta correlação entre si. O Ridge Regression consegue lidar bem com essa situação, distribuindo o peso de forma mais equilibrada entre as variáveis correlacionadas. Já o Lasso Regression conseguiu

identificar e ignorar variáveis menos relevantes, focando nas mais importantes para prever a taxa de engajamento.

Conclusão e Trabalhos Futuros

Conclusão:

O projeto demonstrou como o uso de técnicas de otimização e regularização pode melhorar o desempenho de modelos de regressão linear na previsão da taxa de engajamento. Os modelos de Ridge e Lasso apresentaram o melhor desempenho geral, com menor erro e alta precisão.

Trabalhos Futuros:

Explorar outros modelos de regressão, como regressão polinomial e regressão robusta;

Coletar um conjunto de dados maior e mais diversificado para aumentar a generalização do modelo;

Investigar outras técnicas de seleção de recursos, como a análise de componentes principais (PCA).

Referências

Pedregosa, F., et al. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research, 12*, 2825-2830.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer.

Dataset: Top Instagram Influencers Dataset, disponível em [link do dataset, se aplicável].