Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Autores: Gustavo Cruz, Vinícius Moreira

Resumo

Este projeto tem como objetivo desenvolver um modelo preditivo utilizando a técnica de Regressão Linear para estimar a taxa de engajamento dos influenciadores do Instagram. A análise foi realizada sobre um conjunto de dados contendo informações como número de seguidores, curtidas médias e postagens, com a finalidade de prever a variável "influence_score", que representa a taxa de engajamento do influenciador. O trabalho abrange desde a análise exploratória dos dados, passando pela implementação de modelos de regressão (Linear, Ridge, Lasso e SGD), até a avaliação de seu desempenho por meio de métricas como MSE, RMSE, R^2 e MAE. A validação cruzada foi aplicada para testar a robustez dos modelos, com o uso de regularização (Ridge e Lasso) para evitar overfitting.

Introdução

O uso de algoritmos de machine learning tem se expandido nas mais diversas áreas, e um dos setores que mais se beneficiam dessa tecnologia é o marketing digital. Especificamente no Instagram, as marcas e empresas buscam identificar influenciadores com alto potencial de engajamento, sendo a taxa de engajamento um indicador crucial de influência. Este projeto visa desenvolver um modelo preditivo para estimar a taxa de engajamento dos influenciadores com base em variáveis como número de seguidores, número médio de curtidas, postagens, e outras métricas relacionadas.

Descrição do Conjunto de Dados

O conjunto de dados utilizado neste projeto contém informações sobre influenciadores do Instagram, como o número de seguidores, curtidas médias, postagens e outros dados relacionados ao seu engajamento na plataforma. O arquivo de dados, intitulado `top_insta_influencers_data.csv`, foi processado e preparado para análise com a aplicação de técnicas de limpeza e conversão de dados.

As colunas presentes na base de dados são:

- **followers**: Número de seguidores do influenciador.
- **avg_likes**: Número médio de curtidas por postagem.
- **posts**: Número total de postagens feitas pelo influenciador.
- **total_likes**: Número total de curtidas recebidas pelo influenciador.
- **country**: País de origem do influenciador.
- **influence_score**: A variável alvo, que representa a taxa de engajamento do influenciador.

Metodologia

Análise Exploratória

A análise exploratória dos dados foi realizada para entender melhor a estrutura dos dados e as relações entre as variáveis. Foram aplicadas técnicas de visualização de dados, como gráficos de dispersão e a matriz de correlação, para identificar possíveis correlações entre a variável dependente "influence_score" e as variáveis independentes. Além disso, as variáveis numéricas foram normalizadas para facilitar a convergência do modelo.

Implementação do Algoritmo

A implementação do modelo de Regressão Linear foi feita utilizando a biblioteca Scikit-learn, que oferece um conjunto de ferramentas poderosas para modelagem preditiva. O modelo foi treinado com dados de entrada após a realização do pré-processamento, como o preenchimento de valores ausentes e a conversão de variáveis categóricas em variáveis numéricas.

Os seguintes algoritmos de regressão foram implementados:

- **Regressão Linear (Mínimos Quadrados)**
- **Regressão Ridge (L2 Regularization)**
- **Regressão Lasso (L1 Regularization)**
- **Gradiente Descendente (SGD)**

Validação e Ajuste de Hiperparâmetros

Para otimizar o desempenho do modelo, foram aplicadas técnicas de regularização como Ridge e Lasso, que ajudam a evitar o overfitting e a melhorar a generalização do modelo. Além disso, a validação cruzada foi utilizada para testar a robustez dos modelos, garantindo que o modelo treinado se comporte bem em dados não vistos.

Resultados

Métricas de Avaliação

Os resultados dos modelos de regressão foram avaliados utilizando as seguintes métricas:

- **Mean Squared Error (MSE)**
- **Root Mean Squared Error (RMSE)**
- **R² (coeficiente de determinação)**
- **Mean Absolute Error (MAE)**

A seguir, apresentamos as métricas de desempenho para cada modelo de regressão:

1. **Gradiente Descendente (SGD)**
    - **MSE**: 1,690,284.745
    - **RMSE**: 1,296.34
    - **R²**: 0.2645
    - **MAE**: 1,009.22
2. **Regressão Linear (Mínimos Quadrados)**
    - **MSE**: 1,673,565.134
    - **RMSE**: 1,293.72
    - **R²**: 0.2719
    - **MAE**: 1,003.54
3. **Regressão Ridge (L2 Regularization)**
    - **MSE**: 1,673,102.018
    - **RMSE**: 1,293.57
    - **R²**: 0.2721
    - **MAE**: 1,003.23
4. **Regressão Lasso (L1 Regularization)**
    - **MSE**: 1,686,490.225
    - **RMSE**: 1,296.47
    - **R²**: 0.2652
    - **MAE**: 1,006.57

Validação Cruzada

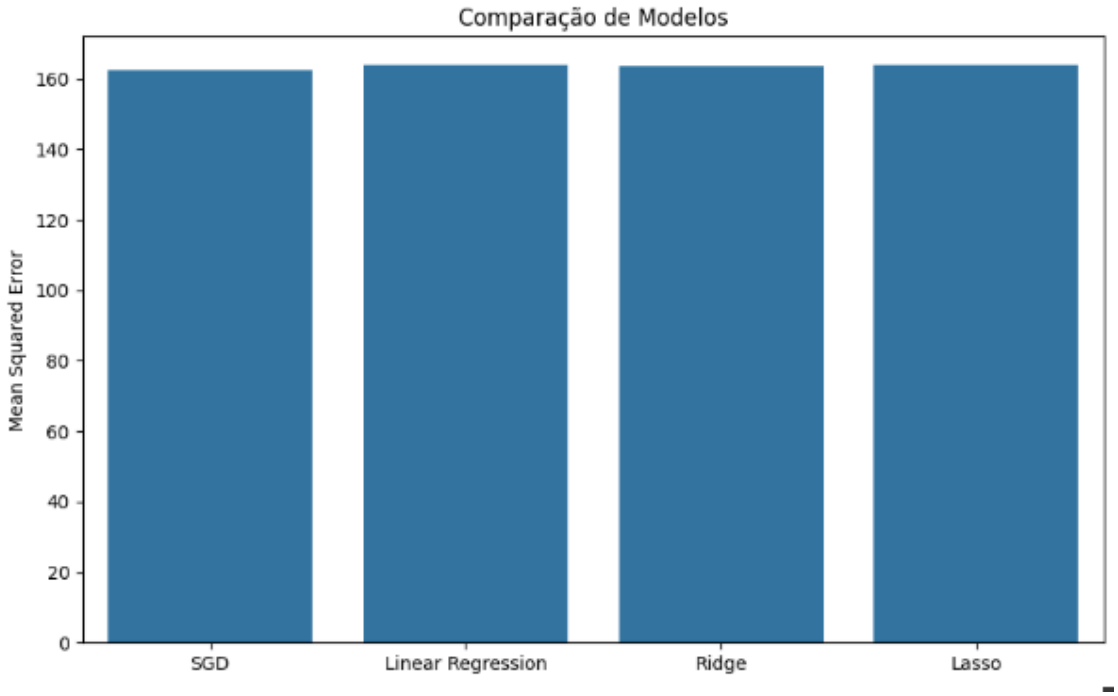
A validação cruzada foi aplicada aos modelos para garantir a robustez dos resultados. As médias das métricas de MSE para cada modelo foram:

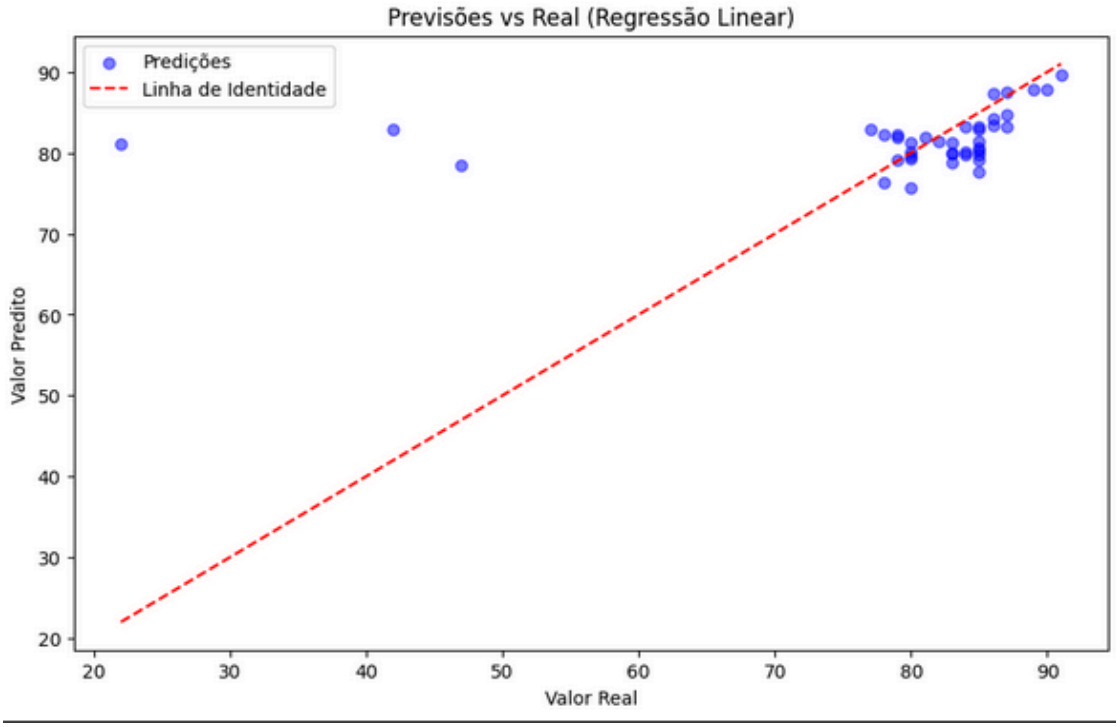
- **SGD**: -1,678,760.028
- **Regressão Linear**: -1,674,764.544
- **Ridge**: -1,674,519.055
- **Lasso**: -1,688,377.456

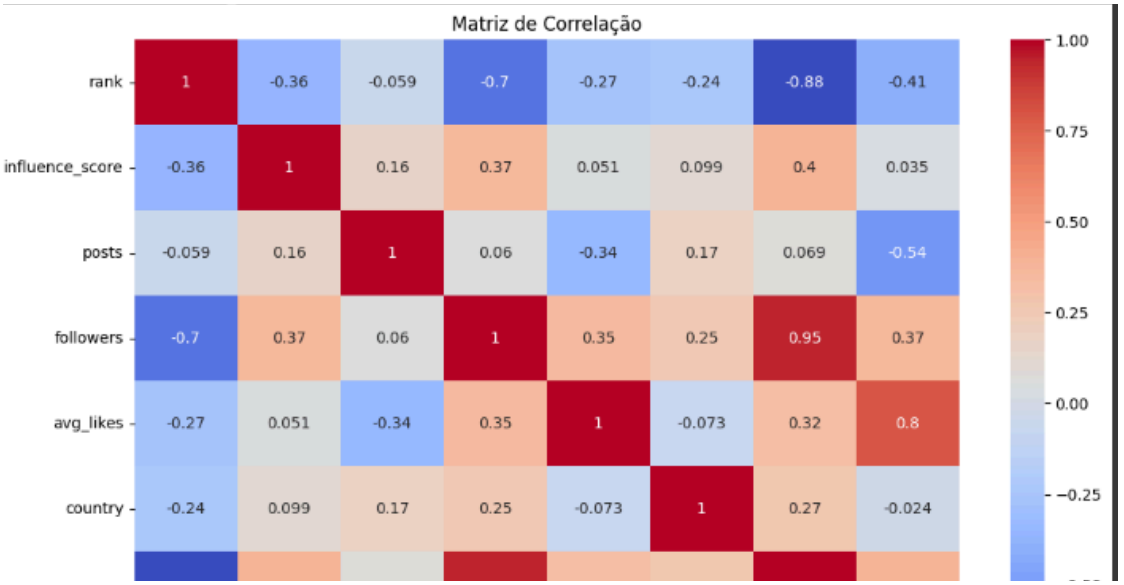
A validação cruzada foi aplicada a todos os modelos, e as comparações de desempenho foram feitas com base no MSE.

Visualizações

O código gerou gráficos para comparar o desempenho dos modelos e visualizou as previsões em comparação com os valores reais. A seguir, apresentamos alguns exemplos de gráficos gerados:


Comparação de Modelos
Mean Squared Error
0
20
40
60
80
100
120
140
160
SGD
Linear Regression
Ridge
Lasso


Previsões vs Real (Regressão Linear)
Predições
Linha de Identidade
Valor Predito
Valor Real
20
30
40
50
60
70
80
90



Discussão

Limitações

Embora o modelo tenha apresentado bons resultados, algumas limitações foram identificadas:

- A qualidade dos dados pode ser melhorada, principalmente com mais registros sobre os influenciadores.
- A normalização foi necessária para garantir a convergência dos modelos, mas pode haver ajustes adicionais para otimizar os resultados.
- O ajuste de hiperparâmetros, como a taxa de aprendizado no Gradiente Descendente (SGD), pode ser refinado para melhorar a precisão.

Impacto das Escolhas

As escolhas de técnicas de regularização, como Ridge e Lasso, tiveram um impacto significativo, ajudando a evitar overfitting e melhorando a performance dos modelos em dados não vistos. A validação cruzada também foi fundamental para garantir a robustez dos resultados.

Conclusão e Trabalhos Futuros

Este projeto demonstrou como a Regressão Linear e suas variações (Ridge e Lasso) podem ser eficazes na previsão de taxas de engajamento de influenciadores no Instagram. A aplicação de técnicas de regularização ajudou a melhorar o desempenho dos modelos, e a validação cruzada garantiu a generalização dos resultados.

Trabalhos Futuros

- Experimentar com outros modelos de machine learning, como Árvores de Decisão e Redes Neurais.
- Explorar técnicas de feature engineering para incluir novas variáveis que possam melhorar a precisão do modelo.

Referências

1. **Scikit-learn Documentation**: https://scikit-learn.org/
2. **Pandas Documentation**: https://pandas.pydata.org/
3. **Numpy Documentation**: https://numpy.org/
