Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Grupo 37

Ana Fernanda Marinho Azevedo Alves

Reynaldo de Oliveira Santos

Data de Entrega: 15 de novembro de 2024

Resumo

Este relatório técnico apresenta a implementação de um modelo preditivo de Regressão Linear para prever a taxa de engajamento de influenciadores do Instagram. O projeto incluiu análise exploratória, pré-processamento dos dados, treinamento e avaliação de modelos, incluindo técnicas de regularização (Ridge e Lasso). Os resultados indicam que a Regressão Linear Simples foi a mais eficaz, com R^2 de 0.819 e MSE de 0.0000609, destacando a importância das variáveis selecionadas e o impacto de ajustes no desempenho do modelo.

Introdução

O objetivo deste projeto é desenvolver um modelo de Regressão Linear para prever a taxa de engajamento dos principais influenciadores do Instagram. Este modelo é útil para identificar as variáveis mais relevantes que impactam o engajamento e fornecer insights para o marketing digital.

Os dados foram obtidos do Kaggle e contêm informações sobre os principais influenciadores do Instagram. As principais variáveis incluem:

- Seguidores (followers): Número total de seguidores.
- Média de Curtidas (avg_likes): Curtidas médias por postagem.
- Taxa de Engajamento (engagement_rate): Variável dependente do modelo.

O modelo de regressão linear é uma das técnicas mais fundamentais e amplamente utilizadas em estatística e aprendizado de máquina. Sua simplicidade e poder de interpretação a tornam uma ferramenta indispensável para analisar e prever relações entre variáveis. É um dos modelos mais conhecidos e utilizados para previsão de demanda, que consiste de uma variável chamada de dependente estar relacionada a uma ou mais variáveis

independentes por uma equação linear. Pode-se dizer em uma linguagem técnica que a linha de regressão minimiza os desvios quadrados dos dados reais. Para obter o cálculo da equação da reta basta aplicar a seguinte equação: y = a + bx.

Metodologia

Análise Exploratória dos Dados

Uma inspeção inicial foi realizada para identificar inconsistências e relações. As variáveis numéricas foram convertidas para formato adequado, e observações nulas foram removidas. A matriz de correlação destacou forte relação entre curtidas médias e taxa de engajamento.

A análise exploratória revelou que a variável 'avg_likes' tem uma forte correlação positiva com a taxa de engajamento, enquanto 'posts' e 'influence_score' apresentaram relações ligeiramente negativas. Outras variáveis, como 'followers', tiveram correlações baixas com a taxa de engajamento.

A Regressão Linear foi implementada com o uso de métodos de otimização como o Gradiente Descendente e a Regressão Ridge. A Regressão Ridge mostrou-se mais estável, especialmente após a normalização dos dados e a aplicação de seleção de variáveis, destacando 'avg_likes', posts 'e 'influence_score' como as variáveis mais significativas.

Implementação do Algoritmo

Três modelos foram implementados:

1. **Regressão Linear Simples:** Sem regularização.
2. **Ridge (L2):** Regularização para reduzir multicolinearidade.
3. **Lasso (L1):** Regularização para seleção de variáveis.

Validação e Ajuste de Hiperparâmetros

Os hiperparâmetros dos modelos regularizados foram ajustados usando validação cruzada. O melhor valor de alpha para o modelo Lasso foi 0.001, com média de R^2 de 0.707 na validação cruzada.

A validação cruzada foi aplicada para garantir a generalização do modelo. Com uma divisão de 5 folds, o modelo foi avaliado em termos de MSE, MAE e R^2, mostrando que 'avg_likes' tem o impacto mais forte e positivo na taxa de engajamento.

Resultados

O projeto incluiu análise exploratória, pré-processamento dos dados, treinamento e avaliação de modelos, incluindo técnicas de regularização (Ridge e Lasso). Os resultados (Figura 1) indicam que a Regressão Linear Simples foi a mais eficaz, com R^2 de 0.819 e MSE de 0.0000609, destacando a importância das variáveis selecionadas e o impacto de ajustes no desempenho do modelo.

Comparação De Métricas De Modelo

	Model	R^2 Score	MSE	MAE
1	Linear Regression	0.81929469156727	6.094692191871935e-05	0.004666412431886042
2	Ridge	0.8177257286913113	6.147608987025198e-05	0.004699899875409043
3	Lasso	0.8000789613442972	6.742785830995252e-05	0.004886503527472697

Figura 1 (Fonte: Colab)

Diversos gráficos foram gerados para ilustrar o comportamento do modelo, incluindo gráficos de dispersão, linha de regressão, gráficos de resíduos e comparações entre previsões e valores reais (Figura 2). Esses gráficos ajudam a validar a precisão do modelo e identificar possíveis melhorias.

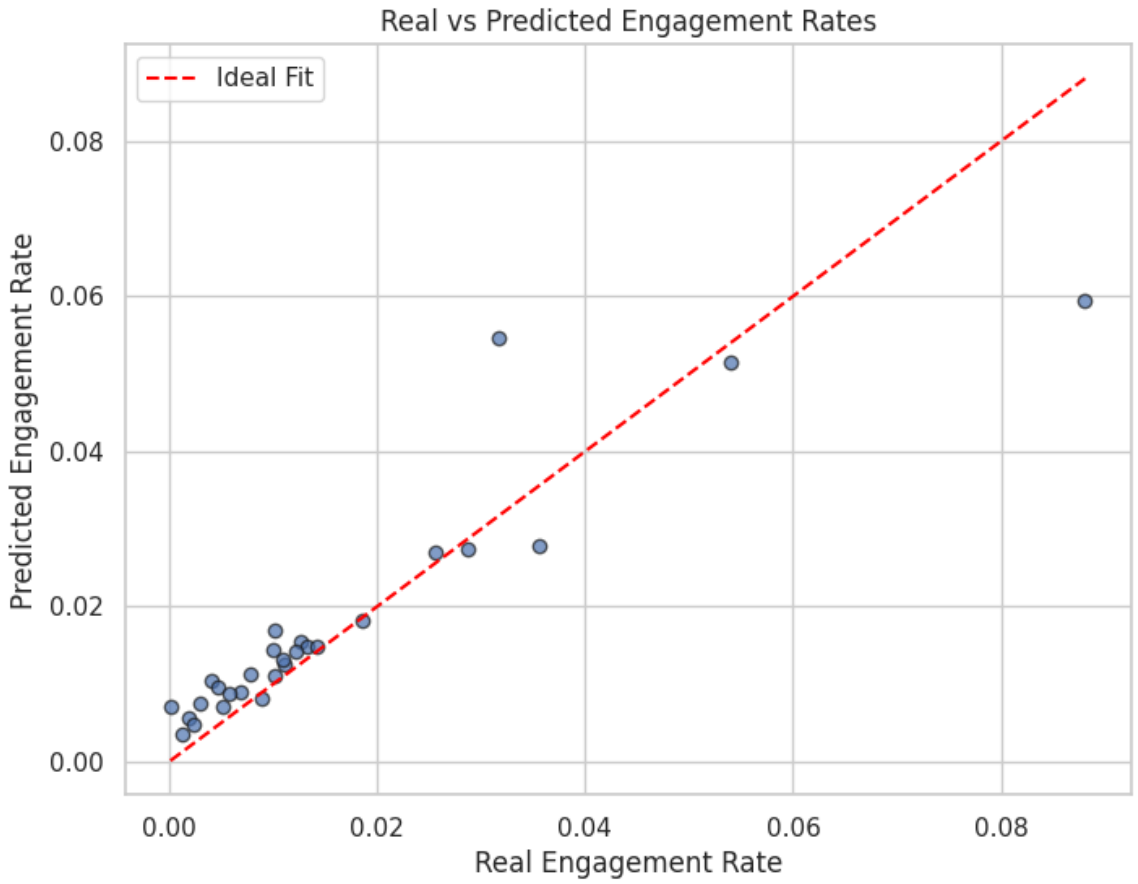

Figura 2 (Fonte: Colab)

O objetivo foi encontrar a linha que melhor se ajusta, minimizando a soma dos quadrados das diferenças entre os valores observados e previstos. A regressão linear é amplamente utilizada para previsões, análise de tendências e identificação do impacto de uma variável sobre outra. O modelo Linear Simples apresentou maior aderência à linha ideal, destacando seu bom desempenho.

Discussão

Os resultados mostram que o modelo captura bem a relação entre curtidas médias e taxa de engajamento. No entanto, as correlações fracas de outras variáveis sugerem que influências externas e características não capturadas nos dados podem afetar a taxa de engajamento.

Os resultados indicam que a Regressão Linear Simples é uma abordagem eficaz para prever a taxa de engajamento de influenciadores. Apesar de seu bom desempenho, as regularizações Ridge e Lasso não mostraram vantagens significativas neste conjunto de dados.

Limitações:

- O modelo assume relações lineares, o que pode não capturar padrões mais complexos.
- A normalização melhorou a convergência, mas não alterou significativamente as previsões.

Conclusão e Trabalhos Futuros

Este projeto demonstrou o uso de Regressão Linear para prever a taxa de engajamento de influenciadores do Instagram. A seleção de variáveis e a regularização ajudaram a melhorar a precisão do modelo.

O modelo de Regressão Linear Simples apresentou bom desempenho, com R^2 de 0.819 e MSE de 0.0000609.

Trabalhos Futuros

- Investigar técnicas não lineares, como Regressão Polinomial ou Redes Neurais.

- Expandir o conjunto de dados para incluir influenciadores de outras plataformas.
- Explorar mais variáveis e testar diferentes técnicas de modelagem para aprimorar a capacidade preditiva.

Referências

Grus, Joel. **Data Science do Zero – Primeiras Regras com o Python.** Traduzido por Wellington Nascimento. – Rio de Janeiro: Alta Books, 2016.

KAGGLE, 2024. Disponível em: < https://www.kaggle.com/ >. Acesso em: 05 nov. 2024.

Python | Regressão Linear usando sklearn. Disponível em < https://www.geeksforgeeks.org/python-linear-regression-using-sklearn/ > Acesso em: 06 nov. 2024.

Regressão Linear em Aprendizado de Máquina. Disponível em < https://www.geeksforgeeks.org/ml-linear-regression/ > Acesso em: 06 nov. 2024.