Centro de Pesquisa e Desenvolvimento Tecnológico em Informática e Eletrônica - CEPEDI

Mateus Lisboa

Solana Bonfim Lemos

Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Ilhéus - BA

2024

Resumo

O projeto tem como objetivo analisar dados de influenciadores do Instagram para identificar padrões e desenvolver modelos que expliquem ou prevejam a taxa de engajamento (60-day engagement rate). A metodologia incluiu o pré-processamento e normalização dos dados, análise exploratória para identificar relações entre variáveis, aplicação de modelos preditivos (como Regressão Linear e Ridge). Como principais resultados, a Regressão Linear alcançou um R^2 de 90,58%, demonstrando excelente capacidade preditiva, destacando a consistência e robustez das análises.

Sumário

1. **Introdução** **4**
2. **Metodologia** **4**
   - **2.1. Análise Exploratória** **5**
   - **2.2. Implementação do Algoritmo** **6**
     - **2.2.1. Regressão Linear Simples** **6**
     - **2.2.2. Gradiente Descendente** **7**
     - **2.2.3. Ridge Regression (Regularização L2)** **7**
     - **2.2.4. Lasso Regression (Regularização L1)** **8**
3. **Resultados** **9**
4. **Conclusão** **10**
5. **Referências** **11**

1. Introdução

A análise de influenciadores digitais tornou-se uma prática essencial no marketing digital, especialmente em plataformas como o Instagram, que desempenha um papel central na promoção de produtos e serviços. No entanto, medir a eficácia de um influenciador vai além do número de seguidores, sendo necessário compreender métricas mais relevantes, como a taxa de engajamento. Essa métrica reflete a interação do público com o conteúdo do influenciador e é um indicador chave para avaliar seu impacto real.

Neste contexto, a utilização de algoritmos de aprendizado supervisionado, como a Regressão Linear e o Ridge, se justifica pela sua capacidade de identificar relações entre variáveis e gerar modelos preditivos confiáveis. Esses algoritmos são especialmente úteis em situações onde o objetivo é prever uma variável-alvo com base em múltiplos fatores, como o engajamento de influenciadores considerando dados como seguidores, postagens, interações e outras métricas.

O conjunto de dados utilizado nesta análise contém informações detalhadas sobre influenciadores do Instagram, incluindo variáveis como o número de seguidores, a média de curtidas e comentários por postagem, a frequência de publicação e a taxa de engajamento observada nos últimos 60 dias. Essa riqueza de dados permite explorar relações complexas entre as variáveis, identificar padrões de comportamento e desenvolver modelos robustos para prever a eficácia de influenciadores em diferentes contextos.

A abordagem adotada contribui para a escolha informada de influenciadores em campanhas publicitárias, otimizando investimentos e ampliando o alcance das estratégias de marketing digital.

2. Metodologia

2.1 Análise Exploratória

A análise desenvolvida neste projeto teve como objetivo principal explorar o uso de diferentes modelos de regressão para prever o comportamento de uma variável dependente a partir de dados históricos. Para isso, foi realizado um extenso trabalho de preparação e tratamento inicial dos dados, a fim de garantir a confiabilidade dos resultados obtidos pelos modelos.

Durante o tratamento inicial, os dados foram convertidos para tipos numéricos, já que a fonte os forneceu como strings, isso foi necessário pois os modelos de Regressão Linear trabalham com dados numéricos . Além disso, foi realizado um diagnóstico de multicolinearidade utilizando o Variance Inflation Factor (VIF), que revelou uma forte correlação entre as variáveis independentes *new_post_avg_like* e *avg_likes*. Essa característica representa um desafio significativo para os modelos de regressão linear tradicionais, pois pode prejudicar a interpretação dos coeficientes e a estabilidade das previsões.

Como etapa complementar, os dados numéricos foram normalizados. Essa padronização foi essencial para melhorar a convergência de métodos como o Gradiente Descendente e evitar que variáveis em escalas diferentes dominassem os resultados, prejudicando o desempenho geral dos modelos. Posteriormente, o conjunto de dados foi dividido em 60% para treinamento e 40% para teste. Essa proporção foi cuidadosamente escolhida para garantir que os modelos fossem treinados em uma quantidade suficiente de dados, enquanto mantinham-se dados representativos para avaliar a capacidade de generalização.

Com esses passos iniciais, a análise pôde avançar para a implementação de diferentes modelos de regressão, buscando identificar qual método seria mais adequado às características do problema e do dataset.

2.2 Implementação do Algoritmo

A implementação do modelo de Regressão Linear no presente estudo foi dividida em várias abordagens, cada uma com um foco específico para lidar com desafios de multicolinearidade, eficiência computacional e a escolha do modelo mais adequado para a

previsão dos dados. A seguir, são detalhadas as configurações e os testes realizados para avaliar o desempenho de cada abordagem.

2.2.1 Regressão Linear Simples

A Regressão Linear Simples foi implementada como o modelo base, utilizando o método tradicional de **mínimos quadrados**. Esse método visa minimizar a soma dos quadrados dos resíduos (a diferença entre os valores observados e previstos), sendo o método mais comum para ajustar um modelo de regressão linear.

O modelo foi treinado diretamente com os dados após a preparação. O modelo de Regressão Linear simples foi ajustado aos dados com o auxílio da função LinearRegression() do Scikit-learn. As variáveis independentes foram normalizadas, e o modelo foi treinado sem a aplicação de regularização.

Resultados

R^2: 0.9058 – O coeficiente de determinação R^2 de 0.9058 indica que aproximadamente 90.58% da variância da variável dependente é explicada pelas variáveis independentes, o que é um bom indicador de que o modelo está capturando a maior parte da informação presente nos dados.

MSE (Erro Médio Quadrático): 0.000121 – O valor muito baixo de MSE sugere que o modelo é altamente preciso na previsão dos valores.

MAE (Erro Absoluto Médio): 0.00657 – Um MAE baixo também indica que a média dos erros absolutos nas previsões é pequena, reforçando a boa precisão do modelo.

2.2.2 Gradiente Descendente

O método de **Gradiente Descendente** foi utilizado como uma alternativa à solução analítica da Regressão Linear. Ao invés de calcular os coeficientes diretamente, o gradiente descendente ajusta iterativamente os coeficientes para minimizar a função de custo

(geralmente, o erro quadrático médio), o que o torna útil para grandes datasets ou problemas de alta dimensionalidade.

O algoritmo foi configurado com uma taxa de aprendizado (**learning rate**) inicial e ajustado por meio de validação cruzada para evitar problemas de divergência ou subajuste. A função de custo utilizada foi a soma dos erros quadráticos, que é minimizada durante o processo iterativo.

Resultados

R^2: 0.8990 – Embora ainda bom, o valor de R^2 foi ligeiramente inferior ao da Regressão Linear Simples, indicando que o gradiente descendente, embora eficaz, pode ser menos eficiente devido à sua natureza iterativa.

MSE: 0.000129 – O MSE foi um pouco maior em comparação ao modelo de mínimos quadrados, indicando que o modelo de gradiente descendente não alcançou o mesmo nível de precisão.

MAE: 0.00742 – O MAE também foi um pouco mais alto, refletindo uma precisão ligeiramente inferior.

2.2.3 Ridge Regression (Regularização L2)

A **Ridge Regression**, que aplica a regularização L2, foi usada para lidar com a multicolinearidade observada no modelo de Regressão Linear Simples. A regularização L2 penaliza os coeficientes de regressão, impedindo que eles se tornem excessivamente grandes, o que pode ocorrer em casos de alta correlação entre as variáveis.

A regularização foi aplicada com um parâmetro de penalização (λ\lambda) ajustado por validação cruzada. O modelo foi treinado da mesma forma que a Regressão Linear, mas com a adição de uma penalização aos coeficientes das variáveis.

Resultados

R^2: 0.9054 – O desempenho do modelo Ridge foi muito próximo ao da Regressão Linear Simples, mas com maior estabilidade devido à regularização.

MSE: 0.0001215 – O MSE foi praticamente igual ao da Regressão Linear Simples, indicando que a regularização não comprometeu a precisão do modelo.

MAE: 0.00659 – O MAE foi muito semelhante ao da Regressão Linear Simples, confirmando que a regularização não afetou negativamente a precisão do modelo.

2.2.4 Lasso Regression (Regularização L1)

A **Lasso Regression**, que utiliza a regularização L1, foi aplicada com o objetivo de penalizar os coeficientes e, potencialmente, levar alguns coeficientes a zero, permitindo assim uma seleção automática de variáveis. Essa abordagem é útil quando se busca um modelo mais esparso, com menos variáveis incluídas.

Similar ao Ridge, o Lasso também utilizou um parâmetro de penalização (λ\lambda) ajustado por validação cruzada. No entanto, devido ao seu efeito mais agressivo sobre os coeficientes, o Lasso pode eliminar variáveis importantes se o parâmetro de regularização for muito alto.

Resultados

R^2: -0.0304 – O modelo Lasso teve um desempenho muito ruim, com R^2 negativo, indicando que o modelo não conseguiu explicar adequadamente a variabilidade dos dados.

MSE: 0.001324 – O MSE foi significativamente mais alto, refletindo a falha do modelo em ajustar bem os dados.

MAE: 0.0185 – O MAE foi também muito maior, indicando uma grande imprecisão nas previsões.

3. Resultados

Os resultados obtidos destacaram o desempenho dos modelos testados, com especial atenção para a Ridge Regression, que se consolidou como a abordagem mais robusta e estável. Essa escolha foi fundamentada em sua capacidade de lidar com a multicolinearidade, controlando a magnitude dos coeficientes e garantindo previsões mais confiáveis. O modelo apresentou métricas próximas às da regressão linear simples, como R^2 elevado e erros reduzidos, mas demonstrou maior estabilidade em cenários de correlação entre variáveis independentes.

A regressão linear simples também apresentou excelente desempenho, com métricas de R^2 e MSE que indicaram boa explicação da variância e baixa dispersão dos erros. No entanto, a sensibilidade desse modelo à multicolinearidade reduz sua aplicabilidade prática em contextos com interdependências significativas entre variáveis, evidenciada pelo diagnóstico de VIF realizado na análise exploratória.

O gradiente descendente, embora eficaz, apresentou resultados ligeiramente inferiores aos métodos analíticos devido à natureza iterativa do ajuste de coeficientes. A taxa de aprendizado foi um parâmetro crítico, com impactos diretos na convergência do modelo, o que reforça a importância de uma otimização cuidadosa dos hiperparâmetros nesse tipo de abordagem.

Por outro lado, o Lasso Regression apresentou desempenho insatisfatório, com métricas significativamente piores em relação aos demais modelos. Esse resultado foi atribuído ao tamanho reduzido do conjunto de dados e ao efeito agressivo da regularização L1, que eliminou coeficientes essenciais para a explicação do modelo.

Assim, os resultados reafirmaram a eficácia da regularização como solução para desafios de multicolinearidade, com o modelo Ridge se destacando como a escolha mais equilibrada entre precisão e confiabilidade. O desempenho inferior do Lasso sugere que sua aplicação pode ser mais adequada em cenários com maior dimensionalidade de dados, enquanto o gradiente descendente e a regressão linear simples permanecem úteis em situações específicas, dependendo da natureza do problema.

4. Conclusão

A análise realizada neste projeto proporcionou aprendizados valiosos sobre a aplicação de modelos de regressão em contextos com multicolinearidade e conjuntos de dados limitados. Um dos principais pontos destacados foi o impacto significativo da multicolinearidade na confiabilidade dos modelos, especialmente na regressão linear simples. Esse problema foi mitigado com sucesso pela utilização de técnicas de regularização, como a Ridge Regression, que se mostrou o modelo mais robusto e estável, equilibrando precisão e controle da magnitude dos coeficientes. Além disso, a validação cruzada desempenhou um papel crucial ao garantir que os modelos apresentassem desempenho consistente, evitando o risco de overfitting.

Outro aprendizado importante foi a limitação do Lasso Regression no contexto deste projeto. Apesar de sua capacidade de selecionar variáveis relevantes, seu desempenho foi comprometido pelo tamanho reduzido do dataset e pela interação entre variáveis, ressaltando a necessidade de ajustes criteriosos em sua aplicação. O tratamento inicial dos dados também foi fundamental para o sucesso da análise, com a codificação adequada de variáveis categóricas e a análise do VIF ajudando a preparar os dados de maneira eficiente.

Para aprimorar futuros estudos, é recomendável expandir o conjunto de dados, explorar modelos mais avançados e sofisticados, e utilizar estratégias de otimização mais robustas, como ElasticNet e métodos de busca para hiperparâmetros. Além disso, a implementação de pipelines automatizados pode aumentar a eficiência do fluxo de trabalho, enquanto a aplicação de técnicas como PCA ou embeddings pode enriquecer a representação das variáveis. Por fim, a inclusão de visualizações interativas pode facilitar a interpretação e comunicação dos resultados, ampliando o impacto da análise em diferentes contextos. Esses avanços podem não apenas melhorar o desempenho técnico dos modelos, mas também aumentar sua aplicabilidade em cenários reais e mais complexos.

5. Referências