RELATÓRIO TÉCNICO: IMPLEMENTAÇÃO E ANÁLISE DO ALGORITMO DE REGRESSÃO LINEAR

MÁRCIO VINÍCIUS MOURA DE SOUSA

17 DE NOVEMBRO DE 2024

RESUMO:

Projeto realizado para a avaliação das unidades 9 e 10 da trilha “Ciência De Dados”, pela ResTic36. Uso esse documento para relatar alguns pontos importantes em relação ao desenvolvimento do projeto.

O principal objetivo foi, através do algoritmo de regressão linear e uma dataset contendo informações sobre os perfis dos influenciadores mais relevantes, fazer uma projeção para o desempenho dessas contas num período de 60 dias. Foquei em tópicos que seriam importantes para a análise de desempenho desse perfil, também filtrando os que eram muito similares.

Os resultados foram bastante interessantes, uma vez que os valores de RMSE e R^2.

INTRODUÇÃO:

Com o crescimento exponencial das redes sociais, os influenciadores digitais estão cada dia mais profissionalizando sua profissão e buscando cada vez mais resultados. Nesse cenário, entender e calcular o andamento de métricas-chave, como a taxa de engajamento, é necessário para marcas e profissionais de marketing ao esboçar estratégias.

A taxa de engajamento é um grande indicador de relevância e impacto, podendo variar amplamente entre influenciadores. Fatores como o número de seguidores, média de curtidas, e frequência de publicações influenciam diretamente no engajamento. Identificar essas relações por meio de análises estatísticas concretas ajuda a otimizar decisões de investimento e permite prever comportamentos futuros.

Para isso, utilizei o algoritmo de Regressão Linear, amplamente empregado para modelar e prever valores numéricos contínuos, devido à sua simplicidade e interpretabilidade.

O conjunto de dados fornecidos pelo dataset que serão usados são: ‘Rank’, ‘Influence Score’, ‘Posts’, ‘Followers’, ‘Avg Likes’, ‘Total Likes’, e ‘New Post Avg Likes’.

METODOLOGIA:

Os dados foram utilizados e manuseados de forma a pensar em quais fatores são mais importantes para serem exibidos, além de tratar de dados que antes eram apresentados seguidos por letras ou símbolos que representam suas medidas. Nesse primeiro caso, não considerei algumas colunas já do dataset, a equivalente a “country” não foi utilizada pois 31% dos influenciadores presentes não haviam um país relacionado, além de alguns que até tinham países atrelados a si, mas estavam errados; outras colunas que desconsiderei ao final, antes da aplicação da Regressão Linear foram: “rank”, “avg likes” e “total likes”, por, ao exibir a matriz de correlação dos dados e realizar medições com o VIF, elas apresentam alta correlação entre si e pouca com o target; além da “Channel Info” que também não era importante para o objetivo principal.

Em relação ao algoritmo, usei, dos dados, 80% (0.8) deles para treinamento e 20% (0.2) para teste e a principal métrica para o desempenho foi o RMSE (Raiz do Erro Quadrático Médio) e R^2 (Coeficiente de Relação) para medir a proporção da variância explicada pelo modelo. Foi aconselhado a utilização do PCA para um melhor desempenho, porém, ao utilizar, o resultado foi o inverso do esperado, uma vez que, sem ele, o valor R^2 estava em 0.94 (quanto mais próximo de 1, melhor) e o RMSE estava em 0.54 (quanto menor, melhor), enquanto com o uso da técnica, os valores foram para 0.33 e 2.04 respectivamente, por isso, o deixei de lado.

Com esses fatores e a boa aplicação da técnica de Regressão Linear consegui um bom resultado para uma relação de Realidade x Previsão e um resultado decente para a amostragem de Resíduos, porém, ainda implementei um modelo de regularização através da técnica de Ridge e Lasso, a primeira ajudando a reduzir o impacto de colinearidade sem zerar coeficientes, garantindo que todas as variáveis contribuam de alguma forma para a previsão enquanto a segunda ajudou oferecendo maior interpretabilidade ao modelo, reduzindo automaticamente o número de variáveis relevantes; com isso obtive um melhor resultado em relação a uma já antes boa previsão Realidade x Previsão e uma boa melhoria em relação a representação da Distribuição dos Resíduos.

Em relação aos hiperparâmetros utilizados, de forma geral, as variáveis selecionadas não seriam operadas de forma direta, a única mudança que elas poderiam sofrer até o final da execução seria a conversão de um valor com uma

string como unidade de medida para sua medida de forma completamente numérica. A etapa e processo de Validação Cruzada foi muito importante para avaliar a robustez do modelo em diferentes divisões do dataset. A utilização da divisão em 5 partes iguais dos dados, onde ⅘ (80%) foram usados para treinamento, enquanto o outro ⅕ (20%) usado para teste foi de suma importância para que o risco de Overlifting fosse reduzido, garantindo que o modelo não dependa excessivamente de uma única divisão de treino/teste, e para que seja possível a implementação do cálculo do RMSE. Como não possui hiperparâmetros ajustáveis. O objetivo foi apenas avaliar o desempenho inicial do modelo sem regularização, fiz isso através dos hiperparâmetros alpha utilizado em Ridge e Lasso, onde na primeira, ele é responsável por controlar a penalização, foi ajustado para 1.0 após testes com valores menores e maiores, observando o equilíbrio entre bias e variância; já em Lasso foi definido como 0.1, pois valores maiores tendiam a eliminar variáveis importantes, reduzindo o desempenho preditivo.

RESULTADO:

As métricas de Avaliação foram adquiridas através do desempenho dos modelos treinados (Regressão Linear Simples, Ridge e Lasso). Em relação a métodos, usei RMSE para penalizar de forma mais consistente os erros, identificar diferenças significativas e retornar o erro médio da comparação. Também utilizei o coeficiente de relação R^2 para medir as variâncias da variável dependente.

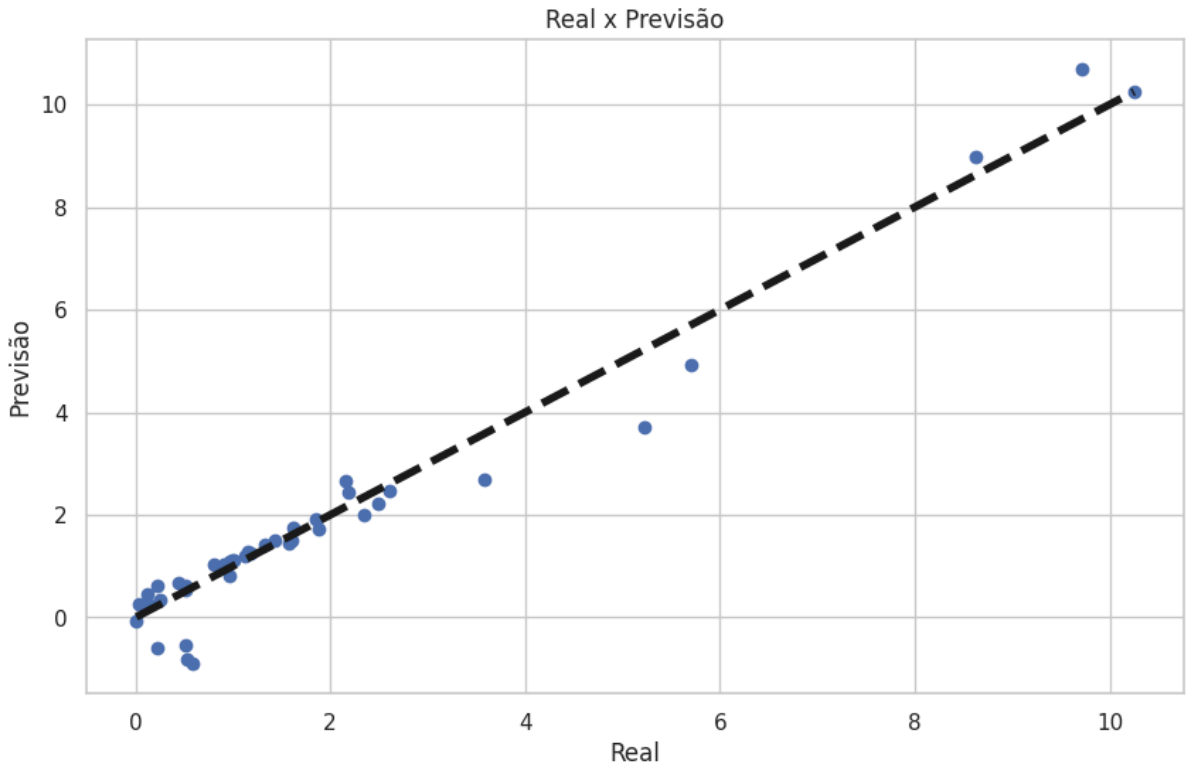
Em relação aos gráficos, tem-se para a visualização o Real x Previsão e a distribuição de resíduos para analisar erros no modelo.

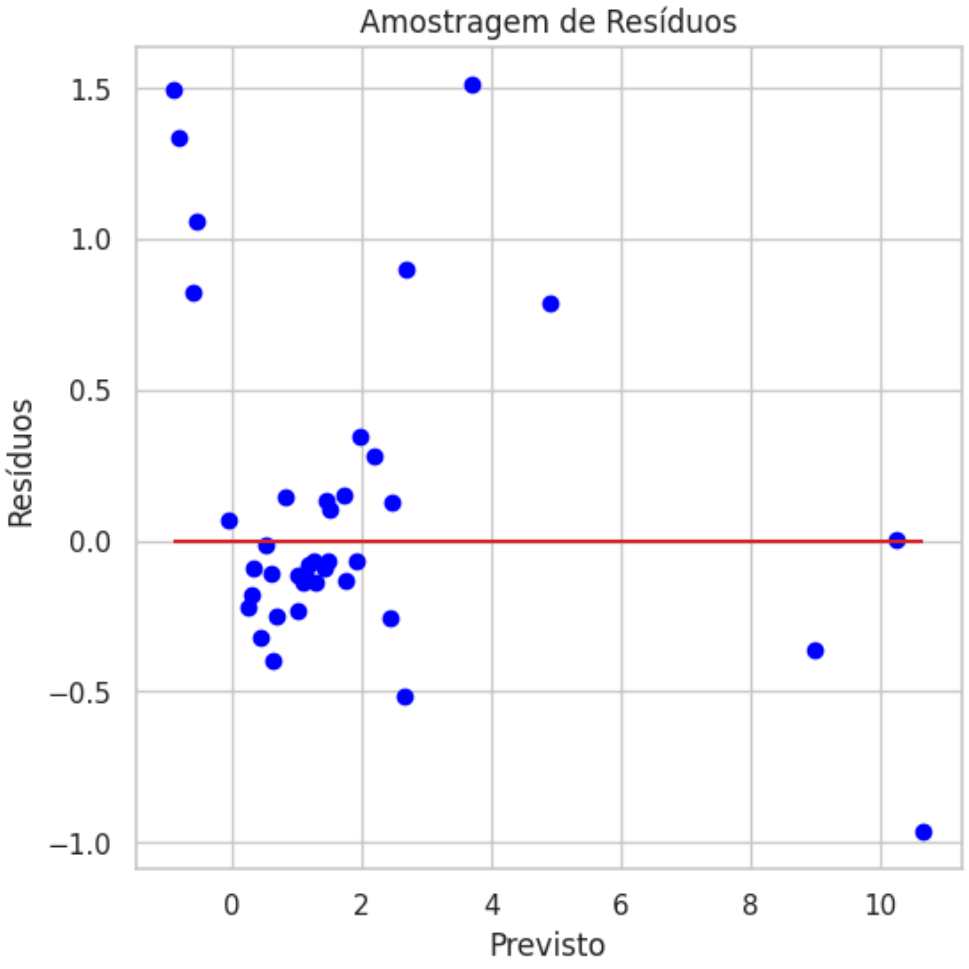
Sobre as métricas, posso dizer:

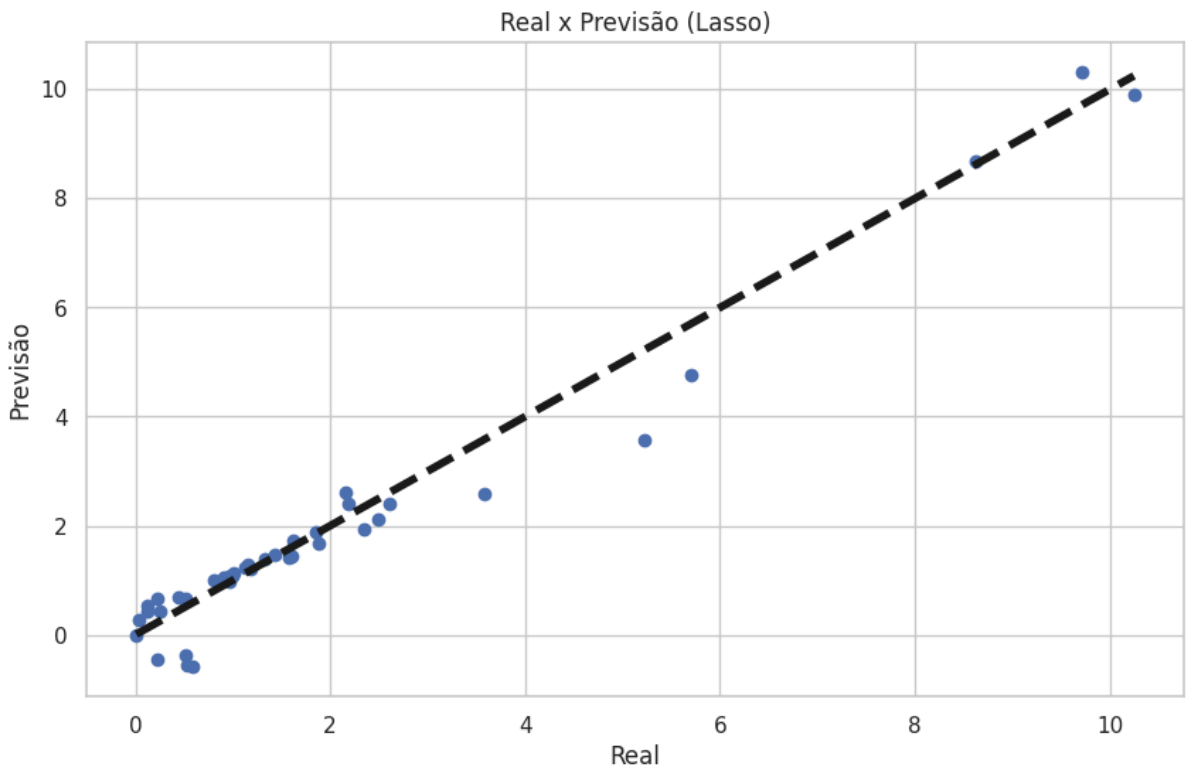
Regressão Linear Simples: Melhor para prever com alta precisão quando não há preocupação com generalização ou multicolinearidade.

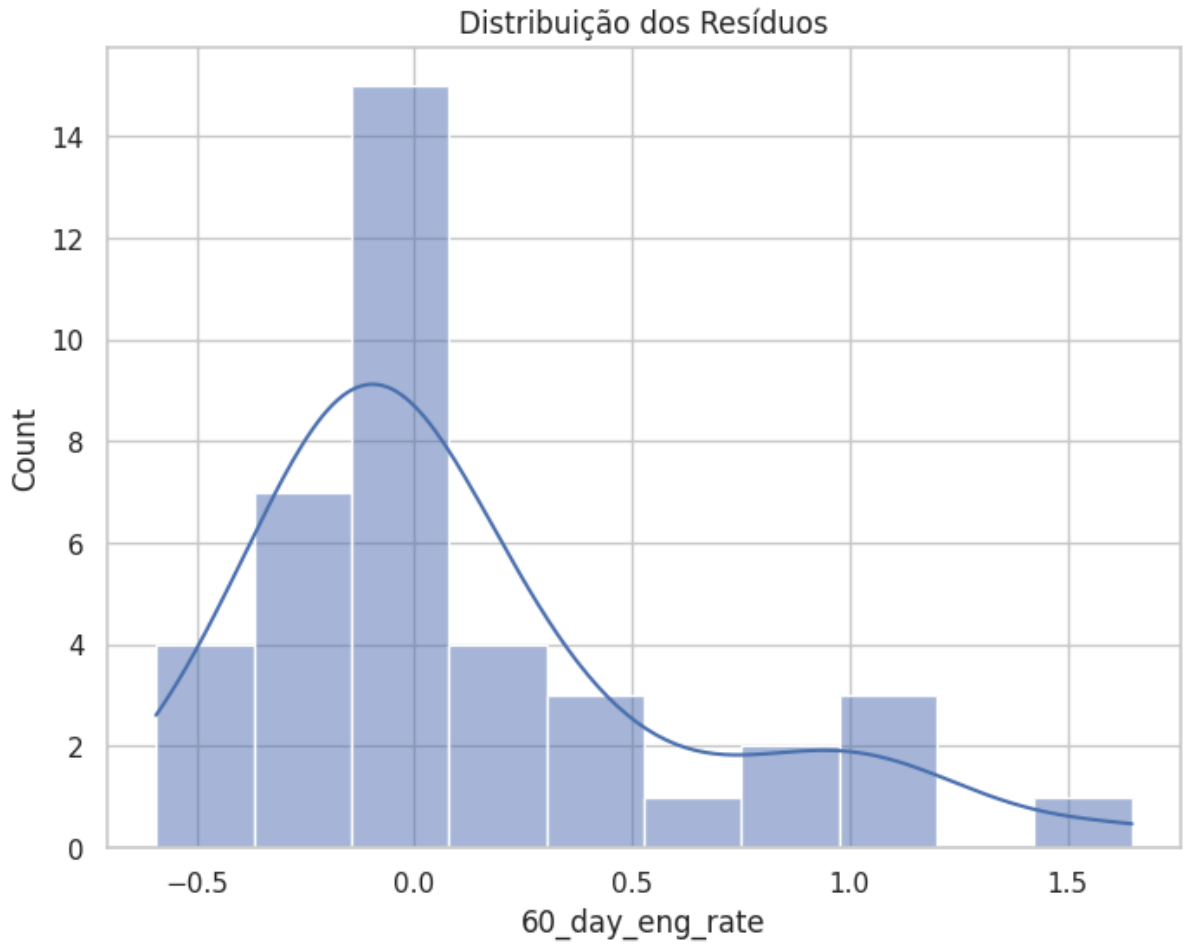
Ridge: Ideal para problemas com múltiplas variáveis correlacionadas, oferecendo robustez adicional.

Lasso: Útil para simplificar o modelo, identificando e descartando variáveis de menor impacto.


Real x Previsão
Previsão
Real
0
2
4
6
8
10


Amostragem de Resíduos
Resíduos
Previsto
1.5
1.0
0.5
0.0
−0.5
−1.0
0
2
4
6
8
10


Real x Previsão (Lasso)
Previsão
Real
0
2
4
6
8
10


Distribuição dos Resíduos
Count
60_day_eng_rate
0
2
4
6
8
10
12
14
−0.5
0.0
0.5
1.0
1.5

DISCUSSÃO:

Com base nos resultados obtidos e os gráficos apresentados, consegui obter uma boa base de como é/será o desempenho desses influenciadores, onde que até superando expectativas, consegui obter uma boa predição dos dados, fazendo o gráfico de Real x Previsão ser muito próximo da linha ideal, já em relação a distribuição de resíduos, tive uma taxa normal, embora ainda apresentasse pequenos desvios.

Tive alguns obstáculos para a construção de um projeto sólido, podendo listar algumas, como o fato do dataset possuir valores nulos para alguns dados, como foi o caso de “country”, outra limitação encontrada foi o fato de que variáveis como “followers”, “avg_likes” e “total_likes” apresentam multicolinearidade que afeta a regressão linear simples, por isso, usei Rigde e Lasso para mitigar esse problema. É válido citar também o fato de que a análise focou em métricas gerais que medem desempenho médio. Não foram utilizadas métricas específicas para outliers ou casos extremos, que podem ter impacto significativo na análise.

Pode-se também falar sobre o impacto de minhas escolhas no desempenho do algoritmo, a seleção inicial das variáveis independentes foi baseada em sua correlação intuitiva e visual com a variável dependente, isso garantiu uma boa precisão, muito por conta também da não utilização de variáveis muito próximas na matriz de correlação, das que não forneciam valores importantes ou que os valores estejam equivocados. O uso da regularização feitas através de Ridge e Lasso foi uma decisão muito assertiva, pois melhorou a robustez do modelo, tornando-o mais adequado para generalização, mesmo com leve perda de R^2, equilibrando desempenho e interpretabilidade, especialmente no caso do Lasso, que eliminou variáveis menos significativas. Outro ponto importante é a divisão do conjunto de dados em 80% para treino e 20% para o teste, garantindo uma avaliação robusta do desempenho do modelo, reduzindo o impacto de variações nos dados.

Conclusão e Trabalhos Futuros:

Foi muito interessante poder aplicar vários fatores da minha jornada na ResTic em um projeto sólido, funcional e bem construído, pude pesquisar mais a fundo fatores que talvez não tivessem sido captados de primeira, além da aprendizagem de novas técnicas e implementações, Regressão Linear foi uma ótima escolha e conseguir ver o projeto rodando foi muito gratificante.

Contudo, sempre há espaço para melhorias, tais como a investigação de relações não lineares, tratar variáveis categóricas ou aplicar técnicas avançadas de seleção de variáveis e explorar outros algoritmos que possam captar padrões complexos nos dados.

REFERÊNCIAS:

PAUL, S. Essentials of Linear Regression in Python. Disponível em: <https://www.datacamp.com/tutorial/essentials-linear-regression-python>. Acesso em: 17 nov. 2024.

WAPLES, J. Simple Linear Regression: Everything You Need to Know. Disponível em: <https://www.datacamp.com/tutorial/simple-linear-regression>.

LinearRegression. Disponível em: <https://scikit-learn.org/1.5/modules/generated/sklearn.linear_model.LinearRegression.html>.