Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Ícaro José Batista de Oliveira
Henrique Leonardo dos Santos Carvalho

Implementação
de um modelo de regressão
para o desafio de Ciência de dados

17/11/2024

1 INTRODUÇÃO

Inferir ou prever determinado dado ou ação - na maioria dos casos - não é uma tarefa trivial e de fácil execução. Apesar disso, com a vanços nos campos da Ciência de Dados e inteligência artificial, modelos preditivos, isto é, de regressão, também surgiram e evoluiram, de maneira que é possível hoje, instanciar - através de bibliotecas - um sistema que, com os devidos cuidados, é capaz de predizer, para uma dada entrada, uma determinada saída.

O projeto atual, vem então, fazer uso desses modelos para inferir a taxa de engajamento de influencidadores do Instagram. Sendo assim, será feito uso de variáveis independentes como "quantidades de post" ou "média de likes para um post novo" para tentar inferir o valor da taxa de engajamento.

O Dataset é composto de 200 entradas (isso é, 200 linhas) e apresenta 10 colunas, sendo elas mostradas na Tabela 1:

Table 1: Descrição dos Atributos dos Influenciadores

Atributo	**Explicação**
rank	Rank do influenciador com base no número de seguidores que ele possui.
channel_info	Nome de usuário do Instagram do influenciador.
influence_score	Pontuação de influência dos usuários, calculada com base em menções, importância e popularidade.
posts	Número de postagens feitas pelo influenciador até agora.
followers	Número de seguidores do influenciador.
avg_likes	Média de curtidas nas postagens do influenciador (curtidas totais / postagens totais).
60_day_eng_rate	Taxa de engajamento nos últimos 60 dias do influenciador, como fração do total de engajamentos realizados.
new_post_avg_like	Média de curtidas nas novas postagens do influenciador.
total_likes	Total de curtidas que o influenciador recebeu em suas postagens (em bilhões).
country	País ou região de origem do influenciador.

2 METODOLOGIA

2.1 Análise Exploratória

Para análise exploratória começou-se primeiro, entendendo como o Dataset estava estruturado, isso é: qual o tipo de dados de cada coluna, quantidade total de linhas e correlatos. Foi nessa etápa que viu-se a necessidade do primeiro ajuste nos dados, visto que para colunas númericas, os valores como mil (1000), milhão (1000000) e bilhão (1000000000) estavam representados por "k", "m" e "b" respectivamente, sendo necessário então, transformar esses valores. A variável Target (taxa de engajamento), como descrita pelo nome é uma taxa, que estava expressa com o simbolo de porcentagem ("1 por cento") sendo necessário transformar (0.01) para aplicar no modelo de regressão.

Depois das transformações de tipo acima, foram usados comando como o "dataframe.info()" para visualizar a presença ou não de valores nulos, e nesse momento foi descoberto, que a coluna "country" tinha muitos valores nulos, optando-se assim por exclui-la, com o entendimento que ela influenciaria pouco na regressão. Outras colunas como o "username" e "rank" também foram excluidas pelo mesmo motivo.

Após isso, analisou-se a distribuição com o "dataframe.describe()", e foi possível ver que existe uma diferença de escala muito grande entre as médias dos diferentes tipos de atributos, indicando uma necessidade de padronização. Além disso analisando o relatório de cada coluna, em alguns casos foi possível analisar como os valores se comportavam por quartis e quão distante estavam da média por exemplo, alguns "acenderam" a luz para possíveis outlieres(alguns valores máximos distantes da média e do quartil 75 por centro), mas decidiu-se manter esses dados visto que são valores reais de influencers, sendo assim possíveis de ocorrer.

Plotou-se após isso, gráficos de dispersão e densidade (confirmando presença de alguns outliers) e, para os gráficos de se dispersão passou-se uma linha de regressão que modela o comportamento das colunas independentes com o dependente (Target), nessa etápa ja foi possível ver possíveis correlções através das linhas.

Por fim, a matriz de correlação confirmou que 2 variáveis ("avg-like" e "new-post-av-like") estavam representando a mesma coisa, sendo assim excluiu-se a coluna "avg-like". Também notou-se que algumas colunas apresentavam baixa relação com o Target, sendo ela a "influence-score", que apresentava quase 0 de correlação com a taxa de engajamento.

2.2 Implementação do Algoritmo

O Algoritmo implementado foi um de Regressão Linear Múltipla (devido a grande quantidade de variáveis independentes), que usa do Gradiente Descendente Estocástico (SGD) como método de otimização para melhor encontro dos coeficientes. Esse método funciona tentando minimizar a função de custo - erro quadrático médio. Além disso, o modelo usa como penalidade o Elastic Net, que combina L1 (Lasso) e L2(Ridge) com base num fator que varia de 0 a 1, nesse projeto, foram feito alguns testes, e o que melhor se adaptou ao modelo foi o Elastic Net priorizando L1 e L2 de maneira igual, ou seja 0.5.

Por fim, os dados foram padronizados (devido a diferença em magnitude) e foi estabelecido um número máximo de iterações de 1000, taxa de aprendizado de 0.01 (a melhor dentre as testadas, não ocasionando em subajuste ou sobreajuste) e foi aplicada uma validação cruzada com 5 folds com uma avaliação baseada no erro médio quadrático (MSE).

2.3 Validação e Ajuste de Hiperparâmetros

Sobre a escolha das variáveis independentes, começou-se, já retirando as colunas do nome de usuário, país (até porque muitos valores eram nulos, e até com um método de povoamento por média, poderia mais atrapalhar que ajudar) e o rank. Essas 3 colunas retiradas, foram vistas com pouco ou nenhuma contribuição possível para o sistema, e foi, assim, retirada.

Outras colunas, como o a "média de likes" ou "influence-score" também foram retiradas.A primeira, o motivo foi a possível redudância que essa coluna e outra "new-post-avg-like" apresentavam, conforme visto na matriz de correlação, elas apresentam correlação entre si de 0.9, e no fim escolheu-se por retirá-la.

Já o "influence-score", analisando tanto os gráficos de dispersão quanto a matriz de correlação, foi possível visualizar, também, que relação a variável Target, a contribuição dessa coluna era quase nula (-0.082), e que não supria o modelo com informações interessantes. Decidiu-se então, pela retirada desta.

Em relação ao parâmetros, não foi usado nenhum método de busca sistemática como o Grid Search ou Random Search, os parâmetros foram sendo testados na mão, com valores árbitrários, e os finais do código, foram os melhores. A escolha se deu, para maior aproximação e entendimento de como o modelo funcionava, não sendo, a aplicação dos métodos de busca por parâmetros ótimos desaconselhada, mas pelo contrário, visto que a aleatoridade e as verificações feitas por esses métodos são, na maioria dos casos, superiores em relação a humana.

A validação cruzada, foi aplicada, como já dito em 5 folds, ela vem para aumentara a confiabilidade do modelo e evitar que overfiting ou underfiting ocorram, foi feita de maneira simples mas eficaz, utilizando o método cross-val-score, que divide o conjunto em diversos conjuntos menores e aplica no modelo, vendo como ele se comporta com dados novos.

3 RESULTADOS

As métricas de avaliação utilizadas foram:

- **Root Mean Squared Error (RMSE):** Mede a diferença entre os valores previstos e os valores reais, sendo amplamente usada em problemas de regressão. No modelo, o RMSE foi de 0.0113 na validação cruzada e 0.0062 no conjunto de teste, ambos valores baixos, indicando que o modelo teve sucesso tanto no treinamento quanto na generalização dos dados. A diferença relativamente pequena entre os valores de validação e teste reforça a capacidade do modelo de generalizar.
- **Mean Absolute Error (MAE):** Avalia o erro médio absoluto entre as previsões e os valores reais. O MAE foi de 0.0041, corroborando que o modelo realiza previsões precisas com baixo erro médio.
- **Coeficiente de Determinação ($\mathbf{R^2}$):** Mede o quanto da variabilidade nos dados o modelo consegue explicar. O valor obtido foi $R^2 = 0.9383$, indicando que o modelo é capaz de explicar 93,83% da variabilidade nos dados, demonstrando um bom ajuste.

O gráfico a seguir relaciona os valores de Y reais com os preditos pelo modelo, mostrando que ele conseguiu se adaptar bem aos dados. A distância para os valores reais é pequena e está dentro da margem, como mostrada pelo MAE e o RMSE:

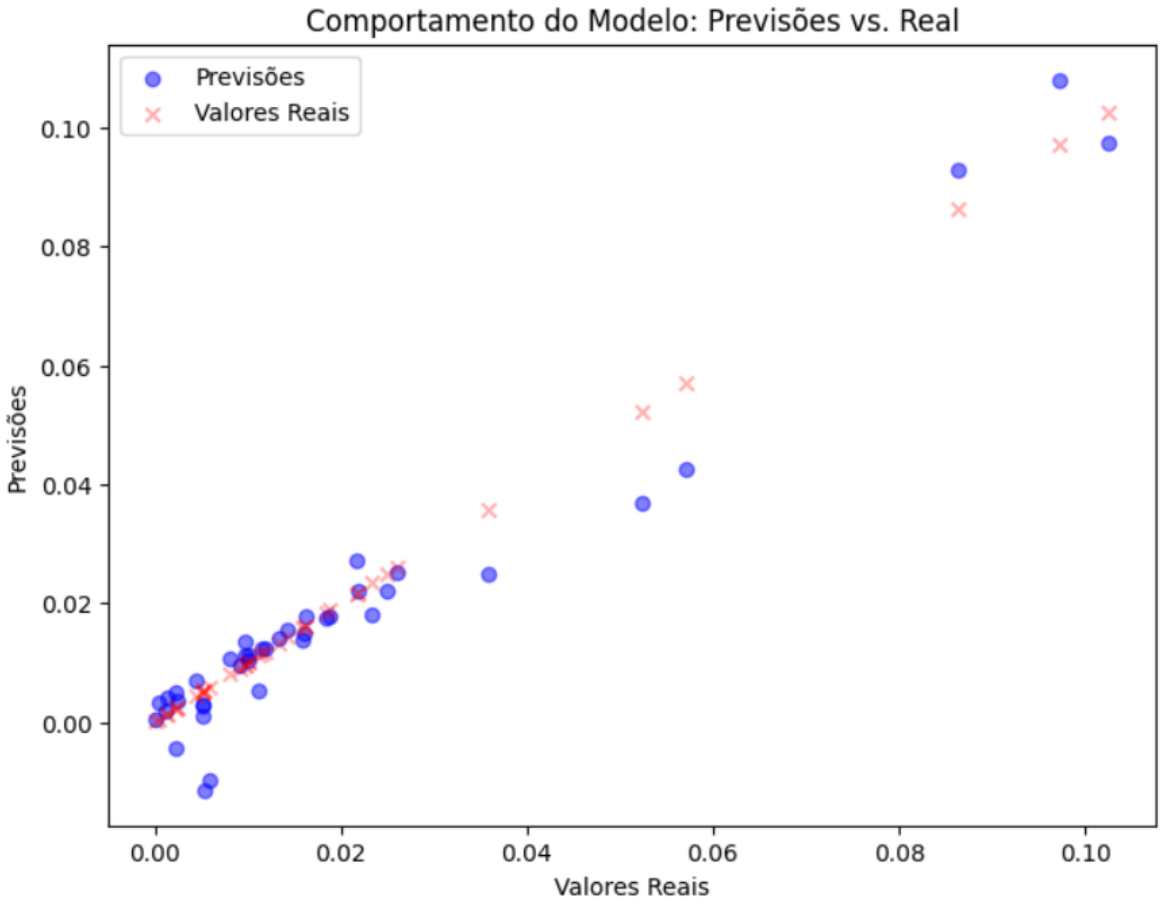

Figure 1: Pontos previstos e reais.

4 DISCUSSÃO

De melhoria, seria entender por que 3 dos valores previstos ficaram abaixo de 0, e tentar aumentar ainda mias a proximidade entre os valores previstos e os reais, para isso aplicar métodos sistemáticos de tunning seriam ideias, tanto para selecionar corretamente quais colunas devem ser selecionados e quais devem ser "excluidas". Para além disso, o tunning dos parâmetros também seria uma boa como já dito anteriormente.

Analisar mais profundamente como os Outliers estão influenciando o aprendizado do modelo também seria interessante, aplicando algum tipo de limpeza de outliers com quartis ou técnica.

5 CONCLUSÃO

Foi possível entender desde a manipulação dos dados, com a análise exploratória (EDA), entendendo como os dados estão distribuidos, quais variáveis influenciam mais ou menos na variável alvo, e se existem valores que podem atrapalhar ou não o modelo. Além disso, foi possível entender mais sobre regularização e como é possível evitar o overfitting utilizando técnicas como o L1 e L2. Por fim, aplicar validação também foi uma mais uma ótima surpresa, pois com a validação é possível ir modificando partes do código de modo a buscar por um desempenho ótimo.

6 Referências

1. Towards Data Science. Step-by-step tutorial on linear regression with stochastic gradient descent. Disponível em: https://towardsdatascience.com/step-by-step-tutorial-on-linear-regression-with-stochast Acesso em: 17 nov. 2024.

2. Scikit-learn. Stochastic Gradient Descent. Disponível em: https://scikit-learn.org/1.5/modules/sgd.htm. Acesso em: 17 nov. 2024.

3. Medium. Turing Talks #20: Regressão de Ridge e Lasso. Disponível em: https://medium.com/turing-talks/turing-talks-20-regress%C3%A3o-de-ridge-e-lasso-a0fc467b5629. Acesso em: 17 nov. 2024.

4. Iá com Café. A importância da validação cruzada em machine learning. Disponível em: https://iacomcafe.com.br/importancia-validacao-cruzada-machine-learning/. Acesso em: 17 nov. 2024.