Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Autores: Aurélio José Ribeiro da Silva, Andressa Luíza Costa de Carvalho

SUMÁRIO

RESUMO

Este relatório detalha a implementação de um algoritmo de Regressão Linear para analisar a taxa de engajamento de influenciadores no Instagram. O objetivo foi realizar predições com base em variáveis como o número de seguidores, engajamento e interações (curtidas e comentários). A metodologia envolveu análise exploratória dos dados, construção e validação do modelo em *Python* com a biblioteca *Scikit-learn*, e visualização dos resultados através de gráficos de dispersão. Os resultados indicam uma correlação significativa entre o número de seguidores, *likes* e o alcance com métricas de desempenho (R^2, MAE e MSE) mostrando um bom ajuste. No entanto, os modelos de regularização apresentam limitações. Conclui-se que a Regressão Linear é útil para captar tendências gerais, pois foi capaz de explicar uma boa parte da variabilidade dos dados.

Palavras-chave: Regressão Linear, Instagram, Análise de Dados.

INTRODUÇÃO

O *Instagram*, uma das plataformas de mídia social mais populares mundialmente, tem sido amplamente utilizado para influenciar decisões de consumo e comportamento social, especialmente por influenciadores digitais. Esses usuários, com grandes números de seguidores, possuem um impacto significativo nas tendências de engajamento, como curtidas, comentários e compartilhamentos. Com o aumento do interesse por marketing digital e análise de redes sociais, surge a necessidade de entender melhor os fatores que determinam o sucesso de uma postagem, especialmente em relação à sua taxa de engajamento.

Neste contexto, o objetivo deste projeto é aplicar o algoritmo de Regressão Linear para modelar e prever o alcance das postagens no Instagram, com base em variáveis como o número de seguidores e interações. A Regressão Linear foi escolhida devido à sua simplicidade e eficácia em problemas de inferência, onde se busca entender a relação entre variáveis independentes e dependentes. O modelo permitirá identificar quais fatores influenciam o engajamento e como essas variáveis interagem para afetar o desempenho das postagens.

O conjunto de dados utilizado neste estudo contém informações sobre postagens de influenciadores, incluindo o número de seguidores, curtidas, comentários e o alcance das postagens. Esses dados fornecem uma visão ampla das interações típicas dentro da plataforma e são essenciais para a construção de um modelo preditivo que possa ser utilizado para melhorar as estratégias de engajamento. O projeto aborda desde a análise exploratória até a validação do modelo, oferecendo insights sobre o comportamento dos usuários do Instagram e o impacto de suas interações nas postagens.

Conjunto de Dados

Os dados foram retirados do link passado na atividade moodle:
Link dos dados

O conjunto de dados utilizado na análise é composto por informações de contas influentes em redes sociais como número de seguidores, posts, curtidas, comentários, entre outros indicadores de engajamento, apresentando uma variedade de métricas que refletem o alcance e o engajamento dessas contas. A estrutura dos dados inclui as seguintes colunas principais:

- *rank*: Classificação da conta com base na sua influência.
- *channel_info*: Nome do influenciador ou canal.
- *influence_score*: Pontuação atribuída que reflete a relevância da conta.
- *posts*: Número total de postagens feitas pelo influenciador.
- *followers*: Quantidade total de seguidores da conta.
- *avg_likes*: Número médio de curtidas por postagem.

- *60_day_eng_rate*: Taxa de engajamento dos últimos 60 dias, em porcentagem.
- *new_post_avg_like*: Média de curtidas em postagens recentes.
- *total_likes*: Total de curtidas acumuladas em todas as postagens.
- *country*: País associado à conta.

Este conjunto de dados é rico em variáveis que permitem uma análise detalhada do impacto das contas em diferentes contextos geográficos e com diferentes padrões de engajamento. Por exemplo, a conta de Cristiano é classificada em primeiro lugar, com mais de 475,8 milhões de seguidores e uma média de 8,7 milhões de curtidas por postagem, destacando-se por um alto nível de influência global. Em contraste, outros influenciadores, como *TheRock*, mostram um menor engajamento médio de curtidas apesar de terem uma base significativa de seguidores. A diversidade de métricas possibilita uma avaliação profunda da correlação entre o número de seguidores, o engajamento e outros fatores relevantes para a modelagem preditiva.

METODOLOGIA

Análise Exploratória

Foi realizada uma análise inicial dos dados para entender as distribuições, identificar possíveis valores ausentes e outliers, e visualizar relações entre variáveis.

Descrição do conjunto de dados

Utilizando o método *info()* do pandas temos as seguintes informações acerca dos dados armazenados no *DataFrame*:

1. Número total de entradas: São 200 entradas que vão do índice 0 a 199, indicando 200 registros únicos.

2. Estrutura das colunas:

    i. *rank*: Tipo *int64*. Possui 200 entradas não-nulas, sugerindo que cada registro tem uma posição ou classificação associada.

    ii. *channel_info*: Tipo *object*. Contém 200 entradas não nulas, indicando informações relacionadas aos canais (possivelmente texto ou identificadores).

    iii. *influence_score*: Tipo *int64*. Com 200 entradas, esta coluna mede a pontuação de influência de cada registro.

    iv. *posts:* Tipo *object*. Esta coluna possui 200 entradas não nulas, possivelmente representando dados textuais ou numéricos sobre postagens.

    v. *followers*: Tipo *object*. Inclui 200 entradas não nulas, provavelmente representando números de seguidores, mas armazenados como *strings* ou texto.

    vi. *avg_likes*: Tipo *object*. Contém o número médio de curtidas como texto, em 200 registros.

    vii. *60_day_eng_rate*: Tipo *object*. Inclui a taxa de engajamento dos últimos 60 dias em formato textual em 200 entradas.

    viii. *new_post_avg_like*: Tipo *object*. Armazena o número médio de curtidas em novas postagens como texto em 200 registros.

    ix. *total_likes*: Tipo *object*. Contém 200 entradas representando o número total de curtidas, armazenadas como texto.

    x. *country*: Tipo *object*. Esta coluna contém 138 entradas não nulas, indicando que dados sobre o país estão ausentes para 62 registros.

3. Tipos de dados: O *DataFrame* possui predominantemente colunas do tipo *object* (8 colunas) o que indica a presença de dados textuais ou numéricos armazenados como texto, além de 2 colunas com dados numéricos inteiros *int64*.

4. Uso de memória: A estrutura usa aproximadamente 15.8K de memória.

Resumo estatístico

1. Rank

- Descrição: Classificação ou posição.
- Análise: Os valores variam de 1 a 200, indicando o intervalo de classificação no conjunto de dados.
- Média: 96,4% sugere que, em média, as entidades estão no meio da classificação.
- Mediana (50%): 93, indicando que metade das entidades estão classificadas abaixo de 93.
- A distribuição de classificações é relativamente ampla (std = 59,6).

2. Influence Score

- Descrição: Uma pontuação que estabelece o nível de influência.
- Análise: Os valores variam de 41 a 93.
- Média: 81,7, mostrando uma tendência para pontuações mais altas no conjunto.
- A diferença entre a média, mediana, e a distribuição estreita (std = 8,7) sugere uma distribuição relativamente uniforme e concentrada.

3. Posts

- Descrição: Número de postagens realizadas.
- Análise: Variando de 20 a 17.500, o que mostra uma ampla variação na quantidade de postagens.
- Média e Mediana são notavelmente diferentes (4.032 vs. 2.900), sugerindo uma assimetria positiva, onde alguns poucos indivíduos postam muito.

4. Followers

- Descrição: Número de seguidores.
- Análise: Intervalo muito grande (32.800 a 475.800.000).
- A média de 83,7 milhões é significativamente influenciada por valores extremos, como sugerido pelo elevado desvio padrão (81 milhões).
- A mediana de apenas 52,7 milhões reforça a presença de alguns valores excepcionalmente grandes.

5. Avg Likes

- Descrição: Média de curtidas por postagem.
- Análise: Os valores variam de 65.100 a 8.700.000.
- A média (1.492.658) e o desvio padrão elevado (1.537.165) indicam grandes variações.
- A mediana de 1.100.000 sugere que a maioria dos valores está abaixo da média, novamente indicando possíveis outliers positivos.

6. 60 Day Engagement Rate

- Descrição: Taxa de engajamento nos últimos 60 dias.
- Análise: Intervalo de 0,01 a 10,25, com variabilidade substancial.
- Média: 1,33%, com um desvio padrão maior que a média (1,80%), indica alta variação e comportamento inconsistente de engajamento.
- A mediana mais baixa (0,68%) sugere que a maioria dos valores é inferior à média, reforçando a presença de outliers.

7. New Post Avg Like

- Descrição: Média de curtidas em postagens novas.
- Análise: Variando de 0 a 6.500.000 curtidas.
- A diferença entre a média (967.334) e a mediana (424.550) demonstra assimetria, com muitos tendo menos curtidas, mas alguns poucos posts com um número extremamente alto.

8. Total Likes

- Descrição: Número total de curtidas recebidas.
- Análise: Enorme variação de 18.300.000 a 57.400.000.000.
- A média (4,275 bilhões) é afetada por grandes números, como sugerido pelo desvio padrão volumoso (6,451 bilhões).
- Valores médios e quartis confirmam que muitos usuários têm menos curtidas do que a média indicada.

Tratamento de dados faltantes, Anomalias e Ajustes

Foi constatado que na coluna country a existência de 62 valores nulos foram removidos com o *dropna()* para que não influencie negativamente nas análises futuras. Para os valores para colunas *'total_likes*', *'posts*', '*followers*', '*avg_likes*', '*60_day_eng_rate*' e '*new_post_avg_like*' foram convertidos para números, pois os mesmos eram apresentados como texto, destaca-se as especificações para que fosse realizada de forma correta: 'b': 'e9', 'm': 'e6', 'k': 'e3', '%': ' '. Além disso, nenhuma anomalia foi encontrada no *dataset*.

Correlação

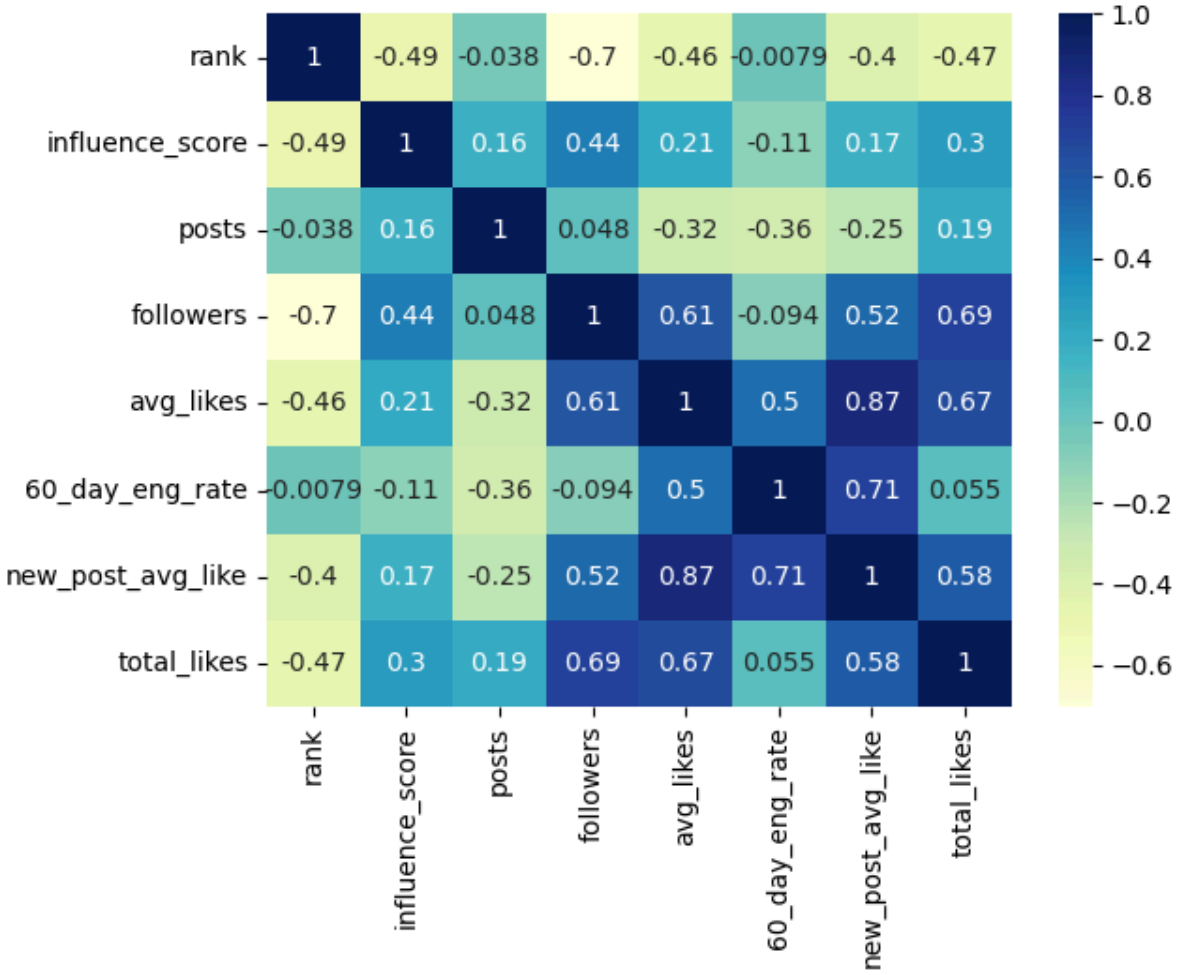

Figura: Gráfico de correlação entre as variáveis.

Para explicar essa matriz de correlação, podemos dividir a análise em alguns pontos principais:

1. <u>Correlação Positiva Forte:</u>
   - *avg_likes* e *new_post_avg_like* têm uma correlação de **0.87**, o que indica que o número médio de *likes* em novas postagens tende a acompanhar a média geral de *likes*.
   - *followers* e *total_likes* têm uma correlação de **0.69**, sugerindo que o número total de *likes* está fortemente relacionado ao número de seguidores.
   - *60_day_eng_rate* e *new_post_avg_like* têm uma correlação de **0.71**, indicando que a taxa de engajamento nos últimos 60 dias está positivamente associada ao número médio de *likes* em novas postagens.

2. Correlação Negativa Forte:
   - *rank* e *followers* têm uma correlação de **-0.7**, o que indica que uma classificação melhor (provavelmente um número de *rank* menor) está associada a um maior número de seguidores.
   - *rank* e *avg_likes* têm uma correlação de **-0.46**, sugerindo que perfis melhor ranqueados tendem a ter um número médio maior de *likes*.
   - *rank* e *total_likes* também possuem uma correlação de **-0.47**, o que reforça a tendência de que perfis com classificação melhor têm mais *likes* totais.

As variáveis *followers*, *avg_likes*, *new_post_avg_like*, e *total_likes* estão positivamente correlacionadas entre si, o que é esperado, já que esses valores refletem a popularidade e engajamento do perfil. A variável rank tem correlações negativas com variáveis de popularidade como *followers*, *avg_likes* e *total_likes*, indicando que perfis mais bem classificados (menor rank) têm mais seguidores e engajamento. A análise de correlação sugere que as métricas de engajamento (como *avg_likes* e *new_post_avg_like*) e o número de seguidores estão fortemente inter-relacionadas. Isso reflete que perfis com mais seguidores tendem a receber mais *likes* e engajamento em suas postagens. Além disso, a classificação (rank) parece ser um indicador inverso da popularidade, onde perfis mais bem classificados (menor valor de rank) tendem a ter métricas de engajamento mais altas.

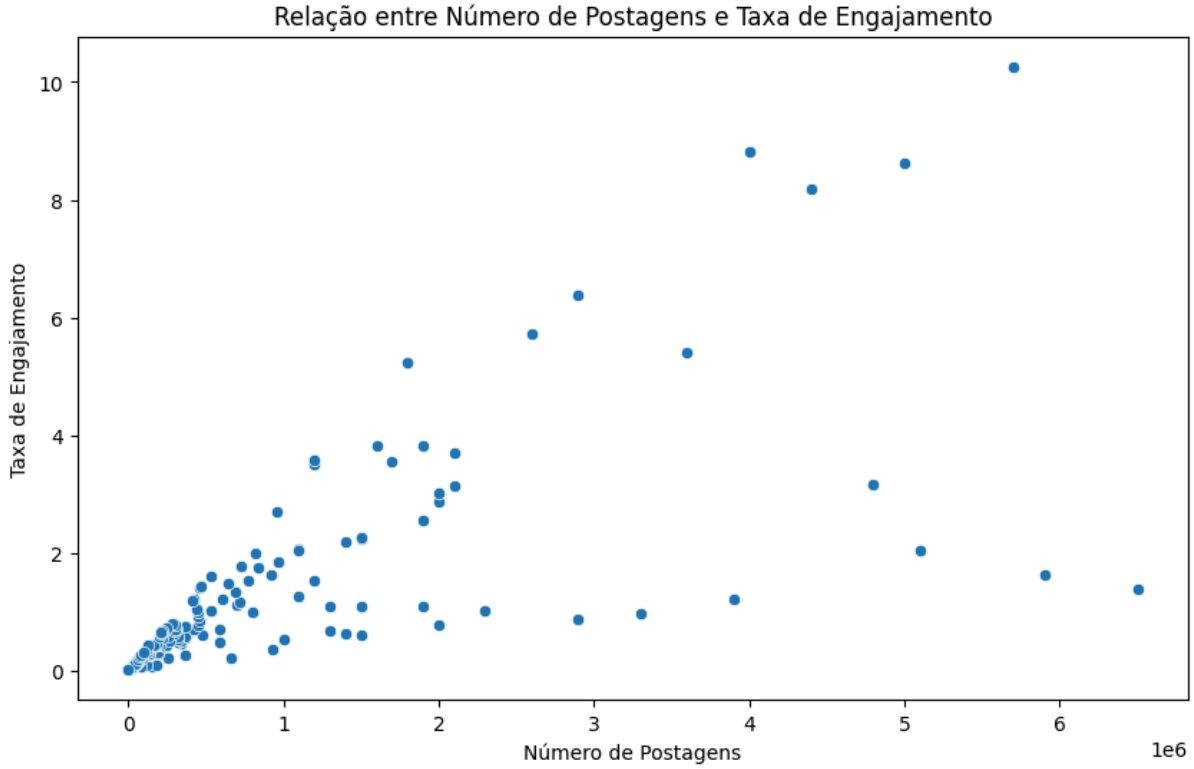

Figura: Gráfico de dispersão postagem/engajamento.

Normalização dos dados

A fim de padronizar os dados para aplicação do modelo preditivo utilizamos uma ferramenta da biblioteca *preprocessing* do *scikit-learn* que normaliza os dados no intervalo específico de 0 a 1, em que para cada valor x_i, temos:

$$x_{scaled} = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$

Essa transformação é feita com o objetivo de garantir a comparabilidade entre variáveis de diferentes proporções como *posts* e *followers*, além de melhorar o desempenho do algoritmo e reduzir o impacto de outliers.

Implementação do Algoritmo

Primeiramente, não foram realizados ajustes explícitos de hiperparâmetros como taxa de aprendizado e número de épocas porque optamos por utilizar a função *LinearRegression* que por sua vez faz o uso do método dos mínimos quadrados ordinários (nativo da função da biblioteca) ao invés do Gradiente Descendente. Isso se dá porque o OLS fornece uma solução analítica exata e que apresentou um bom desempenho o que elimina a necessidade de ajustes mais robustos o que não traria vantagens significativas à aplicação em questão.

O modelo foi configurado com parâmetros padrões. Aqui, detalharemos o processo da implementação da Regressão Linear para prever a variável dependente *60_day_eng_rate* relacionada as variáveis independentes que foram selecionadas com base na relevância que possuem dada a correlação com o engajamento. Para validar essa escolha, aplicamos o método de seleção de recursos RFE (Recursive Feature Elimination). O RFE confirmou que as variáveis escolhidas contribuem significativamente para o modelo, garantindo um conjunto otimizado de preditores:

```
Suporte das variáveis: [ True  True  True]

Ranking das variáveis: [1 1 1]
```

Divisão dos dados

Para tal, temos que:

- Variáveis independentes (X):

- *posts*: Número de postagens.

- *followers*: Quantidade de seguidores.

- *avg_likes*: Média de curtidas por postagem.

- *new_post_avg_like*: Média de curtidas em novas postagens.

- Variável dependente (y):

- *60_day_eng_rate*: Taxa de engajamento dos últimos 60 dias.

Os dados foram divididos em 80% para treino do modelo e 20% para validação e teste de desempenho do modelo, usando *train_test_split* e garantindo que a separação seja aleatória e reprodutível com *random_state=42*.

Configuração do modelo

O algoritmo utilizado foi o de Regressão Linear da biblioteca *scikit-learn*, pois o mesmo é de simples implementação e tem eficácia garantida para problemas de previsão contínua:

```
model = LinearRegression()

model.fit(X_train, y_train)
```

Além disso, foram utilizadas técnicas de regularização (Ridge e Lasso) para fins comparativos com o modelo principal do estudo.

```
ridge_model = Ridge(alpha=1.0)

ridge_model.fit(X_train, y_train)

y_pred_ridge = ridge_model.predict(X_test)

lasso_model = Lasso(alpha=0.1)

lasso_model.fit(X_train, y_train)

y_pred_lasso = lasso_model.predict(X_test)
```

Validação e Ajuste de Hiperparâmetros

A validação do modelo foi feita com o processo de validação cruzada onde o conjunto de treinamento é dividido de forma iterativa de forma a verificar a consistência das previsões durante o treinamento do modelo com o *X_train* e o *y_train*. Vale ressaltar que foi previamente citada a não realização de ajuste de hiperparâmetros.

Validação de Cálculo de Erros

O erro é calculado para treino e validação em cada etapa da validação cruzada.
Assim, os erros calculados para observação do desempenho é dividido em:

- Erro de Treinamento: Avalia o desempenho do modelo nos dados com os quais ele foi treinado.

- Erro de Validação: Mede a capacidade do modelo de generalizar para dados não vistos (dados de teste).

A análise dos erros nos permite identificar se o modelo está sofrendo de *overfitting* (quando o erro de treino é muito baixo, mas o erro de validação é alto) ou *underfitting* (quando ambos os erros são altos). Neste caso, os erros se mostraram baixos tanto no treino quanto na validação, indicando um modelo bem ajustado.

RESULTADOS

Métricas de Avaliação

Após o treinamento, o modelo foi avaliado usando as seguintes métricas:

- **MSE (Mean Squared Error)**: Mede o erro quadrático médio.
- **MAE (Mean Absolute Error)**: Indica o erro absoluto médio.
- **R² (Coeficiente de Determinação)**: Representa a proporção da variação dos dados explicada pelo modelo.

Resultados:

1. Resultados para o modelo de regressão linear (sem regularização):

   - MSE: 0.005939665182447538
   - MAE: 0.04423554964302851
   - R²: 0.8153365216401955

2. Resultados para o modelo Ridge (regularização L2):

   - MSE (Ridge): 0.006084292774353259
   - MAE (Ridge): 0.04527657553923768
   - R² (Ridge): 0.810840067148613

3. Resultados para o modelo Lasso (regularização L1):

   - MSE (Lasso): 0.0329780204237
   - MAE (Lasso): 0.10757089590097402
   - R² (Lasso): -0.025282700926872348

Ao comparar três modelos de regressão para prever a taxa de engajamento de posts em redes sociais, observou-se que a **regressão linear simples** (sem regularização) obteve os melhores resultados, com um **MSE** de **0.0059** e um **R²** de **0.8153**, indicando que o modelo foi capaz de explicar 81,5% da variação na variável dependente. A aplicação da **regularização L2 (Ridge)** resultou em um ligeiro aumento no erro de previsão (**MSE** de **0.0084**) e uma diminuição no **R²** para **73,96%**, sugerindo que a penalização não melhorou a performance do modelo. Já o modelo com **regularização L1 (Lasso)** apresentou uma degradação significativa, com um **MSE** de **0.0330** e um **R²** negativo, o que indica que a penalização excessiva de variáveis não foi benéfica neste caso. Esses resultados sugerem que, para este conjunto de dados, a regularização não trouxe benefícios substanciais, e o modelo simples sem regularização foi o mais eficaz.

Visualizações

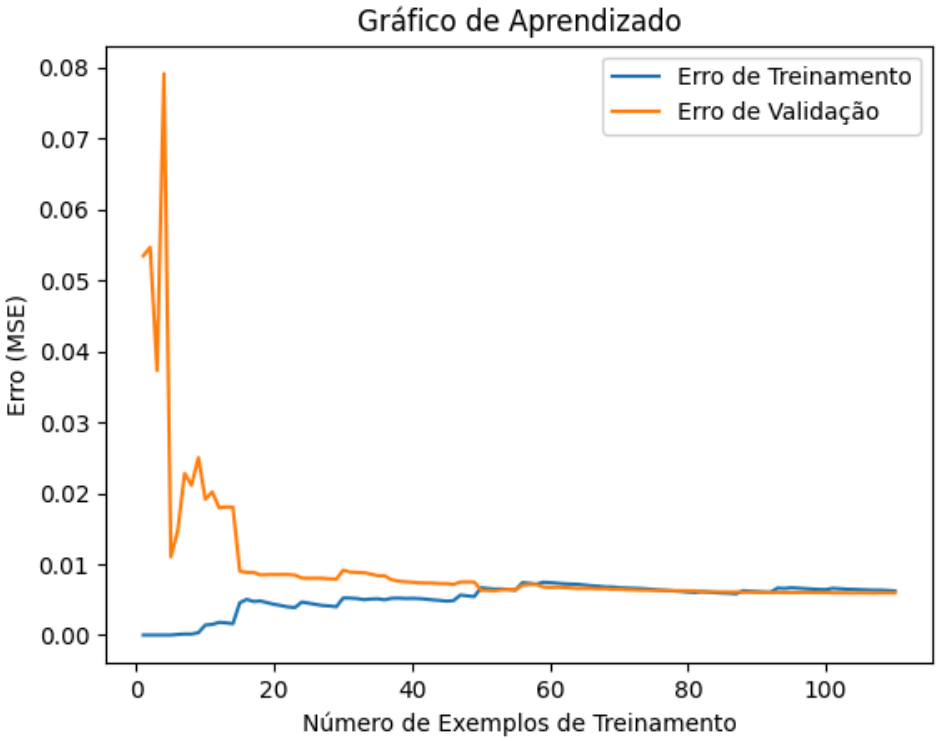

Figura: erros de treinamento e validação.

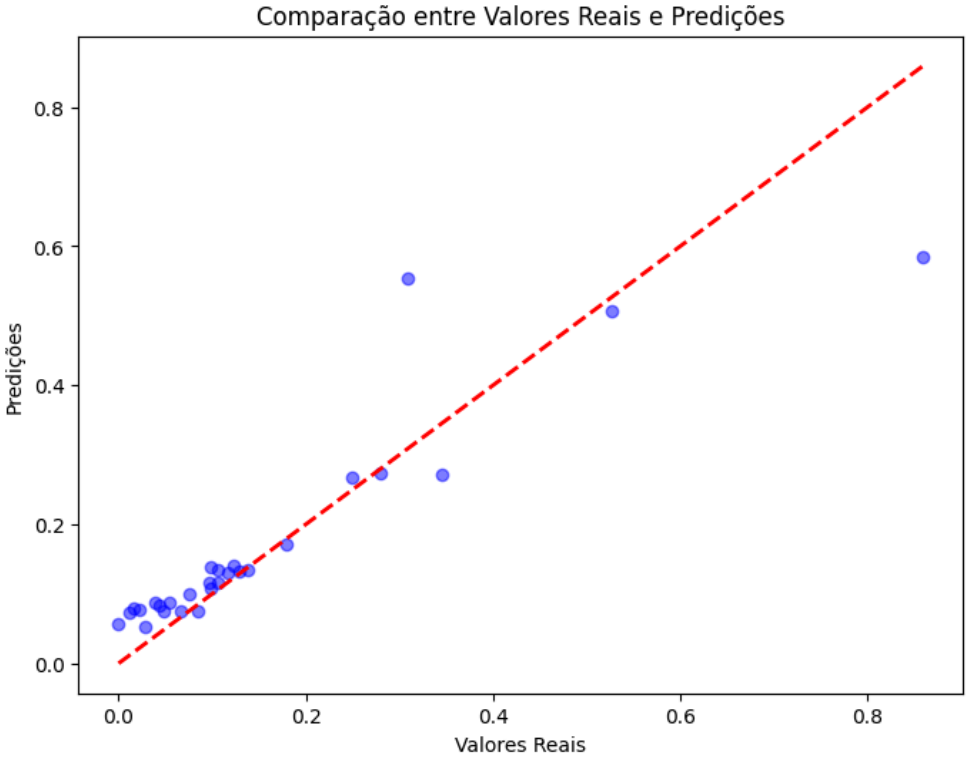

Figura: Gráfico de dispersão com linha de identidade.

DISCUSSÃO

A análise dos resultados desses três modelos de regressão revelam informações importantes acerca do seu desempenho e aplicabilidade da regularização em diferentes cenários:

1. Regressão Linear Simples:

Foi o que obteve os melhores resultados em termos de MSE (0.0059) e R^2 (81,5%) o que sugere que para este conjunto de dados o modelo linear simples é capaz de capturar bem a relação entre as variáveis estudadas. Geralmente, a principal limitação desse modelo é a falta de controle do *overfitting* em conjunto de dados com multicolinearidade ou muitas variáveis, porém não é o caso, logo tivemos um bom resultado.

2. Modelo Ridge (Regularização L2):

A normalização ajudou o Ridge a lidar melhor com as variáveis já que ele penaliza a magnitude dos coeficientes, ele obteve um MSE um pouco mais alto (0.0061) e um R^2 de 81,08%. O MSE mais alto que o modelo linear simples indica que a regularização não era tão necessária nesse conjunto de dados. O Ridge é mais eficaz em dados com multicolinearidade, mas em um cenário com variáveis independentes que não apresentam fortes correlações, sua aplicação pode levar a uma perda de desempenho.

3. Modelo Lasso (Regularização L1):

Esse modelo apresentou o pior desempenho com um MSE seis vezes maior que o modelo linear simples e um R^2 negativo. Isso ocorre porque o Lasso força alguns coeficientes a zero, o que pode ser útil para a seleção de variáveis. Contudo, para este conjunto de dados, essa abordagem foi excessivamente agressiva, resultando em um modelo subajustado. Como esse é um modelo mais adequado para realizar seleção de variáveis, ele acabou removendo informações que eram importantes para a previsão do modelo.

A escolha de regularização foi crucial para entender o impacto que diferentes técnicas têm sobre a performance do modelo. O fato de o modelo Ridge ter um desempenho inferior ao modelo linear simples sugere que a regularização não foi necessária para este conjunto de dados específico, o que pode indicar a ausência de multicolinearidade ou outras características que exijam tal abordagem.

CONCLUSÃO E TRABALHOS FUTUROS

Síntese

- A análise dos resultados mostrou que a regressão linear simples é o modelo mais adequado para este conjunto de dados, com desempenho superior ao Ridge e Lasso. Isso sugere que, para este problema específico, a regularização não trouxe benefícios significativos.
- A regularização L2 (Ridge) tem o potencial de melhorar o desempenho de modelos em cenários com multicolinearidade, mas sua aplicação sem necessidade gerou uma leve degradação da performance.
- A regularização L1 (Lasso), focada em seleção de variáveis, resultou em uma perda substancial de desempenho, indicando que a penalização excessiva foi prejudicial neste caso.

Trabalhos Futuros

- Exploração de Modelos Avançados: Investigar técnicas de Machine Learning mais sofisticadas, como regressão por suportes vetoriais (SVR), Random Forest, ou XGBoost, pode ser interessante para melhorar ainda mais as previsões.
- Validação Cruzada: Implementar validação cruzada para obter uma estimativa mais robusta da performance dos modelos e evitar problemas relacionados ao sobreajuste ou subajuste.
- Aprimoramento da Seleção de Variáveis: Experimentar diferentes métodos de seleção de atributos como PCA (Análise de Componentes Principais), para reduzir a dimensionalidade sem perder informações cruciais.

REFERÊNCIAS

OLIVEIRA NETO, Rosalvo Ferreira de; UNIVERSIDADE FEDERAL DO VALE DO SÃO FRANCISCO. *Ciência dos Dados pelo Processo de KDD*. Junho de 2021. DOI: 10.6084/m9.figshare.14850030.v1. ISBN 978-65-00-24528-8.

James, G.; Witten, D.; Hastie, T.; Tibshirani, R. *An Introduction to Statistical Learning: with Applications in R*. Springer, 2013.

SCIKIT-LEARN. Ridge and Lasso. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.Ridge.html. Acesso em: 16 nov. 2024.

SCIKIT-LEARN. Linear Regression. Disponível em: https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html. Acesso em: 16 nov. 2024.

MATPLOTLIB. Matplotlib for plotting. Disponível em: https://matplotlib.org/stable/users/index.html. Acesso em: 16 nov. 2024.

KUHN, M.; JOHNSON, K. *Applied Predictive Modeling*. Springer, 2013.