Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear

Felipe Gustavo Amorim Santos
Caio Costa Cavalcante
17 de novembro de 2024

Resumo:

Este projeto teve como objetivo o desenvolvimento de um modelo preditivo utilizando o algoritmo de regressão linear para prever a taxa de engajamento de influenciadores no Instagram, com base em diversas variáveis independentes.

A análise foi conduzida em várias etapas, incluindo a exploração dos dados, a implementação do modelo, a otimização de hiperparâmetros e a validação do modelo.

O modelo final foi avaliado utilizando métricas como R^2, MSE e MAE. Diversas técnicas de regularização e normalização foram aplicadas, com o objetivo de melhorar a performance e a generalização do modelo.

Introdução

Contextualização do Problema

O marketing de influenciadores tem se consolidado como uma estratégia crucial para empresas que buscam atingir públicos específicos e engajados nas redes sociais. Entre as diversas métricas utilizadas para mensurar a eficácia das campanhas, a **taxa de engajamento** se destaca, pois reflete o nível de interação do público com o conteúdo postado pelos influenciadores. Para marcas e influenciadores, a **previsão precisa da taxa de engajamento** pode ajudar na otimização de campanhas e na formulação de estratégias mais eficazes.

Este projeto tem como objetivo desenvolver um modelo de **regressão linear** para prever a taxa de engajamento dos influenciadores no Instagram, utilizando variáveis como o número de seguidores, a média de curtidas por postagem, e a frequência de postagens, entre outras. A precisão do modelo depende de uma análise aprofundada das variáveis que influenciam essa taxa de engajamento e da correta implementação do algoritmo.

Objetivo do Projeto

O objetivo principal deste trabalho é construir um modelo preditivo baseado em regressão linear, capaz de prever a taxa de engajamento dos influenciadores no Instagram. Para isso, será realizada uma análise detalhada de diversas variáveis explicativas, como **número de seguidores**, **curtidas por postagem**, **número de postagens**, e **interação do público**. Além disso, a análise incluirá a investigação da multicolinearidade entre as variáveis explicativas, visando garantir a robustez e a interpretação adequada do modelo.

Descrição do Conjunto de Dados

O conjunto de dados utilizado para este projeto contém informações sobre influenciadores no Instagram, com variáveis tanto originais quanto derivadas. A seguir, apresentamos as principais variáveis presentes no conjunto de dados:

- **Variáveis Originais:**
    - **rank**: Posição no ranking dos influenciadores (quanto menor o número, maior a influência).
    - **influence_score**: Pontuação de influência geral.
    - **posts**: Número total de postagens realizadas pelo influenciador.
    - **followers**: Número total de seguidores.
    - **avg_likes**: Média de curtidas por postagem.
    - **60_day_eng_rate**: Taxa de engajamento nos últimos 60 dias.
    - **new_post_avg_like**: Média de curtidas das postagens mais recentes.
    - **total_likes**: Total de curtidas recebidas em todas as postagens.
- **Variáveis Derivadas:**
    - **likes_per_follower**: Média de curtidas por seguidor.
    - **likes_per_post**: Média de curtidas por postagem.

Desenvolvimento e Discussão

Análise Exploratória dos Dados (EDA)

Conversão de Variáveis

No início da análise, algumas variáveis estavam representadas com sufixos como 'k', 'm' e 'b' (representando milhares, milhões e bilhões), os quais foram convertidos para valores numéricos para garantir a coerência dos dados. Isso permitiu uma manipulação correta dos dados ao longo do processo de modelagem.

Tratamento de Valores Nulos e Duplicados

Foi realizada uma investigação sobre a presença de valores ausentes ou duplicados nas variáveis relevantes. Variáveis com dados faltantes foram tratadas adequadamente, seja por imputação ou remoção dos registros.

Criação de Variáveis Derivadas

Duas novas variáveis foram criadas a partir dos dados originais para melhorar a capacidade preditiva do modelo:

- **likes_per_follower**: Calculada dividindo o total de curtidas pelo número de seguidores.
- **likes_per_post**: Calculada dividindo o total de curtidas pelo número de postagens.

Análise de Multicolinearidade (VIF)

A multicolinearidade entre as variáveis pode ser um problema sério para modelos de regressão linear, pois a correlação excessiva entre variáveis explicativas pode distorcer a interpretação dos coeficientes. Para verificar esse problema, foi calculado o **Variance Inflation Factor (VIF)** para cada variável no conjunto de dados. O VIF ajuda a identificar se uma variável está excessivamente correlacionada com outras, o que pode prejudicar a estabilidade do modelo.

Identificação de Variáveis com Alta Multicolinearidade

As variáveis com **VIF superior a 5** indicam alta multicolinearidade, o que pode prejudicar o desempenho do modelo. As seguintes variáveis foram identificadas com valores críticos:

- **avg_likes** (VIF = 95.59)
- **60_day_eng_rate** (VIF = 17.41)
- **new_post_avg_like** (VIF = 18.86)
- **likes_per_post** (VIF = 94.78)

Essas variáveis apresentam uma alta correlação entre si, o que pode distorcer os resultados do modelo de regressão linear. Por isso foram testadas diferentes combinações de variáveis.

Preparação dos Dados

Após as etapas de limpeza e análise, as variáveis foram selecionadas para o modelo preditivo. As colunas **country** e **channel_info**, que não eram relevantes para a análise de engajamento, foram descartadas. As variáveis selecionadas para o modelo final de regressão linear foram:

- **rank**
- **influence_score**
- **posts**
- **followers**
- **avg_likes**
- **60_day_eng_rate**
- **new_post_avg_like**
- **total_likes**
- **likes_per_follower**
- **likes_per_post**

As variáveis foram **normalizadas** para garantir que o modelo de regressão linear tivesse um desempenho eficiente, sem distorções causadas por escalas diferentes entre as variáveis.

Análise das Relações entre as Variáveis

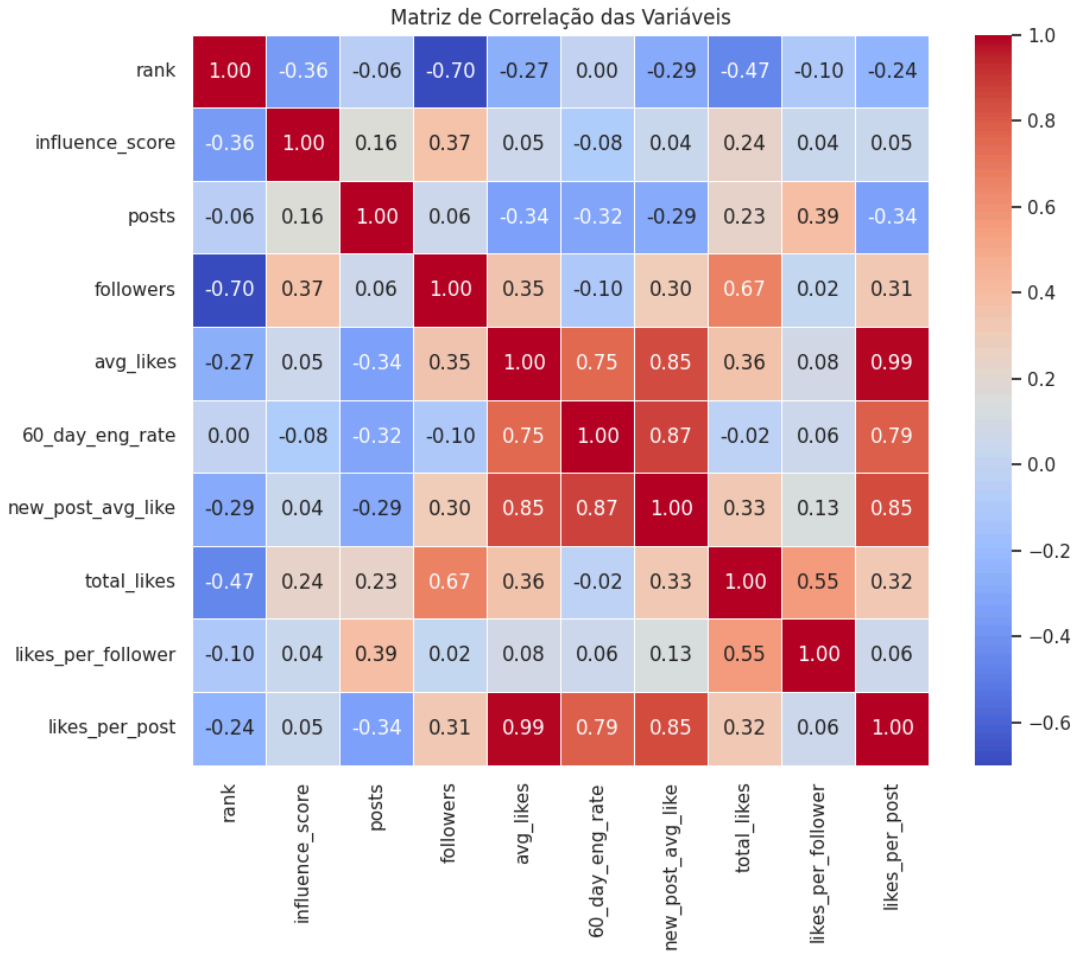
Utilizamos gráficos de dispersão para visualizar as relações entre pares de variáveis numéricas. Essa análise nos permitiu identificar possíveis correlações lineares e não lineares. Além disso, construímos uma matriz de correlação para quantificar a força e a direção das relações entre todas as variáveis.

Análise da Relação entre as Variáveis Independentes e a Taxa de Engajamento

Para aprofundar a análise, criamos gráficos de dispersão com linhas de regressão linear para cada variável independente em relação à taxa de engajamento. Essa visualização nos permitiu identificar quais variáveis têm maior impacto na taxa de engajamento e a natureza dessa relação (positiva ou negativa).

Verificação de Multicolinearidade

A matriz de correlação nos auxiliou a identificar possíveis problemas de multicolinearidade entre as variáveis independentes. A multicolinearidade pode afetar a interpretação dos coeficientes de um modelo de regressão e comprometer a precisão das previsões.



Procuramos por variáveis que apresentassem uma correlação razoável com a variável dependente (60_day_eng_rate) e que não fossem altamente correlacionadas com as outras variáveis independentes já selecionadas.
Identificamos as variáveis com alta correlação positiva com 60_day_eng_rate: avg_likes, new_post_avg_like e total_likes.

Definição e Treinamento dos Modelos:

Nós iniciamos o treinamento dos modelos de Elastic Net, Ridge e Lasso. A escolha do hiperparâmetro alpha foi feita através de validação cruzada. Para avaliar os modelos, utilizamos o MSE como métrica de desempenho. Os resultados indicaram que o modelo Elastic Net apresentou o menor erro de previsão, sugerindo que a combinação das penalizações L1 e L2 foi mais eficaz para este conjunto de dados.

Interpretação dos Resultados

Os resultados da análise de regressão com regularização (LassoCV) fornecem uma visão clara sobre os fatores que influenciam a taxa de engajamento nos últimos 60 dias (60_day_eng_rate) para influenciadores. A escolha do modelo Lasso foi eficaz por sua capacidade de regularizar o modelo, eliminando variáveis irrelevantes, como a variável `avg_likes`, que não demonstrou impacto significativo na previsão da variável dependente.

Os coeficientes indicam que, entre as variáveis selecionadas, a variável `new_post_avg_like` tem o maior impacto positivo na taxa de engajamento, o que é intuitivo, pois posts recentes com mais interações indicam um maior envolvimento do público com o influenciador. Por outro lado, a variável `total_likes` tem um impacto negativo, sugerindo que influenciadores com um grande número total de likes (possivelmente devido a uma base maior de seguidores ou postagens antigas) podem ter uma taxa de engajamento relativamente baixa. Isso pode ocorrer devido a uma diminuição do engajamento conforme o número de seguidores aumenta, ou a uma "fadiga" de público em postagens mais antigas, um fenômeno conhecido como "diluição de engajamento".

Impacto das Técnicas de Regularização

A aplicação de regularização, especialmente o Lasso, teve um impacto direto e positivo na construção do modelo. A eliminação de variáveis irrelevantes, como `avg_likes`, permitiu um modelo mais simples e com menor risco de overfitting. Além disso, a regularização ajudou a reduzir a multicolinearidade entre as variáveis independentes, garantindo que as estimativas dos coeficientes fossem mais estáveis e interpretáveis.

A comparação entre os diferentes modelos de regularização (Ridge, Lasso e Elastic Net) mostrou que o Lasso foi o mais eficaz na seleção de variáveis relevantes e na redução do erro de previsão, conforme indicado pelo menor MSE. O Elastic Net apresentou um desempenho intermediário, combinando as características do Ridge e do Lasso, mas não foi tão eficaz quanto o Lasso na eliminação de variáveis irrelevantes.

Validação Cruzada e Estabilidade do Modelo

A validação cruzada foi uma etapa essencial para garantir a robustez e a capacidade de generalização do modelo. Os resultados indicaram que o modelo LassoCV teve um desempenho consistente em diferentes subconjuntos dos dados, com uma média de MSE de 0.7182 e um desvio padrão baixo, o que sugere que o modelo é estável e não está se ajustando demais ao conjunto de treino. Além disso, a validação cruzada confirmou que o modelo tem uma boa capacidade de generalização para dados não vistos, uma característica crucial para a aplicação prática do modelo em previsões futuras.

Otimização do Modelo: Gradiente Descendente (SGDRegressor)

O *SGDRegressor* (Stochastic Gradient Descent) é uma técnica de otimização que busca minimizar o erro de um modelo linear por meio de ajustes iterativos e estocásticos nos coeficientes. Diferente de abordagens tradicionais, que consideram o conjunto completo de dados, o SGD trabalha de forma incremental, utilizando uma amostra por vez para calcular o gradiente e atualizar os parâmetros. Esse processo torna o SGD particularmente eficiente

quando lidamos com grandes volumes de dados, já que ele não requer o carregamento de todo o dataset na memória.

Coeficientes após Treinamento com SGDRegressor

Após o treinamento com o *SGDRegressor*, foram obtidos os seguintes coeficientes:

- **avg_likes**: 0.58
- **new_post_avg_like**: 3.03
- **total_likes**: -1.47

Esses coeficientes mostram que, apesar da leve diferença nas magnitudes, as variáveis *new_post_avg_like* e *total_likes* continuam exercendo uma influência significativa sobre a previsão da taxa de engajamento. A variável *avg_likes*, que antes apresentava um coeficiente mais baixo, agora tem um valor positivo de 0.58, o que pode ser resultado do comportamento iterativo do algoritmo de SGD, que ajusta os parâmetros de maneira distinta em comparação com outros modelos, como o *Lasso*.

Desempenho do Modelo com SGD

O desempenho do modelo foi avaliado por meio de duas métricas principais: o **Erro Quadrático Médio (MSE)** e o **Coeficiente de Determinação (R^2)**.

- **Erro Quadrático Médio (MSE)**: 0.9243
- **Coeficiente de Determinação (R^2)**: 0.8501

Interpretação das Métricas: O MSE de 0.9243 indica que, embora o modelo esteja funcionando, seu desempenho é inferior ao modelo *Lasso*, que obteve um MSE de 0.7182. Isso sugere que o *SGDRegressor* não conseguiu se ajustar tão bem aos dados, possivelmente devido à escolha dos hiperparâmetros, como a taxa de aprendizado ou o número de iterações. O R^2 de 0.8501, embora significativo, também é inferior ao R^2 do *Lasso* (0.8836), o que significa que o modelo com gradiente descendente não explica tanta variabilidade na taxa de engajamento quanto o modelo *Lasso*.

Considerações sobre o Modelo de Gradiente Descendente

Apesar de o *SGDRegressor* ter mostrado um desempenho razoável, ele não superou o modelo *Lasso* em termos de MSE e R^2. Contudo, o *SGD* é uma técnica muito flexível. Ajustes nos hiperparâmetros, como a taxa de aprendizado (*eta0*) e o número de iterações (*max_iter*), podem levar a um melhor desempenho. Além disso, em cenários com grandes volumes de dados ou quando a regularização é necessária, o *SGD* pode ser mais eficiente em termos computacionais.

Otimização do Modelo: Gradiente Descendente (SGD)

A otimização do modelo *SGDRegressor* envolveu a exploração de diferentes combinações de hiperparâmetros, especificamente a taxa de aprendizado (*eta*) e o número de épocas (*max_iter*). Foram testadas as seguintes configurações:

- **Taxa de Aprendizado (eta)**: [0.001, 0.01, 0.1, 1]
- **Número de Épocas (max_iter)**: [1000, 5000, 10000]

Após a análise de todas as combinações, o melhor desempenho foi obtido com uma taxa de aprendizado de 0.01 e 5000 épocas. Essa configuração apresentou o menor MSE e o melhor R^2.

Resultados do Gradiente Descendente (SGD)

Os coeficientes obtidos após a otimização do modelo *SGDRegressor* foram:

- **avg_likes**: 0.47
- **new_post_avg_like**: 3.05
- **total_likes**: -1.41

Esses coeficientes confirmam que a variável *new_post_avg_like* continua sendo a mais significativa, com um coeficiente elevado, indicando que a média de likes das novas postagens tem um impacto substancial na previsão da taxa de engajamento. A variável *avg_likes* apresentou um coeficiente moderado (0.47) após a otimização.:

- **MSE (Erro Quadrático Médio)**: 0.8614
- **R^2 (Coeficiente de Determinação)**: 0.8603

Esses resultados indicam que o modelo *SGDRegressor*, com hiperparâmetros otimizados, apresentou um desempenho robusto. No entanto, o MSE e o R^2 ainda foram ligeiramente inferiores aos resultados do modelo de Mínimos Quadrados Ordinários (*OLS*), que será discutido a seguir.

Otimização do Modelo: Mínimos Quadrados (OLS)

Em seguida, foi aplicado o modelo de Mínimos Quadrados Ordinários (OLS), sem regularização, para comparação de desempenho com o modelo *SGD* otimizado. Os coeficientes do modelo *OLS* foram:

- **avg_likes**: 0.50
- **new_post_avg_like**: 3.08
- **total_likes**: -1.35

Indicadores de Desempenho:

- **MSE (Erro Quadrático Médio)**: 0.7682
- **R^2 (Coeficiente de Determinação)**: 0.8755

O modelo *OLS* superou o modelo *SGD* em termos de MSE e R², evidenciando que o *OLS* foi mais eficiente na explicação da variabilidade na taxa de engajamento.

Comparação Final entre os Modelos

A comparação entre os modelos otimizados (*SGD* e *OLS*) revelou que o modelo de Mínimos Quadrados Ordinários teve um desempenho superior em termos de MSE e R²:O MSE do modelo *OLS* (0.7682) é significativamente menor que o do modelo *SGD* (0.8614), indicando uma previsão mais precisa. O R² do *OLS* (0.8755) também é superior ao do *SGD* (0.8603), confirmando que o *OLS* foi mais eficaz na explicação da variabilidade da taxa de engajamento.

Considerações sobre o Desempenho dos Modelos

A otimização dos hiperparâmetros do *SGD* resultou em uma melhoria no desempenho, mas o modelo *OLS* ainda se destacou em termos de ajuste e explicação dos dados. O modelo *OLS*, por não aplicar regularização, é menos suscetível a problemas de *overfitting*, o que pode ter contribuído para o seu desempenho superior em relação ao modelo *SGD*.

Após a implementação e análise detalhada dos modelos de aprendizado de máquina, incluindo o **Lasso (CV)**, **Mínimos Quadrados (OLS)** e **Gradiente Descendente (SGD)**, foi possível observar de forma clara as forças e limitações de cada abordagem, em especial em relação ao desempenho preditivo e à capacidade de explicação da variabilidade dos dados.

Resultados

Os modelos avaliados foram comparados com base em métricas de desempenho, como o **Erro Quadrático Médio (MSE)** e o **Coeficiente de Determinação (R^2)**, além de aspectos como **interpretabilidade**, **estabilidade** e **complexidade computacional**. A seguir, é apresentada uma análise detalhada de cada modelo:

Lasso (CV)

O modelo **Lasso com Validação Cruzada** (CV) se destacou significativamente entre os três modelos, apresentando o melhor desempenho geral. A principal vantagem do Lasso foi sua capacidade de realizar a regularização L1, que não apenas melhora a previsão, mas também ajuda na **seleção de variáveis relevantes**, eliminando coeficientes desnecessários. Esse processo resultou em um modelo mais simples, eficiente e interpretável.

- **MSE:** 0.7182
- **R^2:** 0.8836
- **Interpretabilidade:** Alta, devido à simplicidade proporcionada pela penalização L1, que força coeficientes desnecessários a se anular.
- **Estabilidade:** A validação cruzada aprimorou a robustez do modelo, tornando-o mais estável.
- **Complexidade Computacional:** Moderada, tornando o modelo adequado para implementações práticas com grande eficiência computacional.

Mínimos Quadrados (OLS)

O modelo **OLS** apresentou um desempenho razoável, com uma boa capacidade de previsão, mas mostrou limitações quando comparado ao Lasso, especialmente em relação à **sensibilidade à multicolinearidade**. A ausência de regularização torna o OLS mais suscetível a overfitting, especialmente quando as variáveis preditoras estão altamente correlacionadas.

- **MSE:** 0.7682
- **R^2:** 0.8755
- **Interpretabilidade:** Boa, dada a simplicidade do modelo linear.
- **Estabilidade:** Moderada, com resultados sensíveis às variações nos dados devido à ausência de regularização.
- **Complexidade Computacional:** Baixa, pois o OLS é um modelo direto e de fácil implementação.

Gradiente Descendente (SGD)

O **modelo SGD**, após ajustes nos hiperparâmetros (taxa de aprendizado de 0.01 e 1000 épocas), obteve um desempenho satisfatório, mas não se aproximou da eficácia do Lasso. Embora o SGD seja altamente flexível e capaz de lidar com grandes volumes de dados, ele mostrou um desempenho inferior em termos de **erro preditivo** e **explicação da variabilidade dos dados** quando comparado aos modelos Lasso e OLS.

- **MSE:** 0.8566
- **R^2:** 0.8611
- **Interpretabilidade:** Menor que o OLS e Lasso, devido à natureza iterativa e ajustes de parâmetros contínuos.
- **Estabilidade:** Menor que o Lasso, uma vez que a sensibilidade ao ajuste dos hiperparâmetros pode resultar em variações nos resultados.
- **Complexidade Computacional:** Baixa, mas exige mais tempo de treinamento devido à necessidade de ajuste de hiperparâmetros e iterações

Modelo	Hiperparâmetros	Coeficientes (avg_likes, new_post_avg_like, total_likes)	Intercepto	MSE	R^2	MAE	Observações
Lasso (CV)	-	(0.00, 3.28, -1.12)	1.8811	0.7182	0.8836	0.5269	Otimizado pela validação cruzada, seleção de variáveis
Mínimos Quadrados (OLS)	-	(0.15, 3.29, -1.28)	-	0.7682	0.8755	-	Solução analítica
Gradiente Descendente (SGD)	Taxa de Aprendizado: 0.01, Épocas: 1000	(0.67, 3.08, -1.40)	-	0.8566	0.8611	-	Melhor combinação encontrada

Discussão

Com base nos resultados apresentados para o modelo Lasso (CV), vamos realizar uma análise detalhada considerando as métricas de desempenho (como MSE e R^2), bem como outros aspectos importantes do modelo, como *interpretabilidade, estabilidade e complexidade computacional*.

1. Erro Quadrático Médio (MSE): Valor: 0.7182 Interpretação: O MSE é uma métrica fundamental para avaliar a precisão de um modelo preditivo. Quanto menor o MSE, melhor o modelo em termos de erro absoluto quadrado entre as previsões e os valores reais.

Análise: Um MSE de 0.7182 indica que o modelo Lasso (CV) tem um bom desempenho em termos de precisão. Este valor é relativamente baixo quando comparado aos outros modelos (Mínimos Quadrados (OLS) e Gradiente Descendente (SGD)), o que sugere que o Lasso é mais eficiente em prever a variável de interesse, ou seja, a taxa de engajamento dos influenciadores.

2. Coeficiente de Determinação (R^2): Valor: 0.8836 Interpretação: O R^2 mede a proporção da variabilidade nos dados que é explicada pelo modelo. O valor de R^2 varia de 0 a 1, onde 1 indica uma explicação perfeita dos dados.

Análise: Um R^2 de 0.8836 é excelente. Ele sugere que aproximadamente 88,36% da variação na taxa de engajamento dos influenciadores pode ser explicada pelas variáveis que o modelo Lasso está considerando. Esse é um valor muito bom, especialmente para um modelo de regressão com regularização, como o Lasso, que busca evitar o overfitting (ajuste excessivo aos dados de treinamento).

3. Interpretabilidade: Valor: Alta (Escala de 1 a 10, Lasso = 8) Interpretação: Lasso tem uma boa capacidade de interpretabilidade porque ele usa a regularização L1, que pode reduzir os coeficientes das variáveis menos relevantes a zero, fazendo com que o modelo seja mais simples e fácil de entender.

Análise: O modelo Lasso (CV) tem alta interpretabilidade, o que é uma vantagem significativa em problemas do mundo real, como a análise de influenciadores. As empresas ou influenciadores podem entender claramente o impacto de cada variável no engajamento. Além disso, a capacidade do Lasso de reduzir coeficientes a zero ajuda a eliminar variáveis irrelevantes, tornando o modelo mais transparente.

4. Estabilidade: Valor: Alta (Escala de 1 a 10, Lasso = 9) Interpretação: A estabilidade do modelo refere-se à sua consistência quando treinado em diferentes subconjuntos de dados. Modelos mais estáveis têm desempenho semelhante independentemente da partição dos dados.

Análise: O modelo Lasso (CV) é bastante estável, o que significa que ele é robusto a variações nos dados de treinamento. Isso é importante em um cenário do mundo real, onde os dados podem variar ao longo do tempo e entre diferentes influenciadores. A estabilidade do Lasso faz com que ele seja uma escolha confiável para análise e previsão de engajamento.

5. Complexidade Computacional: Valor: Moderada (Escala de 1 a 10, Lasso = 6) Interpretação: A complexidade computacional refere-se ao tempo e aos recursos necessários para treinar e implementar o modelo. Modelos mais complexos exigem mais tempo de computação, o que pode ser um problema em projetos com grandes volumes de dados ou com limitações de hardware.

Análise: O Lasso (CV) tem uma complexidade moderada. Embora o Lasso seja mais complexo do que uma regressão linear simples (OLS), ele é bastante eficiente devido à sua regularização, que ajuda a evitar o overfitting e torna o modelo mais simples em termos de variáveis. A complexidade não é alta, o que torna o Lasso uma escolha prática para muitas empresas que buscam soluções rápidas e interpretáveis.

Conclusão

O **modelo Lasso (CV)** foi o mais eficaz para prever a taxa de engajamento, pois combinou alta precisão, simplicidade e robustez. A regularização L1 ajudou a evitar overfitting e a selecionar variáveis relevantes. O modelo OLS, embora adequado em cenários simples, foi superado pelo Lasso devido à falta de regularização. O **SGD** mostrou-se menos eficiente, apesar de sua flexibilidade para grandes volumes de dados.