"Relatório Técnico: Implementação e Análise do Algoritmo de Regressão Linear"

Nome dos Residentes: Clara, Jemyma
Data de Entrega: 17/11/2024

Resumo

Este texto relata a análise e modelagem de uma base de dados do Instagram retirada do Kaggle, bem como os resultados encontrados a partir desses estudos. Com o objetivo de aplicar alguns dos conhecimentos adquiridos ao longo de meses de aulas de Ciência de Dados, a Regressão Linear foi utilizada para tentar prever um valor numérico com base em um conjunto de variáveis explicativas (como número de seguidores, número de likes, localização, etc). Foi buscado obter a preparação de dados relevantes e a limpeza dos mesmos para garantir a qualidade da inspeção. A Análise Exploratória dos dados foi feita para identificar padrões, relações entre as variáveis e possíveis problemas e então houveram construções de diferentes modelos, escolhas e ajustes, considerando as características dos dados e o objetivo da análise. Na avaliação dos modelos, foram utilizadas métricas como RMSE e R^2 para calcular a sua precisão e sua capacidade de generalização, coisas que não ficaram comprovadas devido ao baixo valor dos coeficientes de correlação e da aparente falta de correlação linear entre as variáveis apresentadas. A interpretação dos resultados conclui o relatório, e por fim também é recomendado que essa mesma base de dados seja analisada através de outros métodos estatísticos e computacionais.

Coleta e Preparação dos Dados:

- **Coleta: Obter os dados de fontes confiáveis.**
- **Limpeza: Lidar com valores ausentes, outliers e inconsistências.**
- **Transformação: Normalizar os dados, criar novas variáveis ou transformar variáveis existentes.**

Exploração dos Dados:

- **Visualização: Criar gráficos para entender a distribuição das variáveis e a relação entre elas.**
- **Estatísticas descritivas: Calcular medidas como média, mediana, desvio padrão, correlação.**

Construção do Modelo:

- **Escolha do algoritmo: Regressão linear simples ou múltipla, dependendo do número de variáveis independentes.**
- **Treinamento: Ajustar os parâmetros do modelo aos dados de treinamento.**
    - **Validação: Avaliar o desempenho do modelo em um conjunto de dados de teste.**

Avaliação do Modelo:

- **Métricas: Utilizar métricas como RMSE (Erro Quadrático Médio), MAE (Erro Absoluto Médio), R^2 (Coeficiente de Determinação) para avaliar a precisão do modelo.**
- **Análise de resíduos: Verificar se os resíduos seguem uma distribuição normal e se há heterocedasticidade.**

Objetivos do projeto implementação do algoritmo de regressão linear:

- **construção de um modelo preditivo: Tendo como foco criar um modelo matemático que possa prever o valor de uma variável dependente (y) com base em uma ou mais variáveis independentes (x).**
- **Compreensão do relacionamento entre variáveis: Além de fazer previsões, a regressão linear permite entender como as variáveis independentes influenciam a variável dependente.**
- **Tomada de decisões: Os insights obtidos a partir do modelo podem ser utilizados para obter decisões mais informadas em diversas áreas, como negócios, ciência, engenharia, etc.**
- **Aprendizado : A implementação desse algoritmo de regressão linear é uma excelente forma de aprender sobre conceitos de estatística, machine learning e programação.**
- **Customização: Permite adaptar o algoritmo para necessidades específicas.**
- **Comparação com outras técnicas: Ao implementar a regressão linear, é possível compará-la com outros algoritmos de machine learning e avaliar suas vantagens e desvantagens.**
- **Desenvolvimento de habilidades: A implementação de um projeto de machine learning exige o desenvolvimento de habilidades como: coleta e preparação de dados - como limpeza e transformação . Treinamento e avaliação do modelo - ajustar parâmetros do modelo . Interpretação dos resultados - análises de modelo e desempenho)**

Resultados Práticos para ser obtidos:

- **Modelo preditivo: O resultado mais visível é a construção de um modelo matemático capaz de fazer previsões precisas sobre uma variável dependente com base em outras variáveis independentes. Esse modelo pode ser utilizado:**
- **Tomada de decisões: Prever vendas futuras, estimar custos, avaliar riscos.**

- **Otimização de processos: Identificar os principais fatores que influenciam um determinado resultado e ajustar os processos para otimizá-lo.**
- **Análise de tendências: Identificar padrões e tendências nos dados.**
- **Compreensão dos dados: A regressão linear ajuda a entender a relação entre as variáveis, permitindo identificar quais fatores são mais importantes para explicar a variabilidade da variável dependente.**

Introdução

O que é Regressão Linear?

A regressão linear é uma técnica estatística fundamental utilizada para modelar a relação entre uma variável dependente (aquela que queremos prever) e uma ou mais variáveis independentes (aquelas que acreditamos influenciarem a variável dependente). Em termos mais simples, ela busca encontrar a melhor reta que se ajusta aos dados, permitindo fazer previsões.

A Equação da Regressão Linear

A equação geral da regressão linear simples (com apenas uma variável independente) é:

$y = \beta0 + \beta1*x + \varepsilon$

Onde:

- **y:** Valor previsto da variável dependente
- **β0:** Intercepto (valor de y quando x=0)
- **β1:** Coeficiente angular (inclinação da reta)
- **x:** Valor da variável independente
- **ε:** Erro aleatório (diferença entre o valor real e o valor previsto)

Propriedades da Regressão Linear

- **Linearidade:** A relação entre as variáveis é assumida como linear.
- **Aditividade:** O efeito de cada variável independente sobre a variável dependente é aditivo.
- **Homocedasticidade:** A variância dos erros é constante para todos os valores de x.
- **Independência:** Os erros são independentes entre si.
- **Normalidade:** Os erros seguem uma distribuição normal.

Quando Usar Regressão Linear?

- **Previsão:** Prever valores futuros de uma variável.
- **Análise de Causalidade:** Entender como uma variável independente influencia a variável dependente.

- **Construção de Modelos:** Criar modelos para explicar fenômenos complexos.

Exemplos de Aplicações

- **Finanças:** Prever o preço de ações, estimar o risco de crédito.
- **Marketing:** Analisar a relação entre gastos em publicidade e vendas.
- **Ciências Sociais:** Estudar a relação entre renda e felicidade, por exemplo.
- **Engenharia:** Modelar processos industriais, prever falhas em equipamentos.

Contextualização do Problema e Justificativa para o Uso do Algoritmo

A escolha do problema selecionado com regressão linear é importante para o sucesso do projeto. A justificativa para o uso desse algoritmo traz clareza e embasamento em conhecimentos prévios sobre o domínio e os dados.

Justificativa para o uso da regressão linear:

- **Relação linear: Quando se espera que exista uma relação linear entre a variável dependente e as variáveis independentes.**
- **Previsão: A regressão linear é uma ferramenta poderosa para fazer previsões numéricas.**
- **Interpretação: Os coeficientes da regressão linear podem ser interpretados como a contribuição de cada variável independente na variável dependente.**
- **Simplicidade: A regressão linear é um algoritmo relativamente simples de implementar e entender.**

Descrição do Conjunto de Dados e Seu Contexto

A qualidade e a quantidade dos dados são fundamentais para o sucesso de qualquer projeto de machine learning. É importante que os dados sejam relevantes para o problema em questão e que estejam em um formato adequado para análise.

Contexto do conjunto de dados:

O contexto do conjunto de dados se refere à origem dos dados e às condições em que foram coletados. É importante entender o contexto para interpretar os resultados da análise. Por exemplo, dados de vendas de imóveis coletados em uma cidade grande podem não ser generalizáveis para cidades menores. A escolha do problema e a qualidade dos dados são os pilares de um projeto de regressão linear.

Validação e Ajuste de Hiperparâmetros

Validação: É o processo de avaliar a performance do seu modelo em dados que ele não viu durante o treinamento. Isso permite estimar quão bem o modelo é generalizado para novos dados.

Ajuste de Hiperparâmetros: São parâmetros do modelo que não são aprendidos pelos dados, mas que precisam ser definidos antes do treinamento.

Detalhes sobre as Escolhas das Variáveis:

- **Relevância: As variáveis escolhidas para o modelo devem ter uma relação clara com a variável alvo (aquela que você está tentando prever). A análise exploratória te ajudou a identificar essas relações.**
- **Colinearidade: Variáveis altamente correlacionadas podem causar problemas de multicolinearidade, afetando a interpretação dos coeficientes e a estabilidade do modelo. É importante avaliar a matriz de correlação e considerar técnicas como seleção de features ou regularização para lidar com esse problema.**
- **Transformações: Dependendo da distribuição das variáveis, pode ser necessário aplicar transformações (como logaritmo ou raiz quadrada) para linearizar a relação entre as variáveis e melhorar a performance do modelo.**

- **Interações: Considere a possibilidade de criar novas features que representam a interação entre duas ou mais variáveis existentes. Sobre o Processo da Validação Cruzada:**

A validação cruzada é uma técnica que divide os dados em diferentes subconjuntos (folds). O modelo é treinado em todos os folds, exceto um, que é usado para avaliação. Esse processo é repetido para cada fold, e a performance média do modelo é calculada.

Tipos de Validação Cruzada:

- **Holdout: Divide os dados em um conjunto de treinamento e um conjunto de testes.**
- **k-fold: Divide os dados em k partes iguais. O modelo é treinado em k-1 partes e testado na parte restante.**
- **Validação cruzada estratificada: Garante que a proporção das classes seja mantida em cada fold, o que é importante para problemas de classificação.**

O processo de utilização da Validação Cruzada

- **Estimação mais precisa da performance: Ao usar a validação cruzada, você obtém uma estimativa mais precisa da performance do modelo em novos dados, pois o modelo é avaliado em diferentes subconjuntos de dados.**
- **Prevenção de overfitting: A validação cruzada ajuda a identificar modelos que estão sobre ajustados aos dados de treinamento.**

Métricas Comumente Utilizadas em Regressão Linear:

- **Erro Quadrático Médio (RMSE): mede a diferença média entre os valores reais e os valores previstos pelo modelo. Quanto menor o RMSE, melhor o modelo.**

- **Erro Absoluto Médio (MAE): similar ao RMSE, mas calcula a diferença absoluta média. É menos sensível a outliers.**
- **Coeficiente de Determinação (R^2): indica a proporção da variância da variável dependente que é explicada pelo modelo. Um valor de R^2 próximo de 1 indica um bom ajuste do modelo.**

Importância da análise dos Resultados

A análise dos resultados envolve a interpretação das métricas de avaliação e a comparação do desempenho do modelo com um benchmark ou com outros modelos.

Interpretação das Métricas:

- **RMSE baixo: Indica que o modelo, em média, faz previsões próximas aos valores reais.**
- **MAE baixo: Indica que o modelo, em média, comete erros pequenos nas suas previsões.**
- **R^2 alto: Indica que o modelo explica uma grande parte da variabilidade da variável dependente.**

Discussão

Após a análise exploratória, construção e avaliação do modelo de regressão linear, é fundamental realizar uma discussão crítica dos resultados obtidos. Essa discussão nos permite compreender melhor as forças e fraquezas do modelo.

Discussão Crítica dos Resultados

- **Qualidade do ajuste: Os valores das métricas de avaliação (RMSE, MAE, R^2) indicam a qualidade do ajuste do modelo aos dados. Um R^2 alto, por exemplo, sugere que o modelo explica uma grande parte da variabilidade dos dados, mas não garante que o modelo seja generalizável para novos dados.**

- **Significância dos coeficientes: A análise dos coeficientes da regressão permite identificar quais variáveis são mais importantes para explicar a variável dependente. No entanto, a significância estatística dos coeficientes deve ser verificada para garantir que a relação observada não seja apenas resultado do acaso.**
- **Resíduos: A análise dos resíduos pode revelar padrões que não foram capturados pelo modelo, como heterocedasticidade ou autocorrelação. Esses padrões podem indicar que o modelo não é adequado para os dados ou que faltam variáveis importantes.**

Limitações Encontradas

- **Linearidade: A regressão linear assume uma relação linear entre as variáveis. Se a relação entre as variáveis for não linear, o modelo pode não capturar essa relação adequadamente.**
- **Multicolinearidade: A presença de alta correlação entre as variáveis independentes pode dificultar a interpretação dos coeficientes e aumentar a variância das estimativas.**
- **Outliers: Outliers podem ter um grande impacto no ajuste do modelo. É importante identificar e tratar os outliers de forma adequada.**
- **Dados faltantes: A presença de dados faltantes pode reduzir o tamanho da amostra e afetar a precisão do modelo.**
- **Overfitting: O modelo pode estar sobreajustado aos dados de treinamento, o que significa que ele se ajusta muito bem aos dados de treinamento, mas não generaliza bem para novos dados.**

Impacto das Escolhas Feitas no Desempenho do Modelo

- **Seleção de variáveis: A escolha das variáveis independentes tem um grande impacto no desempenho do modelo. A inclusão de variáveis irrelevantes pode aumentar a variância do modelo, enquanto a exclusão de variáveis importantes pode reduzir o seu poder preditivo.**

- **Transformações: As transformações aplicadas às variáveis podem afetar a linearidade da relação entre as variáveis e, consequentemente, o desempenho do modelo.**
- **Validação cruzada: A escolha do método de validação cruzada e o número de folds podem afetar a estimativa da performance do modelo.**

Conclusão e Trabalhos Futuros

Principais Aprendizados:

- **Importância da análise exploratória: A análise exploratória foi fundamental para entender a natureza dos dados, identificar padrões, outliers e possíveis problemas de qualidade.**
- **Tratamento de dados: A necessidade de tratar os dados, como lidar com valores faltantes e transformar variáveis**
- **Escolha das variáveis: A seleção das variáveis independentes mais relevantes para a variável dependente é um passo crítico para construir um modelo preciso.**
- **Métricas de avaliação: As métricas de avaliação permitem quantificar a qualidade do modelo e comparar diferentes modelos.**

Sugestões de Melhorias:

Explorar modelos mais aprofundados:

- **Regressão não linear: Se a relação entre as variáveis não for linear, modelos como regressão polinomial ou regressão com base em splines podem ser mais adequados.**
    - **Modelos de machine learning: Modelos como árvores de decisão, random forest.**

Engenharia de features:

- **Criar novas features: Combinar ou transformar as features existentes pode melhorar o poder preditivo do modelo.**
- **Seleção de features: Utilizar técnicas de seleção de features para identificar as features mais relevantes e reduzir a dimensionalidade dos dados.**

Tratamento de outliers:

- **Identificar outliers: Utilizar técnicas como boxplots, z-scores ou métodos baseados em distância para identificar outliers.**
- **Tratar outliers: Remover, transformar ou criar uma categoria separada para os outliers.**

Lidar com dados faltantes:

- **Imputação: Utilizar técnicas de imputação para preencher os valores faltantes.**
- **Modelo de imputação: Construir um modelo para prever os valores faltantes com base nas outras variáveis.**

Analisar a heterocedasticidade:

- **Teste de Breusch-Pagan: Verificar se a variância dos erros é constante ao longo dos valores previstos.**
- **Transformações: Aplicar transformações aos dados, como logaritmo ou raiz quadrada, para estabilizar a variância.**

Considerar a autocorrelação:

- **Teste de Durbin-Watson: Verificar se os erros estão correlacionados no tempo.**
- **Modelos de séries temporais: Se os dados forem temporais, utilizar modelos de séries temporais para capturar a autocorrelação.**

Avaliar a robustez do modelo:

- **Validação cruzada: Utilizar diferentes métodos de validação cruzada para estimar a performance do modelo em novos dados.**

METODOLOGIA

1. Análise Exploratória:

Após a base de dados ser importada, foi fácil perceber que ela contém 10 colunas, sendo algumas delas: nomes de perfis, quantidade de seus seguidores, quantos novos posts são feitos em média, o total de likes da conta, em qual país o perfil se encontra, etc.

Verificando as informações da base de dados (df.info), é possível ver que várias variáveis, apesar de serem numéricas, são interpretadas como 'objetos', algo que precisa ser tratado já que cálculos precisarão ser feitos com esses mesmos valores.

Então foi achado o número de informações faltantes em cada coluna, já que dependendo da quantidade, isso poderia ser um problema.

Na coluna *"country"* foram 62 NaN e em *"60_day_eng_rate"* foi apenas um.

Continuando com a análise, uma a uma, as colunas foram modificadas de modo que letras e números foram retirados, a fim de fazer com que o computador passasse a ser capaz de interpretá-las como *float* ou invés de *object*.

Continuando, o comando "*df.describe()"* foi usado para que o desvio padrão e quantidades mínimas e máximas fossem interpretadas, juntamente com os quartis. Com base nos resultados obtidos, foi possível identificar que o desvio padrão da

maioria das variáveis é alto, com exceção de *‘posts*’, que teve o nome substituído por *‘posts_replace’* pós-tratamento, e a variável *‘ 60_rate_raplace’* (std igual a 3.475828 e a 3.329719, respectivamente).

O comando *sns.pairplot(df, kind = 'scatter')* foi feito para a verificação visual das possíveis correlações entre todas as variáveis entre si. As imagens mostraram baixa correlação entre a maioria delas (figura 1).

2. Implementação do Algoritmo de Regressão Linear

Desenvolvimento:

```
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import mean_squared_error

# Carregar os dados
data = pd.read_csv('dados_instagram.csv')

# Separar as features (X) e o target (y)
X = data[['followers', 'avg_likes', 'total_likes', 'country', ‘new_post_avg_like’]]
y = data['influence_score']

# Dividir os dados em treino e teste
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# Normalizar os dados (opcional, mas recomendado para regressão linear)
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)

# Criar o modelo de regressão linear
```

```
model = LinearRegression()

# Treinar o modelo
model.fit(X_train_scaled, y_train)

# Fazer previsões
y_pred = model.predict(X_test_scaled)

# Avaliar o modelo
mse = mean_squared_error(y_test, y_pred)
print("Erro quadrático médio:",mse)
```

Configurações do Algoritmo:

Ao todo, os dados foram dispostos de 7 formas diferentes, porém apenas 5 modelos foram gerados de fato.

Dois modelos funcionais tiveram ‘influence_score’ como variável independente, ‘rank’ foi utilizado em outros dois modelos, e, por último a variável 'avg_likes_replace' foi o *target*.

3. Otimização e Ajustes

Na Normalização dos dados, como pode ser visto abaixo, foi usado *scaler* tanto no ‘X’ de teste como no ‘X’ de treinamento.

```
# Normalizando os dados
scaler = StandardScaler()
X_train_scaled = scaler.fit_transform(X_train)
X_test_scaled = scaler.transform(X_test)
```

4. Análise e Visualização dos Resultados

Métricas de Avaliação:

Para a avaliação dos modelos foram calculadas as métricas do R^2, MSE e Erro Absoluto Médio (MAE), bem como os coeficientes do modelo.

No modelo 1, foi retirada a variável *‘60_rate_replace’* já que esta se tratava de porcentagem. Suas métricas de desempenho foram as seguintes:

Mean Absolute Error: 315.5798252817553
Mean Squared Error: 134944.06520954616
RMSE: 367.34733592275603

Já no modelo 2, no qual foi retirado também a variável *‘followers_replace’*, os resultados foram surpreendentemente piores:

Mean Absolute Error: 243.8401313799731
Mean Squared Error: 126369.1452290355
RMSE: 355.48438113232976

Como um todo, os valores dos erros cresceram muito, o que mostra um baixo poder de predição (correta) dos modelos apresentados.

Interpretação dos Coeficientes:

Em geral, os coeficientes tiveram valores baixos, o que indica que as variáveis não seriam boas preditoras da variável independente nos casos abordados.

Com exceção do modelo 7, que teve como variável dependente *‘avg_likes_replace'*, onde os resultados abaixo foram obtidos:

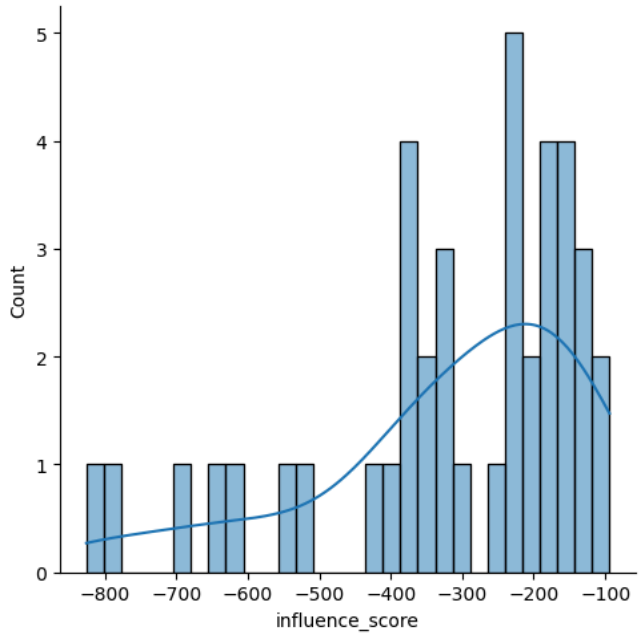
	coef
followers_replace	-57.696956
posts_replace	69.255940
new_post_avg_like_replace	51.700798
total_likes_replace	12.647183
influence_score	8.383834

Os valores acima demonstram que o número de posts está relacionado fortemente com a média de likes que um post recebe, naturalmente.

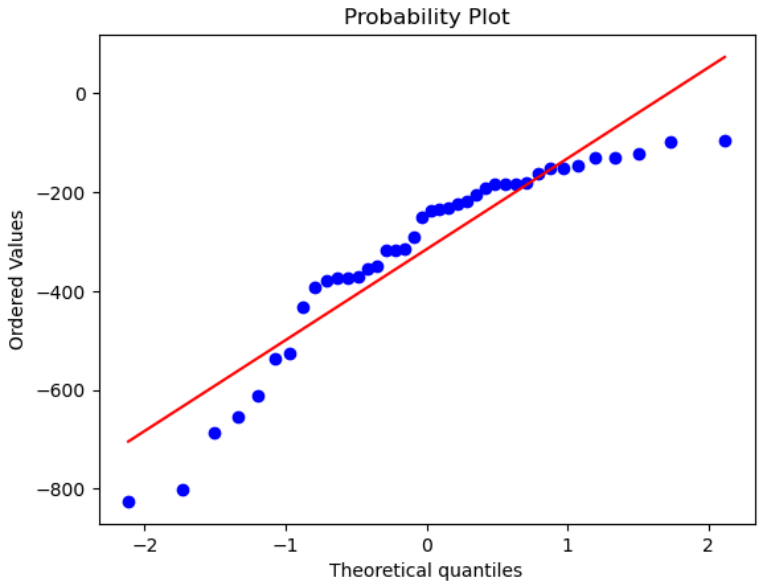
Visualizações Gráficas:

Para uma melhor análise do comportamento dos modelos, foram calculados também seus resíduos e gerado alguns gráficos, já que, caso o comportamento desses resíduos seguisse uma distribuição parecida com a Normal, isso significaria que os modelos estariam bem ajustados.

No primitivo modelo, o gráfico de distribuição dos resíduos foi como demonstrado abaixo:



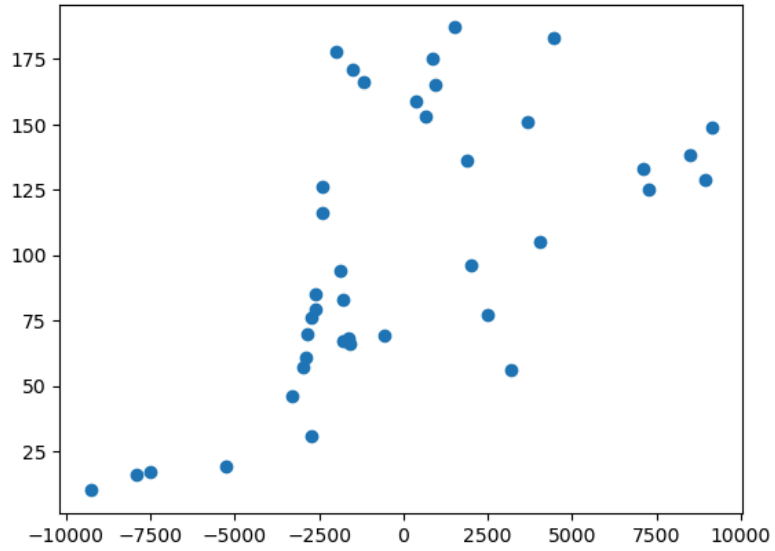
E seu gráfico de verificação da normalidade foi:



Dessa forma, é possível observar que os resíduos não apresentam Distribuição Normal, e o modelo não é muito bom.

No quinto modelo, em que a variável independente era a chamada *rank*, foi também calculado a variável *predictions* usando o modelo da variável de teste X.

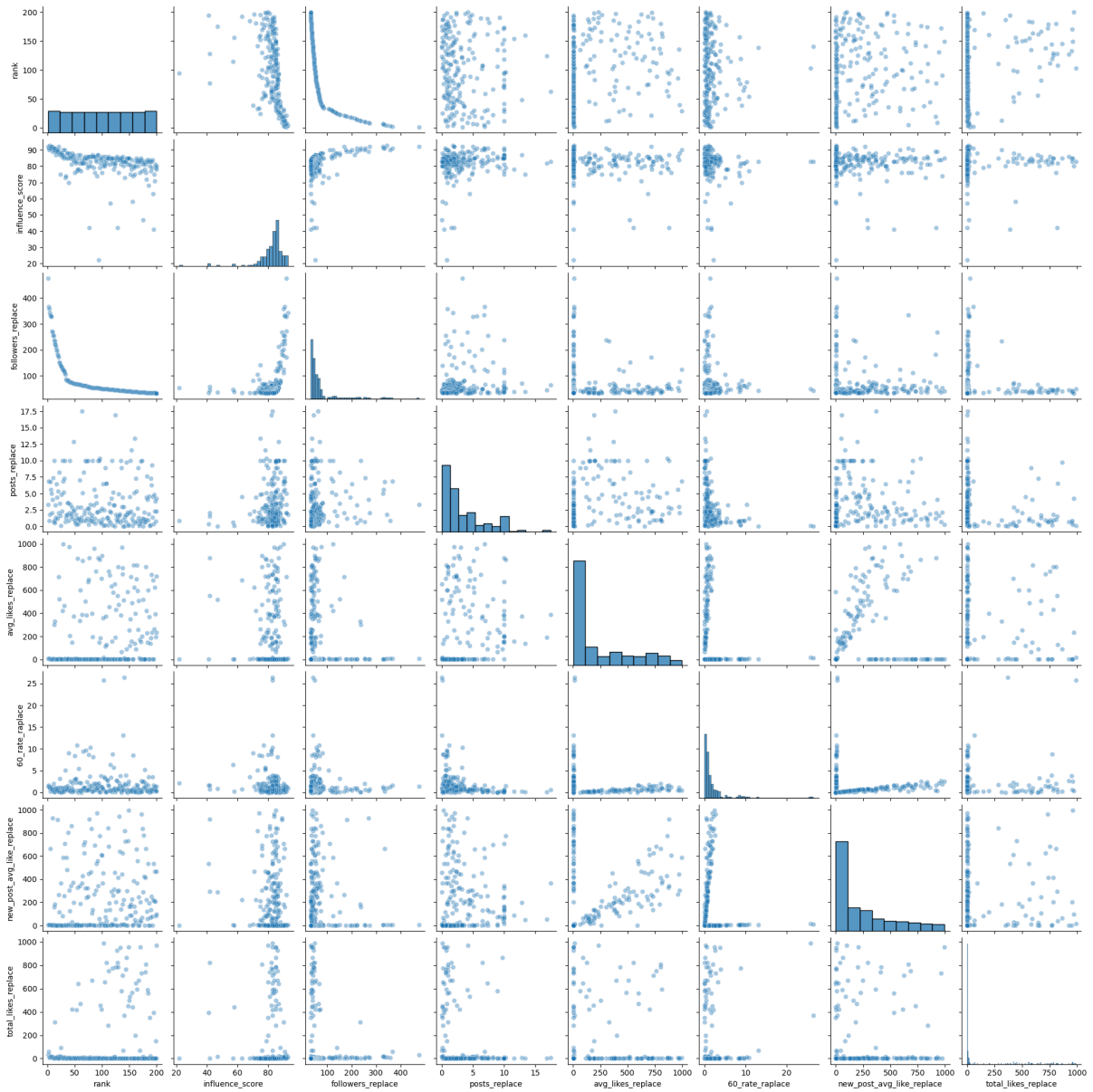
Com o gráfico de dispersão (abaixo) também é possível ver que as variáveis consideradas para tentar predizer o *rank* das contas do Instagram, não são muito boas.



Em conclusão, nenhum dos modelos gerados tiveram desempenho satisfatório, o que provavelmente significa que a Regressão Linear não é a melhor forma de se analisar o banco de dados usado neste projeto, principalmente porque, nenhuma das variáveis apresentaram correlação linear forte entre si, como pode ser visto na análise realizada em Python do notebook Jupyter.

Anexo:

Figura 1:



REFERÊNCIAS

https://repositorio.enap.gov.br/bitstream/1/4788/1/Livro_Regress%C3%A3o%20Linear.pdf

https://www.ime.usp.br/~salles/fatec/estatistica/trabalho/Utiliza%C3%A7%C3%A3o%20da%20regress%C3%A3o%20linear%20como%20ferramenta%20de%20decis%C3%A3o%20na%20gest%C3%A3o%20de%20custos.pdf

https://pepsic.bvsalud.org/scielo.php?script=sci_arttext&pid=S1413-294X2017000200011

https://www.escoladnc.com.br/blog/tuning-de-hiperparametros-como-otimizar-modelos-de-machine-learning/

https://www.datageeks.com.br/ajuste-de-hiperparametros/