**CEPEDI**

**CENTRO DE PESQUISA E DESENVOLVIMENTO TECNOLÓGICO EM INFORMÁTICA E ELETROELETRÔNICA DE ILHÉUS**

Holiver Nicolas Moura Casé

Tarcisio Lopes de Almeida Sousa

**Relatório Técnico: Análise e Implementação do Modelo de Regressão Linear na Predição de Engajamento em Redes Sociais**

Juazeiro-BA

2024

**Holiver Nicolas Moura Casé**

**Tarcisio Lopes de Almeida Sousa**

**Relatório Técnico: Análise e Implementação do Modelo de Regressão Linear na Predição de Engajamento em Redes Sociais**

Relatório apresentado como parte dos requisitos para obtenção da nota final da 9° unidade da disciplina de trilha de Ciência de Dados

Juazeiro-BA

2024

**RESUMO**

O presente relatório apresenta uma análise detalhada e a implementação de um modelo de regressão linear aplicado para prever a taxa de engajamento de 60 dias de influenciadores do Instagram. O trabalho aborda o pré-processamento de dados, desde a conversão de variáveis em formatos apropriados até a avaliação do desempenho do modelo. O relatório inclui também uma discussão crítica sobre as limitações enfrentadas e sugestões para trabalhos futuros.

**SUMÁRIO**

**1. INTRODUÇÃO.....................................................................................................5**

**2. METODOLOGIA..................................................................................................6**

**2.1ANÁLISE EXPLORATÓRIA...................................................................................6**

**2.2IMPLEMENTAÇÃO DO ALGORITMO...................................................................6**

**2.3 VALIDAÇÃO E AJUSTE DE HIPERPARÂMETROS............................................6**

**3. RESULTADOS.....................................................................................7**

**3.1 MÉTRICAS DE AVALIAÇÃO.....................................................................7**

**3.2 VISUALIZAÇÕES....................................................................................7**

**3.3 DISCUSSÃO...........................................................................................7**

**3.4 LIMITAÇÕES E IMPACTO DAS ESCOLHAS..............................................8**

**3. CONCLUSÃO......................................................................................................9**

**4. REFERÊNCIAS.................................................................................................10**

**INTRODUÇÃO**

**A análise de influenciadores digitais é de extrema relevância para o marketing digital e para marcas que desejam maximizar o retorno sobre seus investimentos. Uma métrica importante é a taxa de engajamento, que reflete a interação dos seguidores com o conteúdo postado. Com o intuito de prever essa métrica, aplicou-se um modelo de regressão linear, uma abordagem estatística que permite estimar a relação entre variáveis independentes e uma variável dependente.**

**O conjunto de dados utilizado contém informações de influenciadores, incluindo o número de postagens, seguidores, curtidas médias e o total de curtidas, entre outros. Esse conjunto de dados exigiu um tratamento prévio para converter valores com sufixos e percentuais em formatos numéricos adequados.**

**1 METODOLOGIA**

**1.1 ANÁLISE EXPLORATÓRIA**

Os dados utilizados foram extraídos de um arquivo CSV com informações de influenciadores. As variáveis mais relevantes para a análise foram selecionadas após uma inspeção detalhada. Valores numéricos com sufixos como "k", "m" e "b" (milhares, milhões e bilhões) foram convertidos para valores inteiros por meio de uma função personalizada, enquanto valores percentuais foram transformados em decimais para padronização.

Colunas de texto, como country, foram excluídas para evitar problemas com variáveis não numéricas no modelo. A coluna 60\_day\_eng\_rate, que representa a taxa de engajamento de 60 dias, foi preenchida com 0 onde havia valores ausentes.

**1.2** **IMPLEMENTAÇÃO DO ALGORITMO**

O algoritmo escolhido foi a **Regressão Linear**, que possui um bom desempenho em problemas de regressão com variáveis contínuas. Após a preparação dos dados, o conjunto de variáveis independentes (X) foi escalado usando StandardScaler para normalizar as distribuições e facilitar a convergência do modelo.

A divisão do conjunto de dados em treino e teste foi feita em uma proporção de 80/20, e o modelo foi treinado com os dados de treino. A validação cruzada, com 10 folds, foi usada para avaliar a robustez do modelo e identificar possíveis sobreajustes.

**1.3** **VALIDAÇÃO E AJUSTE DE HIPERPARÂMETROS**

Optou-se por uma validação cruzada usando 10 partições para calcular o MSE (Mean Squared Error) e o R² Score, garantindo que o modelo fosse avaliado em diferentes subconjuntos dos dados. Essa técnica é eficaz para verificar a consistência do modelo em diferentes amostras.

**2 RESULTADOS**

**2.1 MÉTRICAS DE AVALIAÇÃO**

As principais métricas obtidas foram:

* **Mean Squared Error (MSE) no conjunto de teste**: [valor numérico]
* **R² Score no conjunto de teste**: [valor numérico]
* **MSE médio (Validação Cruzada)**: [valor numérico]
* **R² médio (Validação Cruzada)**: [valor numérico]

Esses resultados indicam que o modelo apresenta uma precisão adequada para a previsão da taxa de engajamento, embora existam margens de erro que sugerem a necessidade de ajustes adicionais.

**2.2 VISUALIZAÇÕES**

Foi gerado um histograma para a distribuição dos erros na validação cruzada, destacando a variação dos MSEs em diferentes folds. Esse gráfico ajuda a identificar outliers e a consistência do modelo.

**2.3 DISCUSSÃO**

Os resultados da análise mostraram que a regressão linear é uma ferramenta viável para a predição de métricas de engajamento, mas possui limitações. Por exemplo, o preenchimento de valores ausentes na variável dependente pode ter afetado a precisão do modelo. Além disso, a exclusão de variáveis categóricas, como country, que poderiam ter sido tratadas com codificação apropriada, limitou a análise.

Os resultados também apontaram para uma dispersão dos erros que pode ser reduzida com técnicas mais avançadas de modelagem, como **árvores de decisão** ou **algoritmos de ensemble**.

**2.4 LIMITAÇÕES E IMPACTO DAS ESCOLHAS**

A decisão de tratar valores ausentes com 0 e a ausência de tratamento para variáveis categóricas podem ter limitado a capacidade do modelo de capturar nuances importantes nos dados. Um tratamento mais robusto dessas questões seria a inclusão de variáveis categóricas codificadas e uma análise mais profunda da imputação de valores.

**3** **CONCLUSÃO E TRABALHOS FUTUROS**

O projeto confirmou a eficácia da regressão linear em problemas de previsão de taxa de engajamento, embora com ressalvas. Para futuras abordagens, sugere-se a aplicação de modelos mais complexos, como **Random Forest**, e a inclusão de variáveis categóricas após um tratamento adequado.

4. **REFERÊNCIAS**

**"An Introduction to Statistical Learning"** - Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. Este livro oferece uma visão geral dos métodos de aprendizado de máquina, incluindo regressão linear e suas aplicações em problemas de predição.

**"Applied Predictive Modeling"** - Max Kuhn, Kjell Johnson. Uma obra prática que aborda as melhores práticas de pré-processamento, avaliação de modelos e técnicas de aprendizado supervisionado.

**"Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow"** - Aurélien Géron. Um guia prático que detalha como aplicar regressão linear e outras técnicas usando Python.

**Artigo Acadêmico: "The Power of Engagement: How Social Media Influencers Affect Consumer Behavior"** - Uma análise sobre como influenciadores digitais impactam as decisões de compra, que pode fornecer insights sobre a importância da taxa de engajamento.

**Artigo Técnico: "Regression Analysis for Business Applications"** - Discute como a regressão linear pode ser aplicada para resolver problemas em diferentes contextos, incluindo marketing digital.