Relatório de Análise de Modelos de Classificação

Candidato: Felipe Miguel

Objetivo

O objetivo deste desafio foi aplicar três modelos de classificação no conjunto de dados com o intuito de prever a variável Compra (0 ou 1), que indica se uma compra foi realizada (1) ou não (0), com base em variáveis como Idade, Renda Anual, Gênero, Tempo no Site, entre outras.

Etapas do Código

Carregamento e Exploração dos Dados
O dataset foi carregado usando a biblioteca pandas. Posteriormente, foram verificadas as informações do conjunto de dados, incluindo a existência de dados nulos, outliers e variáveis categóricas. Variáveis como Gênero e Anúncio Clicado foram convertidas para valores numéricos utilizando o LabelEncoder.

Pré-processamento
Foram excluídos valores nulos e outliers das variáveis, como a coluna Tempo no Site (min) que apresentava valores zero. Além disso, foi realizado um gráfico de correlação para verificar possíveis relações entre as variáveis e a variável alvo (Compra).

Divisão dos Dados
O conjunto de dados foi dividido em variáveis independentes (X) e a variável dependente (Compra (0 ou 1), y). Em seguida, os dados foram divididos em treino e teste, utilizando 5% do total para teste.

Criação e Treinamento dos Modelos

Modelo de Regressão Logística: Utilizado para prever a variável Compra com base em uma função linear das variáveis de entrada.

Modelo de Árvore de Decisão: Modelo que divide os dados em subgrupos com base em perguntas binárias, criando uma estrutura hierárquica.

Modelo de Random Forest: Envolve a criação de várias árvores de decisão, com cada árvore sendo treinada com um subconjunto diferente dos dados. A decisão final é tomada pela maioria das árvores.

Avaliação dos Modelos A avaliação dos modelos foi realizada utilizando a acurácia, a matriz de confusão e o F1-score para cada modelo.

Resultados dos Modelos

Modelo de Regressão Logística:

Acurácia: 66,67%

Matriz de Confusão: 6 0

3 0

F1-Score: 0.0
A regressão logística teve uma performance limitada, com um F1-score de 0.0, o que indica uma classificação muito baixa para a classe positiva (Compra = 1).

Modelo de Árvore de Decisão:

Acurácia: 88,89%

Matriz de Confusão: 5 1

0 3

F1-Score: 0.86
A árvore de decisão se saiu melhor que a regressão logística, com alta acurácia e um bom F1-score, indicando uma boa capacidade de distinguir entre as classes, especialmente a classe Compra = 1.

Modelo de Random Forest:

Acurácia: 66,67%

Matriz de Confusão: 5 1

2 1

F1-Score: 0.4
O modelo de random forest teve um desempenho similar ao da regressão logística em termos de acurácia, indicando que o modelo teve dificuldades para classificar corretamente a classe positiva (Compra = 1).

Interpretação dos Resultados

Desempenho do Modelo:
O modelo de Árvore de Decisão foi o melhor em termos de acurácia e F1-score, indicando que ele é o mais eficaz para este conjunto de dados. A regressão logística e o random forest apresentaram desempenhos mais fracos, com o modelo de regressão logística especialmente com dificuldades em identificar a classe positiva.

Análise da Matriz de Confusão:
A matriz de confusão mostra que tanto a regressão logística quanto o random forest falharam em classificar corretamente os exemplos da classe Compra = 1, enquanto o modelo de árvore de decisão teve um desempenho bem melhor, acertando a maioria das classificações.

Conclusão

Os resultados dos modelos variaram, com a árvore de decisão se destacando. No entanto, o desempenho pode ser melhorado com um maior volume de dados e um tratamento mais eficiente dos valores nulos. Mais dados e uma análise mais aprofundada dos hiperparâmetros podem aprimorar as previsões e aumentar a precisão dos modelos.