Interpretações dos resultados obtidos

Resumo Estatístico do Dataset

1. **Idade**:
   - Média: 38.51 anos, com desvio padrão de 12.68.
   - Intervalo: entre 18 e 59 anos.
   - Percentis: 25% têm até 28 anos, 50% têm 38 anos (mediana), e 75% têm até 50 anos.
2. **Renda Anual**:
   - Média: $58.253,96, com desvio padrão de $25.612,06.
   - Valores variam de $30.000 a $100.000.
   - Os quartis indicam que a maior parte das pessoas ganha $30.000 (25%) a $70.000 (75%).
3. **Tempo no Site (min)**:
   - Média: 17.35 minutos, mas com desvio padrão de 7.72.
   - Há um valor atípico: -1.0 minutos no mínimo (possivelmente um erro de dados).
   - A maioria dos clientes gasta entre 10.86 e 23.88 minutos no site.
4. **Compra (0 ou 1)**:
   - Apenas 33% (média de 0.33) dos clientes realizaram uma compra.
   - A classe está desbalanceada (mais 0s do que 1s).

Gráficos de Distribuição

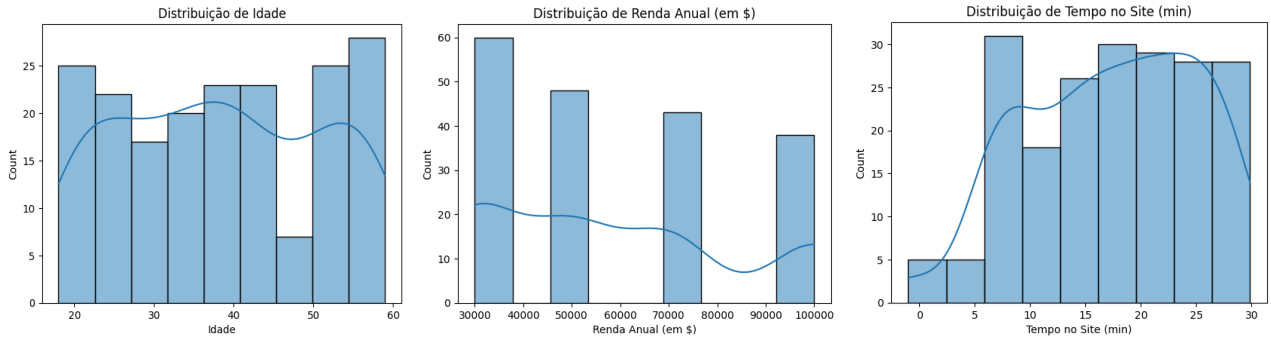

Gráfico 1: Distribuição de Idade

Forma da Distribuição:

A distribuição da variável "Idade" apresenta um padrão quase uniforme, com uma leve oscilação nos intervalos.

A faixa de idade entre 18 e 60 anos é bem representada, sem grandes picos ou quedas abruptas.

Faixas Destacadas:

Há uma leve concentração de indivíduos nas faixas próximas aos 20 anos e acima de 50 anos.

O intervalo dos 40 anos apresenta uma leve redução no número de observações.

Impacto nos Modelos:

A distribuição bem balanceada entre as faixas etárias sugere que a variável "Idade" pode ter impacto nos modelos preditivos, especialmente se houver relação direta entre idade e compra.

Gráfico 2: Distribuição de Renda Anual (em $)

Forma da Distribuição:

A distribuição de "Renda Anual" não é uniforme, com uma concentração evidente de indivíduos com renda em torno de $30.000 a $40.000.

À medida que a renda aumenta, o número de observações diminui, com um leve aumento para rendas próximas a $100.000.

Faixas Destacadas:

A faixa de $30.000 domina a distribuição, indicando que a maioria dos indivíduos tem uma renda mais baixa.

A distribuição apresenta um viés negativo, com a cauda mais longa para rendas altas.

Impacto nos Modelos:

A concentração na faixa de $30.000 sugere que a "Renda Anual" pode ser um fator discriminante no comportamento de compra.

A baixa representatividade de rendas altas pode dificultar a identificação de padrões para indivíduos desse grupo.

Gráfico 3: Distribuição de Tempo no Site (min)

Forma da Distribuição:

A curva de densidade suavizada sobre o histograma mostra uma distribuição aproximadamente unimodal, indicando um comportamento homogêneo entre os usuários no tempo de permanência.

Faixas Destacadas:

A maior concentração de usuários está na faixa entre **10 e 20 minutos**, com o pico maior em torno de 15 minutos. Isso indica que a maioria dos usuários

passa um tempo intermediário no site. Há uma leve diminuição no número de usuários conforme o tempo aumenta além de 20 minutos. No entanto, há uma queda acentuada após 30 minutos.

Impacto nos Modelos:

Poucos usuários passam menos de 5 minutos no site, sugerindo que a primeira interação pode não ser suficiente para engajar o público nessa faixa de tempo.

Gráfico de Pairplot

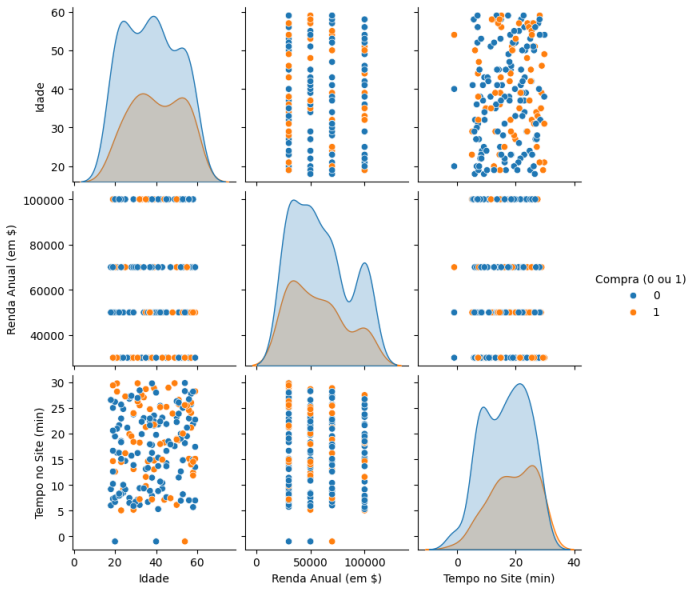

Gráfico gerado:

O gráfico exibe a relação entre **Idade**, **Renda Anual** e **Tempo no Site**, com as cores diferenciando quem realizou compras (laranja, 1) e quem não realizou (azul, 0).

Análise visual:

1. **Idade vs. Compra**:
   - Não há um padrão claro, mas compradores (1) parecem concentrados em faixas intermediárias de idade.
2. **Renda Anual vs. Compra**:
   - Parece haver maior concentração de compradores com rendas anuais entre $40.000 e $70.000.
3. **Tempo no Site vs. Compra**:
   - Quem comprou tende a passar mais tempo no site, enquanto quem não comprou apresenta uma distribuição mais variada.
4. **Densidade Geral**:

    - Para todas as variáveis, a sobreposição das distribuições (azul e laranja) indica que os padrões não são fortemente separáveis.

Modelos de Machine Learning

1. Regressão Logística

Resultados:

- **Acurácia**: 69.1%.
- **Recall e Precision**:
    - Para a classe 0 (não comprou), o modelo foi excelente, com alta precisão (70%) e recall (97%).
    - Para a classe 1 (comprou), o modelo falhou completamente (0% em todas as métricas). Isso ocorre devido ao desbalanceamento das classes.

Conclusão:

- A regressão logística não conseguiu identificar corretamente os compradores (1) porque a classe minoritária é negligenciada.

2. Árvore de Decisão

Resultados:

- **Acurácia**: 55% (pior que a regressão logística).
- **Recall**:
    - Classe 0: 86% (bom).
    - Classe 1: 12% (muito baixo).

Otimização:

- Após a busca por hiperparâmetros, os melhores valores foram:
    - max_depth: 5 (limita a profundidade da árvore para evitar overfitting).
    - min_samples_leaf: 4 (mínimo de amostras em cada folha).
    - min_samples_split: 2 (mínimo para dividir um nó).

Conclusão:

- A árvore de decisão não conseguiu modelar bem a classe 1, mesmo com otimização.

3. Random Forest

Resultados:

- **Acurácia**: 65% (ligeiramente melhor que a árvore de decisão).
- **Importância das Variáveis**:

    - **Tempo no Site** foi a mais relevante (47.2%).
    - **Idade** veio em seguida (30.9%).
    - **Renda Anual** (11.9%) e **Gênero** (5.2%) foram menos importantes.

Otimização:

- Melhores parâmetros encontrados:
    - max_depth: 10
    - min_samples_leaf: 4
    - min_samples_split: 2
    - n_estimators: 200 (número de árvores na floresta).

Validação Cruzada:

- A média da acurácia foi 64.47%, indicando consistência.

Conclusão:

- Random Forest teve um desempenho intermediário, mas ainda não conseguiu lidar bem com o desbalanceamento.

Recomendações

1. **Tratar o Desbalanceamento**:
    - Aplicar técnicas como oversampling (SMOTE) ou undersampling para equilibrar as classes.
2. **Análise de Dados**:
    - Revisar os valores atípicos, como o tempo no site negativo (-1), para garantir qualidade dos dados.
3. **Modelos Alternativos**:
    - Experimentar modelos robustos ao desbalanceamento, como XGBoost ou LightGBM.
4. **Feature Engineering**:
    - Criar novas variáveis ou combinar existentes para capturar melhor as relações, como proporção de "Tempo no Site" pela "Idade".
5. **Avaliação de Métricas**:
    - Usar métricas como F1-score ou AUC-ROC, que são melhores para classes desbalanceadas.
6. **Análise de Gênero**:
    - Considerar o impacto do gênero (a importância é baixa, mas pode ter interações com outras variáveis).