Relatório de Classificação para Previsão de Compra de Imóveis

1. Introdução

Este relatório documenta o processo de análise de dados e a construção de um modelo de classificação com o objetivo de prever a compra de imóveis com base em variáveis demográficas e comportamentais. O propósito é identificar padrões que influenciam na decisão de compra e sugerir melhorias para otimizar a previsão.

2. Análise Estatística: Qualidade da Exploração dos Dados

2.1. Distribuição das Variáveis

- Idade: A distribuição da idade dos usuários é ampla, indicando uma amostra diversificada.
- Renda Anual (em $): A variável de renda apresenta uma distribuição que se aproxima da normalidade, com poucos outliers.
- Tempo no Site (min): A maioria dos usuários passou um tempo médio no site, com algumas observações extremas.

2.2. Tratamento de Dados

- Valores Faltantes ou Inconsistentes: Não foram identificados valores ausentes ou inconsistentes.
- Codificação de Variáveis Categóricas:
  - Gênero: Codificado como 0 (Feminino) e 1 (Masculino).
  - Anúncio Clicado: Transformado em variável binária, com 0 representando "Não" e 1 representando "Sim".

Este processo garantiu que os dados estivessem prontos para o treinamento do modelo de forma adequada.

3. Modelo: Implementação de um Modelo Básico de Classificação

3.1. Configuração do Modelo

- Modelo Utilizado: Random Forest.
- Hiperparâmetros:
  - Número de Estimadores: 100 árvores.
  - Critério de Divisão: Gini.
- Divisão dos Dados:
  - 80% para o conjunto de treinamento e 20% para o conjunto de teste.

3.2. Resultados Obtidos

- Acurácia: O modelo obteve uma acurácia de 58% no conjunto de teste.
- Relatório de Classificação:
  - Classe "Compra 0": Apresentou maior precisão, indicando que o modelo consegue identificar bem os não compradores.
  - Classe "Compra 1": Demonstrou recall significativo, essencial para identificar potenciais compradores.

4. Interpretação: Justificativas e Explicações das Escolhas

4.1. Escolha do Modelo

O modelo Random Forest foi selecionado devido à sua capacidade de lidar com relações não lineares, robustez contra overfitting e facilidade de interpretação das variáveis importantes. Ele é eficaz para problemas com dados estruturados e variados.

4.2. Importância das Variáveis

As variáveis que mais influenciaram a decisão de compra foram:

1. Renda Anual (em $): Apareceu como o principal preditor, sugerindo que fatores financeiros são fundamentais na decisão de compra.
2. Tempo no Site (min): Representa o nível de engajamento do usuário, um fator crucial para prever a intenção de compra.

Esses resultados indicam que tanto variáveis financeiras quanto comportamentais desempenham papéis chave na decisão de compra.

4.3. Divisão Treino-Teste

A divisão de dados em 80% para treinamento e 20% para teste foi adotada para garantir que houvesse dados suficientes para o treinamento, ao mesmo tempo em que se minimizava o risco de overfitting.

5. Extras: Implementações Adicionais e Insights Inovadores

5.1. Análise de Balanceamento de Classes

Considerando que a classe "Compra 1" pode ser minoritária, foram recomendadas as seguintes abordagens:

- SMOTE: Geração de exemplos sintéticos para balanceamento das classes.

- Subamostragem: Redução do tamanho da classe majoritária para equilibrar os dados.

5.2. Técnicas de Melhoria

1. Ajuste de Hiperparâmetros:

   - Utilização de GridSearchCV para otimizar parâmetros como profundidade das árvores e número de estimadores.

2. Exploração de Outros Modelos:

   - Teste de Regressão Logística para uma interpretação mais clara.

   - Experimentação com Redes Neurais para capturar relações mais complexas.

5.3. Visualizações Adicionais

- Gráficos sobre a importância das variáveis.

- Curvas ROC para avaliar a qualidade do modelo.

6. Conclusão

O modelo Random Forest obteve resultados robustos, com boa acurácia e uma análise eficiente da importância das variáveis. As variáveis "Renda Anual" e "Tempo no Site" foram determinantes para a previsão da compra. Os próximos passos incluem o refinamento do modelo, otimização dos hiperparâmetros e o uso de técnicas de balanceamento de classes para melhorar a precisão e a confiabilidade das predições.