Exercício: Análise de Inadimplência no Conjunto de Dados Credit Card Default com Regressão Logística

Objetivos

1. Implementar uma regressão logística para prever a inadimplência (`default`) com base nas variáveis `balance`, `student`, `income` e outras variáveis relevantes.
2. Analisar os coeficientes do modelo e a significância estatística das variáveis.
3. Avaliar o desempenho do modelo usando métricas de classificação e interpretação dos resultados.

Instruções

1. **Carregar e explorar os dados**:
    - Baixe o conjunto de dados **Credit Card Default Data Set**.

```
from ISLP import load_data

# Carregar os dados
df = load_data('Default')
```

    - Carregue os dados em um DataFrame e verifique o resumo estatístico e a distribuição das variáveis principais (`balance`, `income`, `student`, etc.).
    - Verifique valores ausentes e tipos de dados, e faça o tratamento adequado, caso necessário.
2. **Preparação dos dados**:
    - Transforme a variável `default` para uma variável binária, se necessário (e.g., `Yes` para `1` e `No` para `0`).
    - Codifique a variável categórica `student` em uma representação binária (`0` para não-estudante, `1` para estudante).

```
# Preparação dos dados
# Converter a variável 'default' para binária (1 para Yes, 0 para No)
df['default'] = df['default'].apply(lambda x: 1 if x == 'Yes' else 0)

# Converter a variável 'student' para binária
df['student'] = df['student'].apply(lambda x: 1 if x == 'Yes' else 0)
```

3. **Dividir os dados em treino e teste**:
    - Divida os dados em conjuntos de treino e teste (e.g., 70% treino, 30% teste).
4. **Construir o modelo de regressão logística**:

- Implemente a regressão logística usando uma biblioteca como `Scikit-Learn` (`LogisticRegression`).
- Treine o modelo no conjunto de treino usando `balance`, `student`, `income` como variáveis preditoras de forma independente e depois com todas juntas.

5. **Análise dos coeficientes e interpretação**:
    - Exiba os coeficientes do modelo e interprete o impacto de cada variável sozinha na probabilidade de inadimplência e depois para as variáveis em conjunto:
        - Interprete o coeficiente `balance` (qual é o impacto de um aumento de uma unidade no saldo devedor na probabilidade de inadimplência?).
        - Avalie o coeficiente `student`: qual é o impacto do status de estudante na probabilidade de inadimplência?

6. **Avaliação do modelo**:
    - Gere previsões para o conjunto de teste (para as variáveis sozinhas e depois em conjunto).
    - Calcule e interprete as seguintes métricas de classificação:
        - **Acurácia**: Qual é a proporção de previsões corretas?
        - **Matriz de Confusão**: Quantos são verdadeiros positivos, verdadeiros negativos, falsos positivos e falsos negativos?

7. **Ajuste do limiar de decisão**:
    - Experimente alterar o limiar de decisão (e.g., prever inadimplência se `p > 0.3` em vez de `p > 0.5`).
    - Observe e discuta como isso afeta a precisão, a revocação e outras métricas de desempenho do modelo apenas para todas as variáveis juntas.
8. **Relatório e Conclusão**:
    - Resuma as principais conclusões do modelo. Alguma variável tem mais impacto sobre a inadimplência?