Exercício 4: Avaliação de Classificação em Dados Desbalanceados

Exercício 4: Avaliação de Classificação em Dados Desbalanceados

(a) No contexto de detecção de fraude (dados tipicamente desbalanceados), por que a acurácia padrão geralmente não é uma métrica adequada para avaliar a performance do classificador?

Em dados desbalanceados, como os encontrados na detecção de fraude onde transações fraudulentas são raras em comparação com as legítimas, a **acurácia padrão pode ser enganosa**. Um classificador trivial que sempre prevê a classe majoritária (transação legítima) pode alcançar uma alta acurácia (próxima de 100%) simplesmente por classificar corretamente a grande maioria das instâncias, sem identificar nenhuma transação fraudulenta. Isso ocorre porque a acurácia é definida como a proporção de previsões corretas sobre o número total de instâncias, e em datasets desbalanceados, a classe majoritária domina essa métrica. O classificador falharia em seu objetivo principal, que é detectar a classe minoritária (fraude), apesar de apresentar uma alta acurácia.

(b) Defina Precisão, Recall e F1-Score, interpretando-os especificamente em termos da identificação de transações fraudulentas. Em que situações práticas (contexto de negócio) você poderia priorizar Recall sobre Precisão, ou vice-versa?

Em termos de identificação de transações fraudulentas:

- **Precisão** (Precision) é a proporção de transações classificadas como fraudulentas que são realmente fraudulentas. Matematicamente, é definida como:

$$\text{Precisão} = \frac{\text{Número de Transações Fraudulentas Corretamente Identificadas}}{\text{Número Total de Transações Classificadas como Fraudulentas}}$$

  Uma alta precisão indica que, quando o modelo sinaliza uma transação como fraude, é muito provável que ela seja realmente fraudulenta, minimizando os **falsos positivos** (transações legítimas erroneamente marcadas como fraude).

- **Recall** (Revocação ou Sensibilidade) é a proporção de todas as transações fraudulentas reais que foram corretamente identificadas pelo modelo. Matematicamente, é definida como:

$$\text{Recall} = \frac{\text{Número de Transações Fraudulentas Corretamente Identificadas}}{\text{Número Total de Transações Fraudulentas Reais}}$$

  Um alto recall indica que o modelo é eficaz em detectar a maioria das transações fraudulentas, minimizando os **falsos negativos** (transações fraudulentas que não foram detectadas).

- **F1-Score** é a média harmônica entre a Precisão e o Recall. É uma métrica que busca um equilíbrio entre as duas:

$$\text{F1-Score} = 2 \times \frac{\text{Precisão} \times \text{Recall}}{\text{Precisão} + \text{Recall}}$$

  O F1-Score é útil quando você deseja uma métrica única que leve em consideração tanto os falsos positivos quanto os falsos negativos.

  Em situações práticas:

  – Você poderia priorizar o **Recall** sobre a Precisão quando o **custo de não detectar uma transação fraudulenta é muito alto**. Por exemplo, em cenários de segurança de alto risco ou em transações financeiras de grande valor, perder uma fraude pode resultar em perdas financeiras significativas, danos à reputação ou outras consequências graves. Nesses casos, é preferível sinalizar mais transações como suspeitas (mesmo que algumas

sejam legítimas - falsos positivos) para garantir que a maioria das fraudes seja detectada. A triagem manual adicional dos casos sinalizados pode ser aceitável para minimizar os falsos negativos.

- Você poderia priorizar a **Precisão** sobre o Recall quando o **custo de investigar falsos positivos é muito alto**. Por exemplo, se cada transação sinalizada como suspeita requer uma investigação manual dispendiosa e demorada, ou se a sinalização incorreta de transações legítimas causa grande inconveniência ao cliente e pode levar à insatisfação e perda de clientes. Nesses casos, é preferível ter um modelo que seja altamente confiável em suas previsões de fraude, mesmo que isso signifique perder algumas transações fraudulentas.

(c) Descreva como você usaria os scores de um modelo probabilístico (e.g., $p(\text{fraude}|\mathbf{x})$ da Regressão Logística) e os rótulos verdadeiros para construir uma curva Precisão-Recall (PR). O que a Área Sob a Curva PR (AUC-PR) representa? Por que ela é frequentemente preferível à curva ROC para datasets muito desbalanceados?

Para construir uma curva Precisão-Recall (PR) usando os scores de um modelo probabilístico e os rótulos verdadeiros:

(a) Para cada instância no dataset de teste, o modelo probabilístico fornece um score que representa a probabilidade de ser uma transação fraudulenta, $p(\text{fraude}|\mathbf{x})$.

(b) Ordenam-se todas as instâncias do dataset de teste com base nesses scores em ordem decrescente.

(c) Itera-se através de diferentes limiares (thresholds) de probabilidade. Para cada limiar:

- Consideram-se todas as instâncias com score acima do limiar como classificadas como "fraude" (positivo previsto).
- Consideram-se todas as instâncias com score abaixo do limiar como classificadas como "não fraude" (negativo previsto).
- Calculam-se a Precisão e o Recall com base nas verdadeiras etiquetas de fraude para essa classificação com o limiar atual.

(d) Plota-se um gráfico com o Recall no eixo x e a Precisão no eixo y para cada limiar considerado. A curva resultante é a curva Precisão-Recall (PR).

A **Área Sob a Curva PR (AUC-PR)** representa a **performance média do classificador em diferentes trade-offs entre Precisão e Recall**. Um valor de AUC-PR mais alto indica que o modelo tem um bom desempenho tanto em termos de precisão (baixa taxa de falsos positivos) quanto de recall (baixa taxa de falsos negativos) em uma variedade de limiares de classificação.

A AUC-PR é frequentemente preferível à curva ROC (Receiver Operating Characteristic) para datasets muito desbalanceados porque a **curva PR se concentra mais na classe positiva (minoritária)**, que é o foco em problemas como detecção de fraude. A curva ROC plota a Taxa de Verdadeiros Positivos (TPR ou Recall) contra a Taxa de Falsos Positivos (FPR). Em datasets desbalanceados, mesmo um classificador ruim pode ter um FPR baixo porque o número total de negativos é muito grande. Pequenas mudanças no número de falsos positivos podem resultar em grandes mudanças na Precisão, tornando a curva PR mais sensível ao desempenho do classificador na classe minoritária. A AUC-PR, portanto, fornece uma avaliação mais realista do quão bem o modelo está performando na detecção da classe de interesse (fraude) em cenários desbalanceados.