1. 히트맵(상관계수 행렬) 기반 인사이트

- **Class(사기 여부)와 가장 강한 상관관계를 가진 변수(V14, V17 등)**를 즉시 파악할 수 있었음.
    - 예) V14, V17이 음의 상관관계가 매우 강함.
- **다른 피처(V1~V28)**는 대부분 Class와 약한 상관관계.
    - → **사기 탐지에선 V14, V17이 "핵심 단서"로 작용**할 가능성 큼.
- **V14, V17 등에서 이상치(outlier)를 검출**하면,
  "정상과 다른 비정상적 거래" 패턴을 효과적으로 포착 가능.

---

2. Precision-Recall 임계값 곡선 기반 인사이트

- **임계값이 낮을 때:**
    - 거의 모든 거래를 "사기"로 예측 → Recall(재현율)은 거의 1에 가깝게 유지
    - Precision(정밀도)은 매우 낮음 → 실제 사기는 소수인데 너무 많이 잡아내서, 정상 거래도 사기라고 많이 오탐지
- **임계값을 높일 때:**
    - Precision은 오르고 Recall은 급격히 떨어짐
    - **실제 사기만 골라내는 능력(정밀도)은 올라가지만, 진짜 사기도 일부 놓칠 수 있음**

---

3. 히트맵+PR그래프 결합 실무 인사이트

- **사기 탐지의 '핵심 단서'(V14, V17)를 중점 활용**하여
  모델의 성능(특히 recall)을 극대화할 수 있다.
- 하지만 **데이터 자체가 극도로 불균형**(사기:정상=1:500 이상)이므로
  임계값을 낮추면 실제 사기는 거의 다 잡지만,
  정상 거래를 사기라고 오탐지할 확률이 매우 높음(precision↓).
- **V14, V17 등 중요 피처의 이상치(outlier)는 실제 사기 거래의 중요한 패턴**일 수 있으므로
  이상치를 "무작정 제거"하기보다는,
  별도 변수로 관리하거나, 사기탐지 특화 특성으로 모델에 반영하는 게 효과적일 수 있음.
- **임계값 선택은 "비즈니스 위험 허용도"에 따라 달라야 한다.**
  - "실제 사기 거래를 절대 놓치면 안 된다"(Recall 중시)면 임계값 낮추기
  - "고객 불편을 최소화"(Precision 중시)면 임계값 높이기
  - 실제 업무에서는 둘의 균형점에서 threshold를 선정

결론

1. 신용카드 사기 데이터에서 상관계수 분석 결과, V14·V17이 핵심 신호 역할을 한다는 점이 뚜렷하며, 실제 이상치 탐지로 중요한 사기 패턴도 추출할 수 있다.

2. Precision-Recall 곡선으로 threshold를 조정하면, recall을 높이면 실제 사기는 거의 모두 잡지만 오탐이 매우 많아지고, precision을 높이면 실제 사기를 놓치게 되므로,

3. **비즈니스 목적에 맞는 임계값 선택이 필수**다.
   이상치 행을 별도 특성으로 적극 활용하고, V14·V17 등 주요 변수에 대한 추가적인 탐색 및 피처 엔지니어링이 모델의 실전 성능 향상에 핵심 역할을 할 것이다."