반도체 칩의 불량 예측

반도체 공정에서 생산성을 올리기 위하여 불량을 미리 탐지하는 것은 매우 중요하다. 본 프로젝트에서는 반도체 칩 공장의 생산 라인에서 수집된 센서 데이터를 바탕으로 불량제품을 탐지하는 모델을 개발하고 평가한다.

A사의 반도체 공장에서는 웨이퍼가 여러 단계를 거쳐 가공되면서 최종적으로 칩이 생산된다. 가공 단계에서는 다양한 장비가 웨이퍼에 필요한 처리를 해주고, 각 장비에 설치된 센서에서는 가공 환경 및 결과에 대한 정보가 수집되어 중앙 서버에 저장된다. 하지만 각 센서의 특징과 동작 방식은 장비업체의 보안 사항이기 때문에, 생산업체에서는 장비의 유지 관리를 위해 측정된 값만 알 수 있을 뿐 센서 자체에 대한 정보는 알 수 없다. 이러한 과정을 거쳐 생산된 칩은 최종적으로 테스트 단계를 거쳐 정상(normal)과 불량(defect)으로 판단된다. 최종 테스트에 들어가는 비용이 높기 때문에, 가능하면 불량 제품은 사전에 선별하여 정상 제품만 테스트하기를 원한다.

구체적으로, 칩의 생산 비용(P)은 100원이고, 테스트 비용(Q)는 1,000원이다. 하나의 칩을 생산과 테스트를 통해 제품으로 만들게 되면 총 비용(P+Q)은 1,100원이다. 이 칩의 판매가격(S)는 2,000원으로, 하나의 정상 칩을 판매하면 900원의 이익(S-P-Q)을 얻는다. 칩이 테스트에서 불량으로 판정된다면 1,100원의 손해가 나지만, 테스트 없이 폐기 처분한다면 100원의 손해만을 내게 된다. 만일 불량 제품을 테스트 전에 선별하여 폐기할 수 있다면 손해를 줄일 수 있다. 대신 정상 제품을 폐기한다면 900원의 이익을 놓치고 100원의 손해를 보게 된다.

이러한 문제를 해결하기 위해 가공 단계에서 얻어진 센서 정보를 바탕으로 제품의 불량 여부를 사전에 판단하는 기계학습 모델을 만들고자 한다. 주어진 데이터는 다음과 같다.

파일	설명	변수
data98_semi_train.csv	훈련 데이터셋	● Label: 최종 테스트에서 판단된 불량 여부 ● v001~v590: 해당 제품에 대한 센서 데이터
data98_semi_test.csv	평가 데이터셋	● v001~v590: 해당 제품에 대한 센서 데이터

주어진 훈련 데이터를 이용하여 예측 모델의 성능 및 비용절감 효과를 예상하며, 실제 평가 데이터에 대하여 예측을 진행하는 것을 목표로 한다. 아래의 사항을 블랙보드를 통해 프로젝트 결과물로 제출한다.

- 보고서: 아래의 문제에 대한 답변, pdf 형식
- 프로그램 코드: 분석에 사용한 실제 코드, ipynb 형식 혹은 py 형식
- 평가데이터에 대한 예측값: 0은 normal, 1은 defect로 표현하여, 평가 데이터 표본의 순서에 따라 한 줄에 하나의 예측값을 적어 텍스트 파일 (txt) 형식으로 제출. 파일 이름은 "project01_학번.txt"로 정하여 제출.

아래의 질문에 답하여 프로젝트를 수행하시오.

(1) 모델을 이용한 사전 예측없이 모든 칩에 대하여 실제 테스트를 수행한다고 하자. 훈련 데이터 셋을 바탕으로 총 이득을 계산하면 얼마인가? 데이터 분석을 통하여 불량을 완벽히 판별하는 이상적인 모델을 만들었다고 가정했을 때의 총 이득은 얼마인가? 만일, 모델이 90%의 Recall과 90%의 Precision을 가질 때의 총 이득은 얼마인가?

(2) 필요한 EDA를 수행하고 훈련 데이터를 분석에 사용할 수 있는 데이터로 전처리하시오. 각 변수의 분포 확인 및 시각화, 데이터의 변환, 이상치 탐색 및 결측치 처리, 통계 분석을 통한 변수의 선별 등을 포함한다. EDA와 전처리 과정을 설명하고, 특이 사항을 보고하시오. 최종적으로 전처리된 데이터의 표본의 수와 변수의 수를 보고하시오.

(3) 전처리된 데이터를 바탕으로 불량에 대한 예측 모델을 만드시오. 다양한 모델을 시도하고 적절한 검증 과정을 거쳐 최종적으로 하나의 모델을 선택하시오. 어떠한 모델을 시도하였고, 어떠한 검증 방식을 사용하였는지 설명하시오. 특히, 모델 선택을 위한 성능 지표를 어떻게 결정하였는지 설명하시오. 선택된 모델에 대하여 훈련 데이터에서의 성능과 검증 데이터에서의 성능을 보고하시오. (중요: 대부분의 이진 분류에서는 확률값 0.5를 기준으로 하여 분류하지만, 실제로는 꼭 그럴 필요는 없다. 다른 값을 사용하여 분류하여 원하는 성능을 달성할 수도 있다.)

(4) 위의 모델을 이용하여 평가 데이터를 분류하고 그 결과를 파일에 저장하여 제출하시오. (주의: 평가 데이터도 전처리가 필요한데, 이때는 평가데이터의 요약 정보를 이용할 수 없고 훈련 데이터의 정보를 이용해서만 전처리를 수행하여야 한다. 평가 데이터는 실제로 공장에서 하나씩 발생하기 때문에 특정 변수의 평균이나 분산을 구할 수 없다.)

(5) 위의 모델을 훈련 데이터에 적용하여 총 이득을 산정하고, 모델이 효용성을 갖는지 확인하시오. 모델을 사용하지 않는 경우에 비해 얼마나 이득이 증가하였는가? 이상적인 모델에 비하여 얼마나 이득을 올릴 수 있는가?

(6) 마케팅 부서에서는 반도체 가격 경쟁으로 인하여 내년도 판매가격(S)이 1,200원이 될 것으로 예상하였다. 이 때, 모델을 쓰지 않는 경우, 이상적이 모델을 쓰는 경우, 위에서 개발한 모델을 쓰는 경우 각각의 총 이득은 얼마인가? 모델을 수정하여 이득을 더 높일 수 있는 방안이 있는지 확인하시오. 전체 훈련 데이터를 이용하여 확인하시오.

(7) 어제 밤 경쟁 회사의 공장에서 대규모 화재가 발생하였다. 내년에 반도체 품귀 현상이 발생할 것으로 예상되었다. 마케팅 부서는 판매가격(S)을 5,000원으로 인상하였다. 이 때, 모델을 쓰지 않는 경우, 이상적이 모델을 쓰는 경우, 위에서 개발한 모델을 쓰는 경우 각각의 총 이득은 얼마인가? 모델을 수정하여 이득을 더 높일 수 있는 방안이 있는지 확인하시오. 전체 훈련 데이터를 이용하여 확인하시오.