2024. 12. 07

고등지능정보공학II

Take Home Exam

1. 문제 및 데이터 개요

메르세데스 벤츠 자동차 회사는 자동차를 생산한 후 신뢰성과 안정성 검증을 위해 테스트를 실시한다. 자동차의 테스트 시간은 각종 구성 요소(옵션으로 4WD, 에어 서스펜션, head-up display 등)에 따라 달라진다. 본 시험은 자동차 구성요소에 따른 테스트 시간을 예측하는 회귀 문제이다. 테스트 시간을 정확히 예측하면 total test time을 줄이는데 도움이 된다.

학습 데이터는 3,600개이며, 각 데이터 포인트는 376개의 입력변수와 1개의 출력변수로 구성된다. 테스트 데이터는 609개이며, 각 데이터 포인트는 동일한 입력변수를 갖지만 출력변수 값은 없다. 출력변수는 테스트 시간으로 연속형 변수이며, 입력변수는 8개의 범주형 변수와 369개의 이진 변수로 구성되어 있다. 예측 성능은 R^2이다.

2. 분석 절차

A. 아래 절차에 따라 문제를 푸시오 (**60점**).

(1) 회귀 문제에서 이상치는 예측 성능을 왜곡시킬 가능성이 매우 크다. 따라서, 출력변수의 히스토그램를 바탕으로 99.99 percentile이상인 출력변수 값을 이상치로 판단하여 모델링에서 제외하시오.

(2) 입력변수에서 분산이 작은 변수는 예측에 도움이 되지 않는다. 범주형 변수에 대해 box plot[1]을 그리고 변수를 제거하시오. 이진 변수에 대해서는 표본분산을 계산하고, scatter plot(x축: 변수, y축: 표본분산)을 그린 후 분산이 0 이거나 동일 분산을 갖는 변수를 제거하시오.

(3) 범주형 변수에 대해 label encoding[1]을 통해 전처리를 실시하시오.

(4) 원 입력변수들의 교호작용 변수인 X_iX_j를 구성한 후 일반 선형회귀를 실행하고, 회귀계수의 p_value가 0.05보다 큰 교호작용 변수를 모두 찾아 제거하시오.

(5) Feature engineering을 통해 주요 특징을 추출하는 것은 예측 성능을 높이는데 매우 중요

[1] 본 강의에서 설명은 안 했지만 스스로 공부하여 실행하시오.

하다. 다음 특징을 고려하시오.

- (1), (2), (3), (4) 스텝에 따라 제거되지 않은 원변수 집합
- 제거되지 원변수 않은 원변수 집합에 대해 linear PCA
- 제거되지 원변수 않은 원변수 집합에 대해 kernel PCA
- 제거되지 원변수 않은 원변수 집합에 대해 Gaussian random projection[2]★.

(6) 예측 모델로는 random forest, PLS, Kernel ridge, Gaussian process regression을 사용하고, 각 feature engineering과의 조합에 따른 최종 모델 조합을 선정한 후 테스트 데이터에 대한 예측값을 1번.csv 파일로 제출하시오 성능을 높이기 위해 사용한 기법이나, 사용한 이유 등에 대해 자세히 설명하시오. (하이퍼파라미터 튜닝을 해야 함)

(7) 이외의 자신만의 모델과 전처리 방법을 자유롭게 사용하여 최종 모델 조합을 선정한 후 테스트 데이터에 대한 예측값을 2번.csv 파일로 제출하시오. 성능을 높이기 위해 사용한 기법이나, 사용한 이유 등에 대해 자세히 설명하시오.

3. 제출기한

12월 15일 23시59분 (LearnUS에 테스트 데이터에 대한 예측 csv 파일들, 소스 코드, 보고서를 업로드하시오.)

4. 유의사항

1번, 2번의 결과에 따라 최대 60점을 부여함 (등수에 따라 성적 부여)

보고서 점수로 40점을 부여함. 보고서 내에는 분석에 대한 논리적 이유를 최대한 상세히 작성해야 함

제출 파일은 예측 파일들(1번.csv, 2번.csv)과 보고서임

다른 학생의 코드를 일부, 전부 표절한 경우 두 학생 모두 0점 처리함.

늦게 제출할 경우 10분당 5점씩 감점함.

---

[2] ★ Gaussian random projection은 차원축소 방법으로 본 강의에서 설명은 안 했지만 스스로 공부하여 실행하시오.