💯

ADP ML 기출

30회

회귀 - 차원축소, 회귀분석 가정, k-fold 교차검증

첫번째 문제 : 혈압 관련 데이터, 종속변수 DBP

- 1-1) EDA를 시행하라.
- 1-2) 데이터 전처리가 필요하다면 수행하고 이유를 작성하라.
- 1-3) train test set을 DBP컬럼 기준으로 7:3 비율로 나누고 잘 나뉘었는지 통계적으로 나타내라.

- 2-1) 독립변수의 차원축소의 필요성을 논하고, 필요에 따라 차원을 축소하고 불필요하다면 그 근거를 논하시오.
- 2-2) 2-1 작업 후 데이터가 회귀분석의 기본가정을 따르는지 설명

- 3-1) 회귀분석 알고리즘 3개를 선택하고 선정이유와 장단점 비교
- 3-2) 1-3에서 구분한 데이터를 기준으로 3개의 회귀 분석 모델링을 진행하고 평가지표 rmse로 가장 최적화된 알고리즘 선정
- 3-3) 3-2에서 가장 성능 좋은 알고리즘을 이용하여 K-Fold 교차검증을 수행하시오.

분류 - 통계 검정, SMOTE, 로지스틱회귀 & XGB

두번째 문제 : 자전거 사고 관련 데이터

- 4-1) 발생시각을 통해 평일인지 주말인지를 구분하는 '주말여부' 범주형 변수 추가하고 데이터 분포를 확인하라. (월 ~ 금은 평일, 토요일과 일요일을 주말)

- 4-2) 사고내용에 따라 각 독립변수들(가해자성별,가해자연령,가해자차종,사고유형,기상상태,주말여부 등)이 유의한지 **통계적 검정**하라.
- 4-3) 4-2 에서 유의한 변수들만 가지고 SMOTE 오버샘플링을 수행하고 범주형변수는 변수별 빈도를 나타내고 연속형이면 평균을 나타내시오.
- 4-4) 4-3 데이터를 가지고 사고내용을 종속변수로 하여 로지스틱회귀분석, XGB 분류모델을 만들고 성능 비교를 하고 영향력 있는 변수를 확인하라.

29회

분류 & 회귀 - 전처리, 변수생성

첫번째 문제 : 아파트 데이터

- 1-1) 계약자고유번호를 기준으로 거주연도 별 여러개의 데이터가 쌓여 있다. 각 계약자 고유번호에 대해 가장 최신의 거주연도 행만 남겨라.
- 1-2) 결측치 처리
- 1-3) 이상치 처리

- 2-1) 재계약 횟수의 중앙값을 기준으로 중앙값보다 크거나 같으면 '높음', 작으면 '낮음'으로 재계약 횟수 이분 변수를 구성하시오.
- 2-2) 차원축소의 필요성을 논하고, 필요에 따라 차원을 축소하고 불필요하다면 그 근거를 논하시오.

- 3-1) 재계약 횟수 이분변수를 기준으로 세그먼트를 구분하고 각 세그먼트의 특징을 분석하시오.
- 3-2) **재계약횟수 변수를 종속변수로 하는 회귀 분석**을 두 가지 이상의 방법론을 통해 수행하고 최종 모델을 결정하시오. **재계약횟수 이분변수를 종속변수로 하는 분류 분석**을 두가지 이상의 방법론을 통해 수행하고 최종 모델을 결정하시오.
- 3-3) 최종 채택한 모델에서 각각 유의하게 작용하는 변수를 확인 하고 설명하시오.

- 3-4) 해당 데이터 분석결과로 얻을 수 있는 점 제시

회귀계수 검정, 로지스틱 회귀, SMOTE, XGB

두번째 문제 : 야구 데이터

- 4-1) 각 회차별로 1번 타자의 출루 (1,2,3루타와 사사구(볼넷, 몸에 맞는 공))가 있는 경우에 대해 득점이 발생 했는지 확인하고자 한다. 이를 위한 전처리를 수행하라. (단, 첫 번째 혹은 두 번째 타자가 홈런을 친 경우 해당 회차 데이터는 제외한다.)

```
조건1 : 득점여부를 범주형 종속변수로 한다. (1점이상 득점 :1, 무득점 :0)
조건2 : 각 회차 2번 타자의 데이터는 원핫 인코딩한다.
조건3 : 학습에 적절하지 않은 데이터는 제외한다.
```

- 4-2) 4-1 데이터에 대해 Logistic Regression을 적용하고 2번타자의 희생번트 여부에 대한 회귀 계수 검정을 하라.
- 4-3) SMOTE (random_state =0 지정)를 적용하여 data imbalance를 해결하라.
- 4-4) 4-3 구성 데이터에 Logistic Regression을 적용하고 결과를 분석하라.
- 4-5) 4-3 구성 데이터에 XGB 적용하고 결과를 분석하라.

28회

분류 - 과적합 방지, 오버샘플링(SMOTE), 랜포, 인공신경망, 보팅

- 학생 결석일수 예측 데이터

- 1-1) 데이터 EDA & 차원축소가 필요한지?
- 1-2) 1-1에서 찾은 문제 파악후 처리하기
- 1-3) 과적합 문제가 있다고 가정하고 해결하는 방법 2가지 이상 제시/실행/결과

- 2-1) **랜덤포레스트, 인공신경망(neuralnetwork), LightGBM 모델링** 수행 후 **f1-score** 비교
- 2-2) **Soft-voting, Hard-voting** 수행 후 f1-score 비교
- 2-3) 5가지 방법 중 최적의 방법 선택 및 이유 설명
- 2-4) 만든 모델을 학교 정보시스템에서 활용하려면 어떻게 해야하는지 적고 설명

27회

분류 - 차원축소, 오버/언더샘플링, 이상탐지

- 신용카드 이상탐지 데이터

- 1-1) EDA 데이터 탐색
- 1-2) 변수간 상관관계를 시각화하고 전처리가 필요함을 설명하라

- 2-1) 차원축소 방법 2가지 이상 비교하고 한가지 선택 (종류와 장/단점)
- 2-2) 추천한 한 가지를 실제로 수행하고 선택한 이유 설명

- 3-1) **오버샘플링과 언더샘플링** 장단점 비교 및 선택 구현
- 3-2) 구현 및 알고리즘 2가지 이상 비교, 성능 측정
- 3-3) 현재까지 전처리한 데이터를 통해 모델 수행 후 결과 분석
    - SVM, 나이브 베이즈, lgbm 비교 (+로지스틱 회귀, 랜포, xgboost도)
      Confusion Matrix
      분류평가 지표
      ROC Curve 그리는 것까지 !

- 4-1) **이상탐지 모델** 2가지 이상 기술, 장/단점 설명

- 4-2) 2번에서 만든 데이터로 한 가지 이상탐지 모델을 구현하고, 3번에서 만든 모델과 비교
- 4-3) 데이터분석 관점에서 3번에서 만든 모델과 4번에서 만든 모델 설명

26회

군집분석(K-means, DBSCAN), 추천

- 제품 주문 데이터

- 1-1) 결측치를 확인하고, 결측치 제거할 것
- 1-2) 이상치 제거하는 방법을 설명하고, 이상치 제거하고 난 결과를 통계적으로 나타낼 것
- 1-3) 전처리한 데이터로 Kmeans, DBSCAN 등 방법으로 군집을 생성할 것
  - 최적의 군집 개수를 판단하는 방법 - **엘보우 기법**

- 2-1) 위에서 생성한 군집들의 특성을 분석할 것
- 2-2) 각 군집 별 대표 추천 상품을 도출할 것
- 2-3) CustomerID가 12413인 고객을 대상으로 상품을 추천할 것

25회

군집분석(K-means, DBSCAN), 시계열 SARIMA

- 제품 주문 데이터, 관광객 시계열 데이터

군집분석(K-means, DBSCAN)

- 1-1) F(소비자별 구매빈도), M(소비자별 총 구매액) feature를 새로 생성해서 그 결과값으로 탐색적 분석 실시
- 1-2) F, M feature 기반으로 군집분석 실시, 필요시 이상값 보정
  - IQR (사분위수 범위)를 활용한 이상치 제거
- 1-3) 군집 결과의 적합성을 **군집 내 응집도, 군집 간 분리도**의 개념을 사용해서 서술
  - 실루엣 계수를 이용한 군집분석 평가 : 실루엣 분석
- 1-4) 적합된 군집 별 특성에 대한 의견과 비즈니스적 판단 제시

관광지 매달 평균 이용객 : 시계열 SARIMA

각 row는 관광지 A의 1990년 1월 부터 25년동안의 매달 평균 이용객 숫자

- 2-1) EDA와 시각화를 진행하라.
- 2-2) 결측치 처리와 해당 **결측치 처리 방식**에 대한 논리적 근거를 제시하라.
- 2-3) **계절성을 반영한 시계열 모델**을 제시하고 정확도 측면에서 모델 성능 평가 할 것
- 2-4) 분석결과 활용 가능 여부에 대한 분석 전문가로서의 제안

24회

회귀 - 결측치, 예측 모델 2개 제시, 모델 최적화 방안

- 학생 결석일수 예측 데이터

- 1-1) 데이터 EDA 및 시각화
- 1-2) 결측치 처리 및 변화 시각화, 추가 전처리가 필요하다면 이유와 기대효과를 설명하라
- 1-3) 결석일수 예측모델을 2개 제시하고 선택한 근거 설명

- 1-4) 선정한 모델 2가지 생성 및 모델의 평가 기준을 선정하고 선정 이유 설명
- 1-5) 모델이 다양한 일상 상황에서도 잘 동작한다는 것을 설명하고 시각화 하라
- 1-6) 모델 최적화 방안에 대해 구체적으로 설명하라

23회

이진분류 - 결측치 대체, 오버/언더샘플링

- 객실 사용 여부 데이터

온,습도,조도,CO2농도에 따른 객실의 사용유무 판별

종속변수 Occupancy, 0: 비어있음 , 1: 사용중

- 1-1) 데이터 EDA 수행 후, 분석가 입장에서 의미있는 탐색
- 1-2) 결측치를 대체하는 방식 선택하고 근거제시, 대체 수행
- 1-3) 추가적으로 데이터의 질 및 품질관리를 향상시킬만한 내용 작성

- 2-1) 데이터에 불균형이 있는지 확인, 불균형 판단 근거 작성
- 2-2) 오버샘플링 방법들 중 2개 선택하고 장단점 등 선정 이유 제시
- 2-3) 속도측면, 정확도측면 모델 1개씩 선택, 선택 이유도 기술
- 2-4) 분석결과 활용 가능 여부에 대한 분석 전문가로서의 제안

22회 **이진분류**

- 피마 인디안 당뇨병 발병유무
- 23회랑 굉장히 유사

21회

회귀 - 오버/언더샘플링, 릿지 라쏘 회귀, 다항회귀

- 학생 성적 관련 소규모 데이터

- 1) 데이터 8:2로 분할하고 선형회귀 적용하시오. 결정계수와 rmse 구하시오.
- 2) 데이터 8:2로 분할하고 릿지 회귀 적용하시오.
  alpha 값을 0부터 1까지 0.1단위로 모두 탐색해서 결정계수가 가장 높을때의 알파를 찾고, 해당 알파로 다시 모델을 학습해서 결정계수와 rmse를 계산
- 3) 데이터 8:2로 분할하고 라쏘 회귀 적용하시오.
  alpha 값을 0부터 1까지 0.1단위로 모두 탐색해서 결정계수가 가장 높을때의 알파를 찾고, 해당 알파로 다시 모델을 학습해서 결정계수와 rmse를 계산

- 아래와 같은 단순 선형 회귀를 3차 다항 회귀까지 적용시켜 계수를 구하고 각 차수별 데이터포인트 스캐터 플롯과 기울기 선을 그리세요.(12점)
  - 원래 데이터 : 독립변수 하나, 종속변수 하나 소규모 데이터

20회

시계열 히트맵, 데이터셋 분할 및 결과 검증

> 날씨 온도 예측 데이터, 종속변수 : actual (최고온도)
>
> 5분간격의 가구별 전력 사용량의 데이터
>
> 태양광 데이터

전력 사용량 - 시계열 히트맵

- 각 가구의 15분간격의 전력량의 합을 구하고 해당데이터를 바탕으로 총 5개의 군집으로 군집화를 진행한 후 아래의 그림과 같은 형태로 출력하라.

- 군집화를 위한 데이터 구성의 이유를 설명하라 (군집 방식에 따라 Cluster컬럼의 값은 달라질수 있음)
- 위 데이터를 바탕으로 각 군집의 요일, 15분간격별 전력사용량의 합을 구한 후 아래와 같이 시각화 하여라 (수치는 동일하지 않을 수 있음 2-1의 데이터가 정확하게 아래와 같은 이미지로 변환 됐는지 주로 확인)

태양광 - 데이터셋 분할 및 결과 검증

- 데이터셋 7:3 분할
- 데이터 전처리 및 예측 모델 생성
- 모델 성능 검증 : RMSE, R제곱, 정확도(아래 방식으로 연산)로 구하여라
- 정확도의 경우 실제값>예측값인 경우 (1-예측값/실제값), 실제값<예측값인 경우 (1- 실제값/예측값)으로 하고 이것들을 평균낸 후 1에서 뺀값으로 한다.

  분수식의 분모가 0인 경우의 정확도는 0.5로 취급한다.
- 최종 결과 제출 : 소수점 3째자리 반올림