기출 복기

adp 정보모음 사이트

https://datamanim.com

1회

고객세분화 (45점)

가. 세분화 변수의 생성 및 선정

– 요약 변수 및 파생 변수의 생성

예) 카테고리별 점유율, 주 인터넷 사용 시간대, 인터넷 사용 일수, 검색 패턴, 쇼핑단계별 이용 패턴,

주 쇼핑 시간대, 주 관심 상품 카테고리, 구매 상품 가격대

– EDA, 상관분석, Decision Tree 등을 통하여 적절한 세분화 변수 선정

나. **<군집 분석>** 및 최적 세분화 집단 생성

– 적절한 군집 분석 기법 제시 및 분석 수행

– 세분화 집단의 **최적 개수** 결정 및 기준 설명

다. 각 세분화 집단의 **특성 분석**, 정의 및 마케팅 **인사이트**

– 각 세분화 집단에 대한 특성 분석 및 시각화

– 특성 분석 결과를 기반으로 각 세분화 집단에 대한 마케팅 관점에서의 정의

– 세분화 분석 결과를 토대로 타당한 마케팅 인사이트 제시

예측 (45점)

가. 세분화 집단별 예측 모형 (구매,이탈,등급 변동,우수 고객 예측 등) 개발을 위한 **<종속 변수>** 정의 및 **<독립 변수>** 선정

– 해당 예측을 위한 타당한 종속 변수의 정의

– 해당 예측을 위한 독립 변수의 생성 및 선정

요약 변수 및 파생 변수의 생성, EDA/상관분석/Decision Tree 등을 통한 <**적절한 변수>** 선정

나. 세분화 집단별 **예측 모형** 개발

– 종속 및 독립 변수의 성격에 따른 적절한 모델링 기법 제시

– 샘플링, 파티션 등 모델링 데이터 준비 및 모형 생성

– 적절한 평가 기준에 의한 모델 평가 및 최종 모델의 선택

다. 예측 모형 분석 및 마케팅 **인사이트** 제시

– 세분화 집단별 예측 모형의 특성 분석 및 **시각화**

– 세분화 집단간 예측 모형의 비교 분석

– 예측 모형 분석 결과를 토대로 적절한 마케팅 인사이트 제시

텍스트 마이닝 (10점)

가. 포털 사이트 검색 기록을 활용한 토픽 분석

- **한 고객**이 일정 기간 동한 포털 사이트에서 입력한 전체 검색 리스트를 하나의 문서(Document)로 간주

- 위의 문서에 기반을 두어 고객의 **관심 주제**를 파악하기 위한 토픽 분석을 수행

나. 토픽 분석 결과의 **해석** 및 **마케팅**에서의 활용 방안 제시

– 도출된 토픽 리스트의 의미를 마케팅 관점에서 해석

– 분석 결과를 고객 세분화 및 예측 등에서 활용할 수 있는 구체적인 방안 제시

[출처] ADP]ADP 실기 기출 문제 정리 (제1회,2회,3회)|**작성자** 수제비

2회

문제1. data munging

1) data : user_id, usage, 방문 category, 접속유지시간

- usage = userId 당 총 접속시간.
- wd_월요일 ~ wd_일요일 : user당 요일별 총 접속시간/user당 총 접속시간
- 방문 category : 총 22개 category중 user가 방문한 카테고리의 비율.
- 유저 접속일수 : user가 방문한 날짜의 수

2) 문제

user_id, usage, wd_월요일, …, wd_일요일, 방문카테고리 비율, 유저 접속일수의 순서로 데이터를 출력. 단, wd_월요일 ~ 방문 카테고리 비율은 소수점 3자리로 맞추어서 출력하라

(aggregate, plyr, reshape 패키지를 연계해서 잘 다루어야 풀 수 있는 문제)

문제2. data mining

1)data : custid, churn(해지 또는 해약), product_Id

custid, churn, product_001,,,,

2) 문제

- 3가지 이상의 churn 예측모델링을 하고, 각각의 성능을 비교하여, 최적의 모델을 선정하고, 이를 시각화하시오.

(의사결정나무, svm, logistic regression, randomForest등을 예시)

- 문제 : 고객들은 churn고객과 아닌 고객으로 구분한 후, 각각을 대상으로 **<연관분석>**을 실시하고, 그 결과를 가지고 상기 고객군의 특성을 비교하라.

문제 3. 통계분석

1)data

- 일일평균 교통량, 도시인구, 도로의 차선수
- 도로의 종류, 트럭의 종류, 도시/농촌여부

2) 문제

- 주어진 변수를 사용하여 **<변수 선택과정>**을 포함하여, 예측 모델을 만들어라.
- 도출된 모델의 **<잔차 분석>**을 통해 교정을 해서 최적의 모델을 찾고, 최종 모델의 결과로 나온 각 파라미터의 의미를 해석하시오.

문제 4. 텍스트 마이닝

1) data

- 제주관광명소 txt 파일
- 블로그 문서 txt 파일

2) 문제

- 데이터로딩에 문제가 있으면 적절한 형태로 데이터를 조정하고, TITLE과 CONTENT에 “제주”나 “관광”이 들어가 문서를 찾으시오.
- 문서에서 명사를 추출하고, 추출된 단어에서 제주관광명소가 아닌 곳을 필터링하시오.
- 단어가 적용된 문서를 기준으로 빈발단어 10개를 찾고, 이를 그래프로 만드시오.

[출처] ADP]ADP 실기 기출 문제 정리 (제1회,2회,3회)|**작성자** 수제비

3회

멍잉, 데이터마이닝, 텍스트마이닝, 통계분석, 인사이트 도출

1) **멍잉**

멍잉 부분에선 고객 자료주고 고객별 총 구매액 총 거래빈도 최근 구매일자 등을 파생변수로 만들고 0-1로 표준화하여 스코어를 구하고 정렬해서 상위 20명 나타내기

2) **데이터마이닝**

로또 번호를 가지고 연관분석 실시

3) **텍스트마이닝**

뉴스 자료 에서 뉴스 키워드 분석

주간 이슈, 이슈별 빈도 등등

4) **통계분석**

당뇨병 환자 정보 주고 나이와 성별이 사망과 관계 있는지 비율과 그래프 이용해 분석하고 상호작용 판단하는거와 로지스틱 모형을 범주형과 연속형 나이로 각각 만들고 어느 모형이 나은지 그 후에 상호작용항 적용하여 교호작용 판단하고 최종모형 결정 후 최종 파라미터 해석

5) **인사이트 도출**

신규 스타일 컬러가 주어졌을 때 어떻게 유사 클래스 정하는지 ?

군집분석이 어떻게 그룹 지정하는지 ?

실제와 예측 트랜드 유사도 확인 방법과 패턴 불일치 시 고려할 외부요인 등

신규만 판매 늘고 유사 클래스는 감소할 경우 전략 등을 적기

[출처] ADP]ADP 실기 기출 문제 정리 (제1회,2회,3회)|**작성자** 수제비

4회

5회

6회

- 크게 데이터전처리/데이터마이닝, 통계분석, 텍스트마이닝 3문제로 구성됨
- 각 문제는 3~~4개의 서브 문제가 주어짐.
- 좀 더 구체적으로 적으면 아래 처럼 총11개의 서브 문제로 구성됨.

1-(가),(나),(다),(라)

2-(가),(나),(다)

3-(가),(나),(다),(라)

- 각 문제별 배점은 얼마인지 표시되어 있지 않음.

문제1. 데이터 전처리 및 데이터 마이닝(R로 QF(?) 알고리듬을 직접 구현하여 단계별로 만들어 보는 문제)

- **데이터**

sales.csv 5천건

cust_id,prod_id,amt

1,prod01,100

1,prod02,200

11,prod02,300

21,prod03,100

...

(가) sales.csv를 읽어서 cust_id 및 prod_id별로 amt가 0보다 크면 1, 아니면 0으로 아래 처럼 만들 것

cust_id prod01 prod02 prod03 prod04 prod05 prod06 ...

1 1 0 1 1 0 0

11 0 1 0 0 0 0

21 0 0 1 1 1 0

31 1 1 0 1 1 1

...

(나) 위의 결과를 이용하여 cust_id 별로 상호 인접정도를 파악하기위해 cust_id별로 피어슨상관계수 행열을 아래와 같은 모양으로 만들것

1 11 21 31...

1 1.00 0.72 0.61 0.83

11 0.72 1.00 0.75 0.88

21 0.61 0.75 1.00 0.93

31 0.83 0.88 0.93 1.00

...

- (가)의 결과를 그냥 cor함수에 넣으면 오류남

(다) 특정 user의 유사도가 높은 15 user 구해서 다음 형태로 만들기? (행과 열이름은 cust_id였고 각 행렬의 값은 같은 상품을 구매한 amt 였던 것으로 추정됨)

1 11 21 31...

1 1200 100 200 400

11 100 2300 500 900

21 200 500 1500 800

31 400 900 800 1000

...

(라) 특정 user 당 5개씩 추천 아이템 생성하기? (잘 기억 안남)

prod01,prod03 ,prod06....

2. 통계분석

- **데이터 :** risk_proj.txt (성별,인종별,나이,활동성,risk)

(가) 성별에 따라 risk가 차이가 있는지 분석할 것

나이등 다른 변수와 교호작용 있는지 알아볼 것(가정사항등 정의)

(나) 인종별 risk가 차이가 있는지 분석

나이등 다른 변수와 교호작용 있는지 알아볼 것(가정사항등 정의)

(다) 인종및 성별에 따라 risk가 차이가 있는지 분석. 교호착용 분석

3. 텍스트 마이닝

- 데이터 : location.txt (UTF-8, 헤더 없음)

가평,ncn

가야,ncn

남이섬,ncn

....

- 데이터 : blog.txt (TAB문자로 구분됨)

DATE TITLE CONTENT

20150101 제목 봄관련 내용....

....

(가) 자료읽기

(ㄱ) location.txt를 읽어서 사용자사전에 등록하기

(ㄴ) blog.txt를 다음 형식으로 읽을 것

DATE : numeric

TITLE : character

CONTENT : character

(나) blog.txt에서 봄여행,벚꽃축제,봄나들이 등 봄과 관련된 문서만 추출하기

(가)에서 읽은 사용자 사전에 들어있는 지명이 들어있는 문서만 추출

(다) 위에서 추출된 문서에 대해 명사추출 및 출현 빈도 10위 추출

(라) 봄과 관련된 지명 출현 빈도 10위까지 추출하여 시각화

10. 기존 기출문제 분석

- 일단 제1회,2회,3회 기출문제는 제 블로그 아래에 하나로 정리해두었습니다.

http://blog.naver.com/sujebee/220694267110

4,5회 기출문제는 아무리 검색을 해봐도 찾을 수가 없었습니다.

- 먼저 기존 기출문제들을 보고 오시면 출제 경향에 대한 감을 잡을 수 있습니다.
- 1회 시험은 아직 시험이 현재와 같은 틀을 갖추기 이전에 출제된 것이라서 지금의 경향과 많은 차이가 있습니다.

1회 시험은 그냥 이런 문제도 있었구나 수준으로만 보면 됩니다.

- **기존 출제 영역 : 데이터전처리, 데이터마이닝, 통계분석, 텍스트 마이닝**

1) 데이터 전처리(멍잉)

- R 스크립트 작성 능력

- ddply, aggregate, merge, melt/cast, sqldf 등을 이용한 처리

- NA 결측치(imputation),이상치 처리.

2) 데이터 마이닝 : 분류, 연관, 군집분석 수준 정도만 출제됨

- 분류예측 : 해지여부 예측

- 연관 : 로또번호분석

- 군집 : 색상 스타일별 군집. 최적 k값 찾기

- 변수 선택 능력 : nearZeroVar, findCorrelation, step(), PCA(princomp, prcomp, biplot)

3) 통계 분석 능력 : 주로 분산분석(ANOVA),로지스틱분석 에서 출제됨

- 교호작용, 잔차분석, 시각화 능력

- 일원/이원분산분석 : aov(), 결측치, 이상치 고려

4) 텍스트 마이닝 : 주로 KoNLP를 이용한 빈도 분석

https://m.blog.naver.com/PostView.nhn?blogId=sujebee&logNo=220699299147&targetKeyword=&targetRecommendationCode=1

7회

집단간 의료비 지출 차이 분석 : t-test, ANOVA, 회귀분석

출처:https://didalsgur.tistory.com/87

8회

폐활량(Fev) 예측 문제 (나이, 키, 성별, 흡연 유무)

1. EDA 및 상관관계 분석
2. 적절한 회귀모형 선택
3. 회귀모형 해석 (변수 별 증가에 따른)

4. 평균 키, 나이 (여자, 흡연자) 일경우 폐활량 예측

출처:https://didalsgur.tistory.com/87

9회

10회

11회

특징

사실 1,3번의 난이도는 기존의 실기 출제 문제보다는 쉬운편이 아니었나 싶다.... 왜냐하면 전처리가 그렇게 까다로운 Data도 아니었고,

특히 3번 문제같은 경우, ADP를 공부하는 많은 수험생이 참고하는 'R을 이용한 Data처리 & 분석실무 (길벗)' 책에 나오는 대표적인 예시 Data이기 때문이다. (사실 문제보고 좀 당황했음... 설마 이게 나올까 했는데..)

물론 Data 일부를 조금 변형해놔서 추가적인 전처리가 필요했지만, 적어도 정확도가 어떻든 Model을 만들고, 예측 결과를 뽑는데까지는 크게 어려운 단계가 없었던 것 같다.

다만 이번에 멘붕이 왔던 문제는 2번 Text Mining이였다.

나는 배점이 높은 1,3번을 먼저 풀고 딱 1시간이 남은 상황에서 2번 Text Mining문제로 진입했는데,

결국 세부 문제 한문제도 제대로 풀지 못했다.

첫번째로 Text Data(Text file이였음)를 불러오는 부분에서 구분자를 가지고 Parsing을 하는데 문제가 발생하여 많은 시간을 잡아먹었고,

(파일 용량 자체가 커서 한번 불러오는데 분단위 시간이 걸림)

문제가 '형용사'를 추출하는 건데 여기서 결국 막힌 것이다.

사실 R을 활용한 Text Mining 대부분의 예시는 '명사'를 추출하는 형태로 되어 있고, 명사 추출은 extractNoun 함수로 쉽게 수행을 할 수 있다.

그런데 형용사 같은 경우는.... SimplePos22 함수를 써야하는데 해당 부분을 제대로 숙지하지 못하고 가서, 결국 for문을 만들다 시간이 부족해 실패를 해버린 것이다.

결국 2번문제를 아예 통으로 못풀다시피 하고 시험을 마쳤다.

1,3번 문제도 완벽히 맞았다는 보장이 없기때문에... 또 망한건가...라는 생각이 들고 있다.

일단 결과가 나오면 또 포스팅을 하도록 하겠다.

떨어지면 또... 준비를 해서 재응시를 해야겠지..

출처 :https://0dood0.tistory.com/33

문제유형은 회차를 거듭하면서 기본적인 골자는 비슷한 형태에서 살짝살짝 detail만 바뀌어 왔는데, 일반적으로 아래와 같다.

1. 통계학 기반 분석

2. Text Mining을 적용한 분석

3. Data mining(ML) 학습을 통한 결과 도출

상기 3개의 주제를 잘 분석하려면 당연히,

분석 언어를 통한, **'Data 전처리', 'Model 생성', '분석 Model별 검증', '결과 해석'** 역량이 있어야 한다.

거기다 2번문제를 잘 풀기 위해, **Text 전처리, KoNLP 패키지를 다양한 방식으로 사용해본 경험**이 있어야한다.

(wordcloud 만드는 수준으로는 해당 문제를 풀 수가 없다...)

금번 회차에도 위와 같이 크게 세가지 주제의 문제가 나왔고 각 큰 단위 문제 안에 세부적으로 3~4문제씩 세부문제가 있는 형태로 출제가 되었다.

금번에는

출산률 분석 : 독립/반응변수 관계를 회귀분석으로 정의 및 결과 해석

1. 통계학 기반 분석 (40점)
- 각 설명변수들과 출산률(종속변수)의 관계를 회귀분석으로 정의 및 결과를 해석하는 문제

2. Text mining (20점)
- 영화평 Data를 전처리 후,'형용사'를 추출 하여 감성 분석 하는 문제

3. Data Mining/ML (40점)
- (R을 공부하는 많은 사람들이 익숙한) 타이타닉 생존자 Data를 Data mining 학습하여, 생존

여부 예측을 하는 문제
- 금번회차의 경우 분석 과정은 상관없고, 오직 제공된 Test Data의 정답만을 제출해서, 예측

한 정답의 적중률이 얼마나 높은지로 채점을 함

출처 https://0dood0.tistory.com/33

12회

- 회귀 분석 문제

1. 변수 시각화(변수간 상관관계, 변수별 이상치 파악)
2. 회귀모형 적합과 유의성 검정
3. 회귀 계수에 대한 standard error가 가지는 의미
4. 회귀분석에서 잔차 분석 및 시각화
5. 회귀분석에서 영향력 관측치와 그 영향 분석

출처:https://didalsgur.tistory.com/87

13회

- 거래내역 데이터를 통해 부정사용여부 탐지 모델 개발

1. 타겟변수 불균형 문제 처리 : resampling, undersampling, oversampling(SMOTE) 특징(장단점) 서술 및 적용(패키지 활용가능)
2. 간단한 시각화
3. 불균형 문제가 해결된 resampled data로 binary classification model 생성
4. Confusion matrix와 AUC 등 다양한 성능 지표를 적용하여 결과 해석
    - 분류문제에서 어떤 것을 주로 봐야하는지 판단 필요

출처:https://didalsgur.tistory.com/87

14회

1. Machine Learning(Regression)

이런 식의 데이터 셋이 주어지고 목표는 Target의 값을 예측하는 것. Regression 문제임. 약 20,000 개 정도 였던 걸로 기억. 출제 측에서 매 시험마다 전형적으로 내는 유형임. Machine Learning 모델

을 만들어 데이터 전처리 하고 예측 값을 뽑아내는 문제.

보통 큰 문제 한 개에 새끼 문제3~5개 정도 나오고,새끼 문제는 대부분은아래 절차들을 요구함.

데이터 전처리->데이터 시각화->예측 모델 설계->테스트 및 검정->결과 시각화

2. 주성분분석(PCA) & ML(Classification)

1번문제랑 거의 똑같은데 차이점이라면 PCA + 분류 문제였음. Target은 2개 이상의 Type이였음. 따라서 다항 분류가 가능한 알고리즘을 사용 했어야 함. 기본적인 로지스틱 회귀분석은 이항이기 때문에 사용하기 어려웠음.

근데 이 문제는 데이터 분석에서 Feature 추출이 얼마나 중요한지를 알게 해줬던 문제임. PCA로 변수를 줄였을 때와 그냥 변수를 사용했을 때의 모델의 성능이 어떻게 달라지는지 확인 할 수 있었음. 풀면서도 재밌는 문제이고 굉장히 잘 만든 문제라고 생각했음(님이 뭔데요) 이외에 나머지는1번이랑 거의 똑같았음.

15회

1. 기계학습(철강회사 data)
1-1)EDA(탐색적 데이터 분석)을 하고 상관분석을 실시하고 분석에 필요한 파생변수를 선별하시오.
1-2)Test, Train data나누고 시각화할 것
1-2-1)목표변수가 1이 나올지 안 나올지에 대하여 알고 싶다. 목표변수를 이항변수로 바꾸고, 로지스틱 회귀분석을 실시하고 ConfusionMatrix를 확인하고 최적의 cut-off value를 구하여라.
1-2-2)로지스틱 분석을 제외하고, SVM을 포함하여 3개의 다항분류모형을 선정하고 ConfusionMatrix를 확인하고 최적의 cut-off value를 구하여라.
1-2-3)위에서 실시한 4가지 모형 중 가장 적합한 모형을 활용하여 군집분석을 실시하고, F1 score값을 구하시오.

2. 통계분석과 전처리(전력데이터)
Usage.csv
timestamp컬럼:타임스탬프 별 전력 사용량, 여기서 타임스탬프는 15분 간격으로 측정된 자료이다.(900단위로 증가하는걸 보아하니 단위는 초로 예상)
usage컬럼: 방금 15분 동안에 사용된 사용 전력
usage_history.csv
time컬럼: xx시 xx분의 단위를 가짐.
Wclass: a는 최저기온,b는 최고기온,c는 상기온, d는 저기온(4개 요인으로 구성)
A,B,C,D,E: 각 유형의 전력 누적사용량
Weather.csv
Date:-년 –월 –일
Avg_temperature: 그날 하루의 평균기온
2-1)usage의 총 사용량을 A,B,C,D,E 유형별 사용량으로 구별하고 아래와 같은 모양으로 연월별 평균값을 계산하여 CSV파일로 작성하시오
(Usage_history&Weather data 활용)

Month	A	B	C	D	E
yyyy-xx					
yyyy-xx					

2-2)가로축을 요일, 세로축을 평균사용량으로 하여 요일별, 유형별 평균 사용량값의 시각화 그래프를 작성하시오. (코드도 함께 제출할 것)
2-3)각 요일별 사용량에 대한 차이가 있는지 분석하시오.
2-4)각 날짜별 전력사용량이 weather의 평균 기온과 어떠한 관계가 있을지 분석하시오.

——

1. 기계학습

steal.csv 데이터 약 1900 x 27

> 1개의 목표변수와 26개의 설명변수

> 목표변수는 1,2,3,4,5,6,7 7개의 요인을 가지고 있다.

1. EDA(탐색적 데이터 분석)을 하고 상관분석을 실시하고 분석에 필요한 파생변수를 선별하시오.

% 시각화와 통계량을 제시할 것.

2-1. 데이터를 훈련용 테스트용 평가용으로 나누고 시각화 할 것.

% 시각화와 통계량을 제시할 것.

2-2. 목표변수가 1이나올지 안나올지에 대하여 알고싶다.

목표변수를 이항변수로 바꾸고 로지스틱 회귀분석을 실시하고

confusionMatrix를 확인하고 최적의 cut off value 정하여라.

% 시각화와 통계량을 제시할 것.

2-3. 로지스틱 분석을 제외하고 SVM을 포함하여 3개의 다항분류모형을 선정하고

confusionMatrix를 확인하고 최적의 cut off value 를 정하여라.

% 시각화와 통계량을 제시할 것.

2-4. 위에서 실시한 총 4개의 모형중에 가장 적합한 모형을 활용하여

군집분석을 실시하고 F1 스코어값을 구하시오.

% 시각화와 통계량을 제시할 것.

모든 문제마다 아래 작은 글씨로 시각화와 통계량을 제시할 것 이런 문구가 있었습니다.

이거 역시 모든 문제마다 다 시행을 해주어야만 하나 하고 감독관님께 여쭤보았습니다만...

하라고 말씀하시더군요..

이것이 참 무슨 의미인지 알쏭달쏭하더라구요

2. 통계분석, 데이터 전처리

usage.csv 데이터 약20000 x 2

usage_history.csv데이터 약 100,000 x 7

weather.csv 데이터 약 600 x 2

usage.csv

timestamp컬럼 : 타임스탬프 별 전력 사용량 여기서 타임스탬프는 15분간격으로 측정된자료이다.(900씩 증가하는 걸로보아서 단위는 초)

usage컬럼 : 방금 15분동안에 사용된 사용전력

usage_history.csv

time컬럼 : xx시 xx분 의 단위를 가짐. 1분마다 찍혔다고 문제에 적혀있었던걸로 기억 그러나 왜 각 값이 2개씩 존재하는지가 의문이였다.

wclass : A는 최저기온 B는 최고기온 C는 상기온? D는 저기온? - 4개의 요인으로 구성되어있다.

A : A 유형의 전력 누적사용량

B : B 유형의 전력 누적사용량

C : C 유형의 전력 누적사용량

D : D 유형의 전력 누적사용량

E : E 유형의 전력 누적사용량

위 데이터에서 A,B,C,D,E는 누적사용량이기 때문에 시간의 지남에 따라 아주 근소하게 증가하는 값들을 가지고 있습니다.

위 자료에선 wclass의 존재의미를 모르겠더군요. 데이터를 이해하고자할때 전력 유형과 같은 값을 가지고 있기에 너무 헷갈렸었습니다.

시간별로 정리된 자료였음에도 불구하고 A가나왔다가 B가나왔다 다시 A가나오는 이상한 데이터.

weather.csv

date : 년 - 월 - 일

avg_temperature : 그날 하루의 평균기온

제가 대충만든 데이터 생김새 모양입니다.

```
> head(usage)
   timestamp     usage
1 1543000000  982.3839
2 1543000900  973.5152
3 1543001800 1036.9795
4 1543002700  946.8271
5 1543003600 1012.3105
6 1543004500  985.5250
> head(usage_history)
   time wclass      A      B      C      D      E
1 00:00      A 1444.4 1234.2 1336.4 1001.1 1212.0
2 00:00      A 1444.4 1234.2 1336.4 1001.2 1212.1
3 00:01      A 1444.5 1234.3 1336.4 1001.3 1212.1
4 00:01      A 1444.6 1234.4 1336.4 1001.5 1212.1
5 00:02      A 1444.6 1234.6 1336.5 1001.6 1212.2
6 00:02      A 1444.7 1234.6 1336.5 1001.7 1212.2
> head(weather)
        date   avg_temp
1 2018-01-01 -0.2383023
2 2018-01-02  6.2317971
3 2018-01-03  0.8308363
4 2018-01-04 11.8630559
5 2018-01-05  2.9087014
6 2018-01-06  7.2640750
```

모두 똑같이 시간이라고하는 공통컬럼이 다른 모양으로 존재할 수 있으니 이것도 조심해야겠더군요.

1-1. usage의 총사용량을 A,B,C,D,E 유형별 사용량으로 구별하고 아래와 같은 모양으로 연월별 평균값을 계산하여 CSV 파일로 작성하시오.

(usage_history데이터와 weather데이터를 활용)

month A B C D E

yyyy-xx .. ..

yyyy-xx .. ..

> ...

1-2. 가로축을 요일 세로축을 평균사용량으로 하여 요일별 유형별 평균 사용량값의 시각화 그래프를 작성하시오. 코드도 함께 제출할 것.

1-3. 각 요일별 사용량에 대한 차이가 있는지 분석하시오.

1-4. 각 날짜별 전력사용량이 weather의 평균기온과 어떠한 관계가 있을지 분석하시오.

시험 후기 입니다.

ADP실기는 이번이 처음이었는데

처음에 10분동안 데이터가 import가 되지 않아서 애먹었습니다.

가장 이해할수 없는 부분이 왜 굳이 시험환경을 리눅스에서 VM웨어를 돌려

디렉터리 설정도 못 하게 해놓고 R을 사용하는데 library로 패키지 로딩도 너무 느려서

답답했습니다.

파일명을 vm 윈도우의 절대경로로해서 read.csv 해서 하는 일반적인 방법이

위치를 못잡기에 절대 안먹습니다.

그러니 처음보시는 분들은 겁먹지말고 R의 오른쪽 하단부분에 있는 plot packages help 이부분에있는 곳에서

files를 클릭하여 활용하시기 바랍니다.

솔직히 1번 기계학습은 어렵지 않게 딱 2시간에 끝냈습니다.

다만 F1 값을 구하는 것에 2-4번에서 모형선정해놓고 뒤로 밀어 두었습니다. 그렇기에

남은 두시간동안 제일 자신있는분야인 전처리를 하려고 가벼운마음으로 넘어갔습니다.

근데 국어가 딸리는 걸까요? 제가 주어진 세개의 데이터셋을 이해하지 못하겠더라구요

usage_history에서 wclass라고하는 최저 최고 저기온 상기온? 이런애들은 왜 존재하는건지도...

이부분에서 데이터를 이해하고자 하였으나 1시간가까이 잡아먹고 이해하지 못한채로

억지로 하라는데로 하기는 하였으나,,, 원하는 모양의 csv 파일은 아닌 것 같습니다.

그리고 가장 알수 없었던 존재가 timestamp...

저는 이 timestamp의 존재를 오늘에서야 알았습니다..

변환하지도 못할뿐더러...그냥 시작값을 weather데이터와 같은 2018년 1월 1일로 잡고 진행하였습니다.

2번 3번 4번은 모두 분산분석과 단순 회귀분석으로 가볍게할 수 있는 문제들이었으나

데이터이해가 되지 않기에 제가 한게 맞는지 의심스럽게 풀면서 넘어갔습니다.

데이터분석 실기시험은 개인적으로 보았을 때, 분석언어의 능숙함보다는 얼마나 주어진 데이터셋을 이해를

잘 하고있는가가 중요한 것 같습니다.

혹시 위 주어진 3개의 데이터셋을 이해하신 분이 계신다면 저좀 이해주시켜주시면 감사하겠습니다.

그리고 그외 추가적인 사항도 제가 답할 수 있는 부분은 답해보도록 하겠습니다.

16회

17회

1. 기계학습 문제(집값 예측)
1-1)EDA&결측값 채우기
1-2)모델 생성
1-2-1)데이터 분할
1-2-2)교호작용을 고려한 다중 선형 회귀 수행
1-2-3)3가지 분류 모델 생성 및 비교, 좋은 모델 선택

2. 시각화 및 시계열 분석(코로나 데이터)
2-1)전체 인구대비 누적 사망률이 가장 높은 5개 국가 추출 후, 국가별 일일확진자, 누적확진자, 일
일사망자, 누적사망자 시계열 그래프 출력
2-2)위험지수 생성 및 해석
2-3)시계열 분석 및 예측 모델 생성

3. 통계 분석(설문 데이터 분석) -> 사전에 역문항들에 대한 처리 필요
3-1)그룹별 평균, 표준편차, 왜도,첨도 산출 및 각 영역별 그룹별 만족도 추세가 어떤지 탐색(EDA)
3-2)응답항목별 차이가 있는지 분석(Anova table)
3-3)탐색적 요인분석 수행(FactorAnalysis)
3-4)신뢰성 지수를 개발 하는 문제 항목별 신뢰성 지수를 구하라.

18회

특징

- 대문제 3 (각각소문제 5) , 데이터 3개
- SOM 모델 추출
- 텍스트 마이닝이 출제 안되다가 다시 출제된 시점 / 영어

1. 고객 등급 예측모형
1)EDA&결측값 채우기
2-1)파생변수 3개 생성& 이유 작성
2-2)Train-Test 분할(7:3)/SOM 군집분석/정오 분류표
2-3)분류분석 4가지

2. 텍스트 마이닝(영어)
-명사 추출&불용어 처리
-빈도 막대그래프

3. 시계열 분석
3-1)평균과 분산 일정 +근거&해석
3-2)ARIMA+근거&해석
3-3)최적 모델 선택 + 근거&해석
3-4)적합도 파악

ㅡㅡㅡ

1. 기계학습 : 고객 등급(1부터 5까지) 분류 예측 모형

- EDA 및 결측치 처리를 포함하여 데이터 전처리

- 파생변수 3개를 생성하고 생성한 근거를 시각화나 통계량으로 제시

- 데이터를 train_test로 나누고 train에 대해 som을 이용해 군집분석을 실시하고 최적화 수행후 confusion matrix를 그리시오

- 랜덤포레스트, 인공신경망(이부분은 기억이 잘 안나네요ㅜㅜ) 을 포함해 4개의 분류 예측 모형을 만들고 각각의 성능을 roc_auc, F1_score로 비교하시오

- 앞에서 정한 모델에 추가로 성능을 높이시오

2. 영어 텍스트

- 영어 문장을 의미없는 단어를 없애고 형태소 분석을 하시오

- 단어 빈도를 시각화하시오 (원래는 워드클라우드도 하는 문제인데 시험 도중에 이부분은 빼주었습니다)

3. 통계분석 : 3~4년치 매출데이터(어떤 데이터인지 가물가물합니다)

- 시계열의 정상성을 만족시키시오

- 모형을 3개 이상 만들어 비교하시오

- 앞에서 정한 모델을 진단(통계적인 진단을 요구했습니다)

- 예측의 정확도를 나타내시오

adp 정보모음 사이트

https://datamanim.com

19회

특징

- 이번 19회에서는 기계학습과, 통계분석만 나오고 텍스트 마이닝은 출제되지 않았습니다.
- **시험문제를 써서 나올 수 없어 기억에 의존하여 복기하였기에, 문제가 정확하지 않을 수 있다는 점 미리 말씀드립니다.**

1. 기계학습(DATA : credit데이터 - 고객이 이탈되었는지 아닌지 분류하는 문제) (총 50점)

- 독립변수로는 성별, 나이, 카드등급, 소득 등의 변수들이 있었습니다.

1-1 : 데이터 전처리 및 시각화(5점) - 연속형변수와 문자로된 범주형 변수를 처리해야합니다.

1-2 : Train과 Test를 7:3으로 나누고 분류분석 3개 실시 및 Confusion Matrix 만들기(15점)

1-3 : 위에서 실시한 분류분석 3개를 앙상블하여 Credit_test를 예측하고(credit_test.csv는 따로 주어짐) result.csv로 만들어서 제출하기 (30점)

- 1-1과 1-2는 기존처럼 코드와 해석결과를 PDF로 만들어서 제출하면 되고 1-3은 CSV파일로 제출하면 됩니다.

2. 통계학습(DATA : Traffic EPS 시계열 분석 - 20년치 데이터이며 1년에 4개씩 데이터가 존재(분기별로 존재)) (총 50점)

2-1 시계열 데이터의 정규성과 이분산성을 분석하기 위해 시각화하고 설명(10점)

2-2 위에서 시계열데이터가 정규성이 아니라면, 고정시계열이 있는지 확인하고 이를 처리(15점)

2-3 SARIMA 분석을 실시, 여러 파라미터를 적용해보고 가장 성능이 좋은 것을 제시(15점)

2-4 위에서 제시한 모델의 잔차와 잡음에 대해 시각화하고 분석(10점)

출처 :https://ysyblog.tistory.com/114

리뷰

1. 기계학습 문제는 무난했다고 할 수 있습니다. 분류문제의 기본적인 전처리, 분류분석, 앙상블 분석 진행을 하면 되기 때문입니다. 다만 1-3은 CSV파일로 만들어서 제출해야하는데, 분석과정을 요구하지 않은 것으로 보아 성능(실제로 맞는지)만으로 평가할 것으로 예상됩니다.

2. 통계학습 문제에서는 진흥원이 '이건 몰랐지' 스킬을 또 실행했다고 할 수 있다. 사실 SARIMA라는 것을 들어본 사람은 거의 없을것입니다.. 필자도 처음 들어본 기법입니다. 주력언어는 PYTHON이지만 시계열 통계분석은 R이 유리하기 때문에 시계열은 R로 준비해갔는데 SARIMA는 처음들어봤기 때문에 패키지 찾는 곳에서 SARIMA를 직접 찾아야 했습니다.

2-1 : 이분산성이 뭔지 몰랐는데, 등분산성과 관련있는것 같아, 등분산성과 관련하여 적었습니다. 그런데 실제로 등분산이 결여된 경우가 이분산이라고 합니다.

2-2 : 고정시계열이 뭔지 모르겠습니다.(찾아봐도 뭔지 모르겠습니다), 따라서 추세나 계절성 등을 처리하여 정상시계열을 만들라고 하는 것이라고 판단하였습니다. 어처피 시계열분석을 하려면 차분을 하던지해서 정상시계열로 만들어야하기 때문이기 때문에 그 과정을 처리하는 부분이라고 생각하였습니다.

2-3 : 위에서 말했던것 처럼 R 패키지 찾는 곳에 검색을 해서 Auto.sarima라는 것을 알아내었고, 이를 활용해서 꾸역꾸역 문제를 풀었는데, 확실하지 않습니다.

2-4 : 시계열의 잔차나, 잡음 분석을 공부해가지 않아 제대로 풀지 못하였습니다.

1번 기계분석은 무난하게 나왔지만, 2번 시계열 통계분석은 정말 할말이 없다. 사실 시계열 분석이 연속으로 나올거라 예상하지도 못했으며, 난생 처음들어보는 분석을 하라고 할 줄은 몰랐다. 진흥원책자에 시계열 분석이 많지 않은데 왜이리 시계열 분석을 좋아하는지 모르겠습니다..(진흥원 책자에 SARIMA라는 것이 있는지도 모르겠다)

만약 이번 실기에 떨어진다면 시계열에 대해 더 공부해야할 것 같습니다. 추가로 SARIMA를 찾아보다가 MARIMA등 다양한 시계열 분석 방법이 있는데 이에 대해 잘 공부를 해야할 것 같습니다.

20회

1. 시계열분석(DATA : 온도예측) (총 50점)

- 1년간의 온도 데이터를 줍니다. 독립변수로는 년,월,일,실제온도(실제값-타겟변수), 지연값1(1일 지연 칼럼), 지연값2(2일 지연칼럼), Friend(친구가 예측한 칼럼 ->필요없는 칼럼), foreast_min(예측최소값), foreast(예측값), foreast_max(예측최대값)등의 칼럼이 있었습니다.
- 보통 모든 칼럼 설명을 해 놓는데 foreast_min(예측최소값으로 추정), foreast(예측값 으로 추정), foreast_max(예측최대값으로 추정)은 칼럼 설명이 없었습니다. 문제에서 설명을 해놓는걸 실수한 것인지, 데이터에 실수로 넣은건지, 의도한건지를 모르겠습니다. 이미 실제값이 있는데 예측값을 같이 넣어놓은 것이 무슨 의도로 넣은 건지 모르겠습니다.

1-1 : 데이터 전처리 (10점)

- 문제 : 결측값 처리, 필요없는 칼럼 처리, 데이터 전처리가 되었다는 증명, train/test셋을 어떻게 나눌지 등를 설명해아합니다.

1-2 : 위 데이터를 RandomForest로 검증하고 분석하기(15점)

- RandomForest 예측한계선을 설정하는 방법들을 말하고 어떤 방법을 써야하는지 논해야합니다..(예측한계선인지 정확히 단어가 기억이 안나는데, 단어의 뜻도 뭔지를 모르겠습니다.)
- RandomForest를 활용해 예측 및 검증하고, 데이터 및 파라미터를 조정하여 성능을 강화해야합니다.
- 칼럼별 중요도를 시각화해야 합니다.

1-3 : 위 데이터를 SVM(서포트 백터 머신)으로 검증하고 분석하기(15점)

- SVM 예측한계선을 설정하는 방법들을 말하고 어떤 방법을 써야하는지 논하라.(예측한계선인지 정확히 단어가 기억이 안나는데, 단어의 뜻도 뭔지를 모르겠습니다.)
- SVM를 활용해 예측 및 검증하고, 데이터 및 파라미터를 조정하여 성능을 강화해야합니다.
- 칼럼별 중요도를 시각화해야 합니다.

1-4 최적의 모델 선택하기

- SVM과 RandomForest의 장단점을 논하기
- 위 두모델 중에서 어떤 모델이 좋은지 고르고 그 이유를 설명하기

- 선택한 모델의 한계점을 이야기하고 한계점을 해결할 방법을 설명하기

문제 해결(1-1)

- 처음 문제를 보았을 때는 단순 회귀분석인줄 알았으나, 지연값 칼럼이 들어있다는 점, 1년치 데이터라는 점, train/test를 어떻게 나누어야 하는지를 물어보는 점을 보고 시계열 분석이라는 것을 깨달았습니다. (아닐 수도 있습니다만 가능성이 높습니다.)
- 따라서 타겟변수가 비정상적이라는 것을 ADF와 KPSS를 통해 증명한 후, 로그변환 및 차분을 통해 정상화 한 것을 보여주었습니다.
- train/test를 나누는 것은 시간순서를 기준으로 하여 나누었습니다.(시계열이기 때문에 train_test_split같이 단순히 랜덤으로 7:3으로 나누면 안됩니다.)
- Friend칼럼은 당연히 필요없기 때문에 삭제하였고, 실제값이 주어져있기 때문에 예측값 관련 칼럼들도 삭제하고 시작하였습니다.
- 1년 데이터를 가지고 검증하는 것이기 때문에 Year(년)칼럼도 삭제했습니다.

추가사항(cafe.naver.com/sqlpd/18900 참고)

https://towardsdatascience.com/random-forest-in-python-24d0893d51c0

- 위 링크와 문제가 굉장히 유사하게 나왔습니다. 참고하면 좋을 것 같습니다.
- actual을 기준으로 lag 1, lag2가 맞는지 검증했는데 몇개가 틀리다고 합니다. 데이터의 무결성을 체크하라는 점이 이부분인 것 같습니다.

2. 군집분석(DATA : 가구별 15분 단위의 전력사용량) (총 25점)

- 년월일, 시간, 가구코드, 전력사용량 등의 칼럼을 줍니다.

2-1 클러스터링(10점)

- 가구별, 15분 단위로(????) 전력사용량의 합을 구하고, 이 데이터를 군집화하고 다음과 같이 표를 완성시켜라

Aa 1열	2열	3열	4열
가구코드	DATE	Total_P(전력사용량의 합)	Cluster

2-2 Heatmap으로 시각화하기(15점)

그룹별로 15분마다의 전력사용량을 요일별로 평균낸것을 Heatmap으로 시각화하라

월요일	Aa 속성	1	2	3	4	5	6	7	8
@2021년 9월 21일	제목 없음								
@2021년 9월 22일	제목 없음								
@2021년 9월 30일	제목 없음								
@2021년 9월 24일	제목 없음								
@2021년 9월 25일	제목 없음								
@2021년 9월 19일	제목 없음								
	1:15	1:30	1:45	2:00	2:15	2:30	2:45	3:00	3:15

위 와 같은 그림으로 각 그룹별로 Heatmap으로 시각화하면 됩니다.(총 5개를 만들어야 합니다.)

문제 해결 방안

- 문제는 가구별, 15분 단위로 전력사용량의 합을 구한 다음, 군집화 하라고 하는데.. 애초에 데이터를 15분단위로 줬으면서 어떻게 하라는 건지 이해를 못했습니다.(문제 설명이 정말 거지 같습니다.) 표에서 시간칼럼이 DATE길래 날짜 단위로 합치는 것이라고 판단하여 가구별/일별 단위로 전력사용량의 합을 구한 다음 이를 클러스터링 하였습니다.(K-means를 사용하였지만 어떤 것을 사용하든 문제없습니다)
- 위에서 구한 클러스터링을 다시 원데이터처럼 15분 단위로 만든 뒤 요일별로 전력사용량을 평균낸 다음 그룹별로 히트맵을 만들어야 합니다. 히트맵예시는 시험지에 주어졌으나, 그것과 똑같이 만드는 것이 꽤 어려웠습니다.

3. 회귀분석 (총 25점)

- 데이터는 기억이 잘 안납니다. 하지만 회귀분석의 Flow를 보여주면 되는 문제였습니다.(전처리, 스케일링, 모델링, 예측, 검증 등..)
- train/test를 7:3으로 나누고, 검증 후 R2 score/RMSE/정확도 3개 지표를 구해야합니다.
- 회귀분석에는 정확도라는 지표가 없는데(분류분석에만 있습니다) 정확도를 구하는 공식을 주었습니다. 실제값>예측값인 경우 (1-예측값/실제값), 실제값<예측값인 경우 (1- 실제값/예측값)으로 하고 이것들을 평균내면 됩니다.(공식이 정확하지 않을 수 있습니다)

- 위 과정들을 담은 코드를 제출해야 합니다.
- 회귀분석 모델링은 어떤 모델을 쓰든 상관없습니다. 3개문제 중에서 그나마 무난한 문제였습니다.

출처 :https://ysyblog.tistory.com/221

21회

1. 학생 성적 예측

- 단계마다 적합한 시각화 제시
- 근거 제시

1-1.

1) 탐색적 데이터 분석 & 시각화

2) 결측치 식별하고 최소 2가지 이상의 결측치 예측 방법 제시

ㄴ 그 중 하나의 방법으로 보정

ㄴ 선택 기준과 타당성 반드시 제시

3) 인코딩이 필요한 항목과 이유 제시, 필요한 인코딩 수행

1-2.

1) 학습용 / 테스트용 분할 2가지 방법 제시

2) 학습용 / 테스트용 생성

1-3.

1) 랜덤포레스트, SVM, XGBoost의 공통적인 특징?

2) 위 3개 구축하고 최적 1개 선정

ㄴ 타당성과 성능 개선을 위해 추가 수행할 것?

ㄴ 현업에서 운영할 때 운영 고려사항?

2. 회귀

1. train-test 8:2 분리 → train으로 선형회귀 모델 & test 사용하여 RMSE, 결정계수

2. train-test 8:2 분리 → train으로 Ridge회귀 모델 & test 사용하여 RMSE, 결정계수

ㄴ alpha 0~1까지 0.1 간격으로 조정하여 가장 좋은 결정계수 갖는 alpha 찾기

3. train-test 8:2 분리 → train으로 Lasso 회귀 모델 & test 사용하여 RMSE, 결정계수

ㄴ alpha 0~1까지 0.1 간격으로 조정하여 가장 좋은 결정계수 갖는 alpha 찾기

3. 다항회귀 그래프 시각화 (12점)

다항회귀를 실시하고 (항의 개수는 3차까지) 각 항 별로 그래프를 그려 코드와 함께 제출

4. 이원분산분석 (9점)

- 독립변수 x1과 x2, 종속변수 y에 대하여 이원분산분석(Two-way ANOVA)를 실시
- 분석결과에 대한 통계량을 표로 제출하고 수행 결과를 해석해서 제시

출처 :https://mizykk.tistory.com/127

22회

1. 기계학습

피마 인디언 당뇨 데이터

특징 : 데이터에 헤더가 없음, 대신 시험지에 변수명 제공

1.1 데이터 탐색

1.1.1 탐색적 데이터 분석 수행하시오(시각화 포함)

- info(), describe(), 독립변수 전체 히스토그램, 타겟 분포 그래프(불균형 확인), 변수 전체 상관관계 및 히트맵, pairplot, 결측치 확인 및 시각화 등

1.1.2 이상치 처리하시오

- descsribe, pairplot 에서 이상치 식별 가능했기 때문에 논리적으로 다음 작업으로 타당하였음. inter quantile range 사용할 수 있었으나 EDA 단계에서 보기에 단일 변수의 분포에서 크게 벗어나는 99999, 999 등의 값이 1개 혹은 2개로 존재하는 경우가 많았고, 이는 의도적으로 이상치를 변수 범위에서 크게 벗어나도록 1개 내지 2개를 심어놓은듯한 모습이었음. 또한 나이 변수의 경우 최대값이 999였으므로 다음 최대값인 81로 상식적인 변경이 가능한 경우가 있었음. 즉 IQR(25%선 밑 75%선 위를 이상치로 보기에)을 적용하기에는 데이터에 대한 너무 큰 변환이라고 판단했고 IQR을 사용할 수 있었음을 설명하고 다음 규칙 기반으로 모든 변수에 대해 이상치 처리함

1) 단일변수 백분위에서 1%단위마다 위치하는 값 표시(변수.quantile(np.arange(0,1,0.01))

2) 단일변수 MIN, MAX값 확인

3) 단일변수 히스토그램 확인(크게 벗어난 위치 확인)

4) scatter plot에서 크게 벗어나는 한 개로 확인되면 다음 MAX값으로 변환

5) 상식적 판단 : age등은 21~81 까지 분포했으므로 999를 81로 변환하는 것으로 마무리

6) 상식적 판단 : 측정 수치등은 0이 나올 수 없으므로 1)확인 후 다음 MIN 값으로 변경

1.1.3 앞선 두 단계에서 얻은 향후 분석시 고려사항 작성

- 결측치 없음, 이상치 있음, 히스토그램 관찰시에 0부분이 모든 독립 변수에서 솟은 형태가 나타나므로 이미 결측치를 0으로 채운 데이터일 수도 있음을 짐작해볼 수 있음. EDA에서 변수에 소수점 값들이 있는 경우가 있으나 큰 의미가 없다고 판단되므로 정수 변환 고려해야함. (152.83, 152.72의 차이에 큰 의미가 없다고 판단함, 물론 0.2345 0.7234의 값을 갖는 변수에 대한 얘기는 아님) 선형 모델 사용시 스케일링 고려해야 함, 타겟 분포가 극심하지는 않으나 불균형함. (대략 우세 클래스 500 열세 238) 상관관계 확인 결과 변수의 효과가 거의 동일한 경우 관찰되어(0.99) 제거 고려해볼 수 있음 등

2.1 클래스 불균형을 처리하시오

2.1.1 업 샘플링 과정 설명하고 결과 작성

- imblearn 사용, EDA에서 이미 문제 알고 있었으므로 자연스러운 흐름임. up sampling은 소수 클래스를 늘려서 다수 클래스 개수에 맞춤. random방식과 smote방식 설명하고, 비교하였음

2.2.2 언더 샘플링 과정 설명하고 결과 작성

- imblearn 사용, 언더 샘플링은 다수 클래스를 줄여서 소수 클래스 개수에 맞춤. random방식과 tomeklink방식 설명하고, 비교하였음

2.2.3 둘 중 선택하고 이유 설명

- 둘 중 선택한다면 tomeklink를 선택한다고 하였음. 랜덤 다운 샘플링 방식은 다수 클래스에서 랜덤 샘플링을 통해 소수 클래스 숫자에 끌어당겨 맞추기 때문에 정보 손실이 많음(500개->238개). 랜덤 업 샘플링은 정보 손실이 없지만 중복 관측치가 생기게 됨. 둘 다 언더 피팅과 오버 피팅에 대한 우려를 만들어내게 됨. 굳이 고른다면 tomeklink를 통한 언더샘플링 방법 택한다고 했음. 이유는 정보 손실 최소화(500개->460개) 하면서 소수 클래스의 비율을 3%가량 증가시킬 수 있었기 때문임. 물론 smote는 knn을 통해 생성한 관측치이기 때문에 중복 생성을 안하므로 사용에 유리하겠지만, EDA에서 타겟과의 상관관계 등을 들어 타겟 불균형으로 인해 예측이 안될 정도(열세 클래스를 늘려야 할 정도로)는 아니라고 판단하였다고 작성함.

그러나 가장 좋은 방법이라면 둘 다 하지 않겠다고 추가 작성하였음. 현실에서는 클래스 불균형이 극심한 경우가 많은데 판단하기에 7:3의 클래스 비율이 학습이 안될 정도는 아니고, 데이터에 임의의 변환을 가하는 것은 정보 손실이나 과적합의 문제를 동반하기 때문에 더욱 위험하다고 생각했다고 하였음. 즉 원 데이터에 파생변수를 생성하는 것을 제외한 변환을 가하는 것이 온전한 분석은 아니기에 지양한다고 함.

3.1 모델링 하시오

3.1.1 최소 3개 이상 알고리즘 제시하고 정확도 측면의 모델 1개와 속도 측면의 모델 1개를 꼭 구현(총 2개 이상)

- 선형 알고리즘, 배깅 알고리즘, 부스팅 알고리즘에 대해 설명하였음.
- 파생변수 생성하였음 사칙연산, 연속형 binning, 범주형 빈도 카운트 변수 등 그냥 반사적으로 생각난 것들만 빠르게 수행함
- 각각의 모델로는 logistic regression, random forest, xgboost 사용하였음.

3.1.2 모델 비교하고 결과 설명

- 타겟 불균형 문제가 계속 언급되고, 전체 데이터 관측치 수가 적다고 판단하여 5fold stratifiedkfold 사용하였음.
- %time으로 fit, predict 시간 각각 측정하고 cross_val_score사용해서 모델별 mean accuracy 출력하였음.
- 정확도 측면에선 xgboost 속도에선 logistic이 좋았음, 그러나 정확도도 속도도 유의미한 차이라고 보기에는 어려움이 있다고 함. 하지만 향후에 데이터 관측치가 더 수집되고, 추가적인 변수들이 생겨서 더 복잡한 모델을 사용하게 된다면 그 때는 유의미한 비교가 될 수 있을거 같다고 씀.

3.1.3 속도 개선을 위한 차원 축소 설명하고 수행, 예측 성능과 속도 비교하고 결과 작성

- pca 개념 설명하고 주로 사용하는 이유 설명함. scaling 필요한 이유 설명하였음, standardscaler 적용하였음.
- scaling, pca 모두 적용시에 test leakage 조심해야된다고 했음. (전체에 fit transform하면 안된다고)
- 설명 분산 누적 90%로 설정하였음.
- 요약된 주성분이 전체 데이터의 93%를 설명한다고 하였음.
- 예측 성능 다소 하락 관찰되고 속도에도 차원축소에 의한 유의미한 차이가 없으나, 향후에 데이터가 크게 늘어난다면 이러한 차원 축소적인 접근이 의미가 있을거라고 작성했음.

2. 통계분석

1.1 금속 성분 함유량 변수 1개. (1열 데이터) 제품에 금속 재질 함유량의 분산이 1.3을 넘으면 불량이라고 보고있는데, 제조사별로 차이가 난다고 제보를 받음 분산에 대해 검정을 수행하시오.

1.1.1 연구가설과 귀무가설 작성

1.1.2 양측 검정 어쩌고

1.1.3 검정통계량, 가설 채택

2.1 Lot별 불량 제품 수량 데이터. lot 번호와 불량제품수 두 개의 열. 각 lot별 200개에 대한 불량제품 수.

2.1.1 p관리도에 따라 관리중심선(center line), 관리 상한선, 하한선 구하시오

2.1.2 관리도 시각화 하시오

- **시각화는 구글에 python control chart 검색하면 많이 나오고, 결국 관리도가 생소한 말이지만, 관리 중심선 = 평균, 관리 상한선과 하한선은 평균 기준 3표준편차씩 거리에 위치한 선입니다.**

3.1 데이터 없음. 표에 제품 1,2를 만드는데 사용되는 재료 a b c 컬럼 있고 재료에 따라 최종 만들어지는 제품 두 개에 대한 수량 있음. 최하단 행에는 수익이 있음. 제품 수량을 최대로 뽑으면서 수익이 최적이 되도록 하라고 함.(10점)

4.1 데이터 없음. 상품 a와 b가 있을 때 다음과 같은 구매 패턴이 있다고 함. aa bb bbbb aa aaa bb bbb aa bb a b 정확히 기억 안나지만 대충 비슷함.

4.1.1 구매하는 패턴으로 봐서 두 상품이 연관이 있는지 가설 세우고 검정하시오

4.1.2 연구가설 귀무가설 세우시오

4.1.3 가설 채택하시오

- **runs test, 무작위성을 검정하기 위한 방법으로 귀무가설은 표본이 무작위로 추출되었다 입니다.**
  statsmodels.sandbox.stats.runs import runstest_1samp, runtest_2samp를 사용할 수 있습니다.

출처

https://cafe.naver.com/sqlpd?
iframe_url_utf8=%2FArticleRead.nhn%3FreferrerAllArticles%3Dfalse%26menuid%3D78%26page%3D1%26searchBy%3D0%2(

adp 정보모음 사이트

https://datamanim.com