APD 실기 정복 프로젝트!

- 접수일 : 06.29 ~ 08.24
- 환불마감 : ~09.02
- 시험일 : 09.19(토)
- 결과발표 : 10.20

참고 사이트

https://didalsgur.tistory.com/25

https://0dood0.tistory.com/33

https://blog.naver.com/sinrakr/221778920182

https://blog.naver.com/PostView.nhn?blogId=sinrakr&logNo=221762948199&parentCategoryNo=&categoryNo=21&viewDate=&isShowPopularPosts=true&from=search

-

https://blog.naver.com/boazzong92/221691195637

https://blog.naver.com/elmidion/221679470227

> 통계분석 / 데이터 마이닝 / 텍스트 마이닝 (14~15회 실기시험에는 텍스트 마이닝 출제 안됨.)
>
> 1. 통계분석
>   - 폐활양 예측문제 (나이, 키, 성별, 흡연유무)
>   - EDA 및 상관관계 분석
>   - 적절한 회귀모형 선택
>   - 회귀모형 해석
>   - 평균 키, 나이 일 경우 폐활양 예측
>   - 통계 분석 문제
>   - 변수 시각화 (변수간 상관관계, 변수별 이상치 파악)
>   - 회귀모형 적합과 유의성 검정
>   - 회귀 계수에 대한 standard error가 가지는 의미
>   - 회귀분석에서 잔차 분석 및 시각화
>   - 회귀분석에서 영향력 관측치와 그 영향 분석
> 2. 데이터 마이닝
>   - 백화점 사용패턴 분석
>   - 파생 변숙 생성 및 EDA
>   - 군집분석 및 적절한 집단 갯수 설정
>   - 세그멘테이션 별 의미 파악 (집단 라벨링)
>   - 세그멘테이션 별 마케팅 인사이트 도축
>   - 고객 구매데이터 분석
>   - 구객 구매데이터에서 이후 1개월 간 실제 고객이 구매할 것으로 예정되는 지점 추천
>   - 5개 지점 추천 후 하나라도 맞으면 맞는 것으로 간주 적중률 66.7%이상인 경우만 채점

    - 타이타닉 데이터 분석
        - 상세 정보 없음
3. 텍스트 마이닝
    - TV PRogram Buz 분석
        - tvpro 파일에 있는 단어들을 사전에 추가하기
        - tvpor 파일을 읽고 데이터 전처리
        - 월별/프로그램 별 나온 단어 분석
        - 월별 프로그램 비율 확인
    - 뉴스 기사 분석
        - 뉴스기사 로딩 및 제공된 긍정/부정 어휘를 통한 감성분석
        - 위에서 구한 긍정부정 score를 통해 N개의 그룹으로 클러스터링
    - 영화 리뷰 분석
        - 상세 정보 없음

14회 문제 복기

1. 기계 학습을 이용하여 집 가격 예측 및 검증
2. 다중 로지스틱 회귀 분석 및 confusion matrix 해석

15회 문제 복기

1. 제조 생산 데이터 분석
    - 데이터 탐색 : EDA
    - 데이터 전처리 : 변수선택 (VIF), 파생변수 생성, 데이터분할(train/validation/test(0.2))
    - 로지스틱 분석 : 분류1을 판단하는 모델 생성(종속변수는 총 7개 분류, 분류1 외의 값은 0으로 치환), confusion matrix 해석
    - 로지스틱 분석 외 3개 이상 분류 모델 생성 및 결과 해석 : SVM필수 포함, Precision/Sensitivity 결과 출력
    - 위 모델 중 최고 모델을 선택하여 최적의 군집 개수를 선택하고 클러스터링 수행 : F-1 score 출력
2. 데이터 처리 및 통계 분석
    - timestamp처리 / data 기중 데이터 병합
    - hh:mm, A/B/C/D/E 전력 사용량 데이터
    - yymmdd, 평균기온
    - 요일간 사용량 분석을 수행하고 가장 차이가 있는 요일 도축

https://cafe.naver.com/sqlpd/10794

17회 문제 복기

1. 보고서 제출
    - `R Notebook`, `word`이용해서 보고서를 작성하고 `pdf`로 제출합시다
2. 데이터전처리/기계학습 문제 (25점)
    - 집값예측
    - EDA
    - 모델생성
        - 데이터 분할
        - 집값에 영향을 미칠만한 컬럼(방개수,부엌,리모델링여부)
        - 시각화/전처리/회귀모델평가/규제/앙상블
        - 교호작용을 고려한 다중 선형 회귀

        - 3가지 분류 모델 생성 및 비교, 좋은 모델 선택
    3. 시각화 및 시계열 분석 (25점)
        - 코로나 데이터 (국가별/일별/인구수/확진자수/사망자수/완치자수/검사자수)
        - 시각화/저처리
        - 인구대비 확진자수 도출 (파생컬럼) > TOP5인 국가를 추출 후 시각화하는 문제
        - 전체 인구대비 누적 사망률이 가장 높은 5개 국가 추출 후 국가별 일일확진자, 누적확진자,일일 사망자, 누적사망자 시계열 그래프 출력
        - 확진자수,사망자수,인구수,검사자수,완치자수 등 변수를 활용하여 `위험지수` 파생컬럼 만들기 > 왜 그렇게 만들었는 지 설명
        - 시계열 분석 및 한국의 확진자수 예측
        - 비시계열 모델로도 모델을 별도 생성하기
    4. 통계분석(50점)
        - 설문데이터 분석 (사전에 역문항들에 대한 처리 필요)
        - 데이터 : 조사번호, 그룹, 문항1-1,1-2.....6-8 만족도 지수를 1~5점 값이 들어있음
        - 1-1~1-x 는 항목1
        - 2-1~2-x는 항목 2
        - 그리고 1-1의 역항목은 1-3임.. 이런식으로 역문항에 대한 처리도 해야함
        - 그룹별 평균, 표준편차, 왜도, 첨도 산출 (기술통계)
        - 요인분석
        - 신뢰성 지수를 구하는 식을 주고 그 값을 구하는 문제
        - 나머지 기억안님..

    문제의 범위가 넓어졌다. 문제를 푸는데 최신기술, 시각화, 기존 전통통계분석 다 잘해야함!!

주로 나오는 내용

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

과목 1. 데이터 전처리

- **EDA**
    - 파생변수 만들거나 시각화 (연속형변수 > 범주화)
    - 데이ㅓ 상관관계를 파악할 수 있는 형태로 요약 (월별+성별 데이터 요약, 시간대별 추이)
- **데이터 전처리**
    - 결측치 처리
    - timestamp 처리
    - 데이터 중복 및 누락 확인
- **변수 선정**
    - 요약변수 및 파생변수 선정
    - 어떤 변수를 독립변수로 선택할 건지, 또 어떤 변수를 종속변수로 선택할 건지

과목2. 머신러닝

- 데이터 분류 및 예측
    - SVM
    - 군집분석(고객세분화)
    - 연광성 분석 (장바구니분석)

    - 의사결정나무
    - 로지스틱 회귀분석
    - 여러 모형을 사용하여 예측력 검정하고, 최적 모형 선정
    - 랜덤포레스트
    - 앙상블

과목3. 통계분석

- **교호작용 분석**
    - 카이제곱검정...
    - 변수 A,B 간의 유의미한 관계가 있나요
    - 교호작용 확인 후 인사이트 도축
- **회귀분석**
    - 변수 선택과정
    - 잔차분석
    - 최적모형 선정 후 인사이트 도출

패키지

ggplot, dplyr

.공부 참고 자료

캘리포니아 집값 예측 : https://didalsgur.tistory.com/37?category=665947

15회 실기 자세한 후기 : https://cafe.naver.com/sqlpd/10794

연습문제 : https://cafe.naver.com/sqlpd/12643 (https://github.com/tshahn)

통계학 관련 : https://statwith.tistory.com/706

R을 이용한 데이터 처리 : https://thebook.io/006723/