Data Analysis Practical Test E15: Test Notebook

TAESEUNG HAHN
https://github.com/tshahn/DataAnalPrac

Question No. 1

주어진 데이터 파일은 특정 철강사의 제품코드, 불량코드, 그리고 공정 과정에서 발생한 데이터를 담고 있다.
해당 데이터를 이용하여 다음 문제의 답변을 작성하시오.

제공 데이터 파일: E15Q1_data_raw.csv

- 1-24번 컬럼: Analog Data
- 25번 컬럼: 제품코드 (Binary)
- 26번 컬럼: 불량코드 (Integer with range 1 to 7)

① EDA를 실시하여 결과값을 제시하고, 상관분석을 시행하여 변수 선택 및 파생 변수 생성과정을 풀이하시오.

Load Data

Check and Transform DType

EDA: Check missing values

EDA: Check Distributions

---

② 전체 데이터를 Train, Validataion, Test 용도로 분할하고 시각화하시오. (각 비율: 50%, 30%, 20%)

---

③ 불량코드 1에 대하여, Logistic Regression을 활용하여 이항분류 모형을 생성하시오.

생성한 모형에 대한 최적의 Cut-Off Value를 선정 후, Confusion Matrix를 제시하시오. (반드시 시각화와 통계량을 포함시킬 것)

④ Logistic Regression을 제외하고 SVM을 포함하여 3가지 다항 분류 모형을 만들어 Precision과 Sensitivity(TPR)를 제시하시오.

또한 모델향상과정과 최적화 과정을 통해 Confusion Matrix를 도출하시오.

⑤ 상기 ③번과 ④번 4가지 모형 중 1가지를 선택하여 최적의 클러스터링 개수(단일집단~5개)를 제시하시오.

또한 군집분석을 이용한 모형성능 향상 과정을 수행하여, 성능향상 전후의 F1 Score와 모형 평가 결과를 제시하시오.

Question No. 2

주어진 3개의 파일들은 한 공장의 전력 사용량에 대한 데이터로써, 각각 날씨와 온도, 용도별 전력량계, 전력 총 사용량을 담고있다.
해당 데이터를 종합적으로 이용하여 다음 문제를 풀이하시오.

데이터 파일 설명:

1. E15Q21_usage.csv
    - 900초마다 기록된 900초 단위 전력 총 사용량
    - 1번 컬럼: Datetime (UnixTimestamp)
    - 2번 컬럼: Usage
2. E15Q22_weather.csv
    - 일자별 평균 기온
    - 1번 컬럼: Date (YYYY-MM-DD)
    - 2번 컬럼: Daily Average Temperature
3. E15Q23_usage_history.tsv
    - 1분에 2번씩 기록된 각 용도별 전력 누적사용량
    - 1번 컬럼: Time (HH:MM)
    - 2번 컬럼: Weather Class (A/B/C/D)
    - 3-7번 컬럼: 각 용도(A/B/C/D/E)별 전력 누적 사용량

LOAD DATA

① 첫번째 제공 파일의 총사용량 컬럼을 용도별로 분류하고, 연월과 사용 목적별로 전력의 하루 평균 사용량을 구하여 도표를 도출하시오.

YYYYMM	A	B	C	D	E
202001	__	__	__	__	__
202002	__	__	__	__	__
202003	__	__	__	__	__

CREATE DIFF FROM CUMSUM

CREATE JOIN KEY

JOIN DATASETS

---

② 요일별 평균 전력사용량을 도출하시오. 또한 가로축을 요일, 세로축을 평균사용량으로 하여 요일별 평균 사용량을 시각화하여 제출하시오.

---

③ 요일별 총 전력 사용량의 평균값의 차이를 분석하여, 가장 큰 차이를 보이는 요일은 어떤 요일인지 제시하시오.

---

④ 각 날짜별 평균 기온과 용도별 전력사용량의 관계를 분석하여, 기온과 가장 밀접한 관계를 지닌 사용 용도의 종류를 제시하시오.