Bigdata Analytics

Final Report

INE5015 – 22057

빅데이터 애널리틱스 8조

| 최준희 | 이강산 | 장혜연 | 황태영 |

Written by python@hanyang.ac.kr

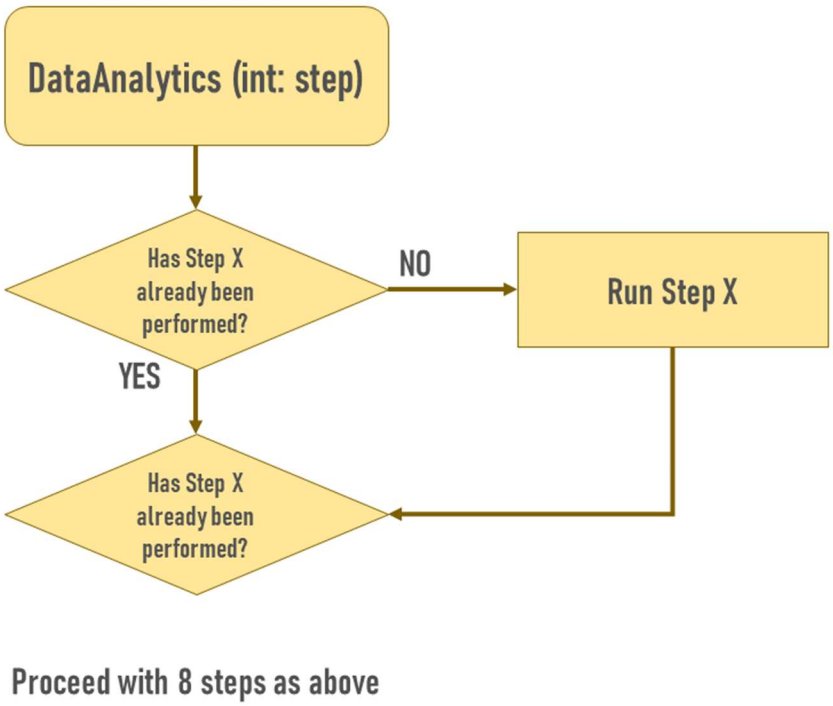
List of Contents

GITHUB : https://github.com/ChoiJunhee/INE5015_bigdata_analytics

Written by python@hanyang.ac.kr

Logic of our Data Processing

전반적인 Preprocessing 과정은 기능, 목적에 따라 Step으로 구별했고, 매 Step마다 시각적, 통계적 데이터를 확인하면서 진행하였다. 아래는 Step에 대한 구성 Logic이다.



DataAnalytics 함수에서 필요한 함수를 호출하여 사용하는 구조이며, 각 단계마다 csv 파일을 저장하도록 하여 변화를 직관적으로 확인하기 용이하도록 디자인하였다.

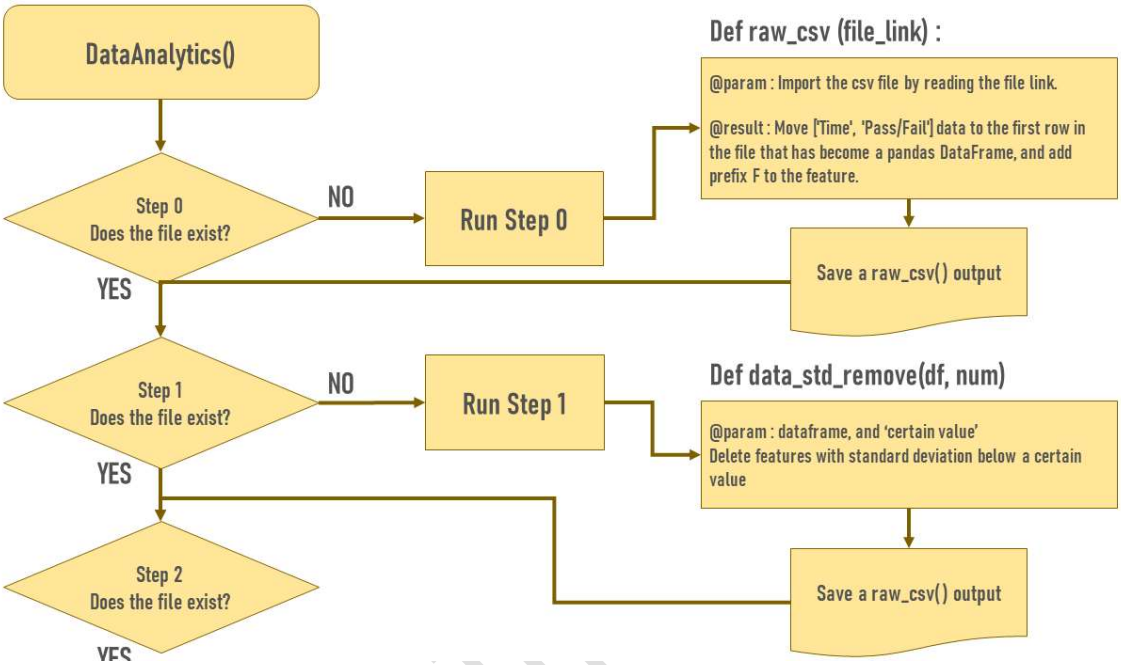
각 Step을 크게 나누어 보면 Raw data를 가공에 용이하도록 조정하는 부분, Pass/Fail 데이터를 분리해 총 3가지 데이터셋으로 나누어 진행할 수 있도록 하는 부분이 있다.

이후 각 Feature (독립변수)간 종속성을 확인하고, 제거하기 위해 진행하는 Correlation 과, 결측치 처리/보정 과정, 이상치 처리/보정 과정. 즉, 데이터 클리닝 과정이 있다.

마지막으로 Feature Selection과 샘플링을 통해 최종 데이터 셋 후보를 선정하고, 퍼포먼스를 확인하는 과정으로 마무리한다.

Step 0 ~ Step 2 : Ready to Preprocess

이 단계는 데이터 파일을 불러오고, 이후 보정에 있어 편의성을 높이기 위해 보정한다.



Explanation of Steps

STEP	STEP DESCRIPTION
STEP 0	raw_csv 함수를 통해, Pandas Read_CSV를 실행하여 데이터 셋을 받아온다. 이후 작업의 편의를 위해 행 위치 변경, Prefix 추가 등의 과정을 거친 Dataframe 파일을 받고, 저장한다.
STEP 1	일정 계수 미만의 표준편차를 가진, 특징 선택의 필요도가 낮은 Feature들을 제거하고, Dataframe 파일을 저장한다. 그리고 데이터 셋을 3개로 나누는 과정을 거친다.
STEP 2	데이터 셋은 Pass Data, Fail Data, Both Data로 나뉘며, 이렇게 데이터 셋을 나누는 것은 프로젝트 초기의 아이디어 중 유용한, 유의미한 결과를 가져오는 계기가 된다.

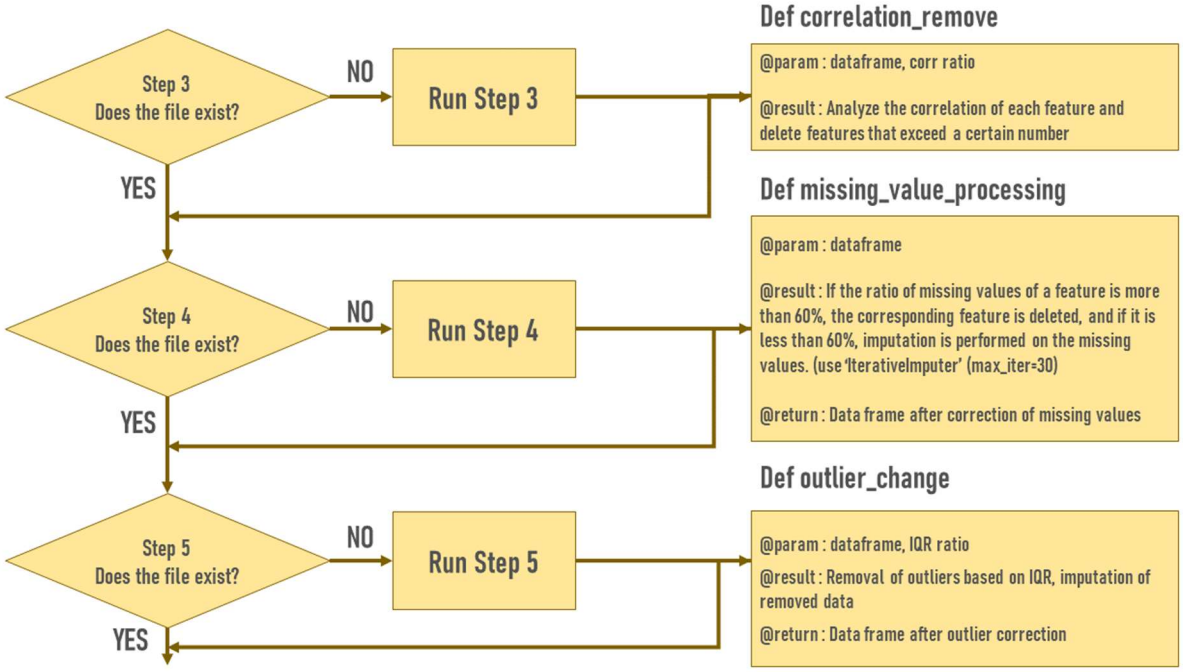
Step 0 부터 Step 2까지의 3 과정은 데이터 전 처리를 위한 준비 단계에 불과하다.

> 위 과정에서 1566개 (1463+104)의 테스트 케이스, 247개의 Feature Data가 남았다.

Step 3 ~ Step 5 : Data Cleaning

이 단계에서는 데이터셋의 노이즈를 제거하기 위해 데이터 클리닝을 진행하는 과정이다.

크게 독립변수 간 종속성 (비율), 결측치 제거 및 보정, 이상치 제거 및 보정으로 나뉜다.



Explanation of Steps

STEP	STEP DESCRIPTION
STEP 3	각 독립변수 (Feature)간 상관성이 높다는 것은 다중공선성 발생 소지가 있고, 정확한 예측을 방해하는 요인으로 작용할 수 있다. 상관관계가 높은 Feature들을 제거하는 과정이다.
STEP 4	데이터에 결측치가 많아도 정확한 예측을 방해한다. 독립변수의 수를 최대한 유지하기 위해, 결측치가 60% 이상인 Feature를 제거하고, 60% 미만은 IterativeImputer를 통해 보정한다.
STEP 5	각 Featuer의 사분위 값을 확인하고, IQR 방식을 통해 양 극단에 있는 이상치들을 제거한다. 이 때, Fail 데이터셋은 오버샘플링을 진행하지 않았기 때문에 가중치를 달리 설정한다.

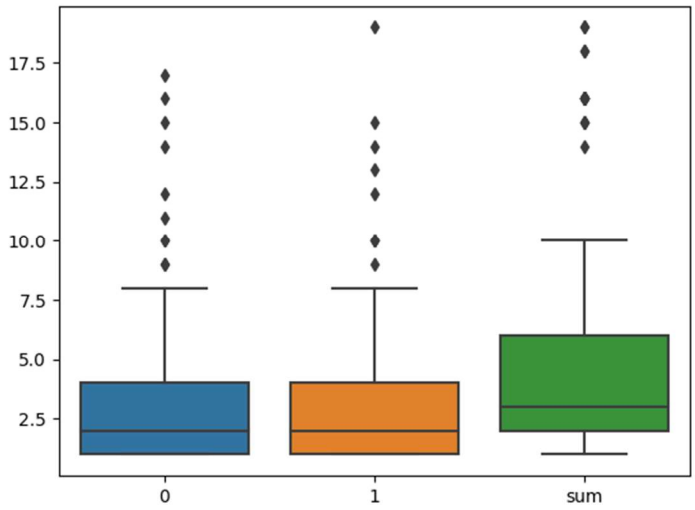
다음 페이지에는 Step3, 4, 5의 구체적인 내용과 편차 제거의 효용성을 비교한다.

Data Cleaning - Overview

데이터 전처리 과정에 있어, 데이터 클리닝 과정은 매우 중요하다. 특히 결측치와 이상치에 대해 처리하는 것은 데이터 분석 및 모델링 결과를 크게 변화할 수 있기 때문에, 데이터의 불안전성과 잡음, 불일치 등을 최대한 효과적으로 처리해야 한다.

Step 3 – Correlation

상관관계를 분석 (Correlation Analysis)을 한다는 것은 통계학적으로 두 변수간 선형적 관계를 분석하는 것이다. 독립 변수 (=Feature)간 상관관계가 높다면 두 변수의 연관성이 높다는 것이고, 필요치 않은 연산을 할 뿐만 아니라 Clustering 에도 방해가 된다.



우리는 이 프로젝트 과정에서 상관관계가 높은 Feature들을 계산했고 아래 예시처럼 선택된 Feature들을 제거했다. (여기서, Fail 데이터와 All/Pass 데이터간 상관관계 분석 내용이 달라 약 20여개의 Feature가 Fail 데이터에는 남아있게 되었다.

(시각화 자료 들어갈 위치)

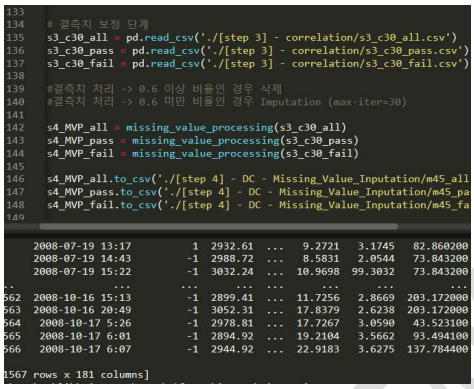
Step 4 – Missing Value

이 프로젝트에 있는 결측치는 패턴이 없는, Random Missing Feature 라 가정하고 진행하였다. 결측치를 처리하는 방법에는 삭제, 대치, 예측이 있는데, 결측치들의 특성이 패턴을 가지고 있다는 가정 하에 예측 모델을 구현해야 하기 때문에 이 프로젝트에서는 Deletion, Imputation 두 가지 방법을 사용하였다.

Scikit Learn에서 제공하고 있는 impute 중 Iterative Imputer를 사용하였다. 다른 모든 특성에서 개별 특성을 추정하는 다변량 대치 방식이며, Round Robin 알고리즘으로 각 Feature를 모델링 하여 결측값을 대치하는 기능을 한다.

또한 Iterative Imputer는 KNN 알고리즘으로 결측 치를 예측하여 채워 넣는 방식인데,

Max-Iter를 30으로 조정함으로써 최종 라운드 동안 계산된 결과를 반환하기전에, 수행할 Round의 수를 늘림으로써 데이터의 완전성이 높아지기를 기대하였다.



```
# 결측치 보정 단계
s3_c30_all = pd.read_csv('./[step 3] - correlation/s3_c30_all.csv')
s3_c30_pass = pd.read_csv('./[step 3] - correlation/s3_c30_pass.csv')
s3_c30_fail = pd.read_csv('./[step 3] - correlation/s3_c30_fail.csv')

#결측치 처리 -> 0.6 이상 비율인 경우 삭제
#결측치 처리 -> 0.6 미만 비율인 경우 Imputation (max-iter=30)

s4_MVP_all = missing_value_processing(s3_c30_all)
s4_MVP_pass = missing_value_processing(s3_c30_pass)
s4_MVP_fail = missing_value_processing(s3_c30_fail)

s4_MVP_all.to_csv('./[step 4] - DC - Missing_Value_Inputation/m45_all
s4_MVP_pass.to_csv('./[step 4] - DC - Missing_Value_Inputation/m45_pa
s4_MVP_fail.to_csv('./[step 4] - DC - Missing_Value_Inputation/m45_fa
```

```
    2008-07-19 13:17         1  2932.61  ...   9.2721   3.1745   82.860200
    2008-07-19 14:43        -1  2988.72  ...   8.5831   2.0544   73.843200
    2008-07-19 15:22        -1  3032.24  ...  10.9698  99.3032   73.843200
..                         ...      ...  ...      ...      ...         ...
562  2008-10-16 15:13        -1  2899.41  ...  11.7256   2.8669  203.172000
563  2008-10-16 20:49        -1  3052.31  ...  17.8379   2.6238  203.172000
564   2008-10-17 5:26        -1  2978.81  ...  17.7267   3.0590   43.523100
565   2008-10-17 6:01        -1  2894.92  ...  19.2104   3.5662   93.494100
566   2008-10-17 6:07        -1  2944.92  ...  22.9183   3.6275  137.784400

1567 rows x 181 columns]
```

ALL

```
...                  ...         ...      ...  ...      ...      ...         ...
1458  2008-10-16 15:13        -1  2899.41  ...  11.7256   2.8669  203.172000
1459  2008-10-16 20:49        -1  3052.31  ...  17.8379   2.6238  203.172000
1460   2008-10-17 5:26        -1  2978.81  ...  17.7267   3.0590   43.523100
1461   2008-10-17 6:01        -1  2894.92  ...  19.2104   3.5662   93.494100
1462   2008-10-17 6:07        -1  2944.92  ...  22.9183   3.6275  137.784400

[1463 rows x 181 columns]
```

PASS

```
..                 ...         ...      ...  ...      ...      ...       ...
99   2008-06-10 15:00          1  2988.39  ...    0.000000  3.0992    0.0000
100  2008-07-10 13:10          1  3052.98  ...   52.701400  3.1106   52.7014
101   2008-09-10 4:34          1  2951.84  ...   93.125224  2.3773   67.7994
102  2008-09-10 15:55          1  3173.18  ...   88.152800  3.8289   88.1528
103   2008-10-15 2:42          1  2903.34  ...   85.312200  3.5250   85.3122

[104 rows x 201 columns]
```

FAIL

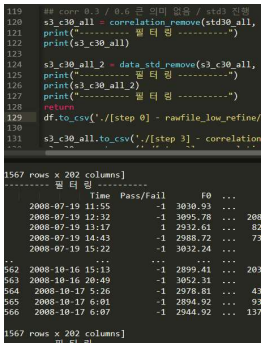
Normal (ALL) 데이터와 Fail 데이터가 다른 이유는, STEP3 에서 약 20여개의 Feature 가 덜 제거된 것으로 보인다. (이 페이지는 수정될 예정입니다)

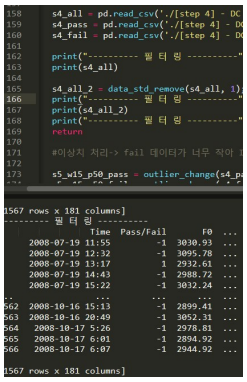
Step 5 – Outlier Value

무야호~

여기서는 Step 3, Step 4와 다르게 편차가 달라지기 때문에, 편차 제거를 한번 더 진행하였다.

Step 3 / 4





STEP 5 (유의미!)

(시각화 데이터 필요)

Feature Selection

여기서는 RFE, KBS 내용 설명

Written by python@hanyang.ac.kr

오버샘플링

STD와 MMS 간 차이가 없음

오버샘플링 SMOTE 랜덤 100회 평균 (test_size = 0.2)

acc, pre, rec, f1, roc_curve

```
RFE_MMS_ALL
0.8061 0.7864 0.8414 0.8128 0.8807
```

```
RFE_STD_ALL
0.8022 0.7848 0.8338 0.8084 0.8793
```

```
KBS_MMS_ALL
0.7999 0.7774 0.8414 0.808 0.873
```

```
KBS_STD_ALL
0.7997 0.7774 0.8408 0.8077 0.8728
```

RFE > KBS, MMS와 STD는 큰 차이 없음.

성능 평가

각 모델에 맞추어 다시 성능을 평가 해야 함.