

Bigdata Analytics

Final Report

INE5015 – 22057

DO NOT LEAK THIS DOCUMENT

빅데이터 애널리틱스 8조

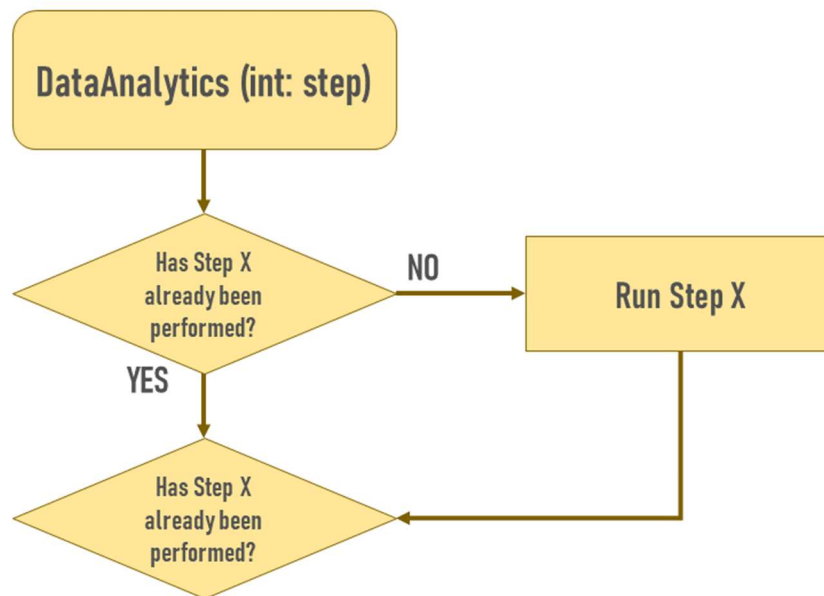
| 최준희 | 이강산 | 장혜연 | 황태영 |

List of Contents

- Logic of Data Preprocessing
- step-by-step process
 - Step 0 ~ Step 2 : raw file refining
 - Data Cleaning Overview (3 Steps)
 - ◆ Step 3 : correlation check and correction
 - ◆ Step 4 : Missing Value Imputation
 - ◆ Step 5 : Outlier corrections
 - Step 6 : Data Scaling
- Feature Selection
- Data over/down Sampling overview
- Performance
 - The performance of two samplings
 - Prediction accuracy according to each algorithm
 - ◆ Logistic Regression
 - ◆ Decision Tree
 - ◆ Random Forest
 - ◆ Boosting

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.



Proceed with 8 steps as above

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

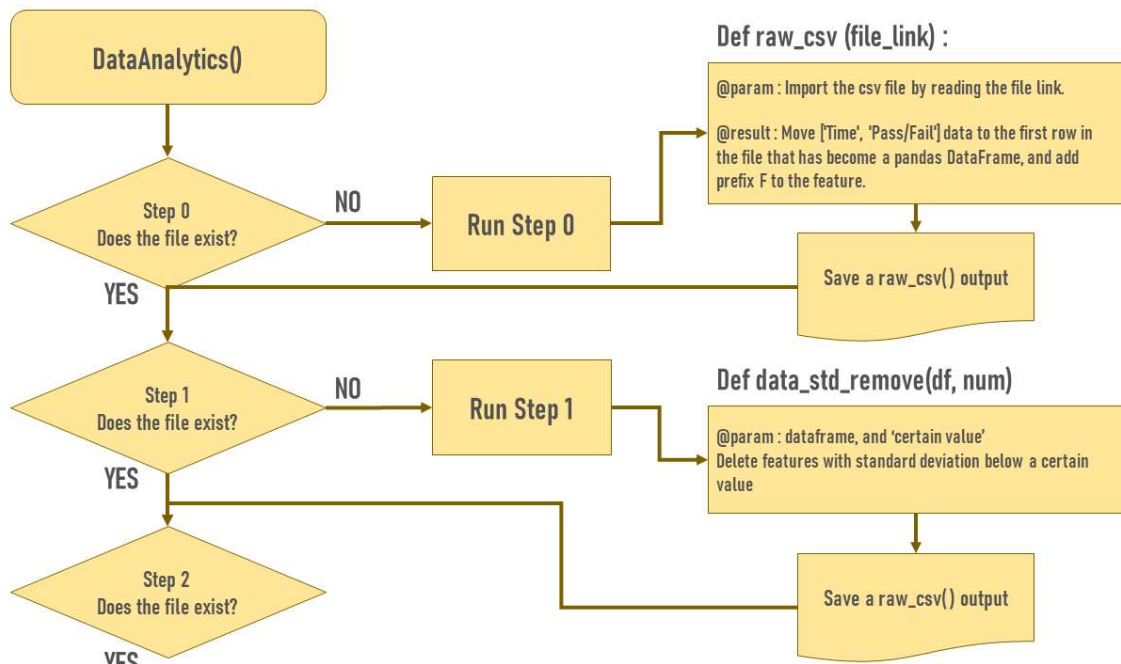
각 Step을 크게 나누어 보면 Raw data를 가공에 용이하도록 조정하는 부분, Pass/Fail 데이터를 분리해 총 3가지 데이터셋으로 나누어 진행할 수 있도록 하는 부분이 있다.

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

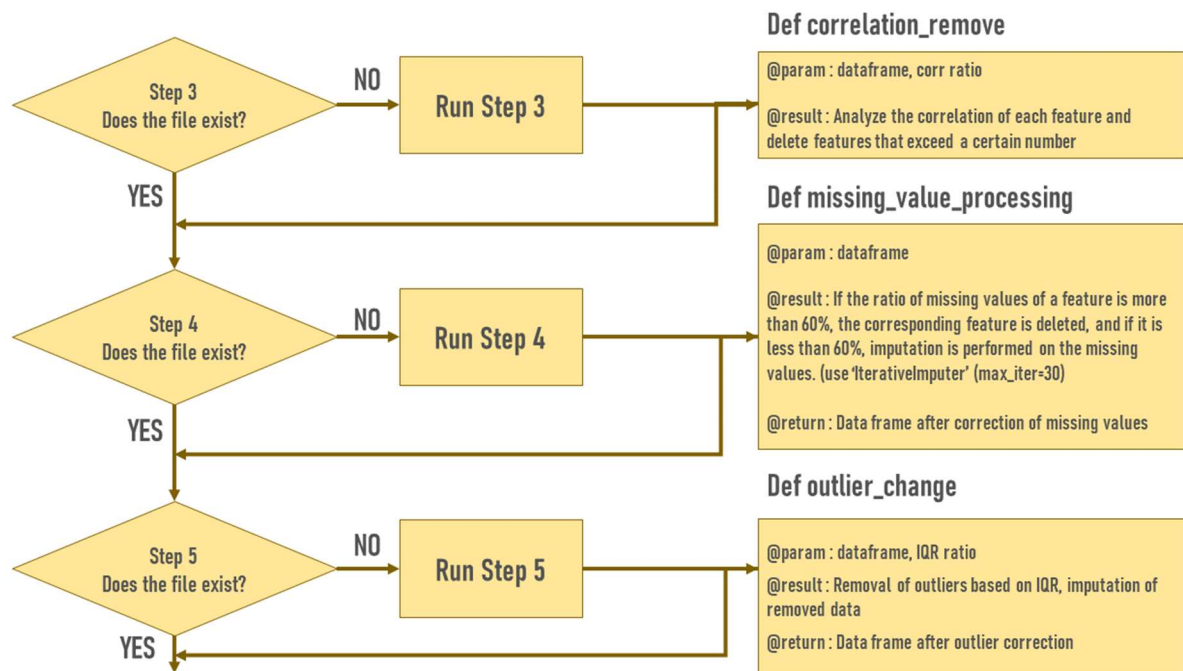
STEP	STEP DESCRIPTION
STEP 0	raw_csv 함수를 통해, Pandas Read_CSV를 실행하여 데이터셋을 받아온다. 이후 작업의 편의를 위해 행 위치 변경, Prefix 추가 등의 과정을 거친 Dataframe 파일을 받고, 저장한다.
STEP 1	일정 계수 미만의 표준편차를 가진, 특징 선택의 필요도가 낮은 Feature들을 제거하고, Dataframe 파일을 저장한다. 그리고 데이터 셋을 3개로 나누는 과정을 거친다.
STEP 2	데이터 셋은 Pass Data, Fail Data, Both Data로 나뉘며, 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 Feature의 사분위 값을 확인하고, IQR 방식을 통해 양 극단에 있는 이상치들을 제거한다. 이 때, Fail 데이터셋은 오버샘플링을 진행하지 않았기 때문에 가중치를 달리 설정한다.

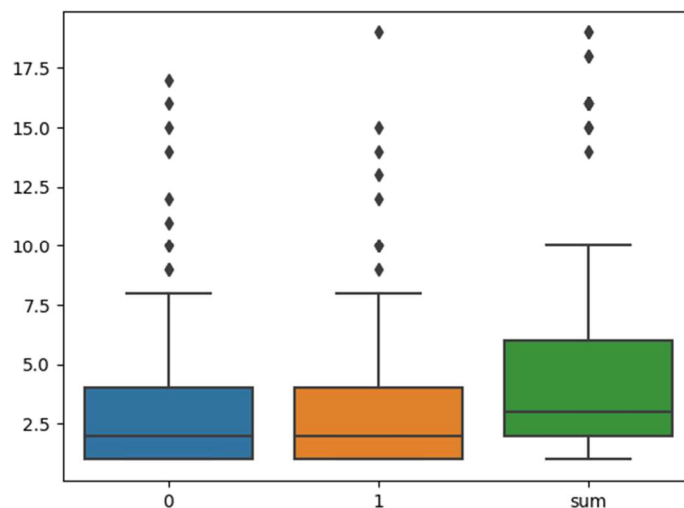
다음 페이지에는 Step3, 4, 5의 구체적인 내용과 편차 제거의 효용성을 비교한다.

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

```

133 # 결측치 보충 단계
134 s3_c30_all = pd.read_csv('./[step 3] - correlation/s3_c30_all.csv')
135 s3_c30_pass = pd.read_csv('./[step 3] - correlation/s3_c30_pass.csv')
136 s3_c30_fail = pd.read_csv('./[step 3] - correlation/s3_c30_fail.csv')
137
138 #결측치 처리 -> 0.6 이상 비율인 경우 삭제
139 #결측치 처리 -> 0.6 미만 비율인 경우 Imputation (max-iter=30)
140
141
142 s4_MVP_all = missing_value_processing(s3_c30_all)
143 s4_MVP_pass = missing_value_processing(s3_c30_pass)
144 s4_MVP_fail = missing_value_processing(s3_c30_fail)
145
146 s4_MVP_all.to_csv('./[step 4] - DC - Missing_Value_Inputation/m45_all')
147 s4_MVP_pass.to_csv('./[step 4] - DC - Missing_Value_Inputation/m45_pa')
148 s4_MVP_fail.to_csv('./[step 4] - DC - Missing_Value_Inputation/m45_fa')
149
150
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
...
2008-10-16 15:13   -1  2899.41    ...  11.7256   2.8669  203.172000
2008-10-16 20:49   -1  3052.31    ...  17.8379   2.6238  203.172000
2008-10-17 5:26    -1  2978.81    ...  17.7267   3.0590  43.523100
2008-10-17 6:01    -1  2894.92    ...  19.2104   3.5662  93.494100
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 Feature가 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

```

110 s3_c3o_all correlation_remove(std3o_all,
111 print("----- 필터링 -----"))
112 print(s3_c3o_all)
113
114 s3_c3o_all_2 = data_std_remove(s3_c3o_all,
115 print("----- 필터링 -----"))
116 print(s3_c3o_all_2)
117 print(s3_c3o_all_2)
118 return
119 dt.to_csv("../[step 0] - rawfile_low_refiner/
120
121 s3_c3o_all.to_csv("../[step 3] - correlation
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1
```

STEP 5 (유의미!)

[illegible]

(시각화 데이터 필요)