백화점 트랜잭션 데이터를 활용한 성별 예측 모델 개발

: 머신 러닝, 딥러닝 기반으로

Kookmin univ

Bigdata Business Statistics

장성민, 송민규, 조민준

Index

Introduction

롯데백화점 데이터로 남/녀를 구분하는 분류분석을 시행하려 한다. 현 데이터는 남성보다 여성이 더 많은 불균형 데이터이다.
따라서 각각 성별에 따른 특징이 보이는 피쳐를 생성하고자 노력하였다.

1. Machine Learning에서는

- 1차적으로는 크게 5가지 관점 (가격, 시간, 장소, 제품, 기타)으로 고객을 파악 및 구분하고자 하였다. 피쳐의 유용성 확인을 위해 countplot과 kdeplot로 EDA 과정을 거쳐서 채택하였고, 이후 5가지 전처리후 모델링을 진행하였다. 해당 데이터는 불균형 데이터이기 때문에 6가지 모델(LogReg, RF, ExtraTree, GBM, XGB, LGBM)을 Stratified K-Fold를 기반으로 하였다. 모델은 BayesianOptimization을 통해서 튜닝하였고 최대한 데이터를 활용하고 성능을 보고자 train_test_split 없이 cross_val_score로만 진행되었다. 이렇게 튜닝된 6개의 모델들로 stacking을 통해 submission을 내었으며, 비슷한 흐름으로 3가지 pipeline을 진행해 총 3가지 submissions로 기하평균을 통해 하나의 앙상블 재료를 만들었다.
- 2차적으로는 Word Embedding을 통해 데이터에 접근하였다. 1차 데이터에서 가장 효과를 보았던 물품의 구매건수를 집중적으로 다뤘으며 크게 2가지 방식으로 진행되었다. 첫번째는 각 고객별 물품의 구매건수를 W2V 학습 이후 Embedding Featurizer 클래스를 통해 W2V이 학습한 피쳐를 다시금 뽑는 것이며, 두번째는 고객이 구매한 물품목록에 여성을 나타내는 0과 남성을 나타내는 1을 넣어 물품과 0 또는 1의 Cosine Similarity를 뽑아내는 것이다. 이러한 2차 방법의 핵심은 oversampling에 변화를 준 부분과 매번 학습시킬 때마다 달라지는 Embedding Featurizer에서 W2V이 잘 학습한 최고의 피쳐를 뽑아내는 부분이었다. 모델링은 1차의 흐름과 같이 6가지 모델 및 BayesianOptimization과 stacking을 통해 진행되었다. W2V은 Pipeline 하나의 submission에서도 1차 pipelines를 웃도는 성능을 내었으며, 이를 앙상블해 두번째 앙상블 재료를 만들었다.

2. 딥러닝 기법에서는

- 크게 3가지 방법(Dnn, Ae, Cnn)으로 모델링을 진행하였다. 기존에 있던 numerical, w2v features가 사용되었으며 출력결과 피쳐가 같음에도 데이터에 대한 접근 및 활용방식이 달라서 머신러닝의 submissions와 상관관계가 낮음이 밝혀졌다. 성능 또한 W2V을 웃돌았기에 고객을 구분하는데 가장 중요한 역할을 하였다. 이 중 개인적으로 가장 효과를 본 것은 AE에 dropout을 추가한 DAE였다. 과적합을 막기위해 dropout, early stopping, L2, BatchNormalization를 사용하였고 그래프의 변화를 보며 계속해서 과적합 없이 가장 성능이 잘 나오는 모델을 만들고자 하였다. 특히 seed의 변화를 통해 매번 달라지는 AE를 거친 피쳐의 변화가 마지막에 앙상블을 할 때 튀는 효과를 방지해주어 성능이 더욱 잘 나왔다.

Round 1 - Numeric & Categorical

Pipeline 1

가격, 시간, 장소, 제품, 기타 피쳐

1600여개 -> Percentile(by LogReg)

Pipeline 2

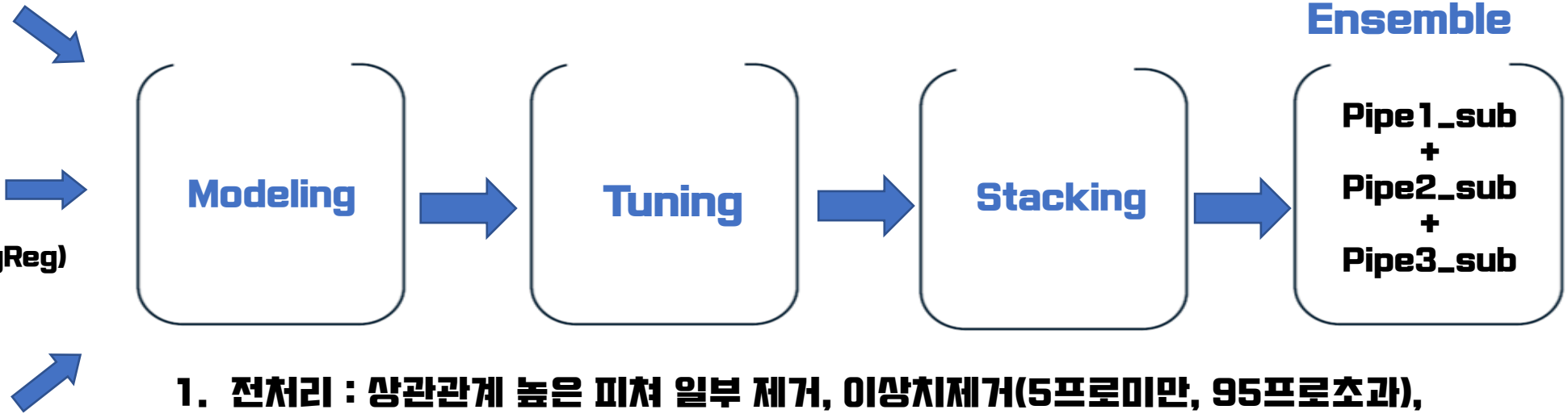
제품 구매건수

380여개 -> Percentile(by LogReg)

Pipeline 3

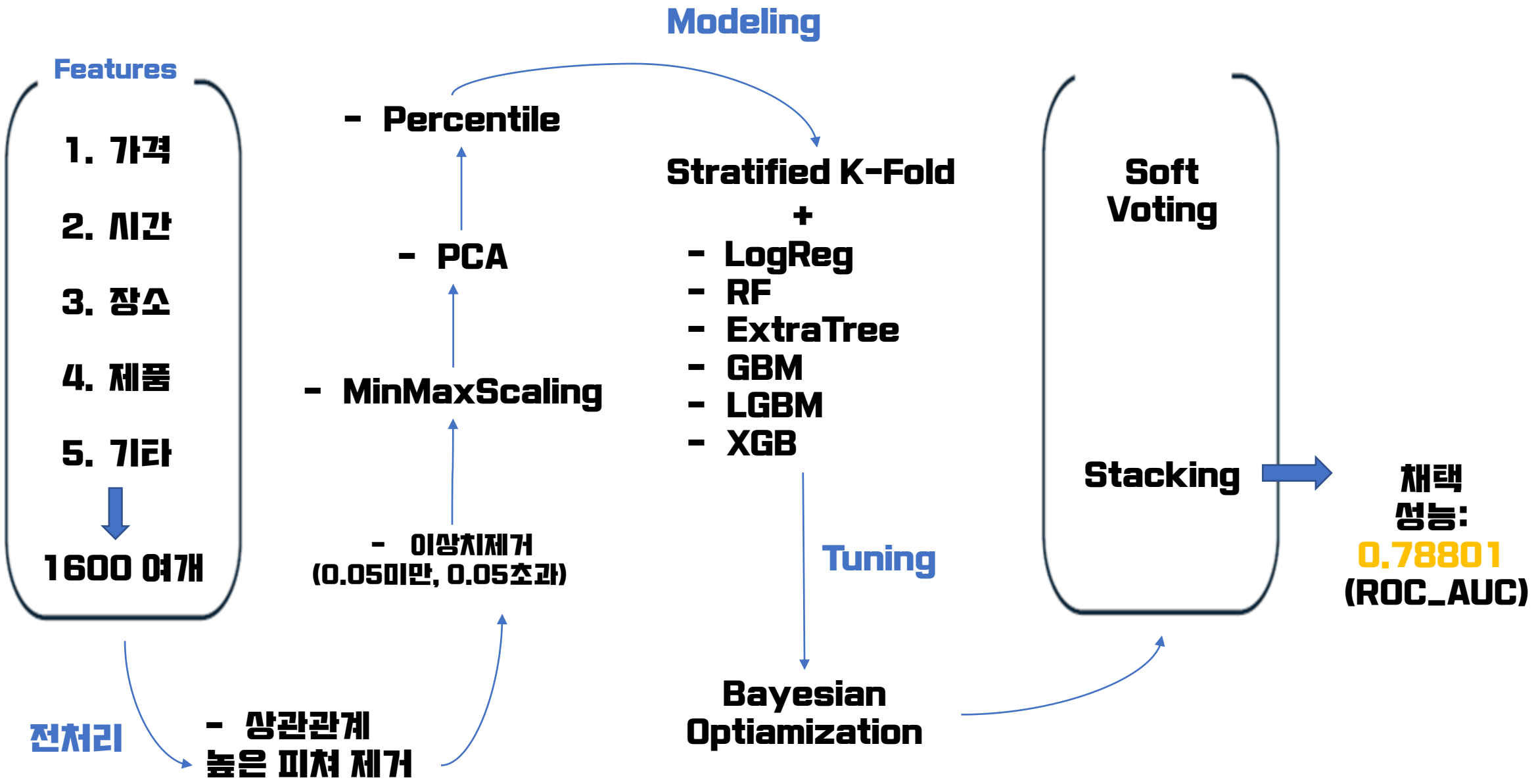
제품 구매건수

380여개 -> KmeansFeaturizier
+
SHAP(by XGB)



1. 전처리 : 상관관계 높은 피쳐 일부 제거, 이상치제거(5프로미만, 95프로초과), MinMaxScaling, PCA
2. 사용한 모델 : LogReg, RF, ExtraTree, GBM, XGB, LGBM
3. 튜닝 : Bayesian Optimization
4. 앙상블 : Stacking (성능이 Soft Voting보다 나았음)
5. 서브미션 앙상블 : Gmean(Pipe1_sub, Pipe2_sub, Pipe3_sub)

Round 1 - Pipeline processing



Round2 - W2V features

Row Data

- gds_grp_nm(331)
- gds_grp_mclas_nm(49)
- Similarity(6)

Corpus

Oversampling

① nunique x,
구매리스트 전체 호출
① 복원추출
② 구매건수에 많을수록 더 oversampling

Word2Vec

〈parameter〉

- Skip-gram 사용
- Num_features(300)
- minword_count(1)
- Window(5)

Modeling

① Percentile사용
② Logreg, RF, XGB 사용
③ Bayesian Optimization 사용
④ Ensemble Stacking 사용

① Best파이브 라인 submission의 성능이 0.79334
② 상관관계에 따라 1,3등의 submission과 우리의 submission을 PM하여 새로운 submission 생성 : score 80219

Round3. Neural Network

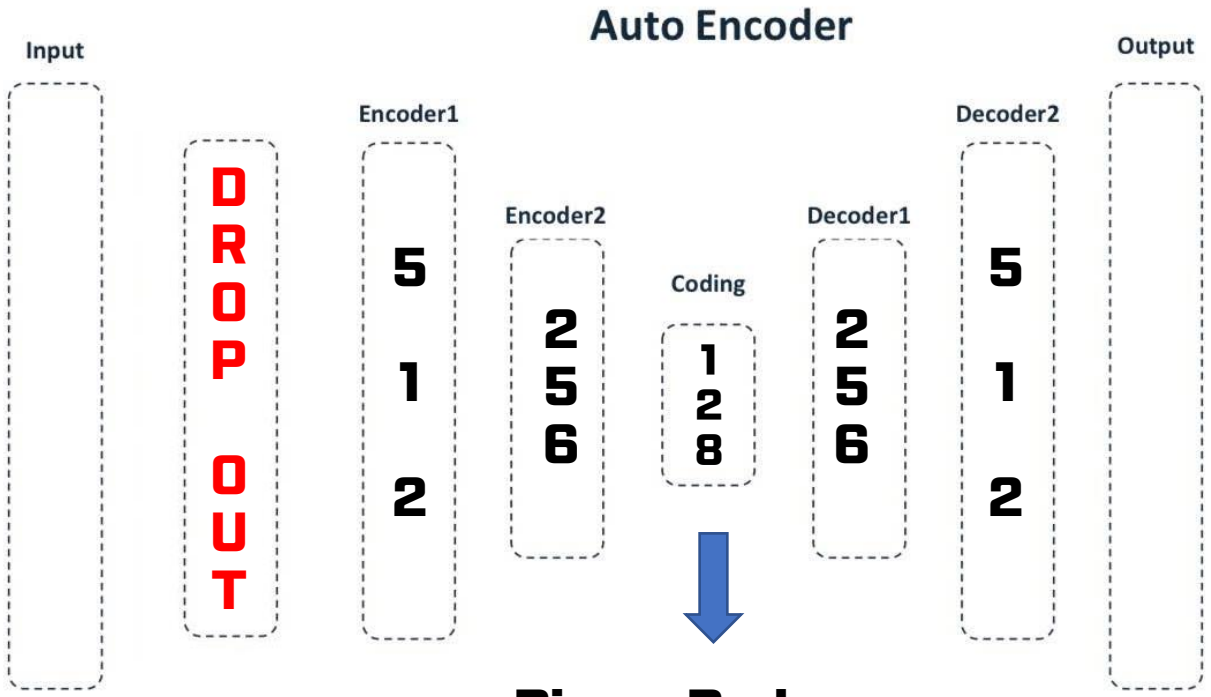
- Dropout Auto Encoder

- 개인2_장성민
  - Feature
  (3500, 143)
- *k-means clustering*
- *Shap*
- *Percentile*

=> (3500, 58)

'대, 중, 소 분류'
- 구매건수
(3500, 4179)
+
(3500, 58)

=> (3500,4237)



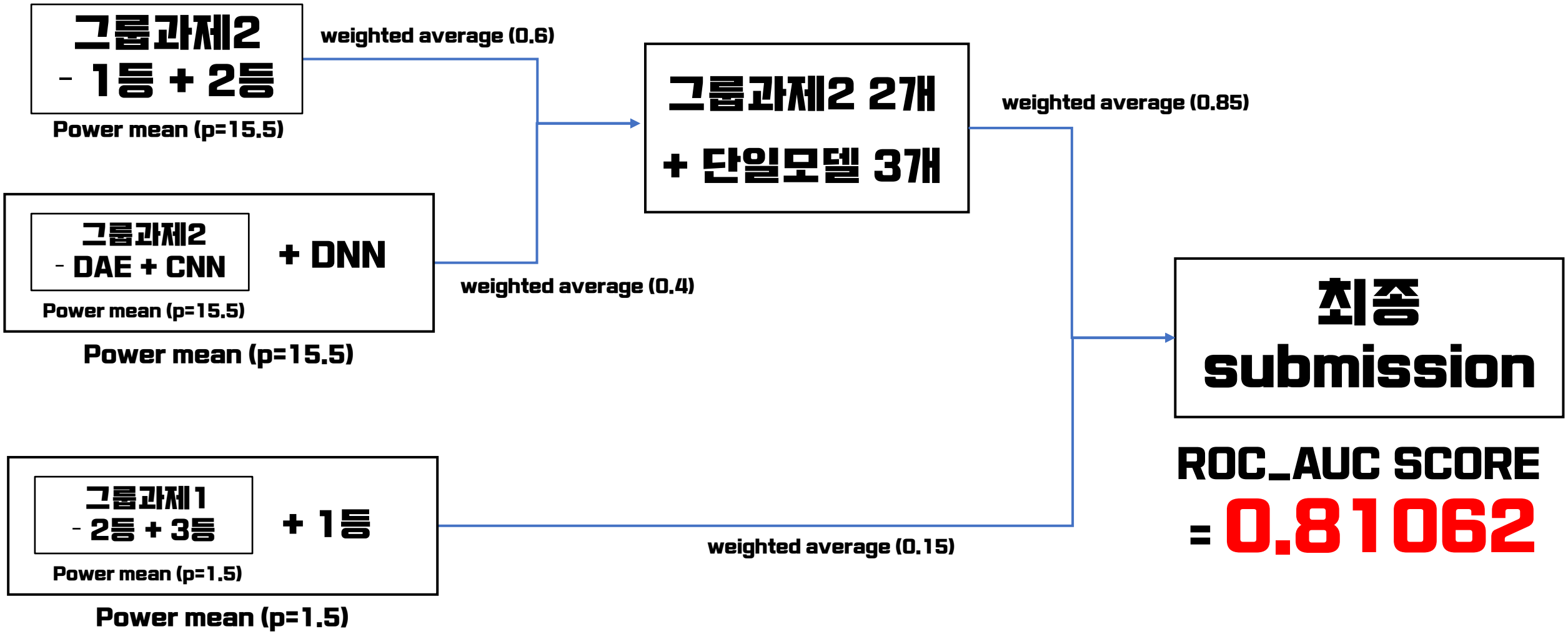
Dimen Reduce

: 4237 -> 128

DNN

Score 0.7945

Final Round - Ensemble all submission



Conclusion

- 머신러닝이든 딥러닝이든 분석과 여러 번의 시행착오를 통해서 높으면서 안정적인 모델을 만들었을 때, 앙상블의 효과가 컸다.

```
0.7815947983121236
0.77009641827271
0.7721088435374149
0.7678446173886904
LogisticRegression
```

→ Score 0.7933

```
Validation Summary:
3    0.767851
0    0.767143
4    0.766841
7    0.766015
8    0.765722
2    0.765514
1    0.765307
9    0.764778
6    0.764618
5    0.762512
dtype: float64
mean=0.76563, std=0.002
```

→ Score 0.7982

- 서브미션 앙상블시 작은 점수의 서브미션끼리 앙상블 할 때, 여러 서브미션을 한꺼번에 앙상블 할 때 보다 성능이 높았다.

그룹과제2
- DAE + CNN + DNN

Power mean (p=15.5)

<

그룹과제2
- DAE + CNN

Power mean (p=15.5)

+ DNN

Power mean (p=15.5)

Conclusion

- Power mean을 사용하는 경우 p값에 따라 성능이 달라지는 것을 확인할 수 있었다. (그러나 p값이 무조건 크다고 좋은 것은 아니다.)

그룹과제2 - DAE + CNN + DNN
Power mean (p=3.5)
Power mean (p=3.5)

<

그룹과제2 - DAE + CNN + DNN
Power mean (p=15.5)
Power mean (p=15.5)

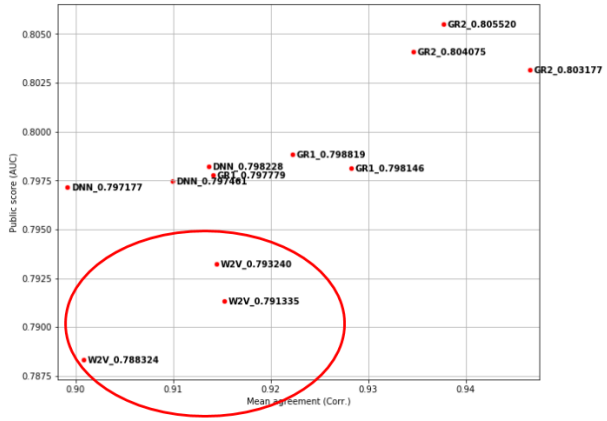
그룹과제1 - 2등 + 3등 + 1등
Power mean (p=1.5)
Power mean (p=1.5)

>

그룹과제1 - 2등 + 3등 + 1등
Power mean (p=3.5)
Power mean (p=3.5)

→ 이 경우에서는 P가 작은 것이 더 나음.

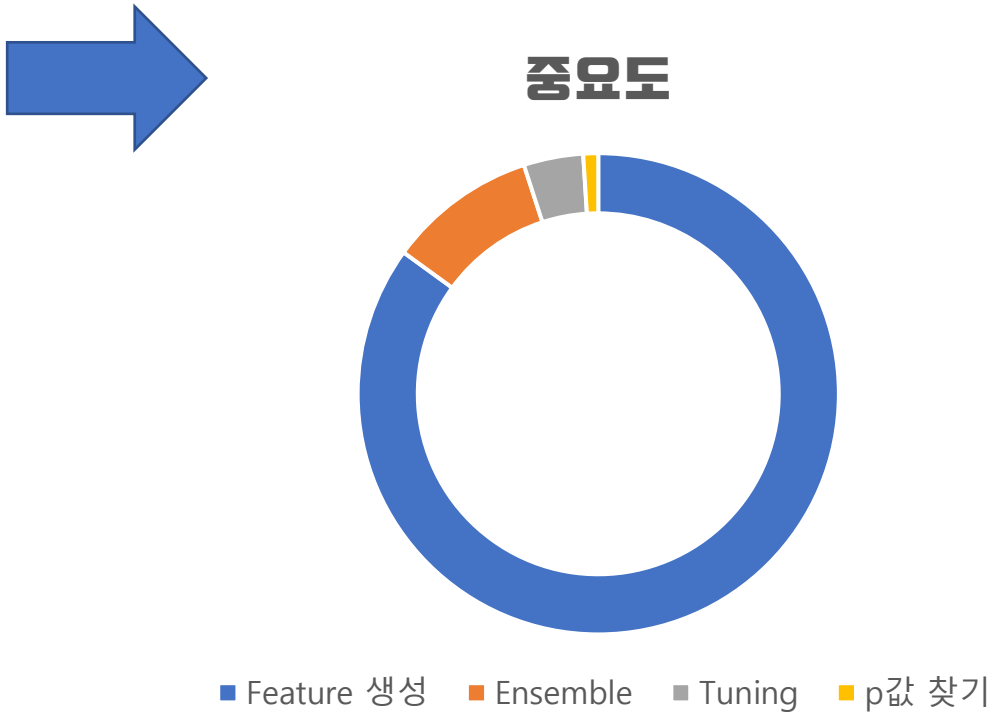
- 모델 간에 상관관계가 낮고 성능이 높은 모델들을 선정하여 앙상블 하는 것이 중요하다.



→ 실제로 상관관계는 낮지만 성능 또한 낮았던 W2V은 앙상블 할 때마다 성능이 안 나옴

Conclusion

- 가장 중요한 것은 성능이 비슷하거나 높은 모델들을 여러 개 만드는 것!



Kaggle에서 공개된 상위권 모델만을 그대로 가져와
여러가지 앙상블 방식을 수행하였다.
상관관계가 낮고, 높은 성능을 내는 서브미션을 생성하기위해
피쳐생성 및 다양한 시도가 가장 중요하지만,
마지막에 좋은 재료를 가지고 어떻게 앙상블을 하는지도 중요하였다.
이번에 주어진 기회안에서 성능을 낼 수 있었던 것은
조원들의 신중한 앙상블 접근 덕분이었다.
멱평균이 일반적으로 성능이 잘 나오는 것은 사실이지만,
이번에는 가장 적절한 가중평균의 가중치를 찾았기 때문이라고 생각한다.
그리고 가장 중요한 것은 좋은 피쳐를 만드는 것이지만,
딥러닝에서는 피쳐를 다룰 모델아키텍쳐의 능력도 중요하다.

Thank You