KML Challenge 2021S

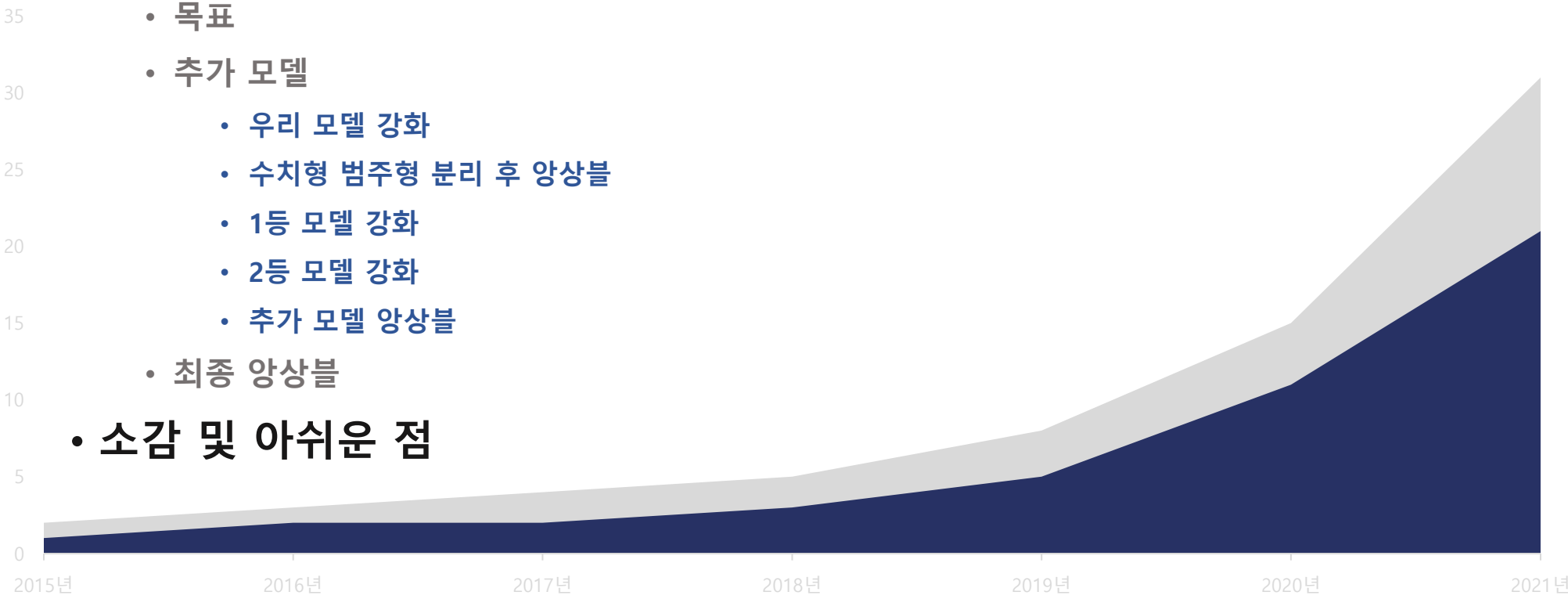
송창용 + 권유진

01

contents

INDEX

- 1ROUND 분석과정
- 2ROUND 분석과정
  - 목표
  - 추가 모델
    - 우리 모델 강화
    - 수치형 범주형 분리 후 앙상블
    - 1등 모델 강화
    - 2등 모델 강화
    - 추가 모델 앙상블
  - 최종 앙상블
- 소감 및 아쉬운 점



01

contents

1ROUND 분석과정

1ROUND 분석과정

1ROUND 목표: 최대한 많은 FEATURE을 만들어 SelectPercentile을 이용해 Feature Selection 한 후, 모델링 해보자!

PCA
94

수치형 변수
140

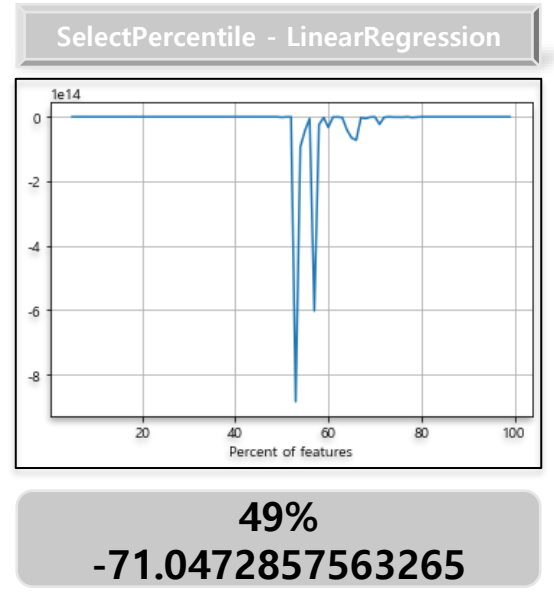
비율 변수
55

Word2 Vector
600

Ont-Hot Encoding
308

이상치 처리: 상하위 5%지점으로 조정
Scaling: RobustScaler

Feature Selection



49%
-71.0472857563265

1ROUND 분석과정

Model Tuning

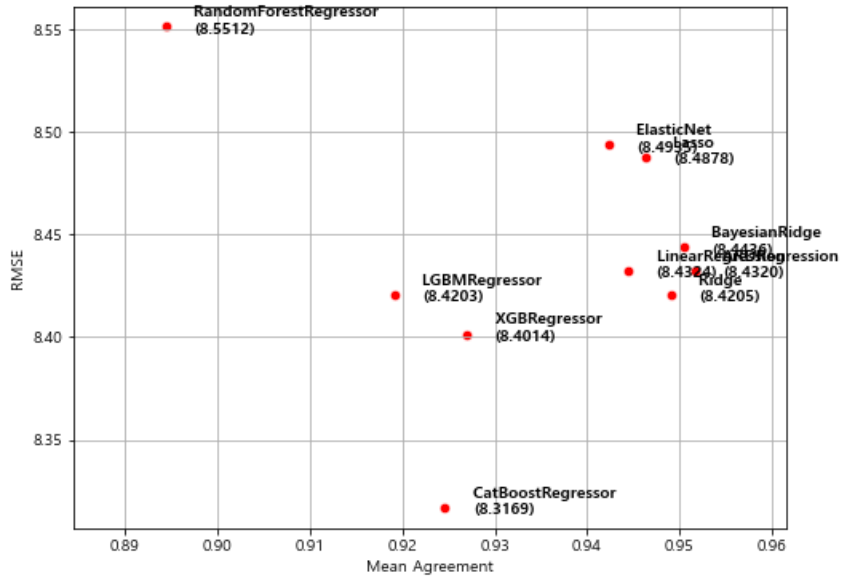
Randomized Search

↓

Grid Search

Ridge:	8.4205
Lasso:	8.4878
ElasticNet:	8.4935
ARDRegression:	8.4320
BayesianRidge:	8.4436
RandomForest:	8.5512
XGBoost:	8.4014
LGBM:	8.4203
CatBoost:	8.3169

Model Ensemble



Ridge+CatboostRegressor 8.263993

∴ **Private Score: 8.21707로 4위 기록**

01

contents

2ROUND 분석과정

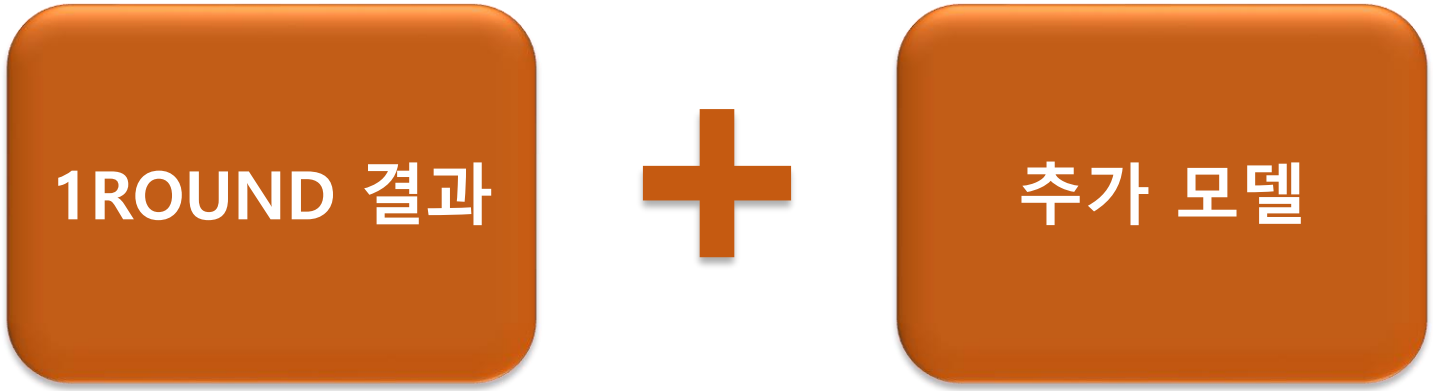
01

contents

2ROUND 분석과정-목표

2ROUND 목표

- 기존 모델 DNN, Ensemble, Stacking 등을 이용해 더욱 강화
- 다양한 Feature을 통해 여러 모델을 생성하여 Submission들을 생성한 후, 이를 Ensemble하여 예측력 상승!



01

contents

2ROUND 분석과정

추가모델 생성

2ROUND 분석과정 – 추가 모델

1. 1ROUND 우리 모델 강화

튜닝방법 변화

Randomized Search → Grid Search

Ridge:	8.4205
Lasso:	8.4878
ElasticNet:	8.4935
ARDRegression:	8.4320
BayesianRidge:	8.4436
RandomForest:	8.5512
XGBoost:	8.4014
LGBM:	8.4203
CatBoost:	8.3169

Bayesian Optimization

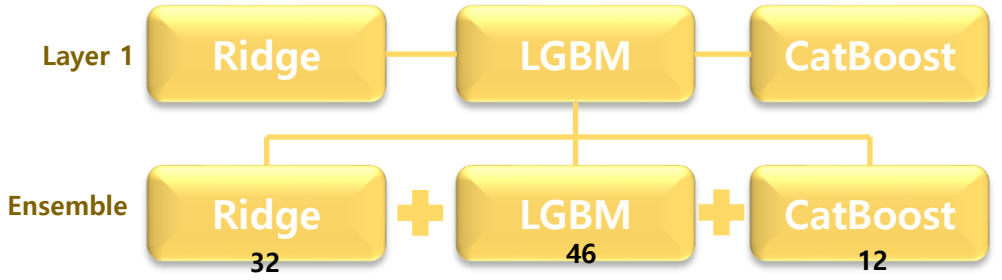
Ridge:	8.4122	BayesianRidge:	8.4371
Lasso:	8.4268	LGBM:	8.2931
ElasticNet:	8.4203	CatBoost:	8.3191
ARDRegression:	8.4142		

Ensemble

Ridge 9 + LGBM 12 + CatBoost 9

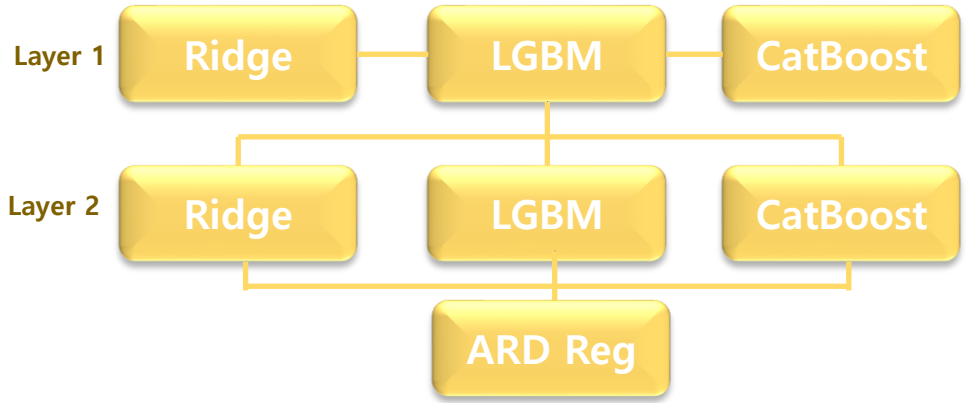
Score: 8.24023568682186

Stacking



Score: 8.222335865797412

3-Layered Stacking

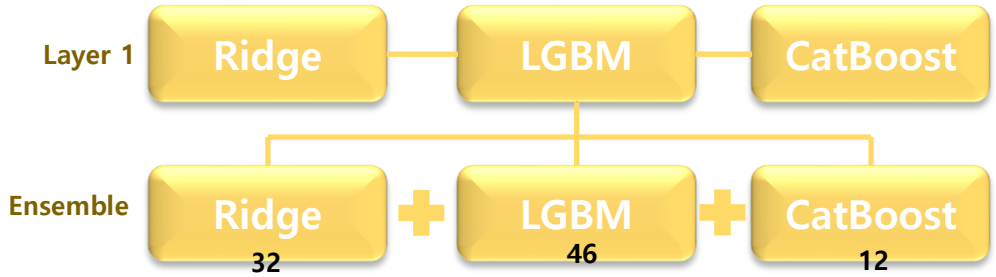


Score: 8.291

2ROUND 분석과정 – 추가 모델

1. 1ROUND 우리 모델 강화

Stacking



Score: 8.222335865797412

Deep Neural Network

Score: 8.317049026489258

77

Ensemble

23

Score: 8.20463

2ROUND 분석과정 – 추가 모델

2. 범주형, 수치형 분리 후 모델링 그리고 앙상블

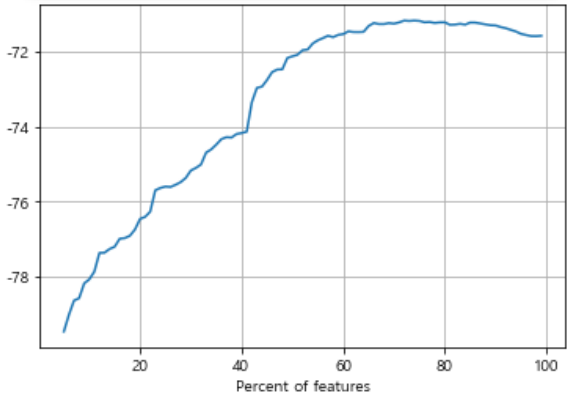
범주형 변수

PCA 94

Word2 Vector 600

Ont-Hot Encoding 308

SelectPercentile - Ridge



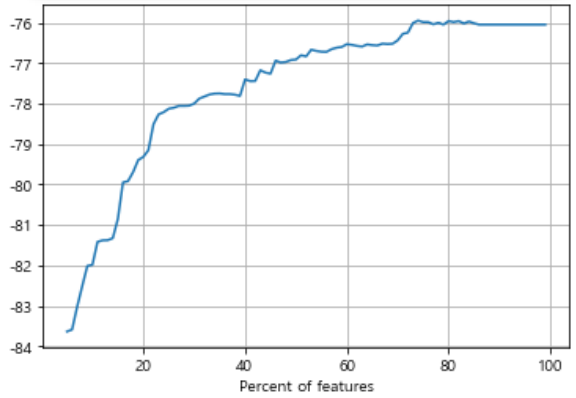
72%
-71.174212754492

수치형 변수

수치형 변수 904

비율 변수 61

SelectPercentile - Ridge



74%
-75.94603981003918

2ROUND 분석과정 – 추가 모델

2. 범주형, 수치형 분리 후 모델링 그리고 앙상블

범주형 변수

Bayesian Optimization

Ridge:	8.4560	BayesianRidge:	8.4851
Lasso:	8.5037	LGBM:	8.3206
ElasticNet:	8.5139	CatBoost:	8.3909
ARDRegression:	8.4719		

Ensemble

Ridge 9 + LGBM 17 + CatBoost 4

Score: 8.285595085499947

Stacking

Layer 1: Ridge, LGBM, CatBoost

Ensemble: Lasso 8 + LGBM 12

Score: 8.269347370883308

13

Deep Neural Network

Input Layer (759)
↓ tanh
Hidden Layer1 (80) — Dropout 40%
↓ tanh
Hidden Layer2 (48)
↓ tanh
Hidden Layer3 (96) — Dropout 10%
↓ tanh
Output Layer (1)

Score: 8.3474

7

Ensemble

Score: 8.244714302499277

2ROUND 분석과정 – 추가 모델

2. 범주형, 수치형 분리 후 모델링 그리고 앙상블

수치형 변수

Bayesian Optimization

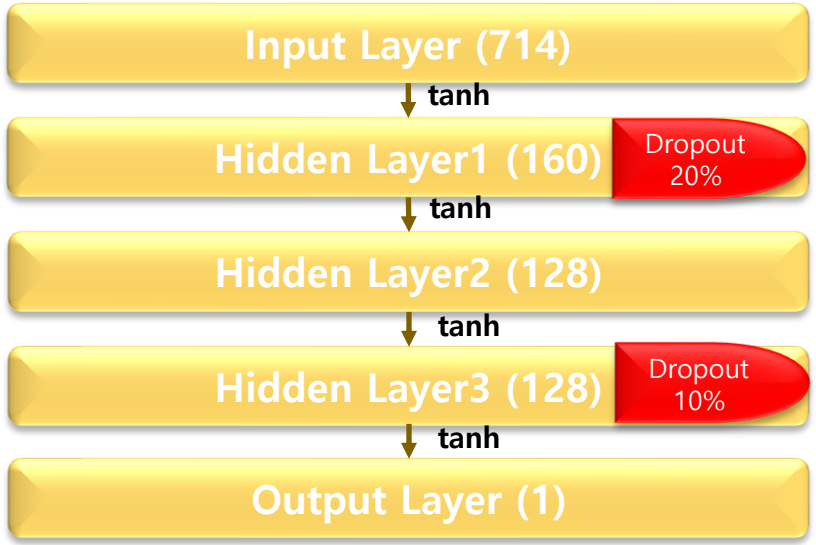
Ridge:	8.6675	BayesianRidge:	8.6644
Lasso:	8.6421	LGBM:	8.5285
ElasticNet:	8.6442	CatBoost:	8.5942
ARDRegression:	8.6346		

Ensemble

ARD Reg (9) + LGBM (17) + CatBoost (4)

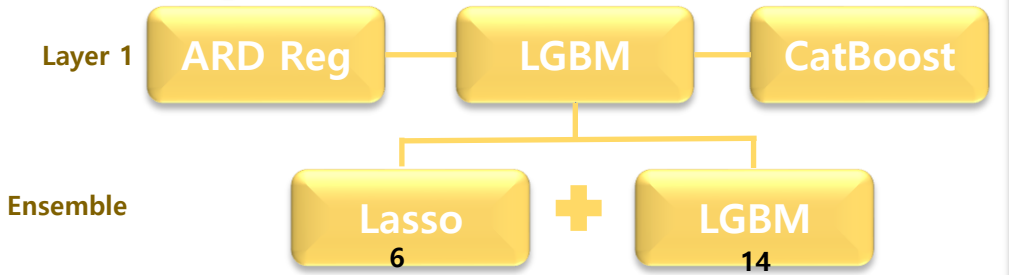
Score: 8.491220383535346

Deep Neural Network



Score: 9.7485

Stacking



Score: 8.460797116616405

2ROUND 분석과정 – 추가 모델

2. 범주형, 수치형 분리 후 모델링 그리고 앙상블

Ensemble

범주형 29 + 수치형 71

Score: 8.195254593392722

DataFrame

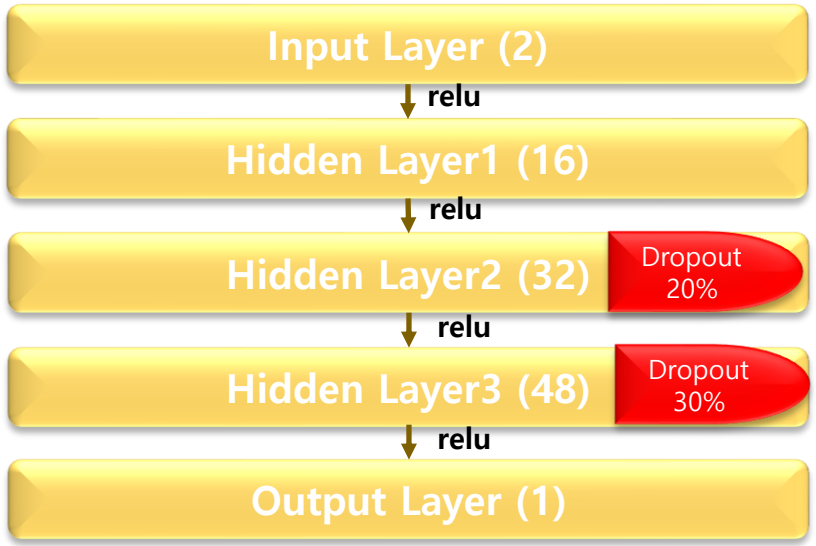
	num	cat
0	10.415650	29.794056
1	11.934349	29.876847
2	8.326676	20.184667
3	9.354029	21.874165
4	7.782973	18.920300

modeling

Bayesian Optimization

Ridge:	7.9512	BayesianRidge:	7..9416
Lasso:	7.9235	LGBM:	8.0078
ElasticNet:	7.9258	CatBoost:	8.2668
ARDRegression:	7.9404		

Deep Neural Network



Score: 8.1544

Ensemble

DNN + LASSO

Score: 7.93704984839293

Lasso Score: 7.9235

2ROUND 분석과정 – 추가 모델

3. 1ROUND 1등 모델 강화

Bayesian Optimization

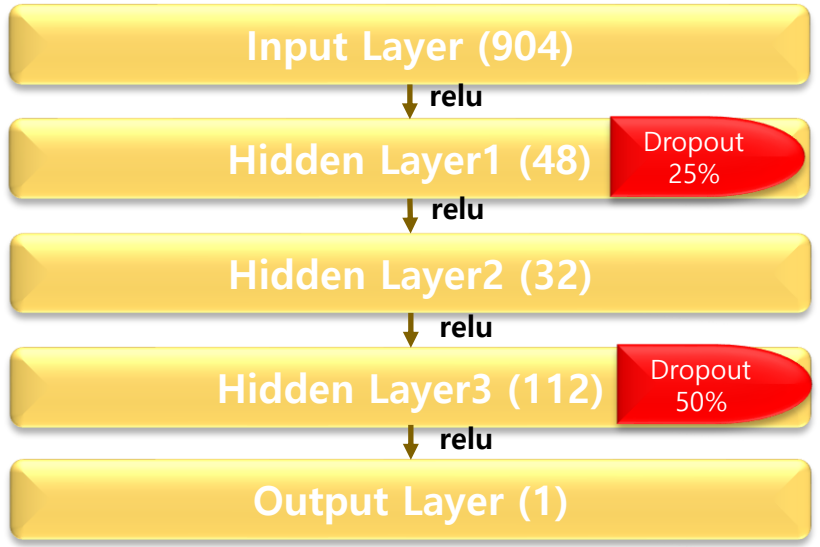
Ridge:	8.4569	BayesianRidge:	8.4204
Lasso:	8.3883	LGBM:	8.0595
ElasticNet:	8.4145	CatBoost:	8.0576
ARDRegression:	8.3914		

Ensemble

ElasticNet 11 + LGBM 40 + CatBoost 39

Score: 8.017669596193688

Deep Neural Network



Score: 8.13182

2ROUND 분석과정 – 추가 모델

4. 1ROUND 2등 모델 강화

Bayesian Optimization

Ridge:	8.3944	BayesianRidge:	8.4089
Lasso:	9.2275	LGBM:	8.1391
ElasticNet:	8.9733	CatBoost:	8.1623
ARDRegression:	8.3913		

Ensemble

ARD Reg 4 + LGBM 15 + CatBoost 11

Score: 8.107152339588835

Deep Neural Network

Input Layer (545)
↓ tanh
Hidden Layer1 (80) — Dropout 25%
↓ tanh
Hidden Layer2 (128) — Dropout 25%
↓ tanh
Output Layer (1)

Score: 8.51854419708252

86 : 14 Ensemble

Score: 8.096641701559442

2ROUND 분석과정 – 추가 모델 앙상블

1Round 우리 모델 강화 0 + Cat_Num 분리 후 앙상블 15 + 1Round 1등 모델 강화 67 + 1Round 2등 모델 강화 11 + 1Round 우리 모델 Feature 변화 7

Ensemble

Score: 7.980374463200102

DataFrame

	1등	2등	round1_수정	범주형,수치형분리	창용
0	40.946605	37.976504	41.032663	40.834154	43.094683
1	25.603701	29.800662	28.773802	27.741242	26.878138
2	26.981058	27.088720	28.180554	27.285695	26.689937
3	32.376266	34.691264	32.630079	32.015374	34.021475
4	35.484068	40.715160	38.291988	39.180976	40.657637

Modeling

Bayesian Optimization

Ridge:	7.8520	BayesianRidge:	7..8431
Lasso:	7.8522	LGBM:	7.9249
ElasticNet:	7.8305	CatBoost:	8.1062
ARDRegression:	7.8445		

Ensemble

ElasticNet 26 + LGBM 4

Score: 7.828767897910982

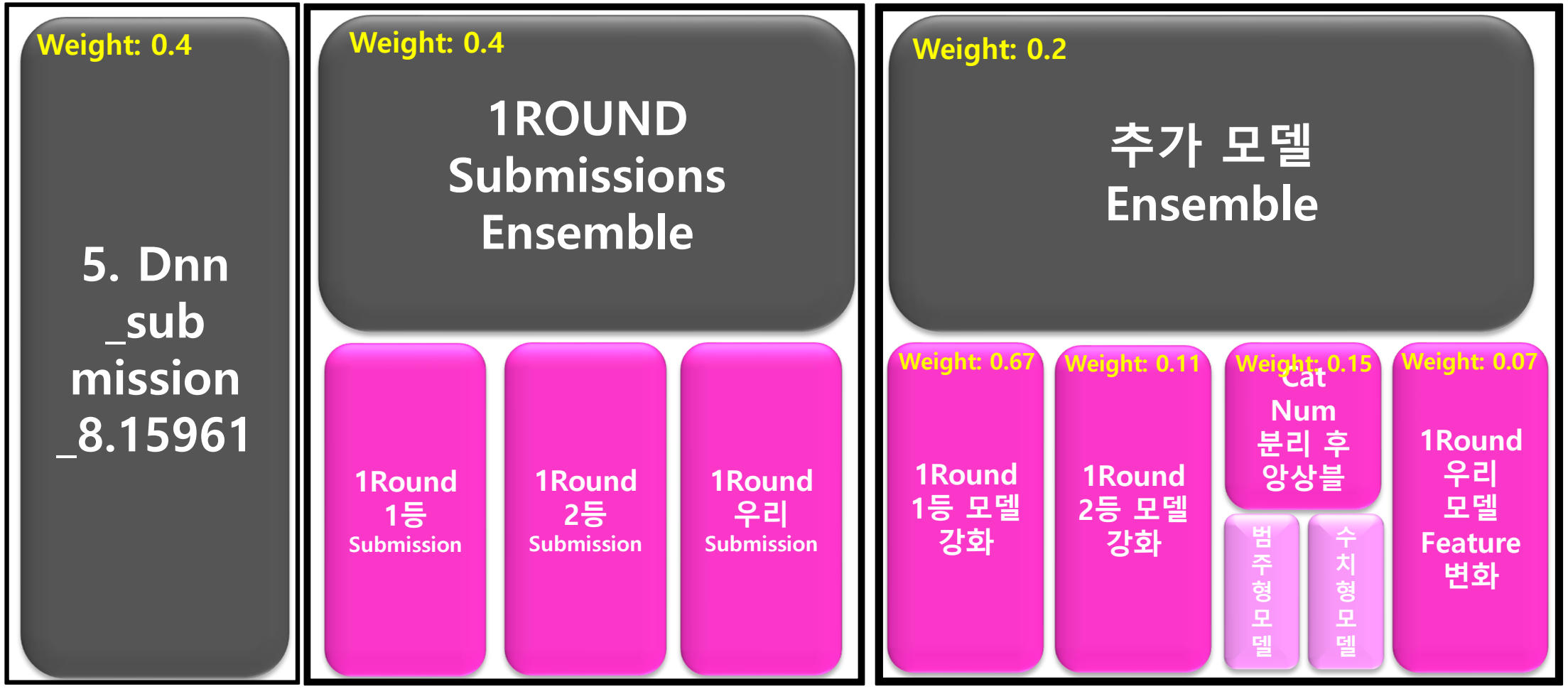
01

contents

2ROUND 분석과정

최종 앙상블

2ROUND 분석과정 – 최종 앙상블



∴ Private Score: 7.99441로 4위 기록

01

contents

소감 및 아쉬운 점

권유진

머신러닝 중간 개인과제를 하면서 한가지 배운 점이 있었다. 바로 feature의 중요성이다. 개인과제의 결과를 결정지었던 것은 바로 feature의 수라고 생각했기 때문이다. 그래서 더욱 이번 competition에서는 많은 feature을 생성하고 이용하려고 했던 것 같다.

무작정 많은 feature을 생성하면서 나는 한 가지 더 깨달았다. 무작정 feature을 늘리는 것이 성능을 향상시키지는 못한다는 것이다. 여기서 교수님께서 해주셨던 말씀이 떠올랐다. 바로 "Garbage in garbage out" 이었다. 기존 Feature에 무작정 Feature을 추가한 다음 Feature Selection을 거친다고 해도 오히려 성능이 떨어지는 것을 발견하였다.

또한 Feature 생성하는 기법(PCA, W2V 등)의 중요성 역시 깨달았다. 이 기법을 사용해 feature을 추가하였을 때 성능이 대폭 상승되는 것을 발견하였다. 개인과제와 이번 competition을 거치면서 교수님께서 해주셨던 '모델의 성능에 영향을 주는 것이 feature, algorithm, hyperparameter이 있는데 그 중 가장 영향을 많이 미치는 것이 feature이다.'라는 말씀에 대해 몸소 체험할 수 있었던 기회였다. 이를 통해 이번 방학을 통해 Feature Generation과 Feature Selection에 관해 매우 자세히 공부하겠다는 다짐을 하였다.

이번 학기에 나 혼자 2학년이고 머신러닝을 이 수업을 통해 처음 접해 큰 걱정이 있었는데 이번 기회를 통해 커다란 성장을 하였던 것 같다.

송창용

중간과제를 할 때까지만 해도 아무것도 몰라 당황했었는데 이번에 프로젝트를 진행하면서 앞으로 할 데이콘이나 캐글 같은 대회에 참가했을 때 어떻게 진행해야 할 지 알게 된게 가장 큰 소득인 것 같다.

이번 프로젝트에서 성능이 잘 나올것이라고 확신하고 시도해본 것이 성능이 안 나와서 실망하기도 하고 힘들기도 했지만 대체로 노력을 들인만큼 성능이 잘 나와주어 재밌게 했던 것 같다.

1차전에서 4등을 한 후 단일 모델의 성능을 올려 1,2,3등 결과와 앙상블 하면 결과가 잘 나올 것이라고 생각하여 3등 피쳐들을 가져와 사용해보기도 하고 변수에 PCA를 적용하여 추가해보기도 해서 성능을 높였지만 앙상블 한 결과가 잘 나와주지 않은 것이 아쉬웠었다.

2차전 기간에 다른 과목의 시험으로 인해 진행과정을 팀원과 잘 공유하지 못한 것이 최종 결과에서 4등을 하게 된 요인인 것 같아 아쉬움이 많이 남는다.