50.007 SUTD
Machine Learning

Presented by **Team Tea**

Andrew Yu
Charmaine Hong
Michelle Halim
Royce Lim
Sean Tan

36 teams

50.007 Machine Learning - Summer 2024

A fun course project for SUTD 50.007 Machine Learning course!

Overview Data Code Models Discussion **Leaderboard** Rules Team Submissions

Leaderboard

Raw Data | Refresh

Search leaderboard

Public **Private**

The private leaderboard is calculated with approximately 80% of the test data.
This competition has completed. This leaderboard reflects the final standings.

#	△	Team	Members	Score	Entries	Last	Solution
1	▲ 11	**Tea**		0.72801	72	2d	
2	▲ 2	Neural Networkers		0.72333	69	3d	
3	—	Blue Line		0.71867	6	3d	

BASELINE

Gaussian NB

Catboost

KNearest Neighbors

Multinomial NB

Adaboost

Logistic Regression

SVM

Voting

XGBoost

Random Forest

Gradient Boosting

Complement NB

Stochastic GD

LGBM

Bagging

And many many more...

BASELINE

KNearest Neighbors

Logistic Regression

SVM

Stochastic GD

Gaussian NB

Multinomial NB

Complement NB

XGBoost

Gradient Boosting

Catboost

Adaboost

LGBM

Random Forest

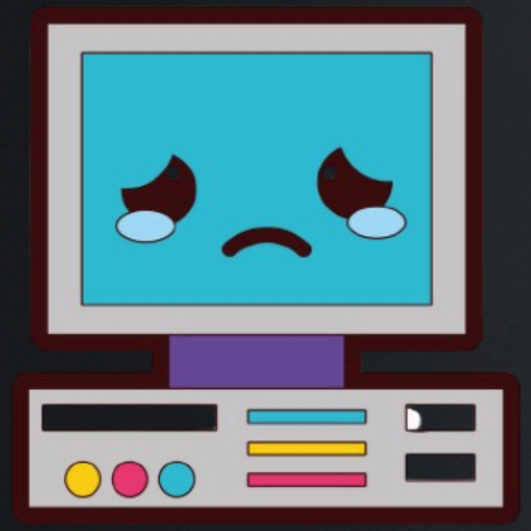
Bagging

Voting

And many many more...

DID YOU KNOW?

“I may not get much sleep but I can’t do this to my laptop too!”



DID YOU KNOW?

rapidsai/**cuml**

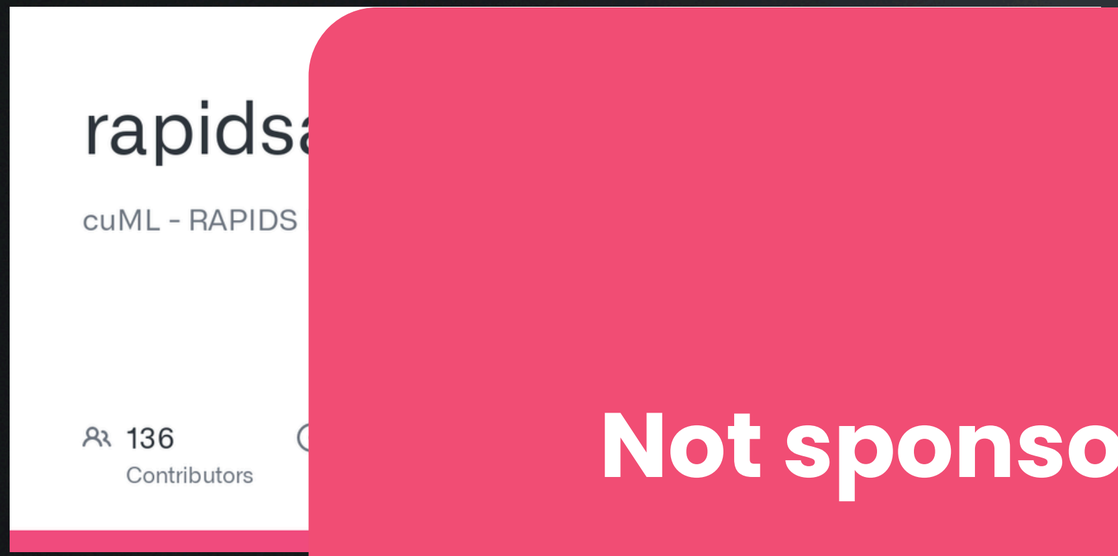
cuML - RAPIDS Machine Learning Library

RAPIDS

136 Contributors | 841 Issues | 4k Stars | 525 Forks

kaggle

DID YOU KNOW?



Not sponsored by Kaggle :(

DATASET CHARACTERISTICS

Shape: (17000, 5000)

Sparsity: 99.85%

Range: [0, 1]

Class ratio: 1.62 $\frac{\text{Class 0}}{\text{Class 1}}$

DATASET CHARACTERISTICS

Correlation

Drop drop drop

Shape: (17000, 5000)

Data Augmentation

- Gaussian Mixture Model

Sparsity: 99.85%

Dimensionality Reduction

- PCA
- SVD

Range: [0, 1]

Gaussian Transformation

- PowerTransform
- QuantileTransform

Class ratio: 1.62

$$\frac{\text{Class 0}}{\text{Class 1}}$$

Sampling

- Undersampling
- Oversampling

DATASET CHARACTERISTICS

Correlation

- Naive Bayes
(Assumption of independence)

Shape: (17000, 5000)

Data Augmentation

- Everyone

Sparsity: 99.85%

Dimensionality Reduction

- Logistic Regression
- Boosting
- Ensemble

Range: [0, 1]

Gaussian Transformation

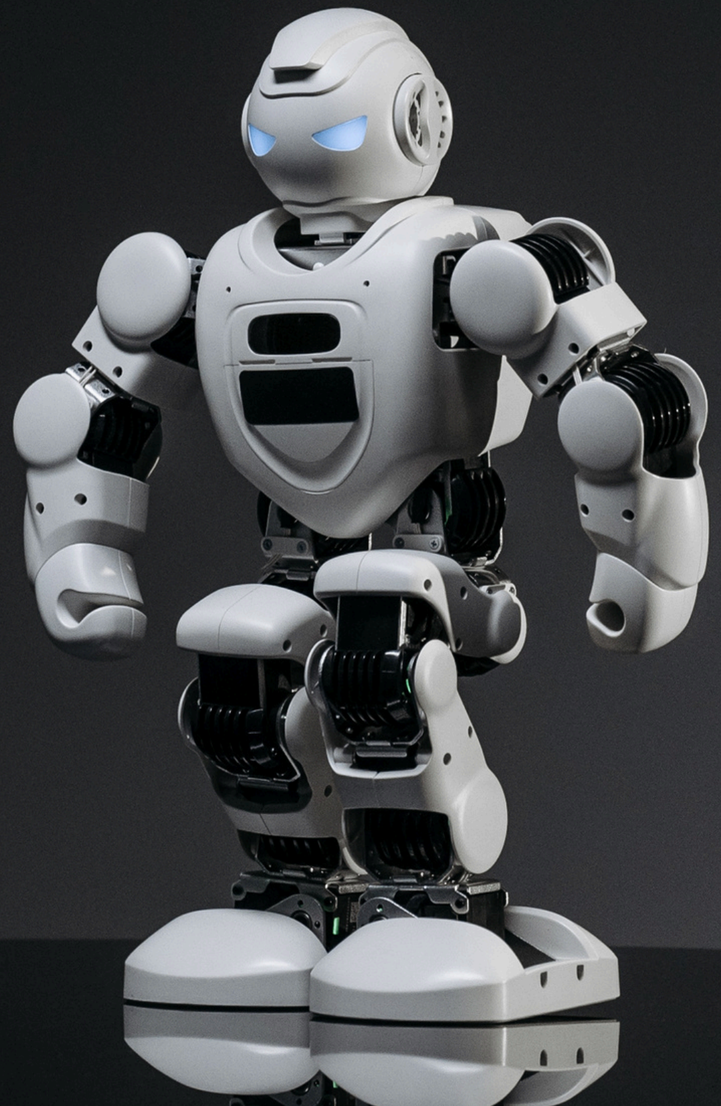
- Gaussian Naive Bayes
- Logistic Regression
- SVM with RBF kernel

Class ratio: 1.62 $\frac{\text{Class 0}}{\text{Class 1}}$

Sampling

- Everyone

FEATURE ENGINEERING

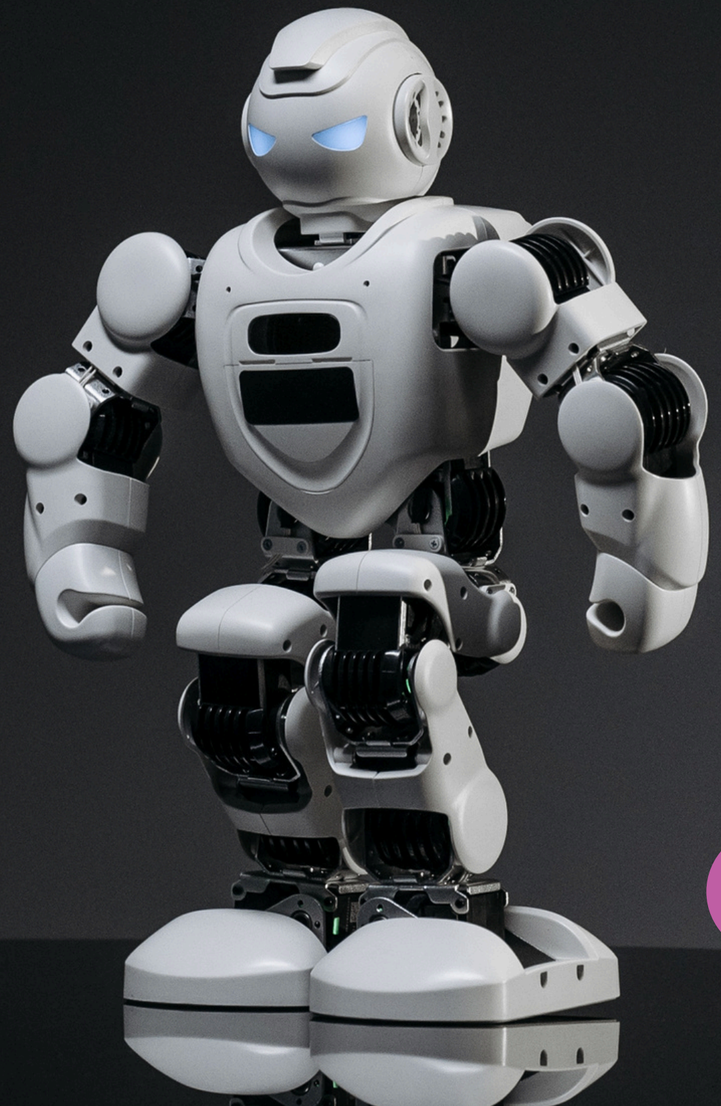


```
enginger_features = {
    'corr': ['yes', 'no'],
    'corr_num': range(0.1, 1, 0.1),
    'dim_red': ['pca', 'svd', 'no'],
    'transform': ['power', 'quantile', 'no'],
    'gmm': ['yes', 'no'],
}
```

👎 GridSearchCV

👍 RandomizedSearchCV

FEATURE ENGINEERING



```
enginger_features = {
    'corr': ['yes', 'no'],
    'corr_num': range(0.1, 1, 0.1),
    'dim_red': ['pca', 'svd', 'no'],
    'transform': ['power', 'quantile', 'no'],
    'gmm': ['yes', 'no'],
}
```

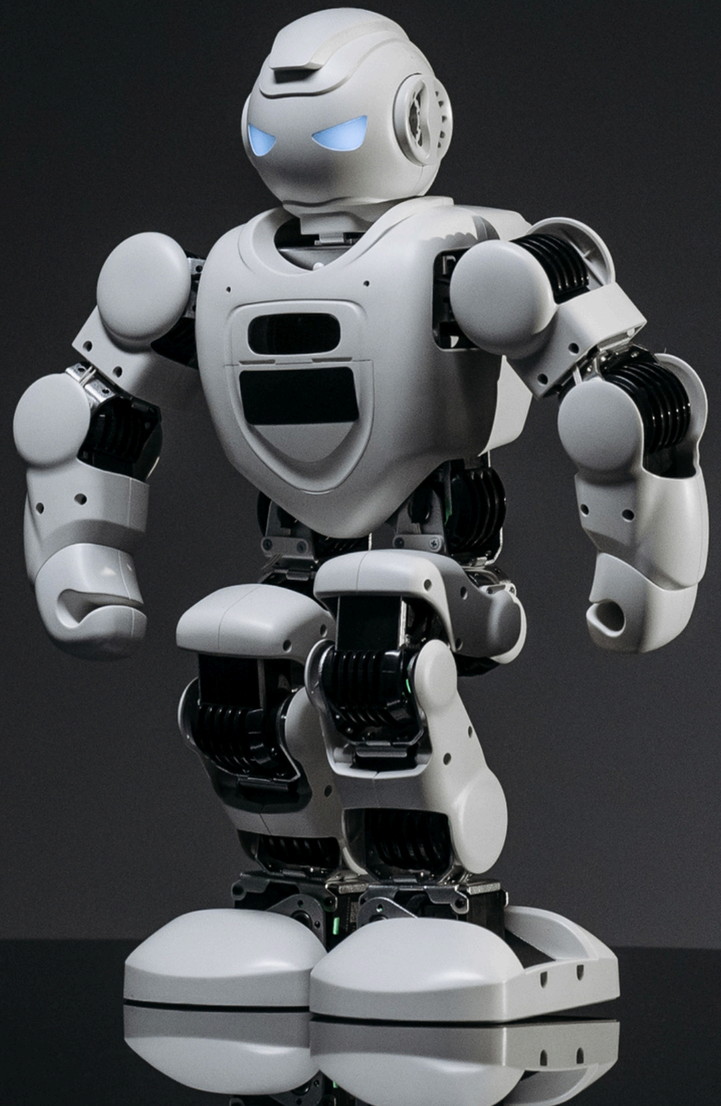
n_iter

no. of hyperparameter combinations

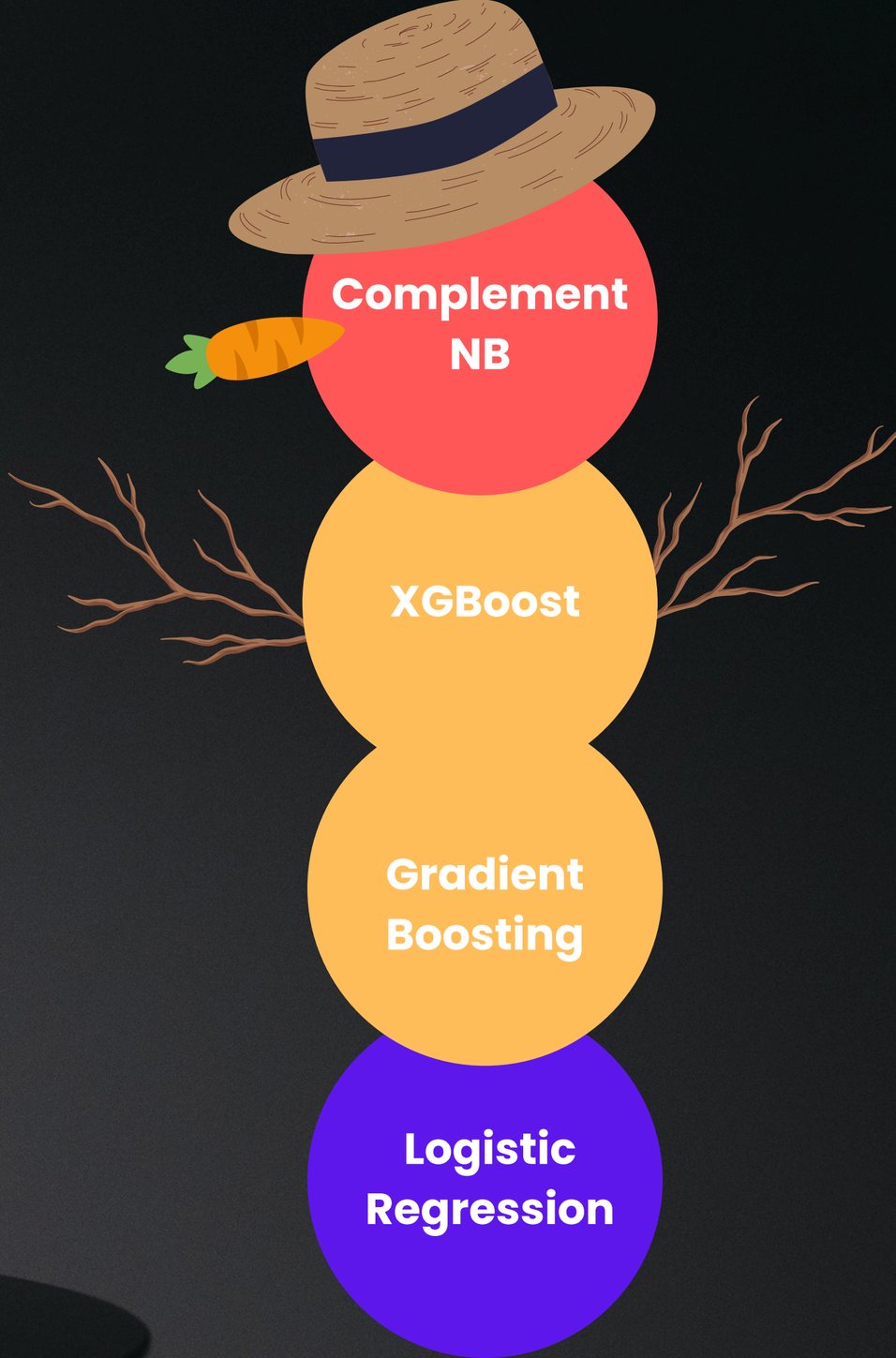
cv

evaluation on different subsets of training data

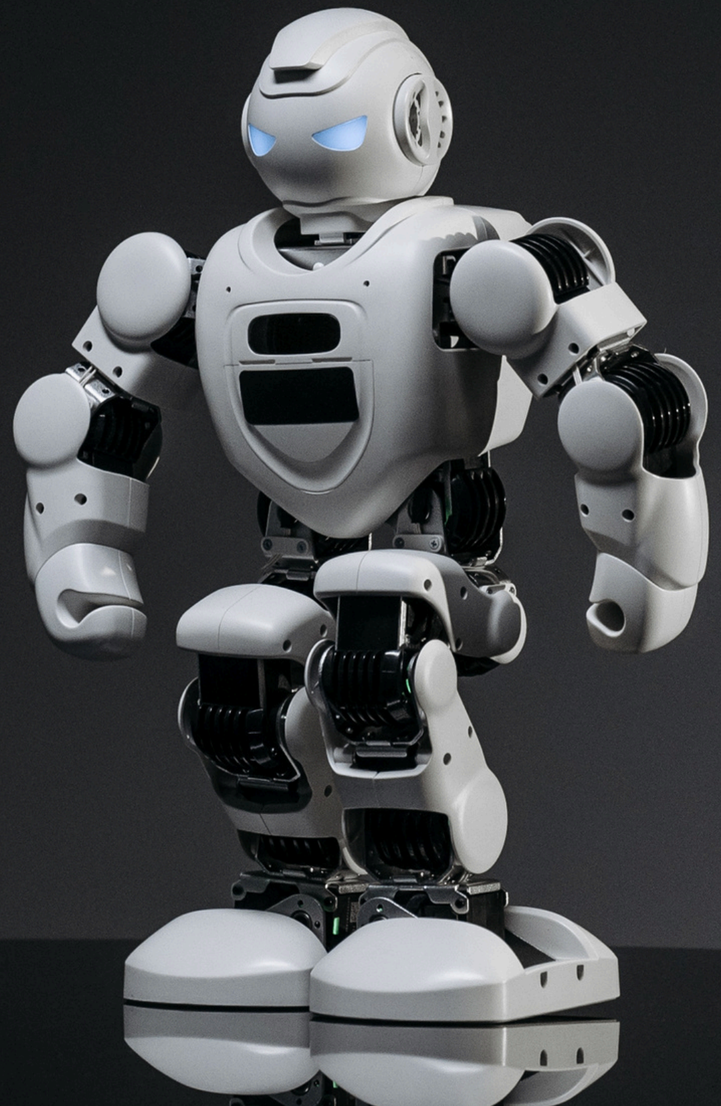
JACKPOT



```
enginger_features = {
    'corr': 'yes',
    'corr_num': 0.4,
    'dim_red': 'no',
    'transform': quantile',
    'gmm': 'no'
}
```



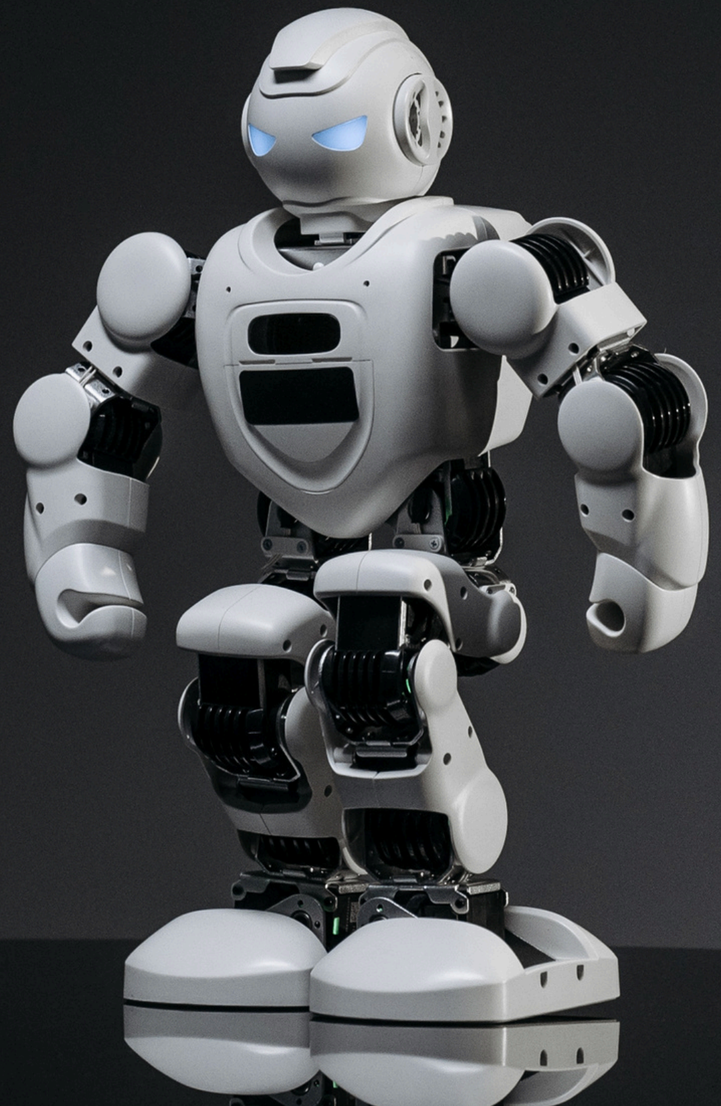
JACKPOT



Voting = soft!

Voting

JACKPOT



Splitting Strategy

Random State

Model Stability

Q & A


THANK YOU