Ensemble Methods

Why Ensemble Methods ?

Kaggle

OSIC Pulmonary Fibrosis Progression Predict lung function decline Featured • 18 days to go • Code Competition • 1776 Teams	$55,000
RSNA-STR Pulmonary Embolism Detection Classify Pulmonary Embolism cases in chest CT scans Featured • a month to go • Code Competition • 118 Teams	$30,000
Lyft Motion Prediction for Autonomous Vehicles Build motion prediction models for self-driving vehicles Featured • 2 months to go • Code Competition • 405 Teams	$30,000
Mechanisms of Action (MoA) Prediction Can you improve the algorithm that classifies drugs based on their biological activity? Research • 2 months to go • Code Competition • 1325 Teams	$30,000
Google Landmark Recognition 2020 Label famous (and not-so-famous) landmarks in images Research • 11 days to go • Code Competition • 667 Teams	$25,000
OpenVaccine: COVID-19 mRNA Vaccine Degradation Prediction Urgent need to bring the COVID-19 vaccine to mass production Research • 17 days to go • 831 Teams	$25,000

Why Ensemble Methods ?

“This is how you win ML competitions: you take other peoples’ work and ensemble them together.” <u>Vitaly Kuznetsov</u> NIPS2014

	Competition	Solutions, interviews and articles
	2019 Data Science Bowl Uncover the factors to help measure how young children learn	Forum • 1st place (zr & oyx) • 2nd place (Fuson) Word2Vec features, • 3rd place (Limerobot) transformer • 4th place (Crystal cave miners) tfidf, rnn

5. Model

1. **Data augmentation** : The model is trained on the full data (full train history and test previous, improve + 0.002).
2. **Loss** : We use rmse loss for training, and weighted rmse loss for validate.
3. **Threshold** : Then use Opitmizer Rounder to optimize thresholds for weighted qwk.
4. **Ensemble** : We just try a simple blending method (0.8 * lightgbm + 0.2 * catboost, the private score is 0.570. Since the cv score is not improved, we do not select it for our final results.

	Competition	Solutions, interviews and articles
	2019 Data Science Bowl Uncover the factors to help measure how young children learn	Forum • 1st place (zr & oyx) • 2nd place (Fuson) Word2Vec features, • 3rd place (Limerobot) transformer • 4th place (Crystal cave miners) tfidf, rnn

Modeling

- For the validation set, we resampled to ensure one sample per one user.
- StratifiedGroupKFold, 5-fold.
- RSA (5 random seed) of LGB, CB, and NN.

Post Processing

- Ensemble = 0.5 * LGB + 0.2 * CB + 0.3 * NN.
- Set the threshold to optimize cv qwk.

Why Ensemble Methods ?

Tree based models

Lightgbm,Xgb, Catboost. (will be soon)

Stack

0 level) NN folds in folds model (5 outer folds, 5 inner folds), lgbm, catboost.
1st level) MLP, Lightgbm.
2nd level) Ridge.
https://www.kaggle.com/c/data-science-bowl-2019/discussion/127312

	long-best-threshold-f1	short-best-threshold-f1
Bert-base	0.618	0.457
Bert-large	0.679	0.541
Albert-xxl	0.700	0.555
ensemble	0.731	0.582

Models

We trained CatBoost, LightGBM, and MLP models on different subsets of the data:

- 1 model per meter
- 1 model per site_id
- 1 model per (building_id, meter)

Why Ensemble Methods ?

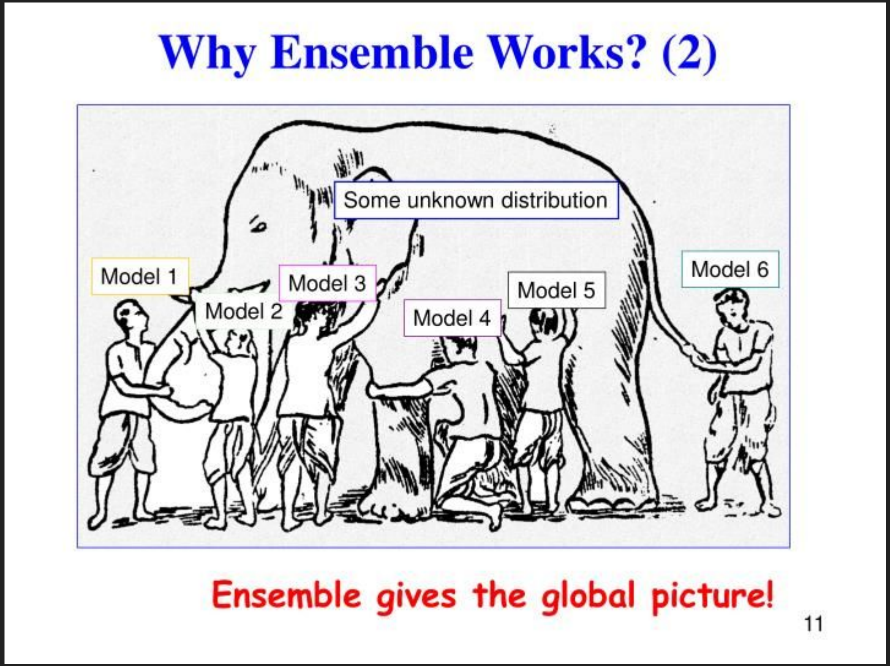
- Create a Baseline model.
- Works well for tabular data (pretty much any)
- Can Ensemble any model

Ensemble Methods ?

"Essentially, all models are wrong, but some are useful."

--- Box, George E. P.; Norman R. Draper (1987). Empirical Model-Building and Response Surfaces, p. 424, Wiley. ISBN 0471810339.

Ensemble Methods ?



The blind men and the elephant.

"Essentially, all models are wrong, but some are useful." and **their combination might be better**

Ensemble thought process

- Combine the predictions of many base estimators **that is combine many machine learning model prediction** **to improve the generalization and robustness of the model.**

Approaches :

1. Different training sets.
2. Different feature sampling.
3. Different algorithms.
4. Different hyperparameters.

CLEVER
AGGREGATION

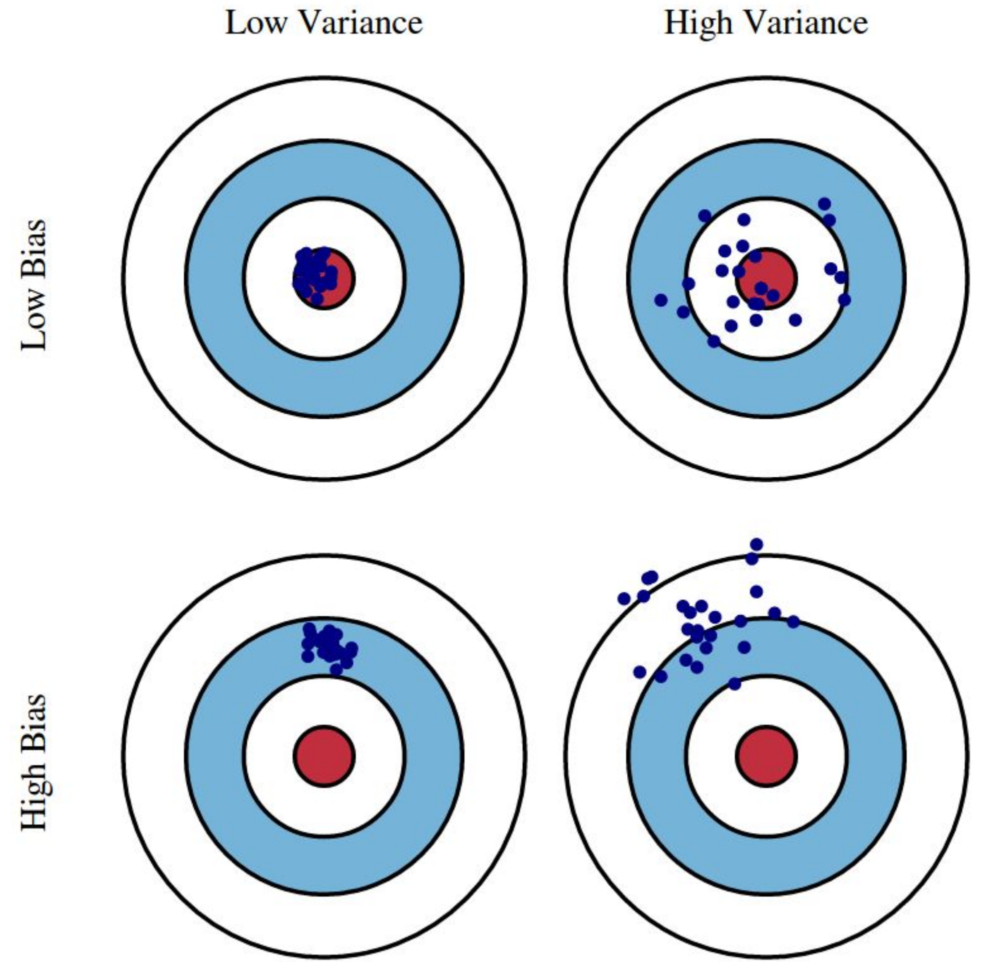
Ensemble thought process

- Combine the predictions of many base estimators **that is combine many machine learning model prediction** **to improve the generalization and robustness of the model.**

Types :

1. Averaging
2. Boosting
3. Voting
4. Stacking

CLEVER AGGREGATION

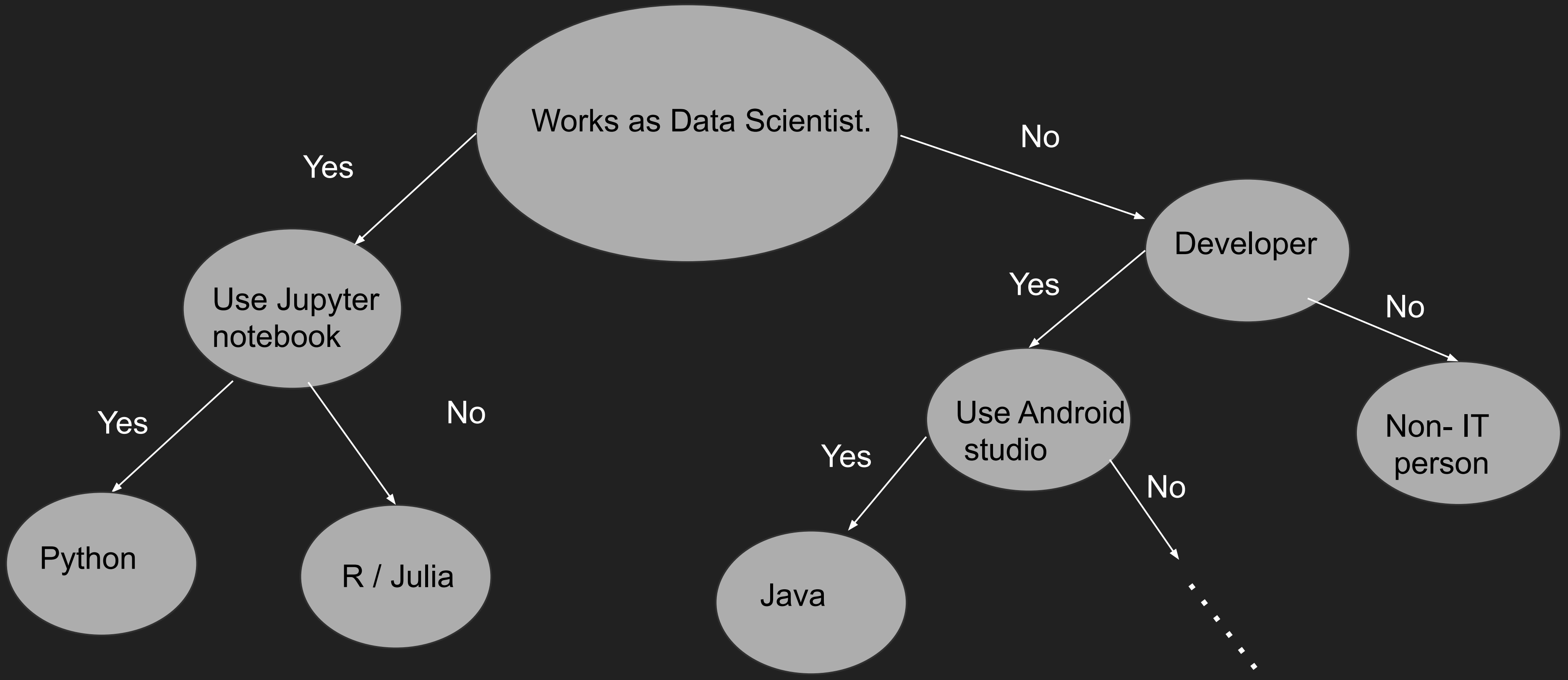

Low Variance
High Variance
Low Bias
High Bias

Ensemble thought process

- Combine the predictions of many base estimators **that is combine many machine learning model prediction** **to improve the generalization and robustness of the model.**

1. Averaging :

   Build different models **independently** and **average** their prediction to **reduce variance.**


Works as Data Scientist.
Yes
No
Use Jupyter notebook
Developer
Yes
No
Yes
No
Python
R / Julia
Use Android studio
Non- IT person
Yes
No
Java
......

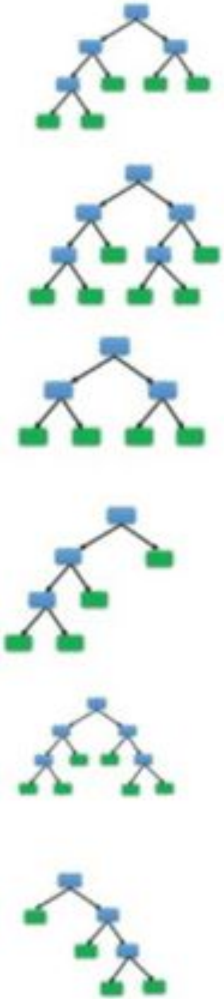
Random Forest

Random Forest :

1. Ensemble of Decision trees :

    a. Create different overfitted decision tree classifiers **using bootstrap aggregation.**

    b. While constructing each decision tree ; randomly choose k features among d.



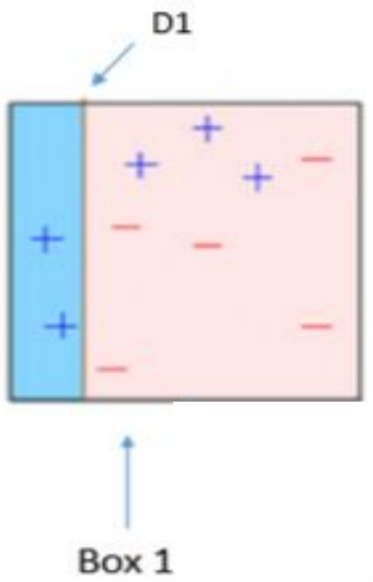
Random Forest in Action!!!

Boosting

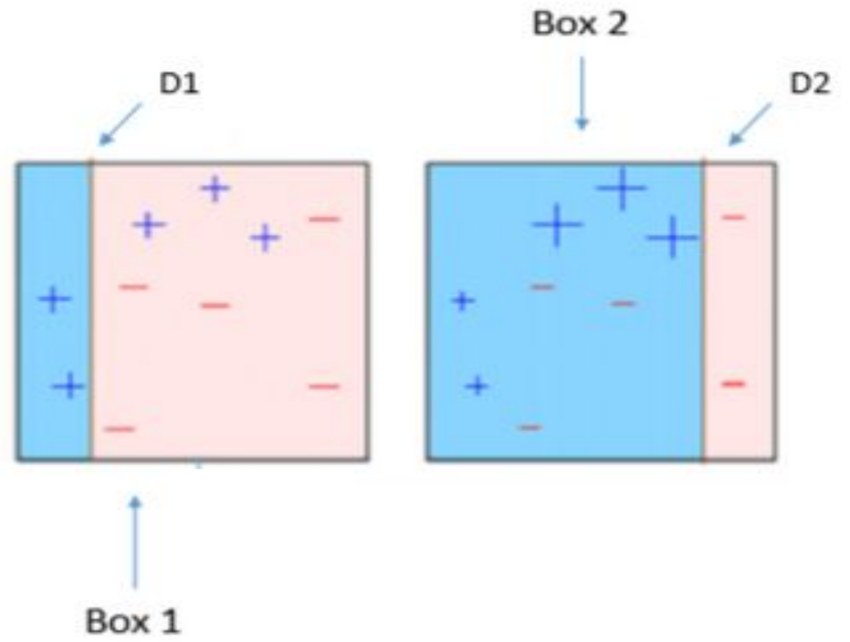
Build different models sequentially and by combining them try to reduce the bias.

- Weighted Boosting
- Gradient based Boosting.

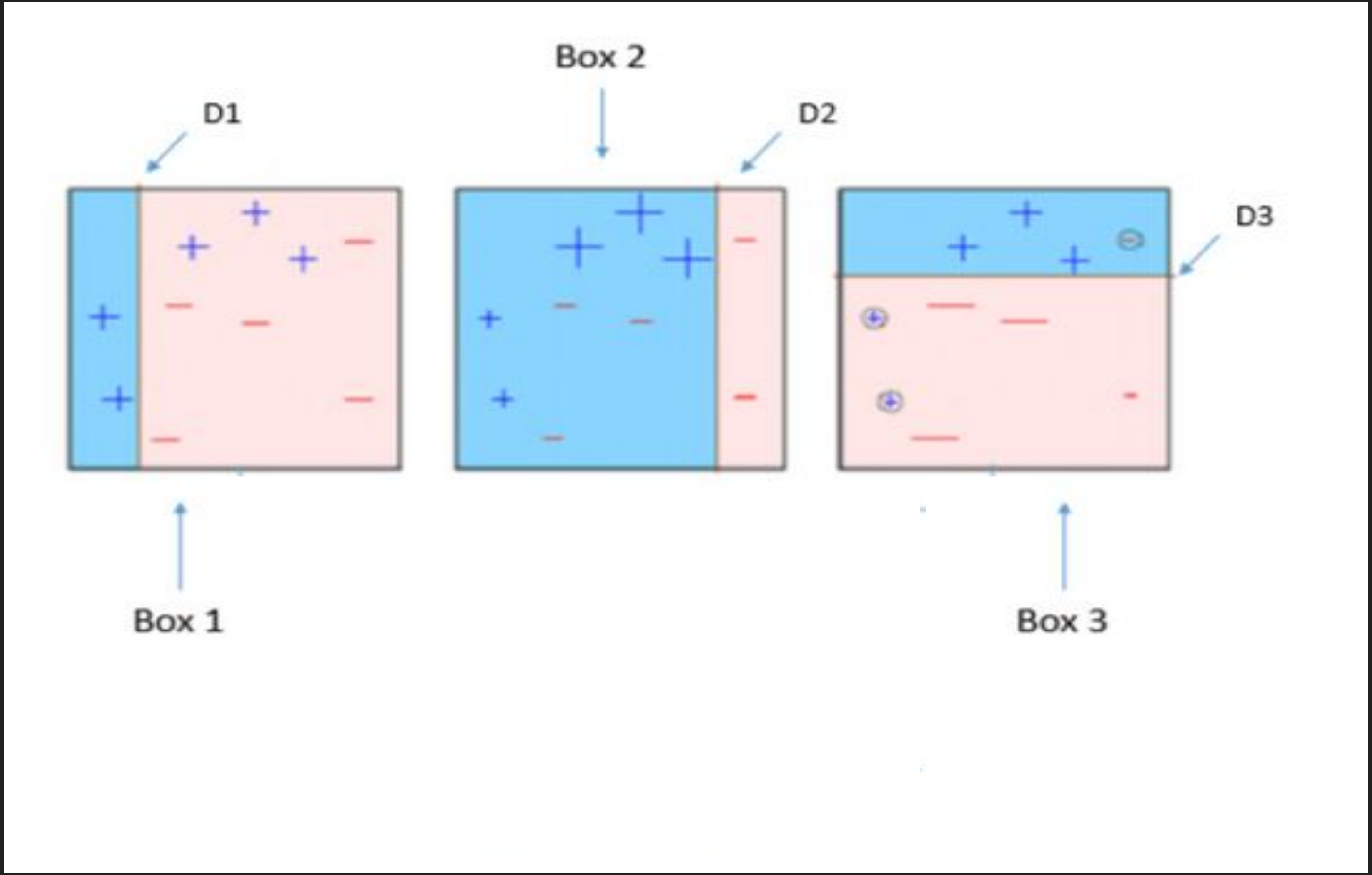
Weighted Boosting (AdaBoost)



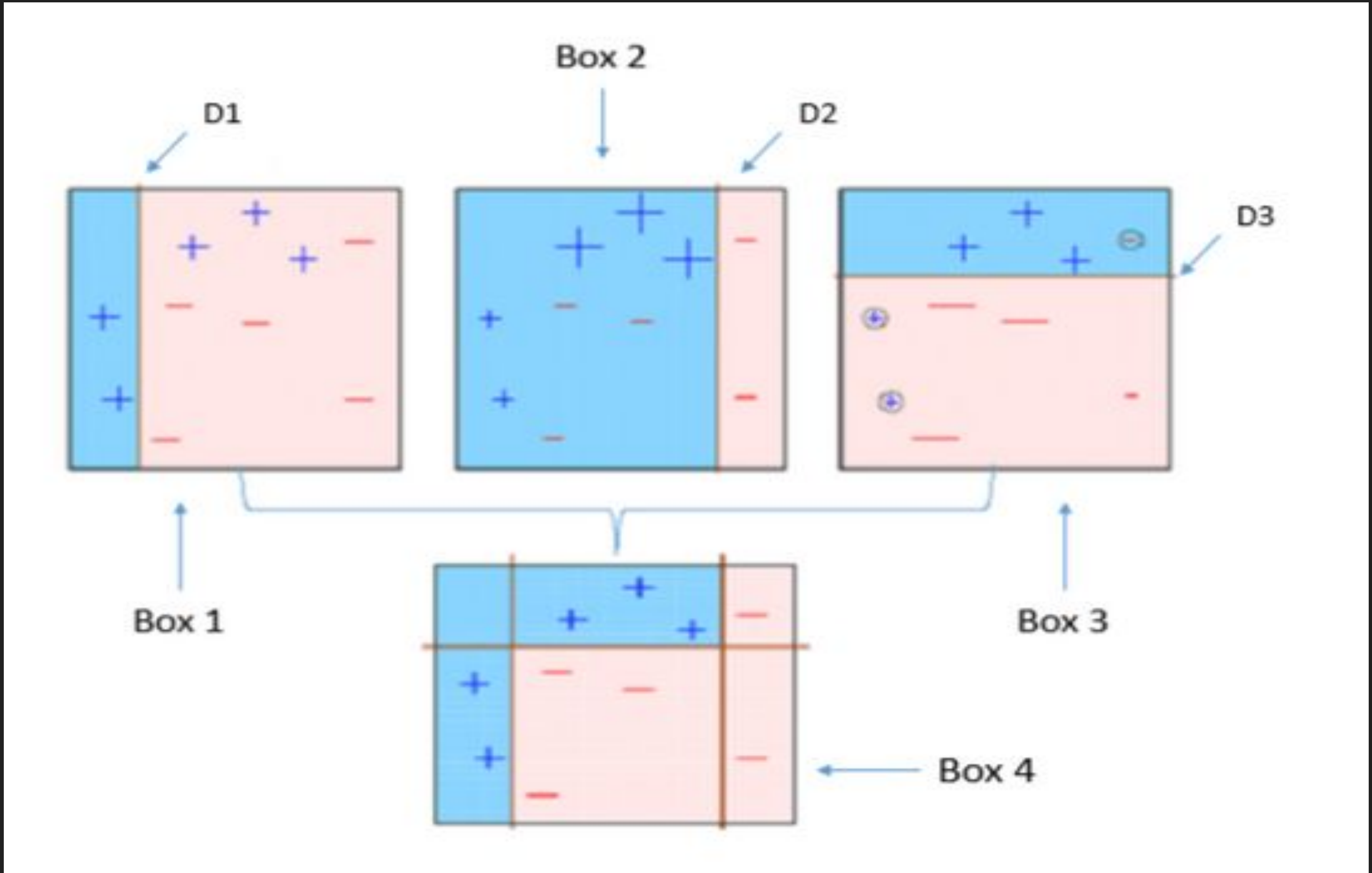
Weighted Boosting (AdaBoost)



Weighted Boosting (AdaBoost)



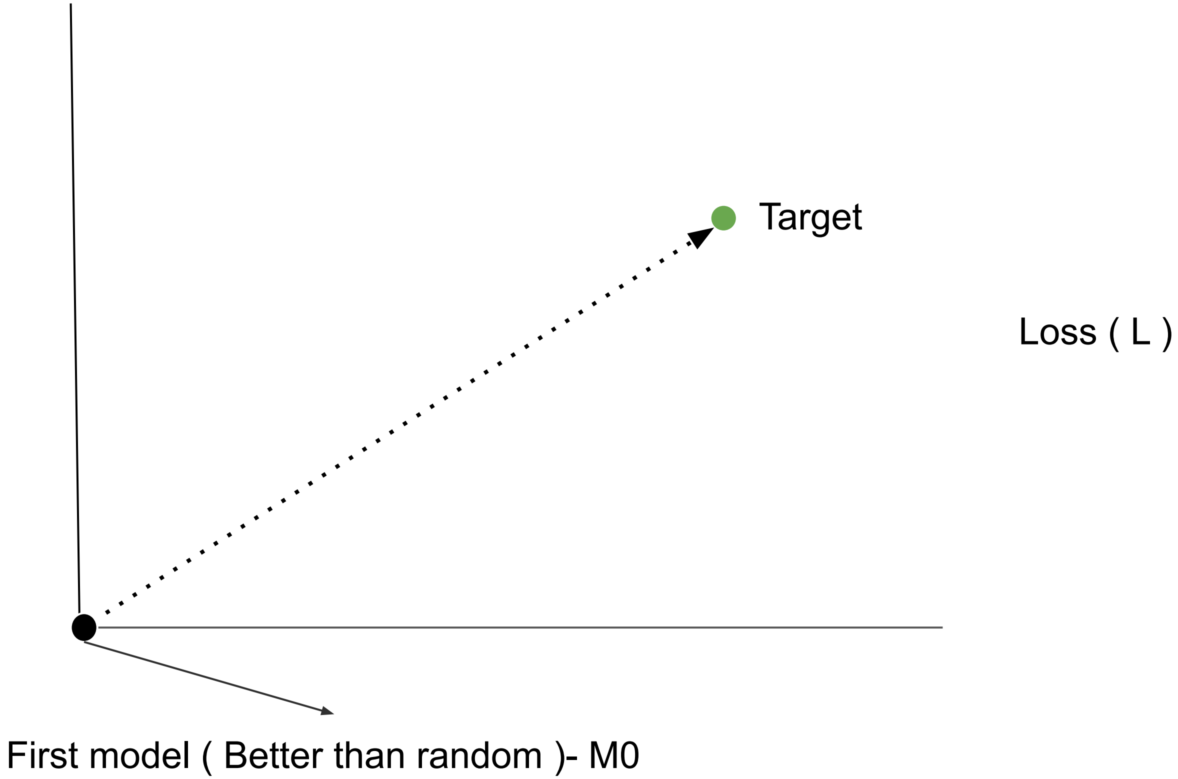
Weighted Boosting (AdaBoost)

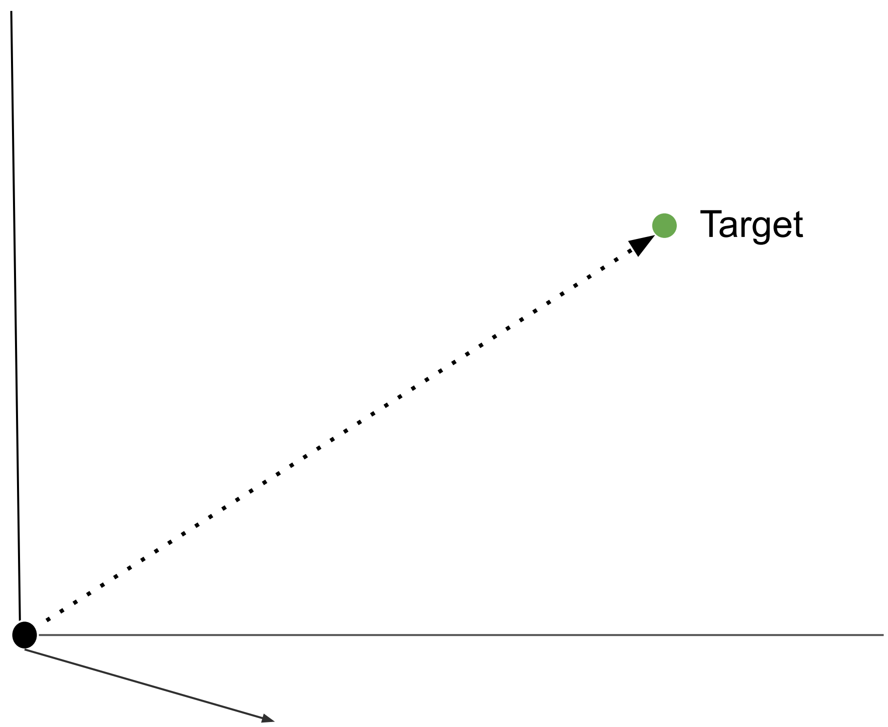


Boosting

Gradient Based Boosting (Gradient Boosting / Xgboost)

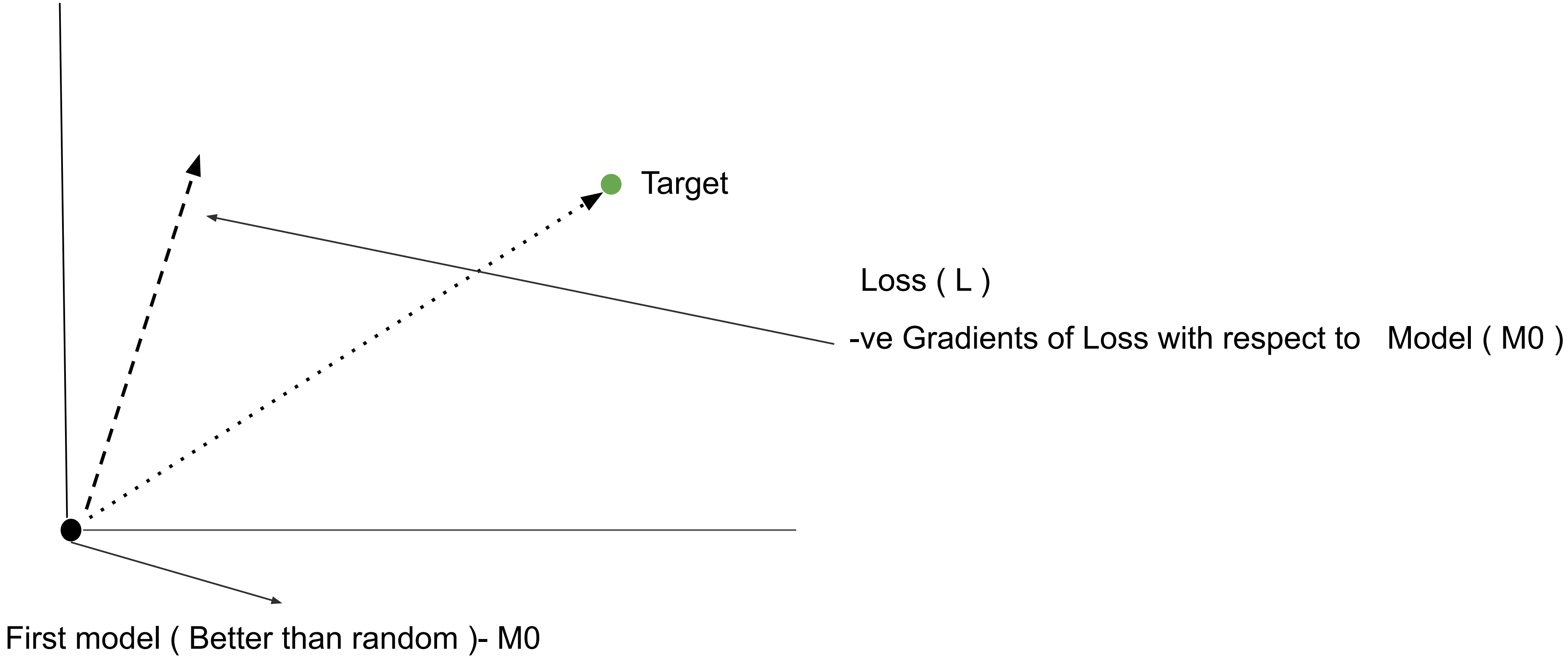
Every model should be better than random guess


Target
Loss (L)
First model (Better than random)- M0

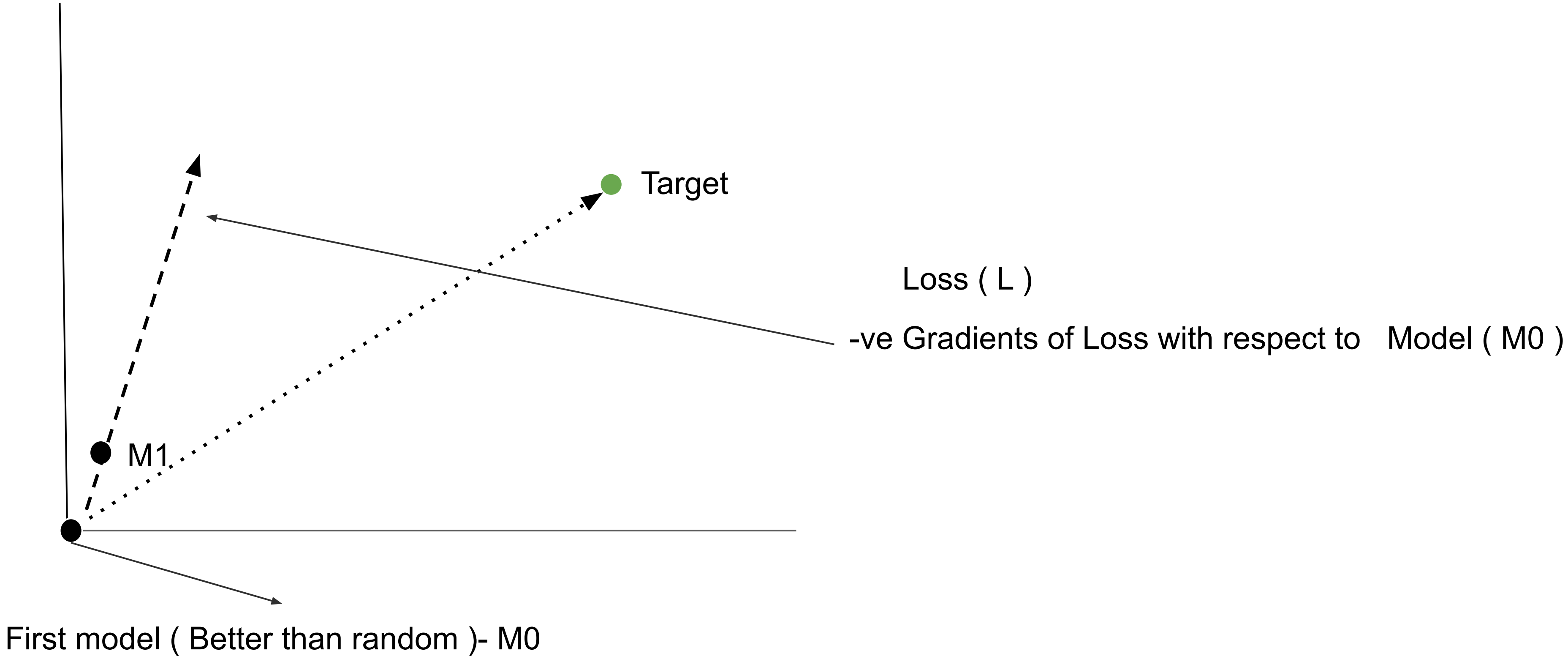


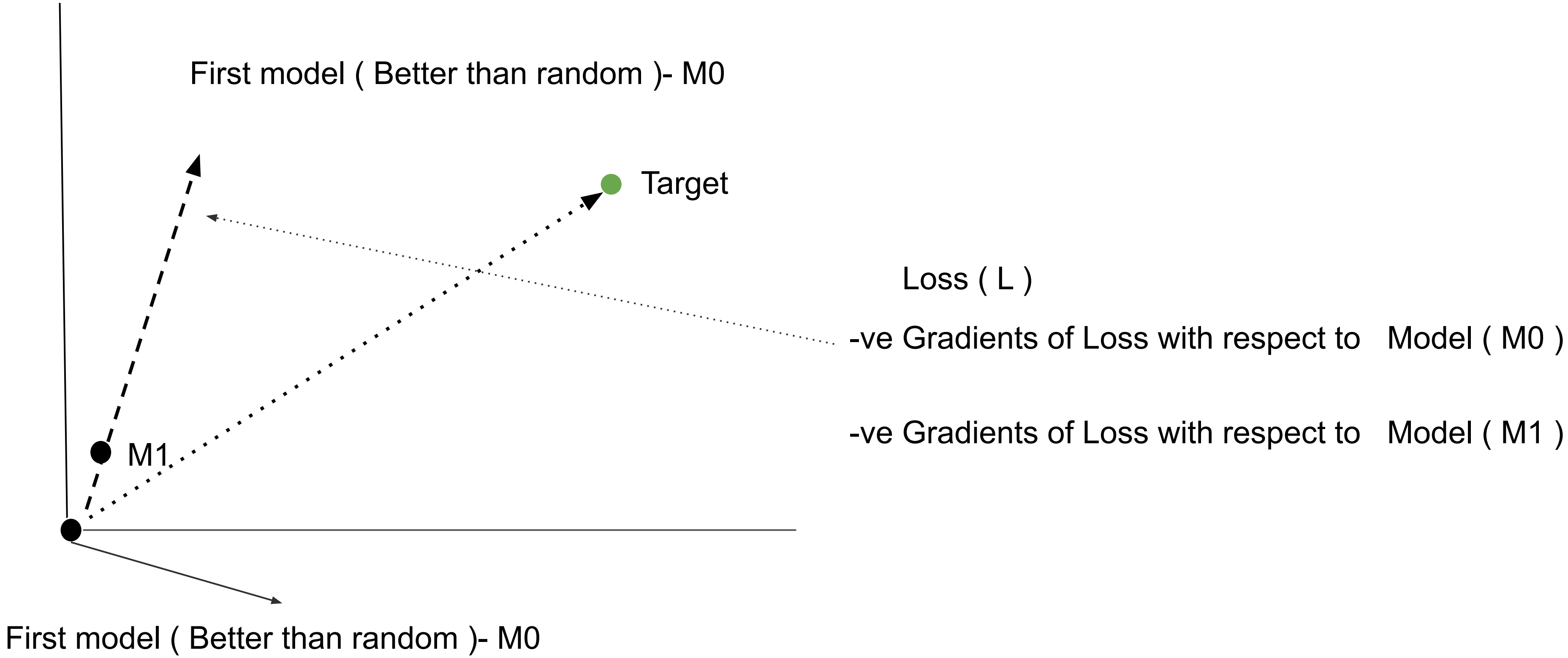
Loss (L)

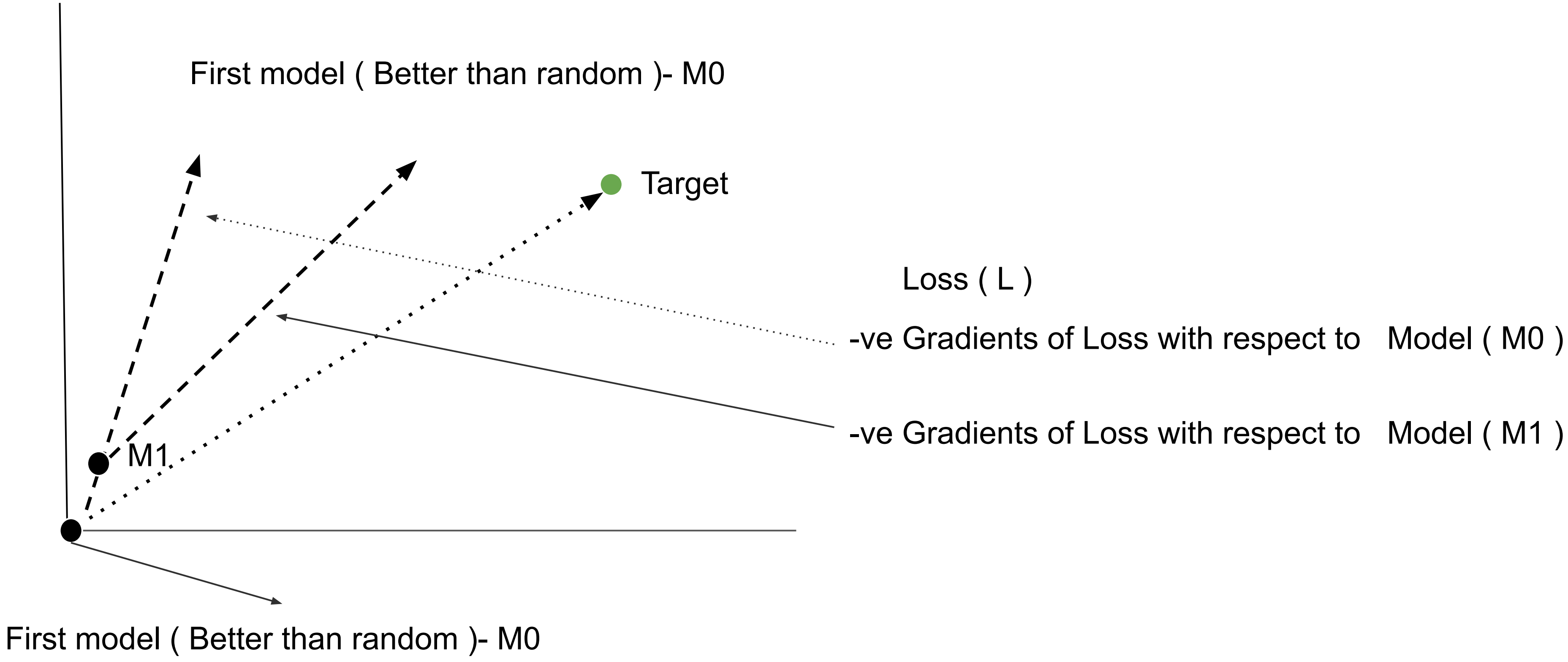
-ve Gradients of Loss with respect to   Model (M0)

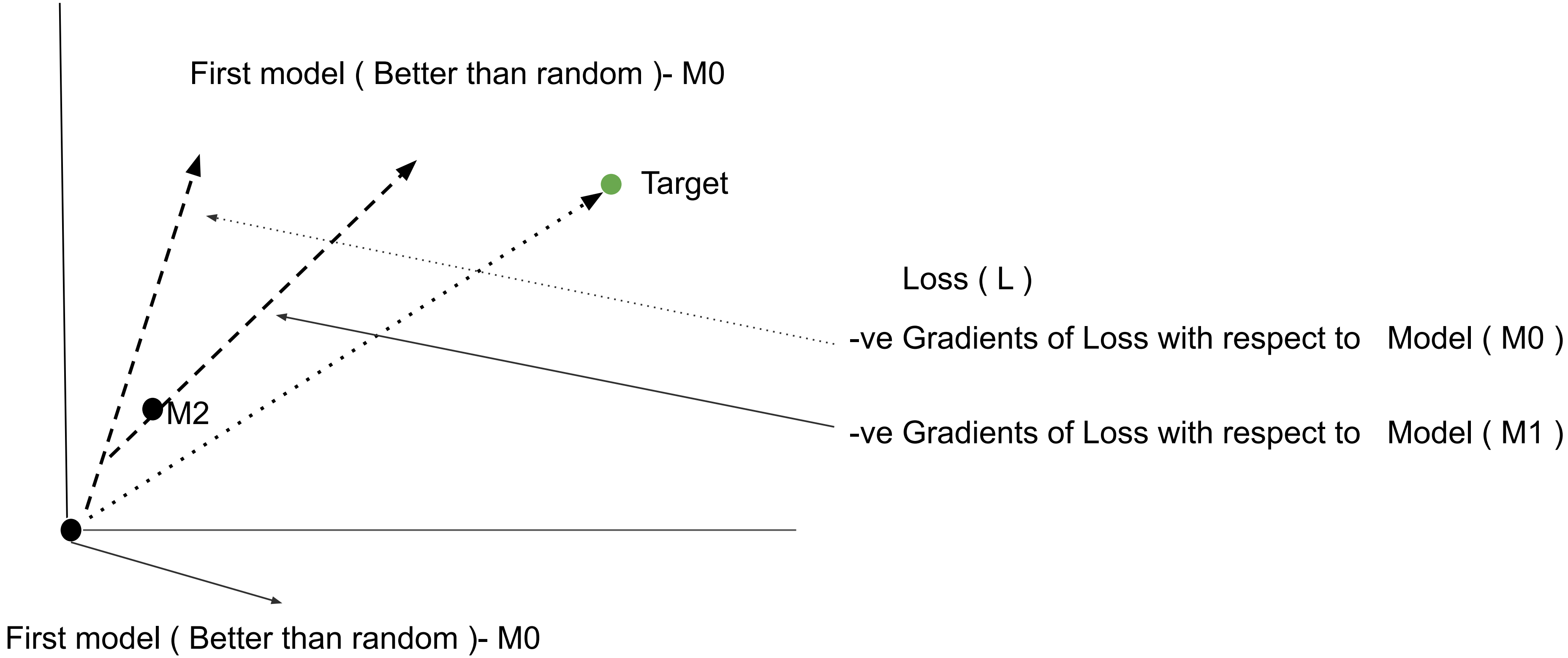


Target
Loss (L)
-ve Gradients of Loss with respect to   Model (M0)
First model (Better than random)- M0


Target
Loss (L)
-ve Gradients of Loss with respect to Model (M0)
M1
First model (Better than random)- M0


First model (Better than random)- M0
Target
Loss (L)
-ve Gradients of Loss with respect to Model (M0)
-ve Gradients of Loss with respect to Model (M1)
M1
First model (Better than random)- M0


First model (Better than random)- M0
Target
Loss (L)
-ve Gradients of Loss with respect to Model (M0)
-ve Gradients of Loss with respect to Model (M1)
M1
First model (Better than random)- M0

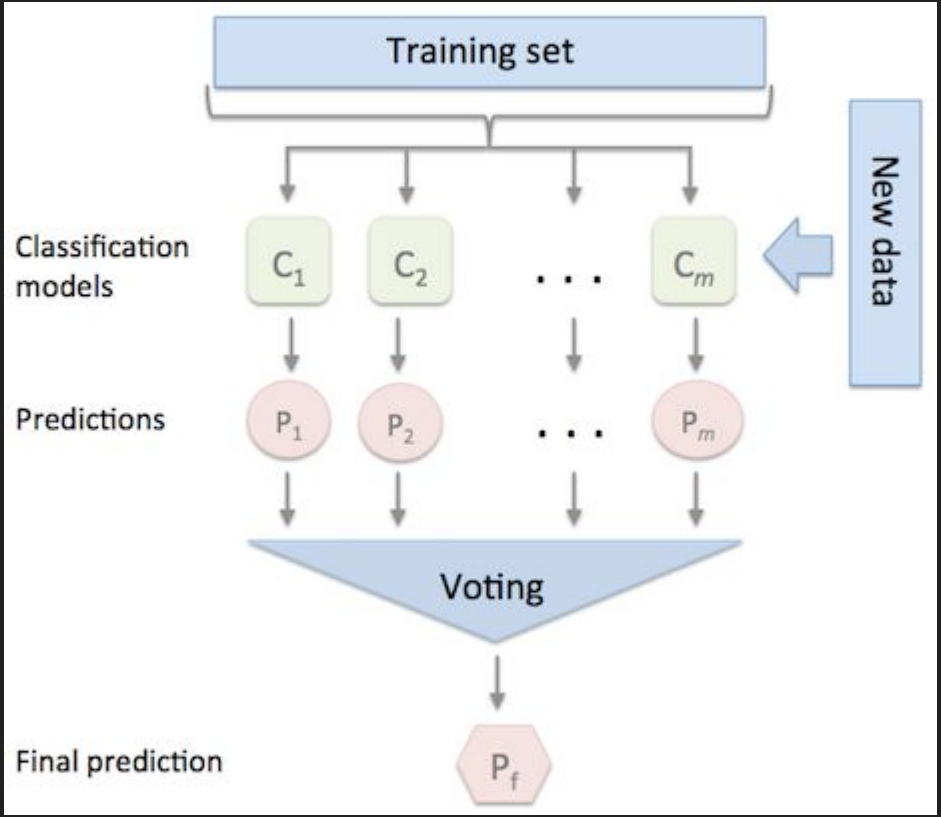

First model (Better than random)- M0
Target
Loss (L)
-ve Gradients of Loss with respect to Model (M0)
M2
-ve Gradients of Loss with respect to Model (M1)
First model (Better than random)- M0

Voting

- Use different classifiers ; and use the majority vote or soft vote to combine.

To ensemble equally well performing model to balance out their individual weakness.

Voting


Training set
New data
Classification models
C_1
C_2
. . .
C_m
Predictions
P_1
P_2
. . .
P_m
Voting
Final prediction
P_f

Voting

Hard Voting : The majority (mode) of the class label is predicted.

Soft Voting : The argmax() of the sum of predicted probabilities is predicted.

Soft Voting and Hard Voting

Classifier 1 predicts class A

Classifier 2 predicts class B

Classifier 3 predicts class B

Final Prediction is
B

Classifier 1 predicts class A with probability 99%

Classifier 2 predicts class A with probability 49%

Classifier 3 predicts class A with probability 49%

(99 + 49 + 49) / 3 = 65.67%

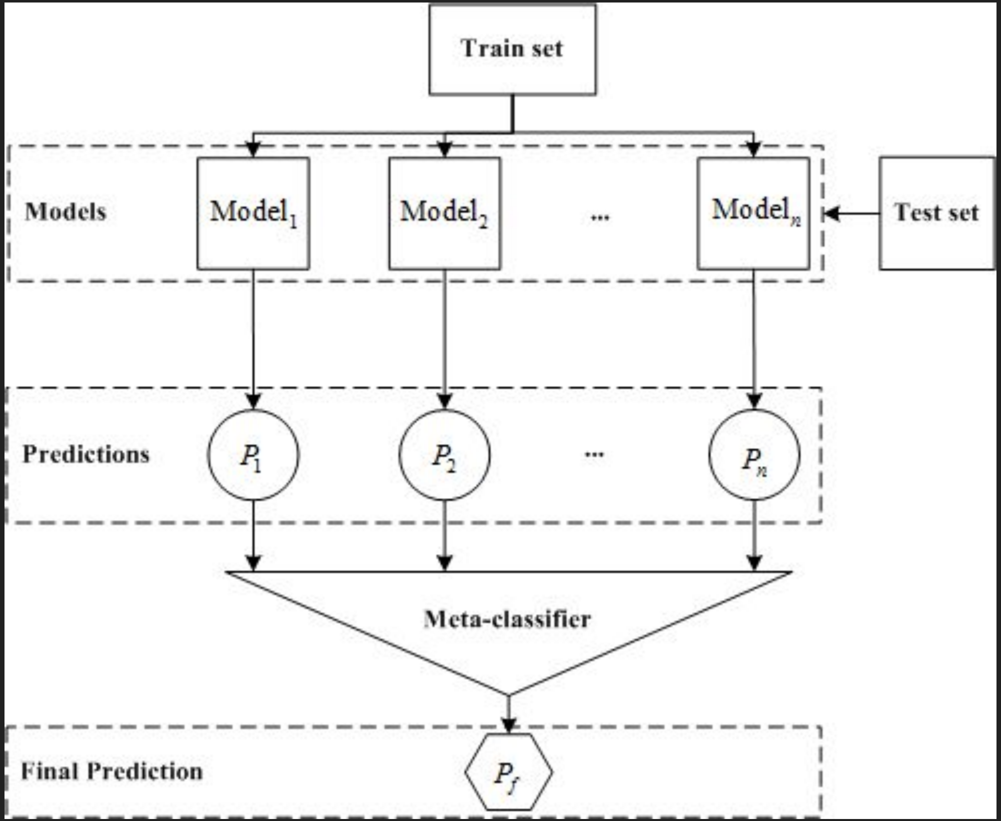
(1 + 51 + 51) / 3 = 34%

Final Prediction is A

Stacking

- To use the outputs of different algorithm as the input feature for the meta-classifier (Final Classifier)

Stacking



Train set
Models
$Model_1$
$Model_2$
...
$Model_n$
Test set
Predictions
P_1
P_2
...
P_n
Meta-classifier
Final Prediction
P_f

Advantages of Ensemble

1. Robustness
2. Good Generalization performance.
3. Parallelization

Disadvantages of Ensemble

1. Human readability / explainability is not good.
2. Takes a lot of time to train.
3. Time - effort off trade off to make accuracy may not make sense.

Resources :

- Kaggle Discussion forum
- Coursera (How to Win kaggle Competitions)
- Book (Ensemble methods , Foundations and Algorithms)
- Practice