Model Ensembling

Bagging and Boosting

Model Ensembling

- One Single model with the training dataset often cannot be an optimal solution
- For instance, one tree fitted to the training dataset, can have error in one or the other way
- Model Ensembling is such as idea in which we combine the predictions from two or more model fits and take an averaging on the predictions or work on errors on process the predictions further

Model Ensembling

- Ensembling is a technique of combining two or more algorithms of similar or dissimilar types called base learners or weak learners.
- This is done to make a system of predictive modelling more robust and the individual algorithms.
- Many times it is observed that the individual algorithms don’t predict in the expected precise way as the group of algorithms do. Hence this technique.

Types of Ensembling

- Categorical Predictions
    - Majority Vote
- Numerical Predictions
    - Averaging
    - Weighted Averaging
- There can be more techniques than mentioned above

Categorical Predictions: Majority Vote

- Suppose that you fitted some 5 models for a classification problem with binary outcomes
- Also consider that you have applied the fitted models separately on the validation set.
- Say, an observation in the validation dataset, which has been predicted as follows: (possible outcomes 1 / 0)

Model 1	Model 2	Model 3	Model 4	Model 5	Majority
1	0	1	0	0	0

Here, the outcome will be considered by majority of vote.

Numerical Predictions: Averaging

- Suppose that you fitted some 5 models for a regression problem with numerical outcomes
- Also consider that you have applied the fitted models separately on the validation set.
- Say, an observation in the validation dataset, which has been predicted as follows:

Model 1	Model 2	Model 3	Model 4	Model 5	Average
102.34	121.39	134.45	129.43	120.34	121.59

Here, the outcome will be considered by averaging. We can adopt any measure of central tendency for averaging.

Numerical Predictions: Weighted Averaging

- Suppose that you fitted some 5 models for a regression problem with numerical outcomes and you intend to put some weights for the findings in the different models
- Also consider that you have applied the fitted models separately on the validation set.
- Say, an observation in the validation dataset, which has been predicted as follows:

Model 1 (w1=0.3)	Model 2 (w2=0.1)	Model 3 (w3=0.4)	Model 4 (w4=0.05)	Model 5 (w5=0.15)	Average
102.34	121.39	134.45	129.43	120.34	121.144

Here, the outcome will be considered by weighted averaging.

Ensembling Techniques

- There can various ways of ensembling. The following are the popularly known techniques:
    - Bagging
    - Boosting
    - Stacking

Bagging

- Bagging is boostrap aggregation
- Multiple bootstrapped samples are drawn from the same data
- With each of these samples, we can fit the same model and ultimately get a majority vote or averaging to get the final prediction.
- Bagging reduces the variance in predictions.
- Random Forest algorithm is a kind of bagging in which with each sample, set of predictors are chosen are at random

Boosting

- Boosting is a sequential technique of fitting a first algorithm and then fitting subsequent algorithms on residuals of the first algorithm by giving higher weight to those observations that had been poorly predicted.
- Boosting makes use of weak learners each of which might not be good for some part of the dataset and ultimately boosts the performance.
- Boosting reduces the bias which may lead to overfitting. Hence parameter tuning is a must for avoiding overfitting.
- ADABOOST, GBM , XGBOOST etc. are the examples of boosting.

Stacking

- This is a technique in which we use more than one algorithms and predict the responses on the train set
- On the top of those responses we build another model.
- On test set, we apply all the models built on the train set.
- This technique has been proved to be one of most promising techniques on various data science projects and also data science competitions.