5 Easy questions on Ensemble Modeling everyone sh know

BUSINESS ANALYTICS (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/BUSINESS-ANALYTICS/) MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/CATEGORY/MACHINE-LEARNING/)

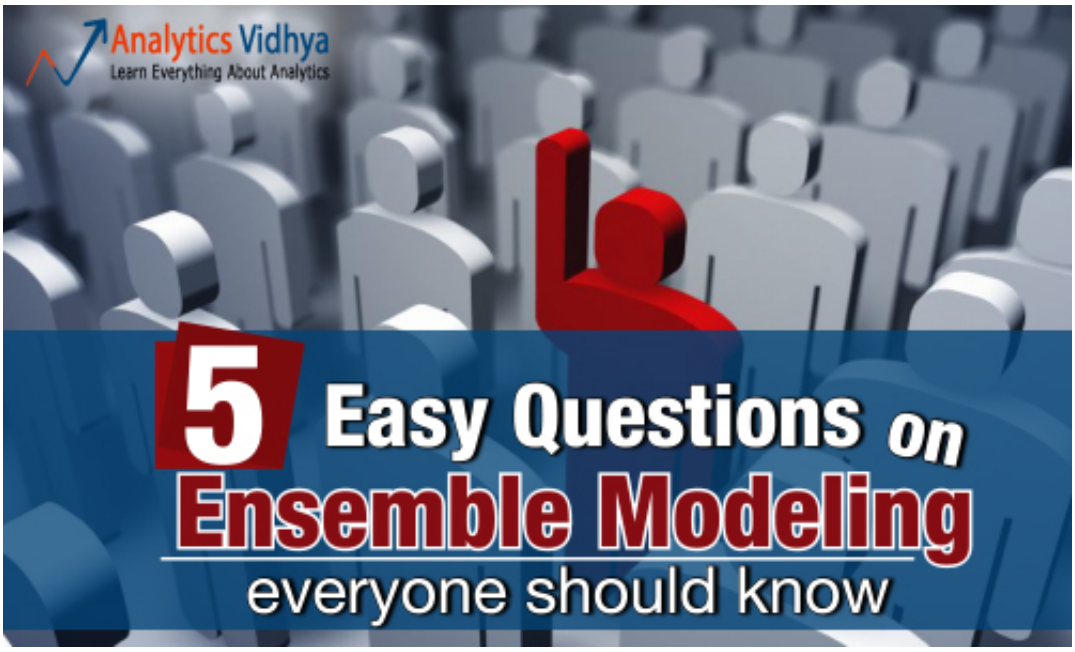
stion2.jpg&description=5%20Easy%20questions%20on%20Ensemble%20Modeling%20everyone%20should%20know)

Introduction

If you've ever participated in data science competitions, you must be aware of the pivot ensemble modeling plays. In fact, it is being said that ensemble modeling offers one c convincing way to build highly accurate predictive models. The availability of bagging ar (http://discuss.analyticsvidhya.com/t/difference-between-bagging-boosting-and-how-the help/4095) algorithms further embellishes this method to produce awesome accuracy leve

So, next time when you build a predictive model, do consider using this algorithm. definitely pat my back for this suggestion. And, if you've already mastered this method, gr to hear your experience about ensemble modeling in the comments section below.

For the rest, I am sharing some of the most commonly asked questions on ensemble mod ever wish to evaluate any person's knowledge on ensemble, you can daringly ask these c check his / her knowledge. In addition, these are among the easiest questions, hence you to get them wrong!



(https://www.analyticsvidhya.com/wp-content/uploads/2015/09/5-question2.jpg)

Which are the common questions (related to Ensemble Model

After analyzing various data science forums (http://discuss.analyticsvidhya.com), I have id
5 most common questions related to ensemble modeling. These questions are highly relev
scientists new to ensemble modeling. Here are the questions:

1. What is an ensemble model?
2. What are bagging, boosting and stacking?
3. Can we ensemble multiple models of same ML algorithm?
4. How can we identify the weights of different models?
5. What are the benefits of ensemble model?

Let's discuss each question in detail.

1. What is an Ensemble Model?

Let's try to understand it by solving a classification challenge.

Problem: Set rules for classification of spam emails

(https://www.analyticsvidhya.com/wp-content/uploads/2015/09/images.jpg)

Solution: We can generate various rules for classification of spam emails, let's look at th
them:

- Spam
    - Have total length less than 20 words

    - Have only image (promotional images)
    - Have specific key words like "make money and grow" and "reduce your fat"
    - More miss spelled words in the email
- Not Spam
    - Email from Analytics Vidhya domain
    - Email from family members or anyone from e-mail address book

Above, I've listed some common rules for filtering the SPAM e-mails. Do you think that all individually can predict the correct class?

Most of us would say no – And that's true! Combining these rules will provide robust pr compared to prediction done by individual rules. This is the principle of Modeling. Ensemble model combines multiple 'individual' (diverse) models together ar superior prediction power.

If you want to relate this to real life, a group of people are likely to make better decisions to individuals, especially when group members come from diverse background. The same machine learning. Basically, an ensemble is a supervised learning technique for multiple *weak learners/ models* to produce a *strong learner*. Ensemble model works bette ensemble models with low correlation.

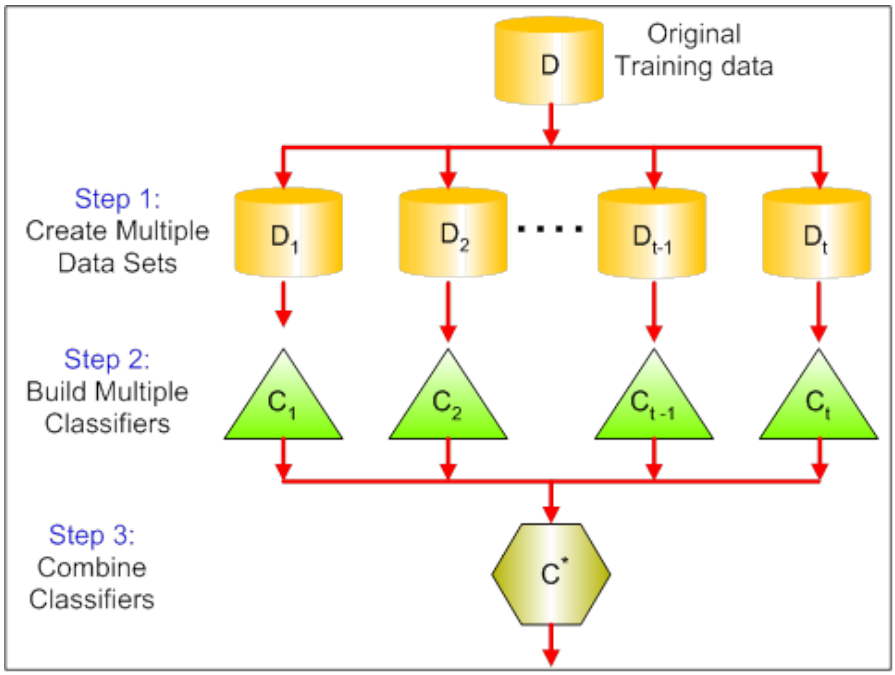
A good example of how ensemble methods are commonly used to solve data science p the *random forest (https://www.analyticsvidhya.com/blog/2014/06/introduction-ranc simplified/)* algorithm (having multiple CART models). It performs better compared to indiv model by classifying a new object where each tree gives "votes" for that class and the fore the classification having the most votes (over all the trees in the forest). In case of regress the average of outputs of different trees.

You can also follow this article "Basics of Ensemble Learning Explained in Simp (https://www.analyticsvidhya.com/blog/2015/08/introduction-ensemble-learning/)" f knowledge on ensemble modeling.

2. What are Bagging, Boosting and Stacking?

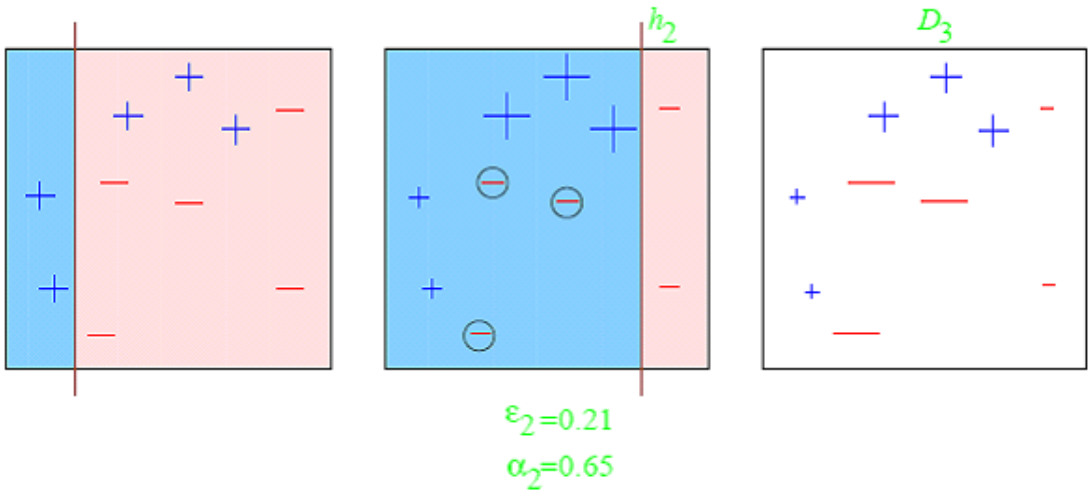
Let's look at each of these individually and try to understand the differences between these

Bagging (Bootstrap Aggregating) is an ensemble method. First, we create random sam training data set (sub sets of training data set). Then, we build a classifier for each sam results of these multiple classifiers are combined using average or majority voting. Baggi reduce the variance error.



(https://www.analyticsvidhya.com/wp-content/uploads/2015/09/bagging.png)

Boosting provides sequential learning of the predictors. The first predictor is learned on data set, while the following are learnt on the training set based on the performance of th one**.** It starts by classifying original data set and giving equal weights to each observatior are predicted incorrectly using the first learner, then it gives higher weight to the misse observation. Being an iterative process, it continues to add classifier learner until a limit is the number of models or accuracy. Boosting has shown better predictive accuracy than b it also tends to over-fit the training data as well.

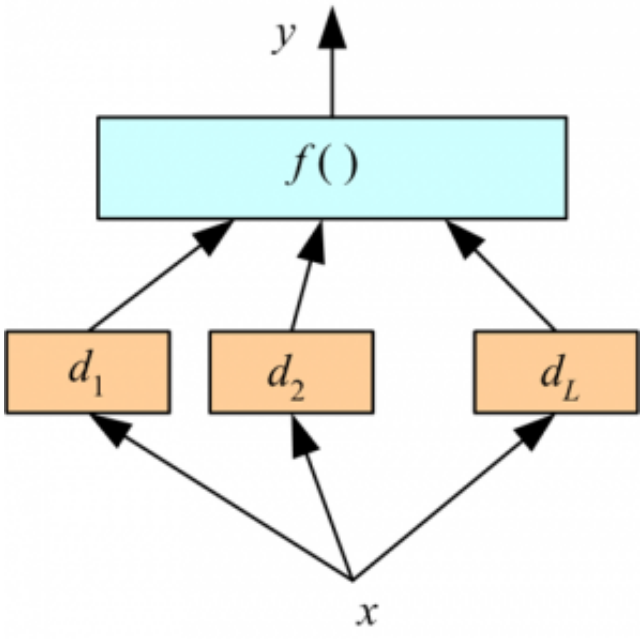


(https://www.analyticsvidhya.com/wp-content/uploads/2015/09/boosting1.png)

Most common example of boosting is AdaBoost and Gradient Boosting. You can also lo articles to know more about boosting algorithms.

- Getting smart with Machine Learning – AdaBoost and Gradie (https://www.analyticsvidhya.com/blog/2015/05/boosting-algorithms-simplified/)
- Learn Gradient Boosting Algorithm for better predictions (with code (https://www.analyticsvidhya.com/blog/2015/09/complete-guide-boosting-metho

Stacking works in two phases. First, we use multiple base classifiers to predict the cla a new learner is used to combine their predictions with the aim of reducing the generalizati



(https://www.analyticsvidhya.com/wp-content/uploads/2015/09/stacking-297x300.png)

3. Can we ensemble multiple models of same ML algorithm?

Yes, we can combine multiple models of same ML algorithms, but combining multiple generated by different algorithms would normally give you better predictions. It is c diversification or independent nature as compared to each other. For example, the pred random forest, a KNN, and a Naive Bayes may be combined to create a stronger final predi compared to combining three random forest model. The key to creating a powerful e model diversity. An ensemble with two techniques that are very similar in nature will perf than a more diverse model set.

Example: Let's say we have three models (A, B and C). A, B and C have prediction accur 80% and 55% respectively. But A and B are found to be highly correlated where as C is correlated with both A and B. Should we combine A and B? No, we shouldn't, because the are highly correlated. Hence,we will not combine these two as this ensemble will not help any generalization error. I would prefer to combine A & C or B & C.

4. How can we identify the weights of different models for ensemble?

One of the most common challenge with ensemble modeling is to find optimal weights to base models. In general, we assume equal weight for all models and takes the average of p But, is this the best way to deal with this challenge?

There are various methods to find the optimal weight for combining all base learners. Thes provide a fair understanding about finding the right weight. I am listing some of the method

- Find the collinearity between base learners and based on this table, then identify the bas ensemble. After that look at the cross validation score (ratio of score) of identified base mo the weight.
- Find the algorithm to return the optimal weight for base learners. You can refer article Finc Weights of Ensemble Learner using Neural (https://www.analyticsvidhya.com/blog/2015/08/optimal-weights-ensemble-learner-neu network/) to look at the method to find optimal weight.
- We can also solve the same problem using methods like:

- Forward Selection of learners (https://www.analyticsvidhya.com/blog/2015/0
techniques-ensemble-modelling/)
- Selection with Replacement (https://www.analyticsvidhya.com/blog/2015/0
techniques-ensemble-modelling/)
- Bagging of ensemble methods (https://www.analyticsvidhya.com/blog/2015/0
techniques-ensemble-modelling/)

You can also look at the winning solution of Kaggle / data science competitions to unders
methods to deal with this challenge.

5. What are the benefits of ensemble model?

There are two major benefits of Ensemble models:

- Better prediction
- More stable model

The aggregate opinion of a multiple models is less noisy than other models. In finance, v
"Diversification" a mixed portfolio of many stocks will be much less variable than just
stocks alone. This is also why your models will be better with ensemble of models
individual. One of the caution with ensemble models are over fitting although bagging take
largely.

End Note

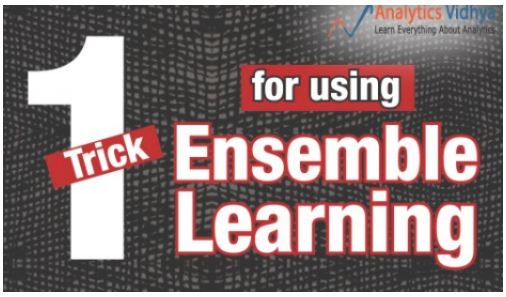
In this article, we have looked at the 5 frequently asked questions on Ensemble mo
answering these questions, we have discussed about "Ensemble Models", "Methods of
"Why should we ensemble diverse models?", "Methods to identify optimal weight for ens
finally "Benefits". I would suggest you to look at the top 5 solutions of data science compe
see their ensemble approaches to have better understanding and practice a lot. It will h
understand what works or what doesn't.

Did you find this article useful? Have you tried anything else to find optimal weights or identify th
learner? I'll be happy to hear from you in the comments section below.

If you like what you just read & want to continue your analytics learning, subs our emails (http://feedburner.google.com/fb/a/mailverify?uri=analyticsvidhy us on twitter (http://twitter.com/analyticsvidhya) or like our facebook page (http://facebook.com/analyticsvidhya).

Share this:

RELATED



(https://www.analyticsvidhya.com/blog/2015/10/trick-right-model-ensemble/)
Powerful 'Trick' to choose right models in Ensemble Learning (https://www.analyticsvidhya.com/blog/2015/10/trick-right-model-

(https://www.analyticsvidhya.com/blog/2015/12/improve-machine-learning-results/)
8 Proven Ways for improving the "Accuracy" of a Machine Learning Model (https://www.analyticsvidhya.com/blo

(https://www.analyticsvidh
g/2017/02/40-questions-t
data-scientist-on-ensembl
techniques-skilltest-soluti
40 Questions to ask a Data
on Ensemble Modeling Tech
(Skilltest Solution)

ensemble/)
October 20, 2015
In "Business Analytics"

g/2015/12/improve-machine-learning-results/)
December 29, 2015
In "Machine Learning"

(https://www.analyticsvidh
g/2017/02/40-questions-t
data-scientist-on-ensembl
techniques-skilltest-soluti
February 12, 2017
In "Machine Learning"

TAGS: BAGGING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BAGGING/), BOOSTING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/BOOSTIN(
MODEL (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/ENSEMBLE-MODEL/), MACHINE LEARNING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/
LEARNING/), STACKING (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/TAG/STACKING/)

Previous Article
Analytics Manager –BI – Gurgaon (3-5 years of experience) (https://www.analyticsvidhya.com/blog/2015/09/analytics-manager-gurgaon-delhi-ncr-3-5-years-experience/)

Next Artic
Analytics Manager – Predictive Modellin and Business Analysis – Gurgaon (3-5 year of experience (https://www.analyticsvidhya.com/blog/2015/1(manager-predictive-modelling-business-analysis gurgaon-3-5-years-experience/

(https://www.analyticsvidhya.com/blog/author/sunil-ray/)

Author

Sunil Ray (https://www.analyticsvidhya.com/blog/author/ ray/)

I am a Business Analytics and Intelligence professional with deep experienc Indian Insurance industry. I have worked for various multi-national Insurance companies in last 7 years.

3 COMMENTS

shanmukh says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/09/QUESTIONS-ENSEMBLE-MODELING/?REPLYTOCOM
SEPTEMBER 30, 2015 AT 5:09 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/09/QUESTIONS-ENSEMBLE-MODELIN
96261)

It's very informative. I like it 🙂

Ashis says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/09/QUESTIONS-ENSEMBLE-MODELING/?REPLYTOCOM
OCTOBER 2, 2015 AT 9:37 PM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/09/QUESTIONS-ENSEMBLE-MODELING/#
96424)

It's very well written..... Really the best article to get introduced to ensemble learning..... I lost in the algorithms before reading this article....Simple description is the best part of t

Chandu says: REPLY (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/09/QUESTIONS-ENSEMBLE-MODELING/?REPLYTOCOM=
OCTOBER 4, 2016 AT 7:36 AM (HTTPS://WWW.ANALYTICSVIDHYA.COM/BLOG/2015/09/QUESTIONS-ENSEMBLE-MODELING/#
116755)

HI ,

It was very Informative . Tanks alot...

Plz clarify , how to check the correlation b/w two models ?referring to Point 3 Example)

Thanks in advance

Chandu

LEAVE A REPLY

Connect with:

(https://www.analyticsvidhya.com/wp-login.php?

action=wordpress_social_authenticate&mode=login&provider=Facebook&redirect_to=https%3A%
ensemble-modeling%2F)

Your email address will not be published.

Comment

Name (required)

Email (required)

Website

SUBM

ANALYTICS VIDHYA

About Us (https://www.analyticsvidhya.com/about-me/)

Team (https://www.analyticsvidhya.com/about-me/team)

DATA SCIENTISTS

Blog (http://analyticsvidhya.com/blog)

Discuss (https://discuss.analyticsvidhya.com)

COMPANIES

Advertise w
(https://ww

Recruit usin
(https://ww

Hackathon
(https://dat