Master of Big Data Analytics

Statistical & Machine learning approaches for marketing

Individual Assignment - Task 1

Andres Olivera

Tree model:

A tree model or called CART as well, is part of the supervised machine learning models and famous modern models like Random Forest and all Bosting use it. (DataCamp, 2020)

As de name implies, it uses a tree like model of decisions. It can be used in daily lives situations to take decisions. It consists on taking binary decisions based on a condition. That is translated in a split, and this process goes on and on. It can be used for regression problems, which handle continuous variables, or in classification problems, handling categorical variables.

Advantages:

- Simple to understand, interpret and visualize
- Performs internally feature selection
- Handles numerical and categorical data
- Low data preparation required
- Handles missing data
- Robust to outliers

Disadvantages:

- Prone to over fitting (High variance)
- If the tree is too big, it becomes hard to read

Tackling overfitting:

- Pruning:
    - Pre-pruning: Set the splitting criteria to stop the tree growing
    - Post-pruning: Grow a very large tree and then prune it.

  Regression: The splitting criteria is the RSS. The lowest the better.

    - You look for the weakest link, the size of the connection between two nodes. To visualize it you can use a scree plot. On the y axis using the RSS and on the x axis the size of the tree.

  Classification:

    - Three splitting criteria:

- Gini: Purity of the node. If all observations are labeled the same or in other words, if all data points are the same in the partition made by the tree, the purity is 100%. This is the most common method. (Augmented Startups, 2020)
- Classification error rate
- Cross entropy: The deviance

ENSEMBLE MODELS:

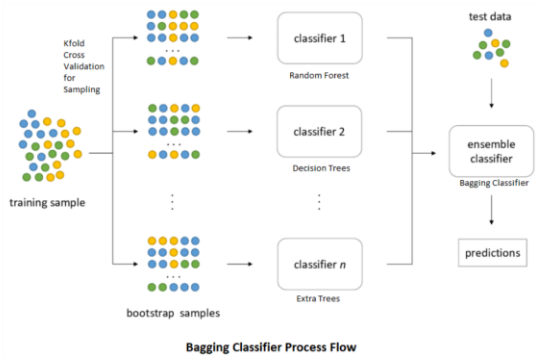
By working together, a group of weak or simple models can create a strong model (learner). Like a single string that by itself is very weak but, by combining many of them, you get a very strong rope.

Bagging:

Uses bootstrap sampling and aggregates the individual models by averaging:

The idea of this model is to create different models and each of them will be a bag. The algorithm will randomly select rows or observation from the training dataset and put them on each bag. While doing this process replacement is allowed, meaning that the same row can be in different bag. Replacement is allowed to have a truly random re-sampling.

Usually the number of samples (rows) taken for each bag (model) is half of the total number of observations (on training set)



https://medium.com/ml-research-lab/bagging-ensemble-meta-algorithm-for-reducing-variance-c98fffa5489f

After the bootstrap, each model will be trained independently and in parallel, how can be seen in the image above. Then, the final prediction, the outcome or the Y, is the mean of all of the other predictions made by each model (or the majority vote). It follows the wisdom of the crowd's saying.

Advantages:

- Reduces variance (Averaging reduces variance) and leaves Bias like it is
- Better performance than a single tree

Disadvantages:

- Black box
- Can lead to high bias
- Computationally expensive

Random Forest:

Random Forest uses the same method than bagging (bootstrap sampling, it trains models in parallel and aggregates all models at the end) but it adds extra randomness when training each tree. Instead of considering all variables to perform the split in the decision tree, the model will select randomly a sample of features and only those will be consider for the split. This increases the model's performance by reducing the correlation between trees.

Improving performance:

Random forest is an easy model in the sense that it has just a couple of hyperarameters to tune. So even for beginners, it can yield good results without tunning it properly.

In R, this are the most common hyperparameters:

Ntree: Is the number of trees the model will evaluate. Usually, it will not overfit even if it has more than necessary trees but it can take way longer.

Mtry: number of variables randomly selected for each split. Controls how much variability or randomness goes into the model

Sampsize: is the number of samples to train on the model

These two control the complexity of the trees.

Nodesize: minimum number of samples in the terminal node.

maxnode: max number of terminal nodes.

Selection criteria:

- Selection: Majority votes for classification or average
- Regression: average

Advantages:

- Classification and regression
- Handles missing and categorical
- Doesn't over fit as much as a simple tree
- Handles large datasets
- Easy to tune (good for beginners)

Disadvantages:

- Regression is not as good as classification
- Black box

Adaboost:

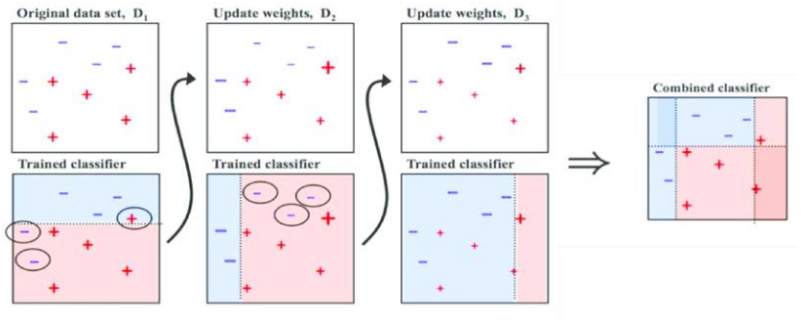
Residual = difference actual value and predicted value

This model uses the *Forward Stagewise Additive Modeling* procedure, which is called Boosting too. *That basically means* that models will be built sequentially, one after the other, rather than

all in parallel like in Random Forest and that each 'new model is influence by the performance of those previously built' (Hamoud & Mahmoud, 2009).

Adaboost works by giving all features an equal weight at the beginning but then, after the first model is trained, the error residuals will be assigned higher weights and the correct classification will have lower weights. The tree won't fully grow, usually has one node and two leafs, like the one below and it's called a Stumb.

Then, each new model will be trained mainly on the residuals (because they have higher weight). This weighted residuals is what allows the next model to improve because it focuses on predicting those *hard* data points that previous models didn't predict accurately. Depending on how much the performance improved, the model itself will be given a weight (Ihler, 2020).



https://medium.com/swlh/boosting-and-bagging-explained-with-examples-5353a36eb78d

By the end, the final model will have learned how to predict both the 'easy' and the 'hard' data points (residuals of previous models), allowing it to achieve a much higher performance than a single model.

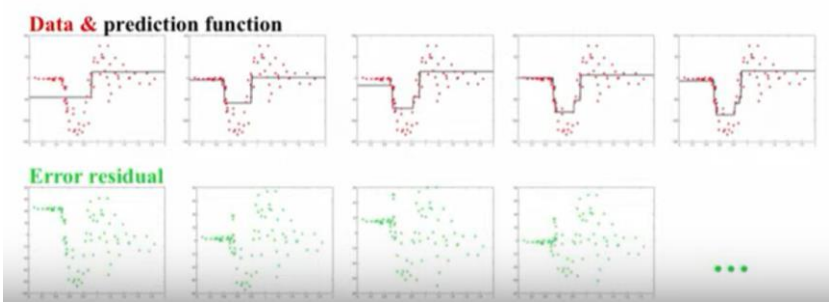
The final prediction will be the weighted sum of the predictions made by all models, and of course, the relevance of each model depends on the weight given before (Starmer, 2020).

Gradient Boosting Machine:

As the name implies, this model uses both boosting and gradient descent to yield the final predictions for either regression or classification problems. This means that GBM fits the first

model with a general decision tree and the next models will be influenced by the performance of the previous, just like AdaBoost. The difference comes when it fits the next model. The subsequent model is fitted on the gradients of the residual instead of the weighted residuals.

Like can be seen on the image below, in this case the residuals were not given more weith like in the other algorithm. In this case, to get to those residuals the algorithm applies a different method so the next model try to fit these residuals as they are.



https://learnersdictionary.com/qa/Some-Uses-of-in-and-on-with-Pictures

GBM is using a gradient decent on a cost function to minimize each residual. In other words, it is 'converting the Tree fitting in to an optimization problem' (Grover, 2020) to find the optimal solution to minimize that residual which will help as input for the next model to make better predictions.

Improving performance:

This model has a wide range of hyperparameters to tune:

Shared many with all the tree-based models: Like min obs in terminal node, observations in each tree, etc.

Additional to those are:

Shrinkage or Learning rate: It is used to reduce the impact of the additional trees added to the ensemble. Given that the model is looking for the optimal value of the residuals, if it makes a mistake and the learning rate is too high, it will have a huge impact. The downside is that the lower it is, the longer it takes to run.

Advantages:

- Flexibility, amount of hyperparameters to tune
- The cost function can be tailor to the problem at hand

Disadvantages:

- Over fits very easily
- Sensitive to extreme values and noise