3.10 Ensemble Methods

3.10.1 Intro

💡 #IDEA **Aim at estimating multiple analytical models instead of using only one.**

Multiple models can:

- *Cover different parts of the data* input space and as such complement each other's deficiencies.
- Be *sensitive to changes* in the underlying data.
  Commonly *used with decision trees*:
- Bagging.
- Boosting.
- Random forests.

3.10.2 Bagging (Bootstrap aggregating) #Bagging

1. Starts by *taking B bootstraps from the underlying sample*. A bootstrap is a sample with replacement.
2. Build a classifier for every bootstrap:
   - For **classification**, a new observation will be *classified by letting all B classifiers vote*.
   - For **regression**, the prediction is the *average of the outcome of the B models*.

The number of bootstraps B can either be fixed (e.g., 30) or tuned via an independent validation data set.

3.10.2.1 Instability

🔑 **Key element for bagging**: *instability of the analytical technique*.

> *If perturbing the data set by means of the bootstrapping procedure can alter the model constructed, then bagging will improve the accuracy*

For models that are robust with respect to the underlying data set, Bagging will not give much added value.

3.10.3 Boosting #Boosting

Estimate multiple models **using a weighted data sample**.

1. Starting from *uniform weights*.
2. Iteratively *re-weight the data according to the classification error*:

- Misclassified cases get higher weights.

💡 #IDEA : **Difficult observations should get more attention.**

> *The final ensemble model is then a weighted combination of all the individual models.*

A *popular implementation* of this is the Adaptive boosting/*Adaboost* procedure.

The number of boosting runs can be fixed or tuned using an independent validation set.

🔑 **Key advantage**: *Easy to implement*.

⤴ **Potential drawback**: *Risk of overfitting* to the hard (potentially noisy) examples in the data, which will get higher weights as the algorithm proceeds. This is *especially relevant in a fraud detection* setting because the target labels are typically quite noisy.

3.10.4 Random Forests #RandomForests

It creates a *forest of decision trees*:

1. Given a data set with n observations and N inputs.
2. m = constant chosen on beforehand.
3. For $t = 1, \ldots, T$
   a. Take a bootstrap sample with n observations.
   b. Build a decision tree whereby for each node of the tree, randomly choose m variables on which to base the splitting decision.
   c. Split on the best of this subset.
   d. Fully grow each tree without pruning.

> ***Random forests** can be used with both classification trees and regression trees.*

🔑 **Key concepts**:

- The *dissimilarity amongst the base classifiers*, which is obtained by adopting a bootstrapping procedure to select the training samples of the individual base classifiers.
- The *selection of a random subset of attributes at each node*.
- The *strength of the individual base models*.

The **diversity of the base classifiers** creates an ensemble that is superior in performance compared to the single models.

3.10.4 Evaluating Ensemble Methods

Various benchmarking studies have shown that *random forests can achieve excellent predictive performance*:

- They rank amongst the **best performing models** across a wide variety of prediction tasks .
- They can *deal with data sets having only a few observations*, but **with lots of variables**.
- They are **highly recommended when high performing analytical methods are needed for fraud detection**.
- They are a *black-box models*.
  - Due to the multitude of decision trees that make up the ensemble, it is very hard to see how the final classification is made.

3.10.4.1 Variable Importance #VI

One way to *shed some light on the internal workings of an ensemble* is by **calculating the variable importance (VI)**.

A popular procedure to do so is as follows:

1. *Permute the values of the variable* under consideration on the validation or test set.
2. *For each tree, calculate the VI value*, the **difference between the error on the original, unpermuted data, and the error on the permuted data**.
   - In a *regression setting, the error can be the MSE*, whereas in a classification setting, the error can be the misclassification rate.
3. *Order all variables according to their VI value*. **The variable with the highest VI value is the most important**.

Next chapter: Evaluating a Fraud Detection Model