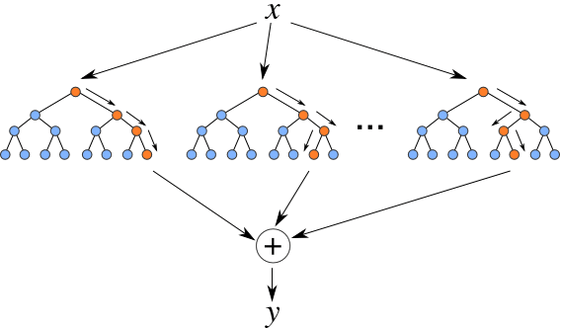
**Model Introduction**

Our mission is to help banks to analyze which groups may go to buy products, so this task is a classification task. On the basis of the decision tree, our group adopted the random forest model which has stronger generalization capabilities and credibility.

First, briefly introduce the decision tree. It is a supervised learning algorithm based on if-the rule and the tree structure. Each internal node in the tree represents a judgment on an attribute, each branch represents the output of a judgment result, and the last leaf node represents a classification result. Then we can get the result of each type of customer.

Random forests are made of many decision trees, and there is no association between different decision trees.



When we conduct classified tasks, new input samples enter, let each decision tree in the forest make judgment and classification, each decision tree, will get its own classification result, which is classified in the decision tree The most, then the random forest will treat this result as the final result.

The first is the process of two random sampling, random forest to the input data to be sampled in rows and columns. For row sampling, the replacement is used. Assuming that the input samples are N, then the samples sampled are also N. The selected N samples are used to train a decision tree as the samples at the root node of the decision tree, while making the input samples of each tree are not all samples during training, making it relatively less prone to over-fitting.

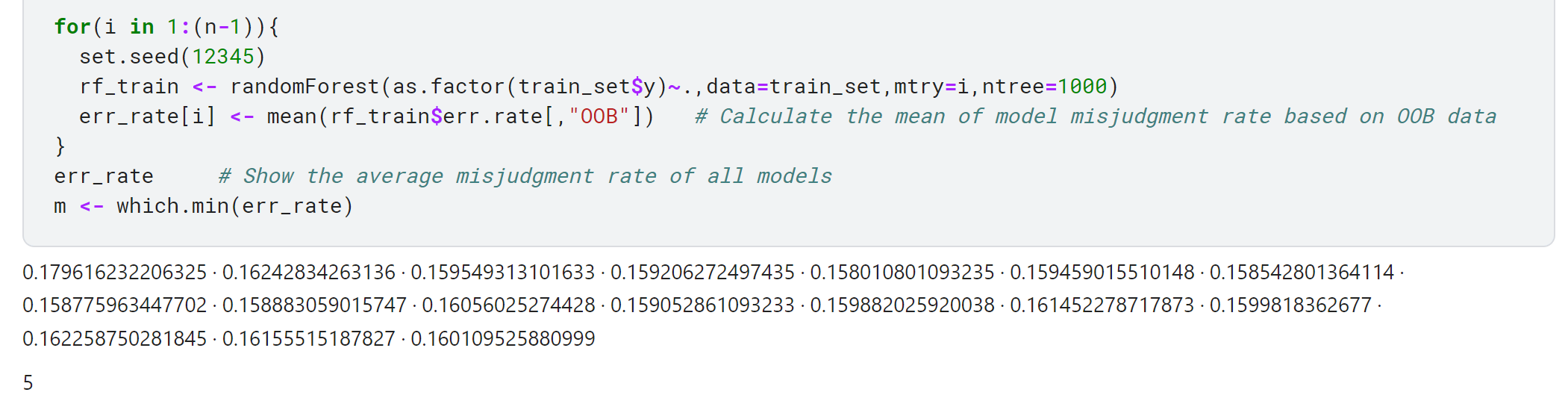
**Why should use random sampling with replacement?**

If random sampling is not performed, the training set of each tree is the same, then the result of the classification of the finally trained tree is exactly the same. In this case, there is no need for bagging. On the other hand, the final classification of the random forest depends on the voting of multiple trees, which should seek the same part, the result that is most likely to represent the model.

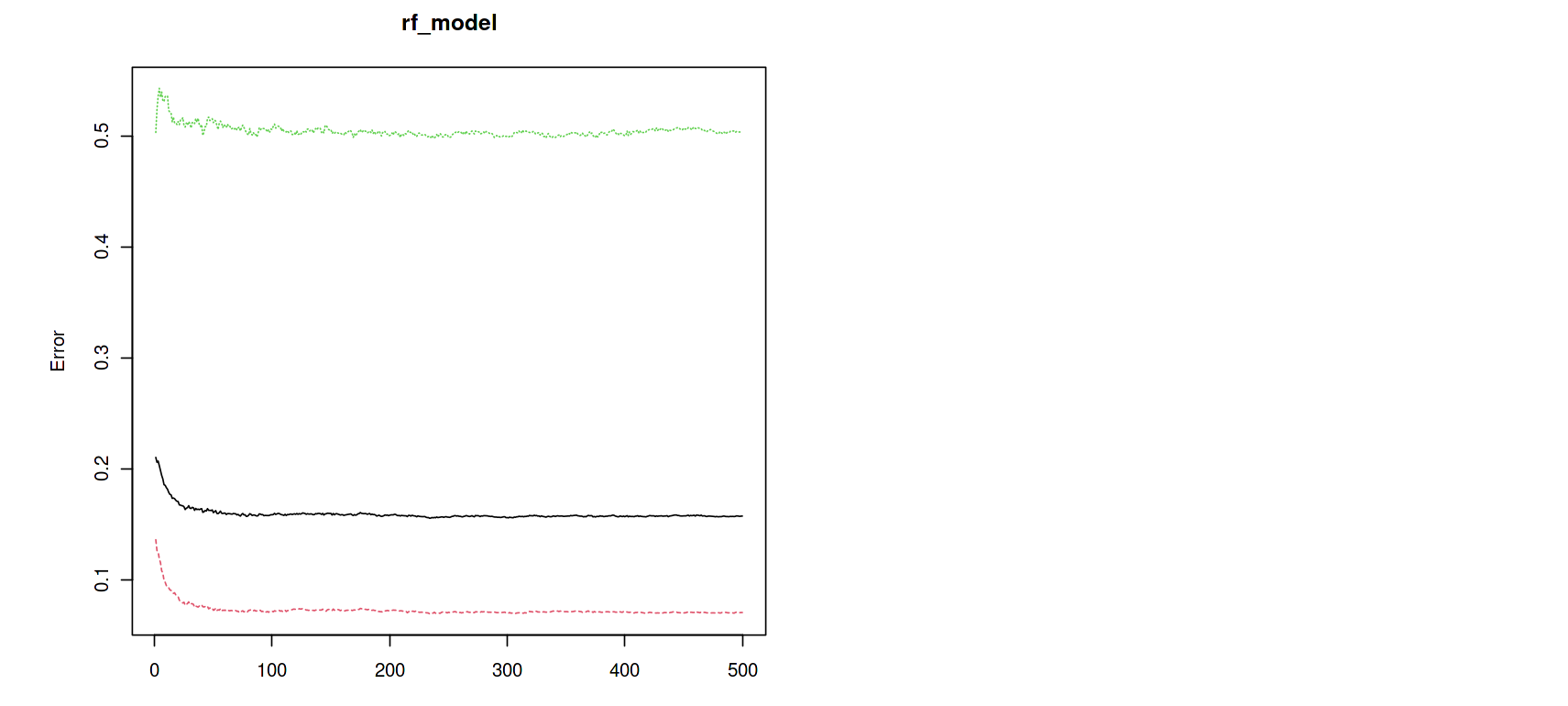
For column sampling, this data sample has 18 attributes. When each node of the decision tree needs to be split, m attributes are randomly selected from these 18 attributes, satisfying the condition . Then a certain strategy (such as information gain, Gini coefficient) is adopted from these m attributes to select 1 attribute as the splitting attribute of this node. In the library of R language, it uses a parameter to choose the number of branch points.

**Fine Tunning**

1. Choose the best number of branch points , we can use the grid search to find the best result which means that using a loop to traverse the number from 1 to the feature numbers, and base on the lowest error to choose the amount.



1. The number of decision tree in the “forest”, we can just plot the model and it will show the three cases of convergency. We can find the after a certain number of trees, the model almost stops converging, so we can just keep the default parameter settings. Too much will cost computing power, too little will fail to converge

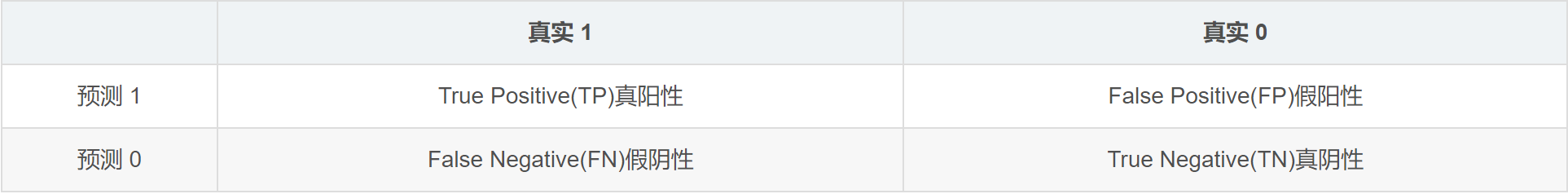


**Model Evaluation**

1. **Out-of-bag error rate**

We mentioned above that in constructing each tree, we used different random and put-back samples for the training set. Thus for each tree, about to of the training instances did not participate in the generation of the tree, they are called the samples of the kth tree. Then use the rest of sample to test the model and we can get the rate of the error.

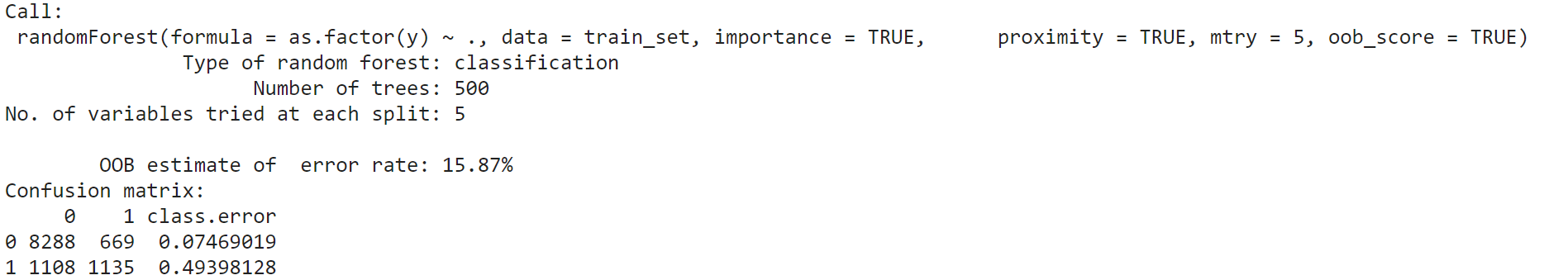
F1 Score is an indicator used in statistics to measure the accuracy of a binary classification (or multi-task binary classification) model. It takes into account both the accuracy and recall of the classification model. The F1 score can be regarded as a weighted average of model precision and recall. Its maximum value is 1 and the minimum value is 0. The larger the value, the better the model.

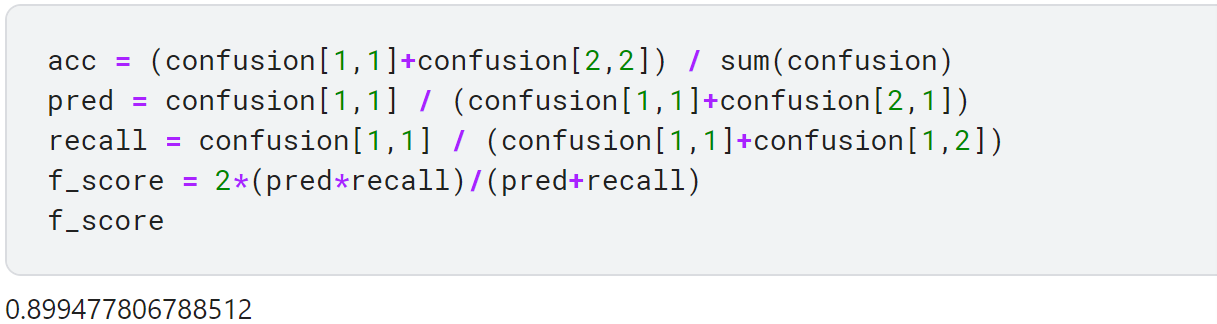


And we can calculate the by the formula:

Where:

After fine tuning, we can find the error just around 15.87% and is about 0.9.





**Optimal solution**

**Promotion time:**

1. Contact customers every three days or twice a week
2. Peddle during the holidays

**Target groups:**

1. Peddle to middle-aged people
2. Administrators and entrepreneurs
3. People who have university degree
4. Married people
5. Non-explanatory factors

The same principal component analysis is performed on the customer data, and among the obtained 6 PC features, the 3 most important ones are selected for auxiliary judgment. If customers’ and 0.6613, then we can filter it out.