Machine Learning In Bank Marketing

Shikan Zheng, Zhi Du, Yihan Cao, and Hui Jiao, Department of MSCS, Marquette University

**Abstract**

One of the most important goals of the financial marketer is to find a connection between customer service experience and expected outcome which is based on predictive analysis. In recent years, more and more analytical techniques have been developed to solve these issues which specifically focused on the predictive analysis of customer side. Machine learning is the study of computational methods to automate the process of knowledge acquisition from examples. [1] Machine learning, as a fast-developing technique, has been used in business data mining for a quite long time. Data analysis and pattern discovery can be achieved by using machine learning techniques and thus it can play such an important role in data mining applications. To understand the strengths and weaknesses of machine learning, the best way is to build some appropriate models in a specific field. During the model building process, we can gain a whole view and a general concept of the machine learning field. In this project, we are trying to apply machine learning method to bank marketing field. Instead of using python, one of the most popular languages to achieve machine learning benefits, we decide to use R, which is also another popular language in statistical field. The problem in this project is mainly about how to use models to predict what kind of features will influence client to make a deposit. The data we used is from UCI, a popular machine learning repository.

**Keywords:** bank marketing, machine learning, data mining, predictive analysis, data analysis.

**Introduction**

More and more marketing activities have reduced their impact on the public with time goes on. In addition, economic pressures and competition are leading a marketing manager to invest in directional campaigns with a strict and rigorous selection of contacts. Such direct campaigns can be enhanced through the use of Business Intelligence (BI) and Data Mining (DM) techniques. [2] For bankers, attracting more clients can increase the deposits, and thus can lead to more cash flow to benefit the banks. The task to predict some patterns in banking data is not a new thing. Basically, it is the bankers’ responsibilities to dig out relevant information regarding customers in order to enhance the service performance. However, with the development of technologies and emerging machine learning in recent years, widespread use of R and python has changed the form of usage of transactional data in banks. R and python have many convenient packages for users to achieve the algorithms and build the models.

Intense competition has forced bankers to look for innovative paths to attract more customers and their deposit to gain more market shares. A number of algorithms have been developed in domains, such as machine learning, statistics, and visualization, to identify or predict patterns in data. [3]

To predict which kind of customers would like to deposit their money, a predictive model is necessary to be implemented. There are many aspects in reality that consists of the factors that can determine the customers’ will. With the data from UCI, we apply several machine learning methods and algorithms including logistic regression, k-nearest-neighbor, decision tree and naïve bayes to predict some relevant decisions. Then we compare these methods and algorithms and come up with the best final model, which has a very good accuracy.

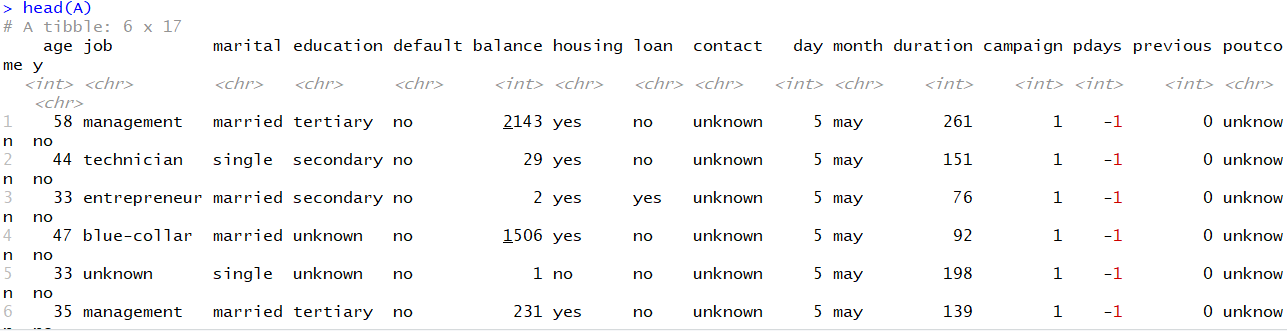
**Related work**

Because this dataset that we used from UCI contains many significant and applicable features, we believe that many projects or researches might have used it. One of the applications is that A Data-Driven Approach to Predict the Success of Bank Telemarketing, which was completed by S´ergio Moro, Paulo Cortez, Paulo Rita. [3] They proposed a data mining (DM) approach to predict the success of telemarketing calls for selling bank long-term deposits. The goal was to model the success of subscribing a long-term deposit using attributes that were known before the telemarketing call was executed. During the modeling phase, and using a semi-automated feature selection procedure, they selected a reduced set of 22 relevant features. Also, four DM models were compared: logistic regression (LR), decision trees (DT), neural networks (NN) and support vector machines (SVM).

The other usage is that Using Data Mining For Bank Direct Marketing: An Application Of The CRISP-DM Methodology. [2] They described an implementation of a DM project based on the CRISP-DM methodology. The business goal is to find a model that can explain success of a contact, i.e. if the client subscribes the deposit. Such model can increase campaign efficiency by identifying the main characteristics that affect success, helping in a better management of the available resources (e.g. human effort, phone calls, time) and selection of a high quality and affordable set of potential buying customers.

**Dataset and features**

The dataset that we used, which was extracted from the UCI Machine Learning Repository, is a dataset with over 45,000 rows and 17 inputs, ordered by date, extracted from [Moro et al., 2014] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014. The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be ('yes') or not ('no') subscribed. There are four datasets: 1) bank-additional-full.csv with all examples (41188) and 20 inputs, ordered by date (from May 2008 to November 2010), very close to the data analyzed in [Moro et al., 2014]. 2) bank-additional.csv with 10% of the examples (4119), randomly selected from 1), and 20 inputs. 3) bank-full.csv with all examples and 17 inputs, ordered by date (older version of this dataset with less inputs). 4) bank.csv with 10% of the examples and 17 inputs, randomly selected from 3 (older version of this dataset with less inputs). What we used is the bank-full.csv with all examples and 17 inputs. The features in this dataset are: age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, previous, poutcome, y. basically, these features can be sorted into two categories: numeric and factor. Because most of them are factors, logistic regression binary analysis would be the best method to predict the outcome.

Since this dataset is very good to use, we need less preprocessing to do. The head output of raw data is: .

Since we need to train out the best model after comparing and contrasting different approaches, we divide our dataset into training set and test set. Then we center and scale the dataset. After that we got 80% of training data and 20% of test data.

**Methods**

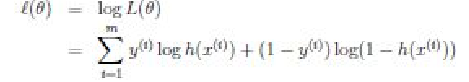
Three primary prediction models are developed in this project: k-nearest-neighbors clustering, logistic regression binary analysis, decision-trees. Each prediction model has different accuracies, but after comparison and contrast, we would use the best accuracy model. The method that we select features would be backward feature selection

1. **k-nearest-neighbors clustering**

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive [integer](https://en.wikipedia.org/wiki/Integer), typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

1. **logistic regression binary analysis**

The binomial logistic regression algorithm works by finding the theta value that maximizes the following log likelihood function:



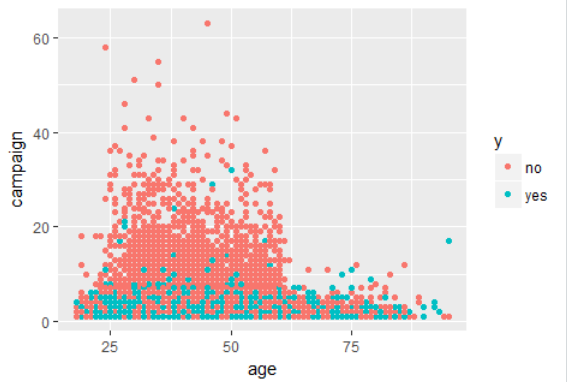
where x represents the features matrix, y represents the labels vector, and h(x) is 1/(1 xp(− x).

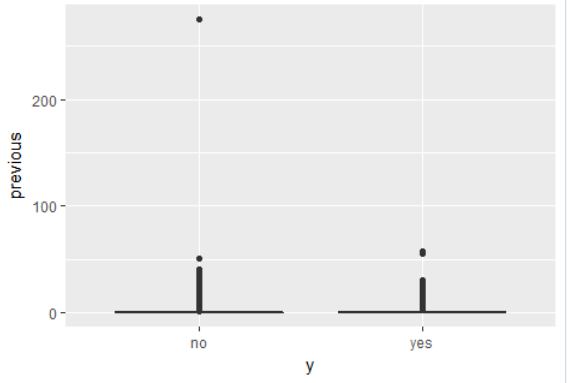
1. **decision-trees**

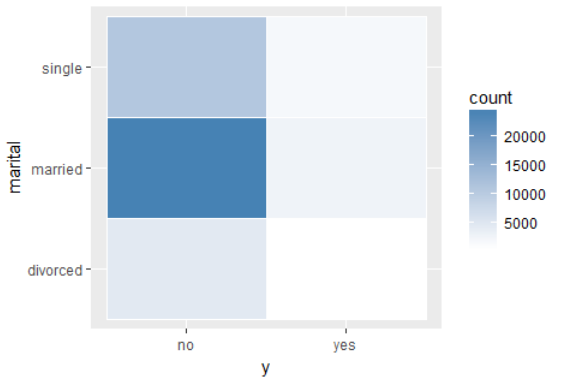
A decision tree is a [flowchart](https://en.wikipedia.org/wiki/Flowchart)-like structure in which each internal node represents a "test" on an attribute (e.g. whether a coin flip comes up heads or tails), each branch represents the outcome of the test, and each leaf node represents a class label (decision taken after computing all attributes). The paths from root to leaf represent classification rules.

In [decision analysis](https://en.wikipedia.org/wiki/Decision_analysis), a decision tree and the closely related [influence diagram](https://en.wikipedia.org/wiki/Influence_diagram) are used as a visual and analytical decision support tool, where the [expected values](https://en.wikipedia.org/wiki/Expected_value) (or [expected utility](https://en.wikipedia.org/wiki/Expected_utility)) of competing alternatives are calculated.

**Results and Discussions**

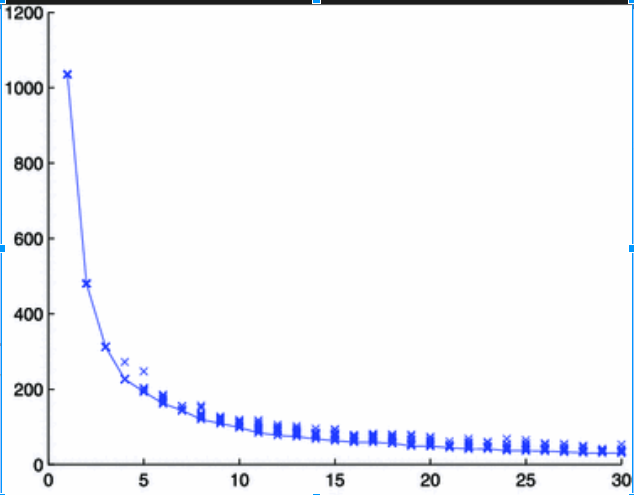




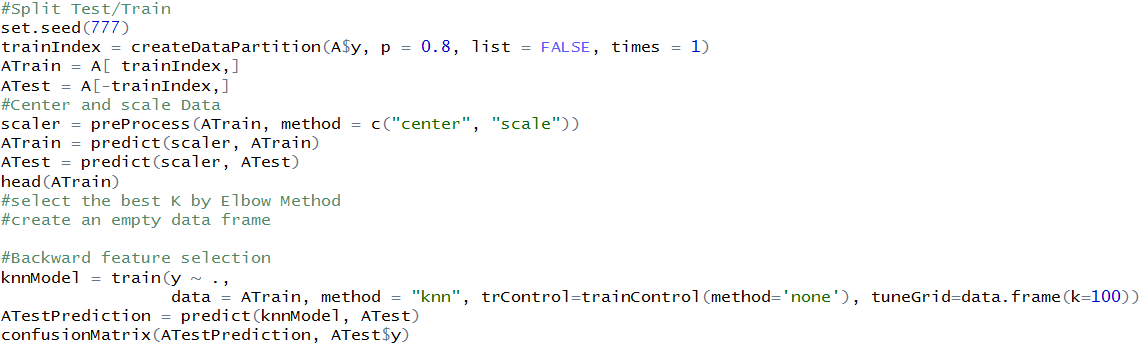


1. **Knn classification**

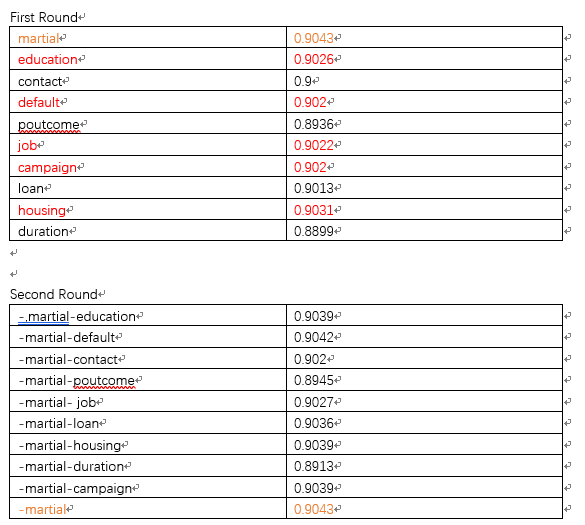
**So basically, the main reason we choose Knn is because this method is very mature and skillful, even though to decide a proper k is painful. Since bad features could casue a remarkable negetive effect, we need to get rid of them and select the best combo.In addition the optimal K value might be different in different features’ combinations, different number of K can cause different feature selecting results, so it is hard to test if a k value is the best for different group. Eventually , the only thing we can do is to infinitely approach the best prediction.**



**We can always use some PCA,K-fold CV,elbow method etc methods to cauculate the K that might approximate the best K at first. For example , in the plot, we need to choose the lowest point and build the model. I choose K=100 since the line has a downward movement,but maximum suppose to be sqrt(n),n equals the numbers of objects, too much tie(K) might cause a error and be foeced to terminate the procedure.**

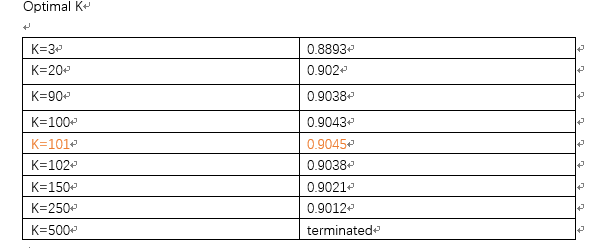


After we build the model by split training and testing data in the whole dataset with a 8:2 proportion, a backward feature selection seems to be more simple than forward.

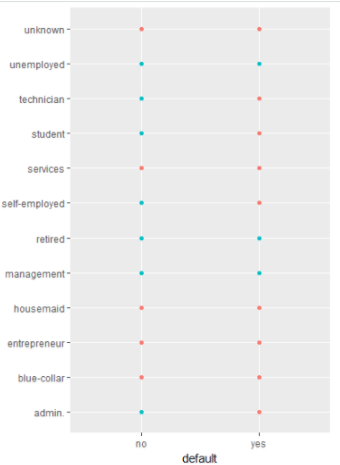


Luckily, after 2 rounds, the second round and third round’s best accuracy is still 0.9043.The best combination is 10 selected feature delete the feature“marital”.

The last step, is to find in this situation, which K is the best. We tried multiple numbers.



Finally, when k=101,the prediction has the best accuracy which is 0.945

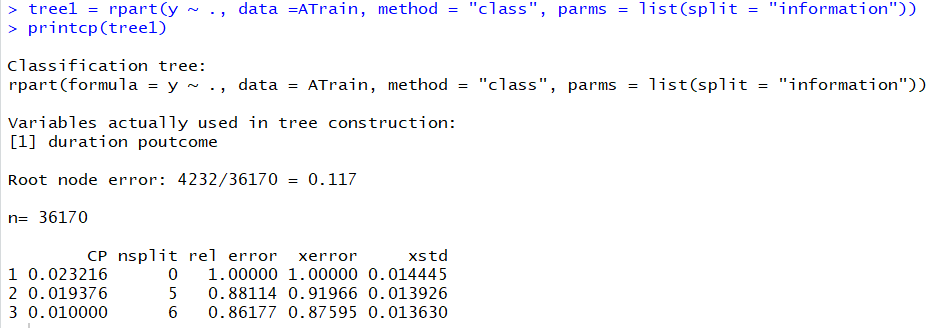


**Overall, KNN maybe not a proper method for this data set , as we see in the upper plot, the most features are factoe type with two varibles. Visually, I would not recommmand knn for this kind of dataset, but it still obtain some good preditions.**

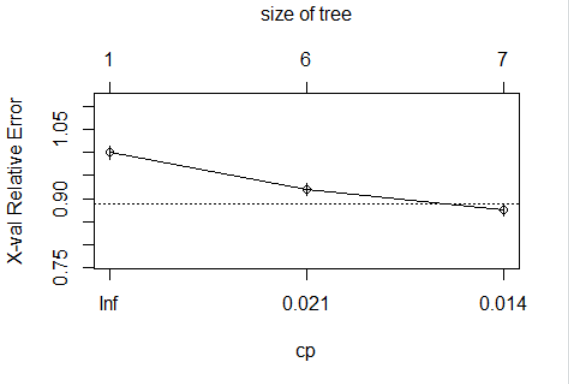
**3.decision-trees**

What we did in decision-trees is that we chose two different methods to split data which are called information gain and gini index. We learned these two methods from data mining class. Basically, these two methods return good results but information gain biased towards multivalued attributes. Besides biased to multivalued attributes, gini index also has difficulty when number of classes is large and tends to favor tests that results in equal-sized partitions and purity in both partitions.

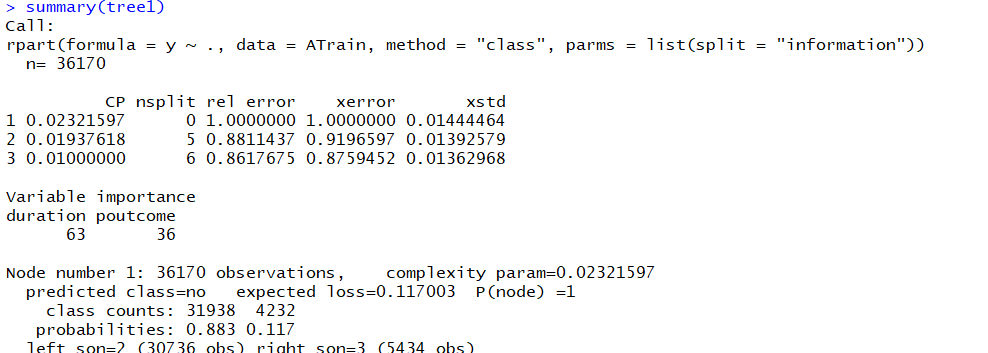
1. Information gain



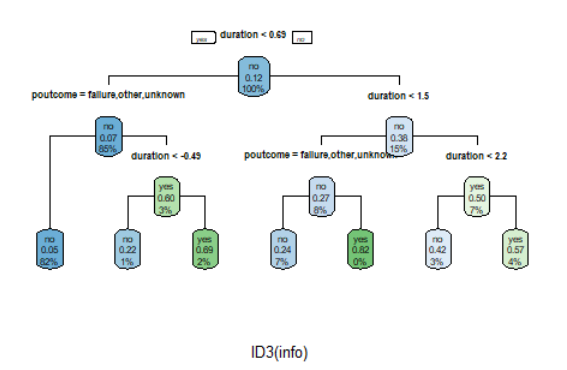
From the output of tree 1, we can easily find out that the root node error is 0.117. The number of total observations that we used is 36170. The variables we actually used in tree 1 construction are duration and poutcome. In addition, we could get some information about CP and relative errors.



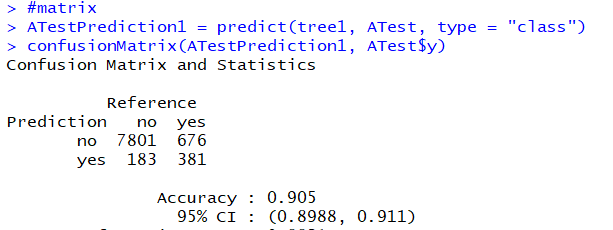
From the cross-validation plot, we can see that the best number of nodes in the lowest layer is 7. The “X-val error” value typically follows a decreasing pattern, from approximately 1.0 at the root node, then it crosses the 1standard error boundary, reaches a plateau, and decreases further when the tree gets complex.



From the summary of tree 1, we can easily get what we need to know for every node in this tree. Apparently, duration has 63 percent of importance in all variables and putcome has 36. Left son = 2 and right son = 3.

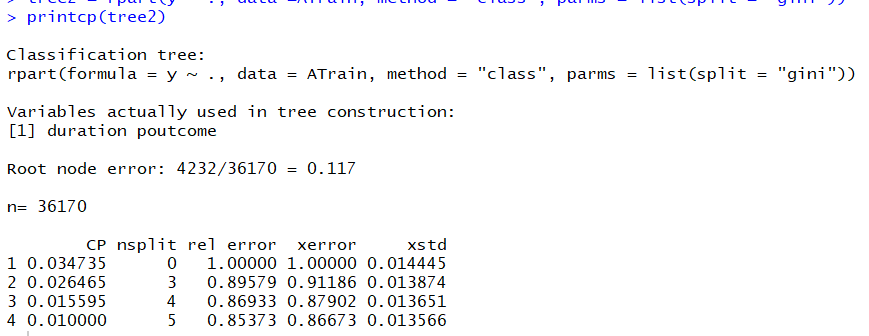


Then we got the confusion matrix of this tree.

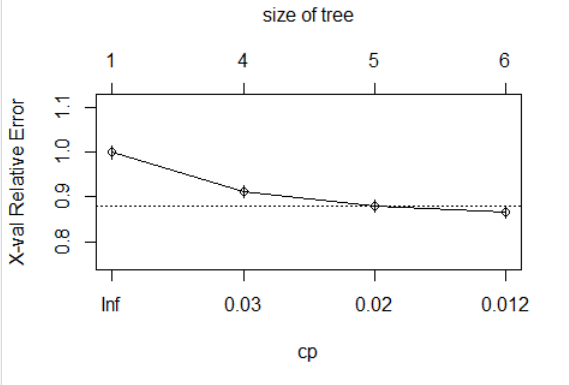


The accuracy is 0.905, which is very good in reality.

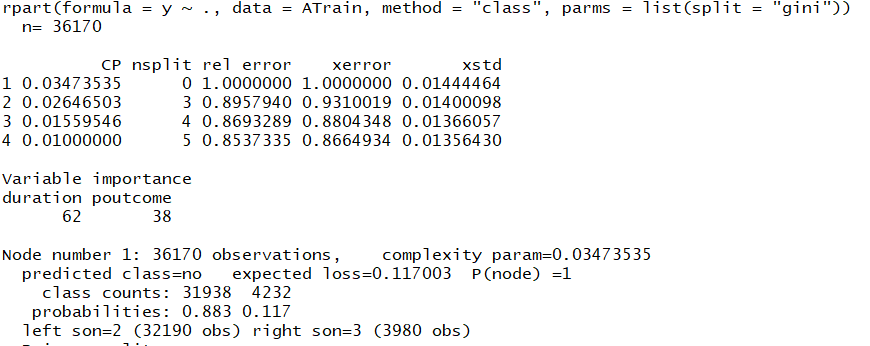
1. Gini index



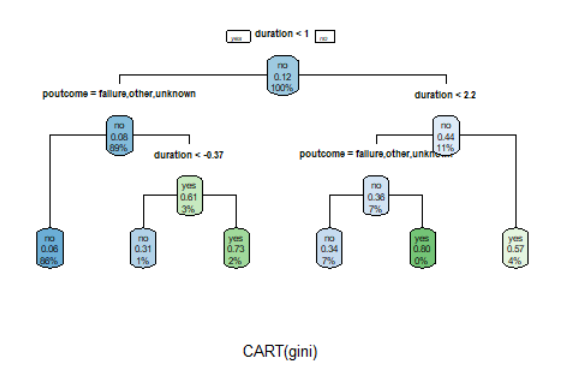
In gini index tree, we got the same root node error 0.117, and the same variables used in tree construction.



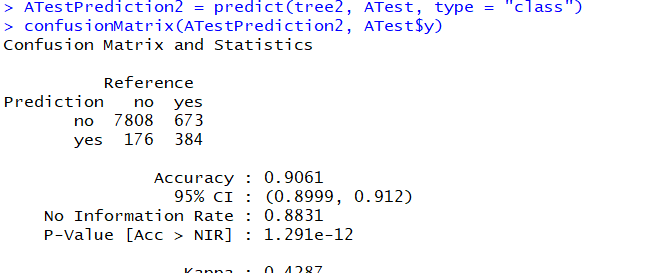
In the cross-validation plot, we can find out that there is a little difference between these two trees. In the gini index tree, we got the best number of nodes in the lowest layer is 6, which is lower than the number in the first tree.



For the summary of this tree, the importance percent of duration is lower than that in the first tree, but they are very close. This tree still has 2 left son and 3 right son.



Last, we got the confusion matrix of this tree.



We can clearly see that tree2 has improved the accuracy a little from 0.905 to 0.9061. So, compared to information gain method to split data, we prefer to use gini index to accomplish the decision tree.

**Conclusions and Future Work**

So far, we have done data preprocessing, features selection, knn-clustering and decision tree. We will be using logistic regression method to do another model in the later time and compare with all the methods and algorithms. Basically, we are expecting the LR model is the best final model because as we mentioned earlier, this dataset has too many factors as predictors and it is more suitable to use logistic regression method.

**Contributions**

Shikan Zheng - data preprocessing & feature selection

Zhi Du - k-nearest-neighbors clustering

Yihan Cao - logistic regression binary analysis

Hui Jiao - decision-trees

In fact, each team member contributed equal parts to this project. We developed every method together and discussed every model together. So, all the members tried their best.

**References:**

[1] Langley, P., & Simon, H. A. (1995). Applications of machine learning and rule induction. Communications of the ACM, 38(11), 54-64.

[2] S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al. (Eds.), Proceedings of the European Simulation and Modelling Conference - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011. EUROSIS. [bank.zip]

[3] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014