**Introduction to Data Science Coursework**

**Siying Xiang**

**16521807**

**2019.12**

**Words count: 1511**

**SECTION A: Summarization**

In this article, the historical data of a similar product of the First Bank’s “1st Platinum Deposit” is analyzed. Through applying the data to decision tree, “duration”, “poutcome”, “contacts” and “month” are evaluated as the most important features. Next, the dataset is applied to random forest, logistic regression, naïve bayes, and effectiveness of the models are evaluated. Random forest outperforms other models and is selected as the model to make prediction. Finally, implementation instruction and recommendations are given.

Firstly, the correlation between features and outcome is analyzed based on empiricism. In terms of age, 84.8% clients are 30-60 years old who have more disposable income than other age groups, enabling them to afford the products. Job is also an indicator of a client’s consuming ability. More than 90% customers with unstable income such as entrepreneurs and self-employed clients rejected the product in previous campaign. The clients who have credit or negative balance have more than 90% rate to reject the product, thus deposit is an important factor to evaluate whether a customer has the capability to purchase the product.

Secondly, all the categorical variables are casted into numerical to calculate the correlations between features in Orange. Spearman correlation is chosen to analyze these neither normal distributed nor linear dependent data. As table 1 shows, correlation between “pdays” and “poutcome” is more than 0.98, and the correlation between “pdays” and “previous” is 0.985.

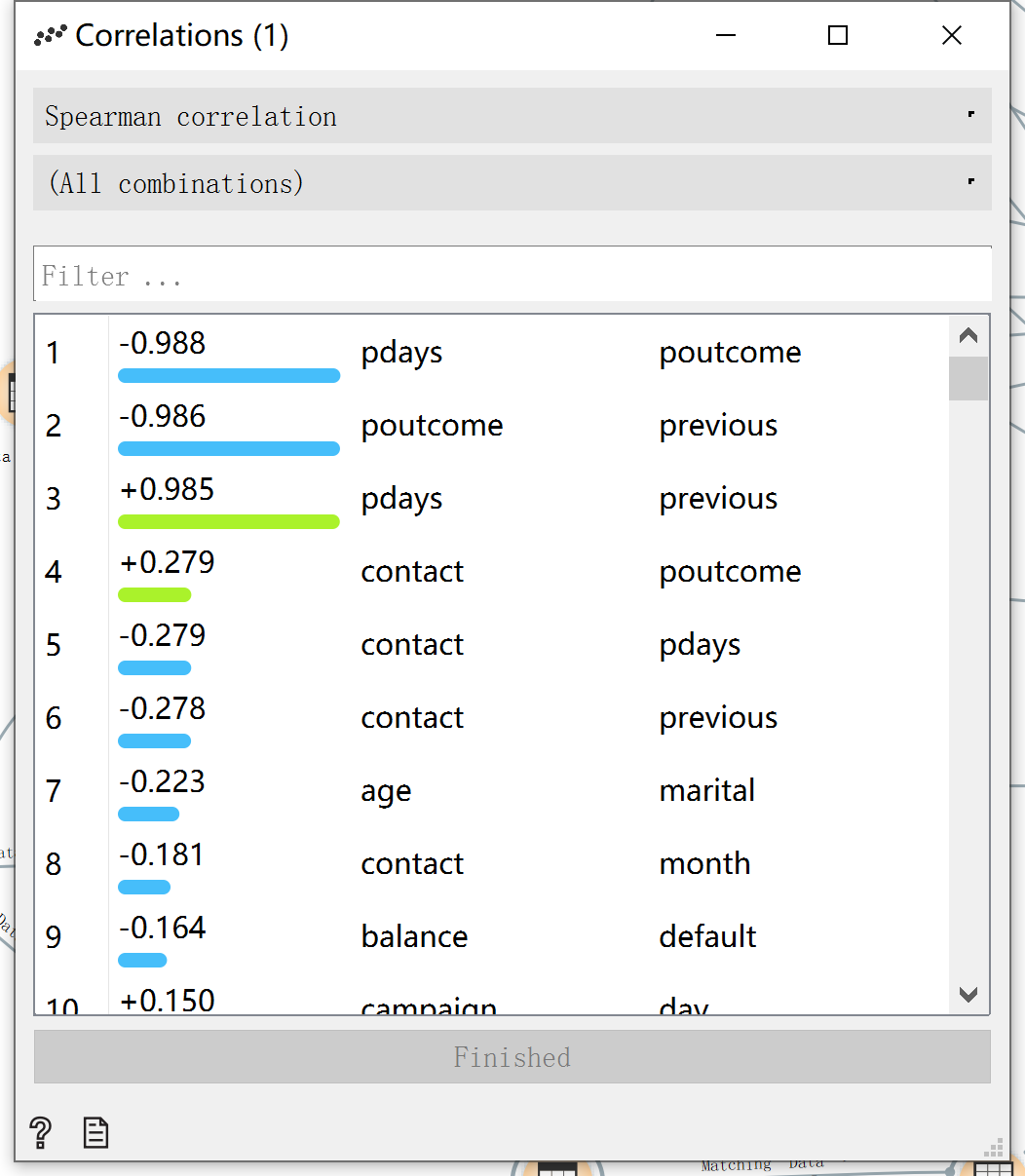


Table 1. Correlation index

**SECTION B： Exploration**

In this part, the importance of features is examined by rank and decision tree, and whether any subgroup exists is discussed. To increase the accuracy of data analysis, the outliers in “campaign” is eliminated through applying statistical adjustments in excel. Information gain shows the expected information a feature brings to prediction, while information gain ratio is more unbiased as it indicates attributes’ intrinsic information (Orange Data Mining, 2015). These two coefficients are utilized to rank the importance of features. It is shown that "duration” and “poutcome” has both the greatest information gain and information gain ratio. This infers that the customers who listen to marketing phone call patiently have higher potential to purchase products. Furthermore, customer loyalty is crucial to the bank. “Pdays” and “previous” reveal the frequency a client is contacted. Thus, a stronger connection between clients and the First Bank tends to bring the company more sales.

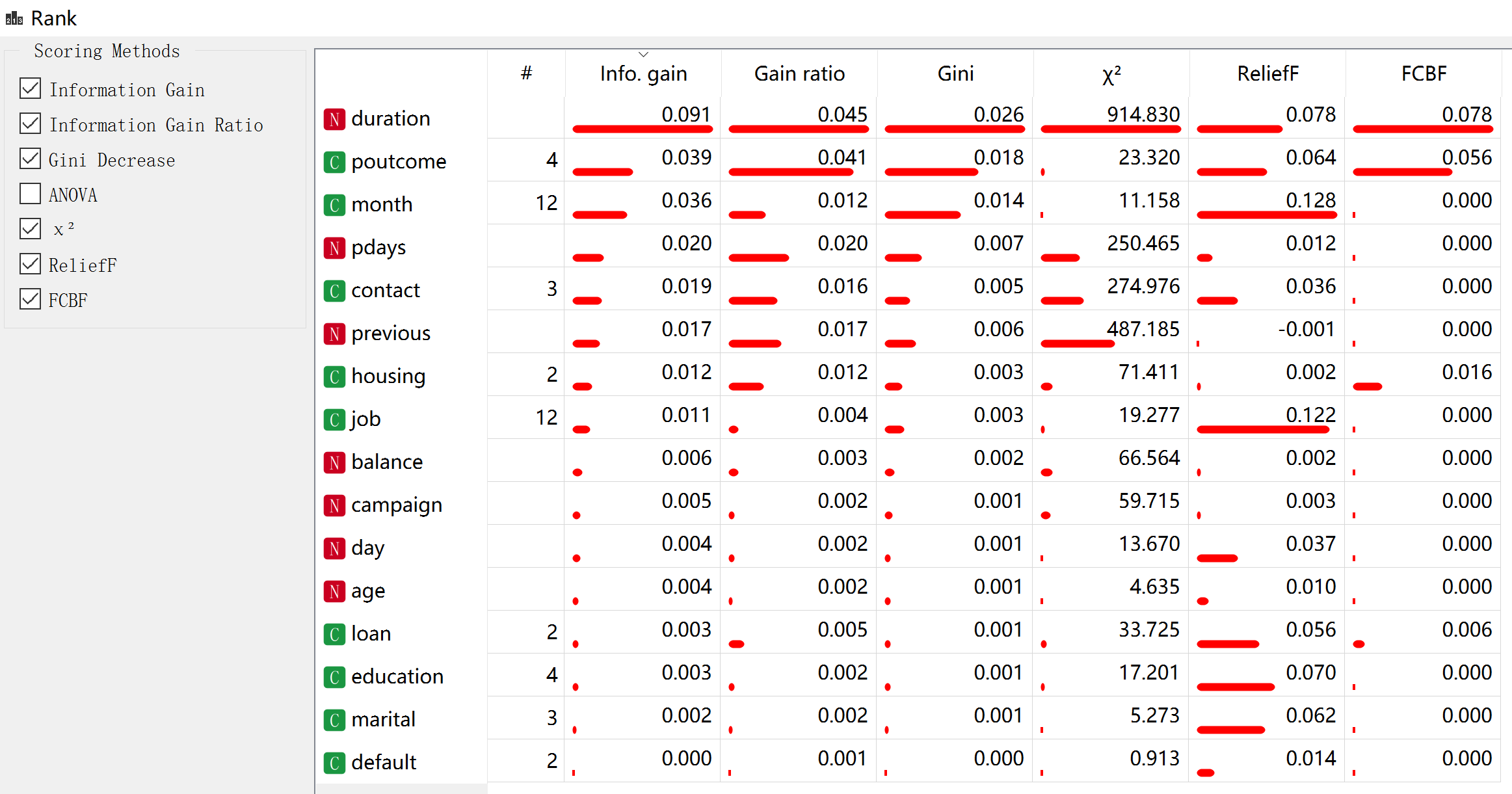


Table 2. Rank index

Next, apply the data to decision tree. Taking the feature “duration” into account, 93.5% of the clients whose duration are less than 405 seconds didn’t purchase the product. In addition, 93.2% of the clients whose contact is unknown and has duration between 402 and 613 seconds didn’t subscribe a term deposit.

However, “duration” will be skipped when making prediction to avoid the leak issue. Thus, another decision tree without “duration’ is developed to better observe subgroups. Under this condition, “poutcome” becomes the variable to separate leaves in the first layer. 89.9% of the clients whose “poutcome” are “failure, other or unknown” didn’t subscribe a deposit. Furthermore, among those clients, 91.7% of those who were contacted in August, February, January, July, June, May or November didn’t make a purchase in this campaign. Hence, “duration”, “poutcome”, “contacts” and “months” are considered as important features. However, the group of “poutcome” is “failure, other or unknown” only takes up 12% of the total dataset, which is considered not large enough to form a subgroup.

**SECTION C: Model Evaluation**

*C.1. Data Preprocess and Model Selection*

In order to avoid the leak issue, the feature “duration” is skipped when making a prediction. Additionally, “pdays” and “previous” are dropped due to their high correlation with “poutcome”. In order to better train the models, the data is balanced by oversampling instances of “y=yes” and “y=no”. 60% of the total dataset is used as training data.

Since the business case involves a classification problem and the data are discrete, random forest, Naïve Bayes, Logistic Regression and Decision Tree are applied for further analysis. K-NN model is not chosen because it is an unsupervised method, while the required dataset has a target variable.

*C.2. Models’ Parameter*

*C.2.1 Logistic Regression*

Logistic regression estimates the probability that a new instance belongs to target class. After trial, it is found that when regularization type is Lasso and strength C=0.180, the model performs the best in AUC and other evaluation coefficients.

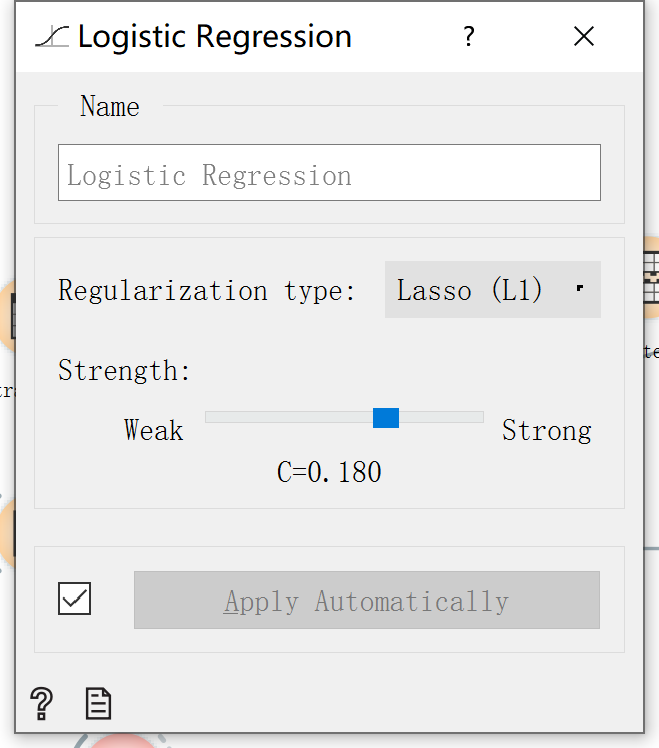


Table 3.1. Logistic Regression’s Parameter

*C.2.2 Naïve Bayes*

Assuming attributes are independent from each other, a Naïve Bayes classifier apply Bayes’ theorem on the classification of clients. Naïve bayes is efficient in terms of both storage space and computation time, non-accurate class probability estimation. However, Naïve Bayes’ parameter cannot be manually adjusted.

*C.2.3 Decision Tree*

We tuned from 1 to 60 and found 12 is the best result for the minimum number of instances in leaves. Using the same method, the decision tree is adjusted to stop split subsets at 60, limit the maximal tree depth to 50-80, and stop when majority reaches 100%.

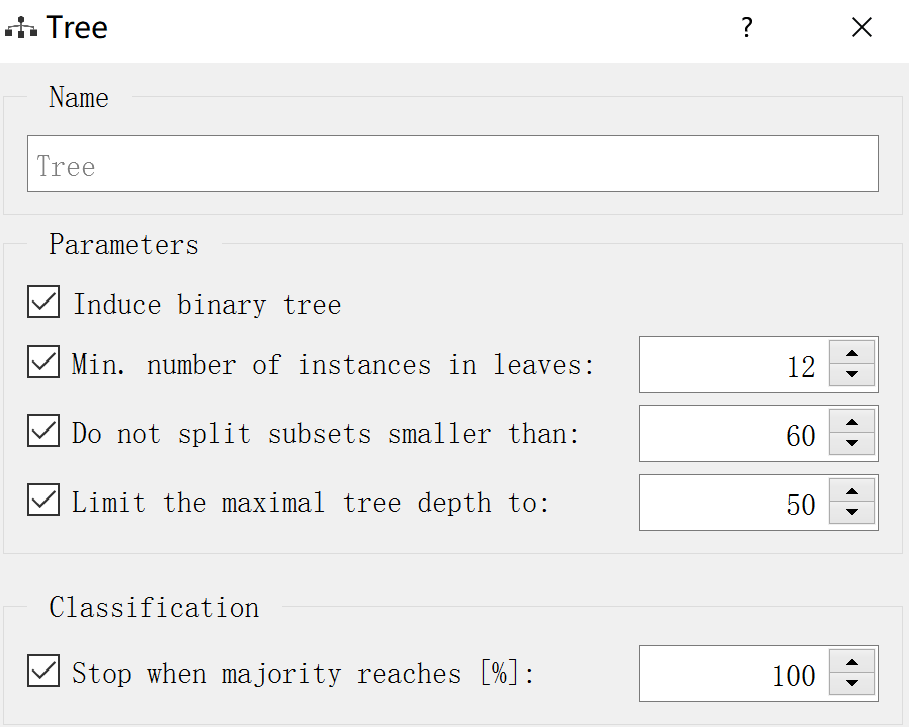


Table 3.2. Decision Tree’s Parameter

*C.2.4 Random Forest*

Using different subset of input variables, Random Forest builds multiple trees. Each tree makes a prediction and the majority class prediction wins. The performance of Random Forest is tested to be the best when the number of trees is configured into 400, and the number of attributes considered at each split is 13. In terms of growth control, the limit depth of individual trees is set to be 31 and the model is not allowed to split subsets smaller than 7.

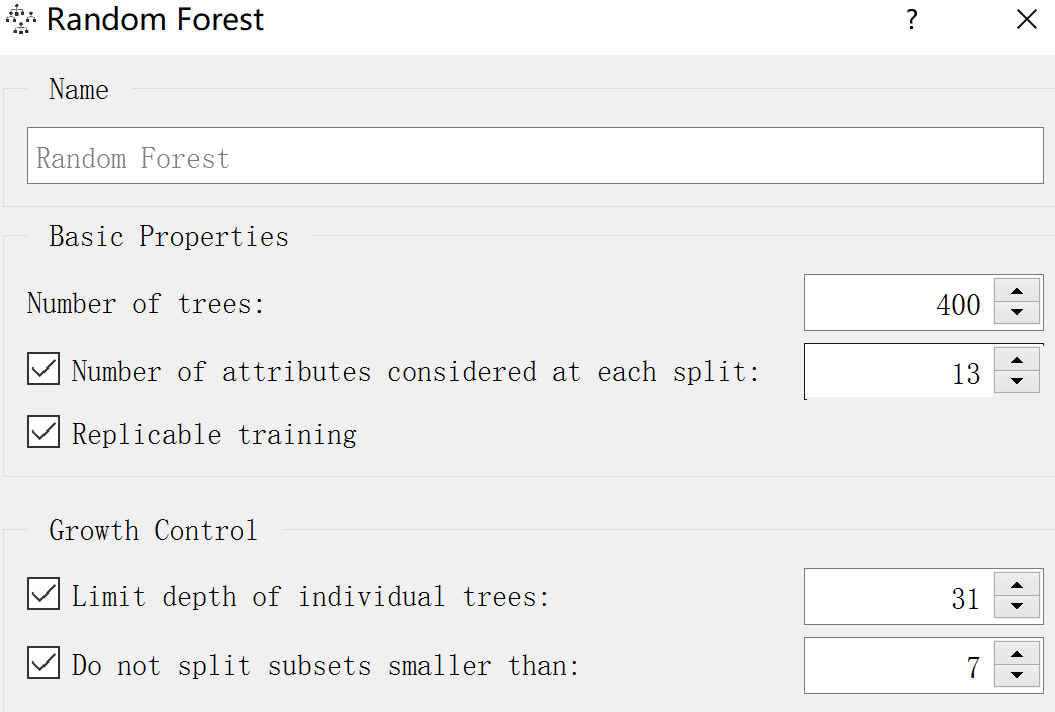


Table 3.3 Random Forest’s Parameter

*C.3 Model Evaluation*

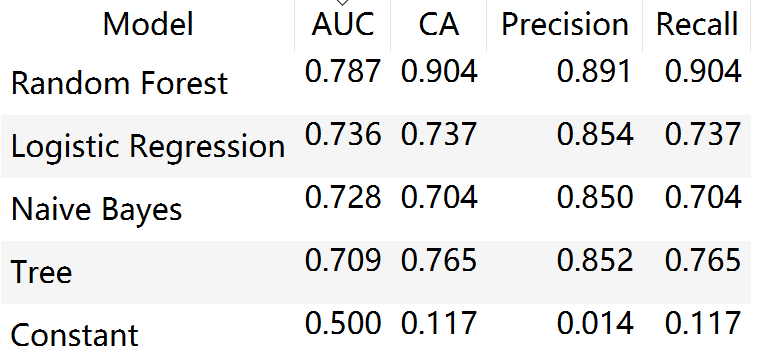


Table 4 Test & Score Evaluation Matrix

The models are evaluated by various evaluation matrix, Receiver Operating Characteristics (ROC) Curves and Lift Curve, and they are compared to the baseline performance benchmark. Area under the ROC curve (AUC) refers to the area under the ROC curve, and it illustrates the probability that a randomly chosen positive instance will be ranked ahead of a randomly chosen negative instance. AUC and CA are taken as the most important coefficients to evaluate the effectiveness of models. As the chart shows, random forest has the best performance in all coefficients.

As the Receiver Operating Curve (ROC) accommodate uncertainty through showing entire space of performance possibilities, it is used to evaluate the models’ performance. (Provost, F. and Fawcett, T., 2013). It has a false positive rate as x axis and true positive rate as y-axis. The trade-offs between benefits and costs of each model is depicted in the ROC graph (ibid.). As it is shown below, the random forest has the best performance across four models as it has the biggest AUC.

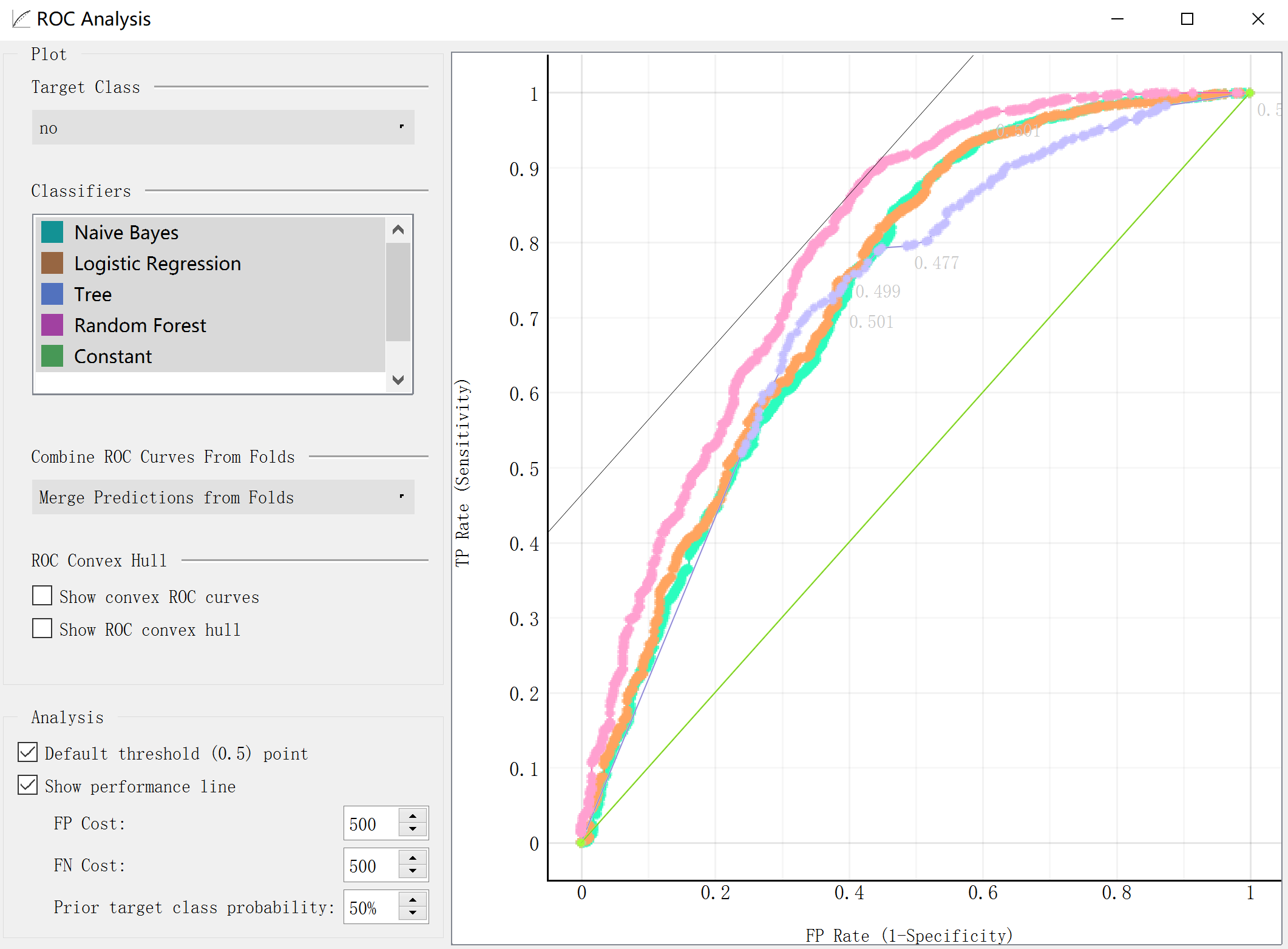


Table 5.1 ROC Curve

As Provost, F. and Fawcett (2013) stated, lifting Curve shows how much the performance of a model is lifted up over the random performance diagonal. In summary, among these four models, random forest has the best performance in AUC, accuracy, precision and recall. Moreover, it also outperforms other models in ROC and Lift Curve.

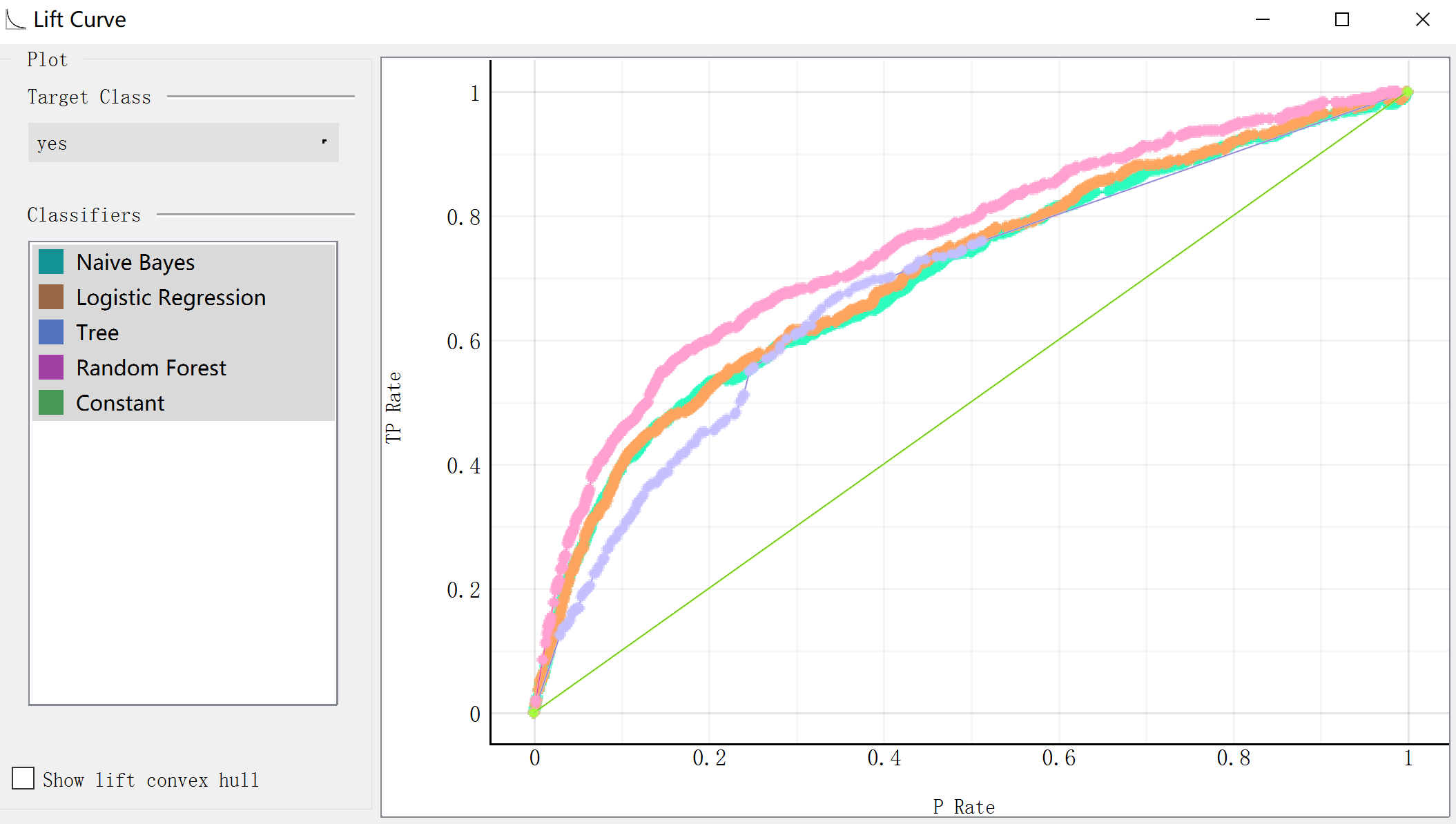


Table 5.2 Lift Curve

**Section D: Final Assessment**

Comparing the confusion matrix of random forest with other models, although the number of its true positive instances is not the greatest, it is very accurate at predicting negative instances. Therefore, it correctly predicts hundreds of negative instances more than other models, saving costs of contacting the customers that are not likely to buy products.

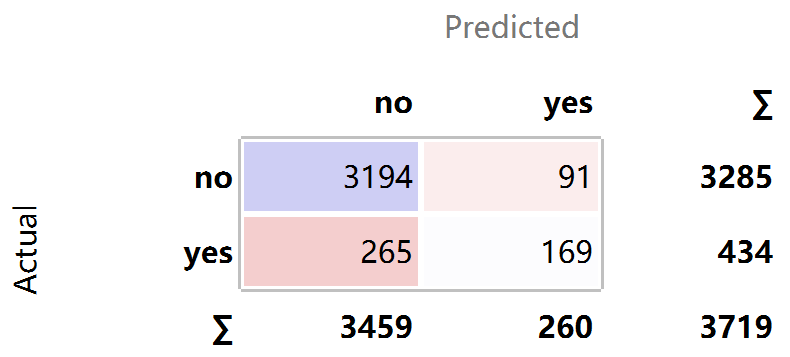
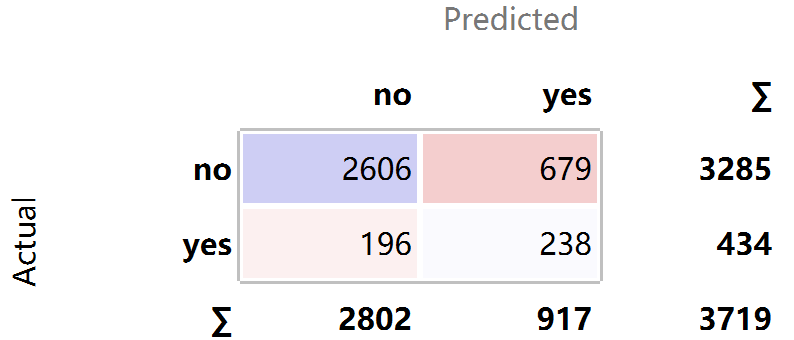
 

Table 6.1 Confusion matrix of random forest Table 6.2 Confusion matrix of decision tree

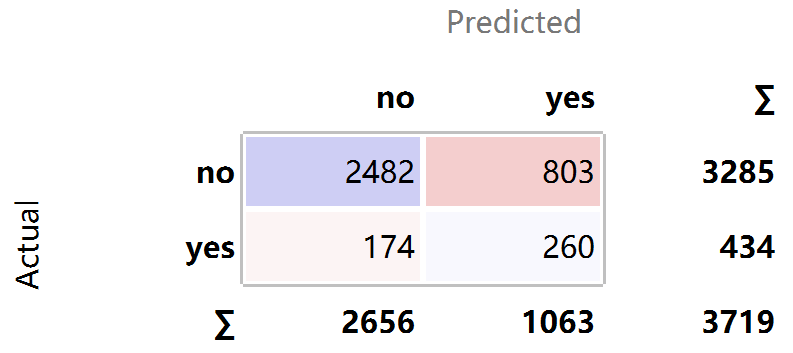
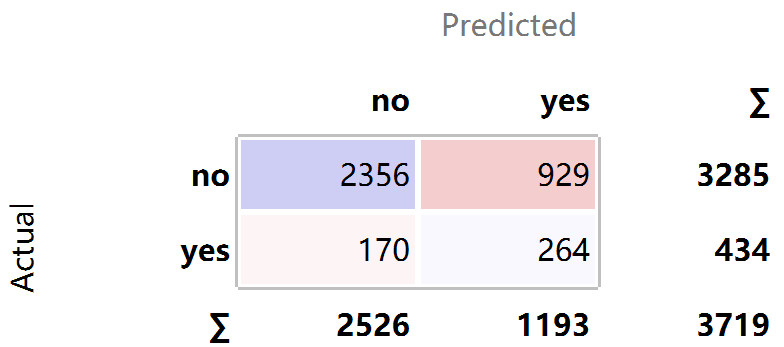
 

Table 6.3 Confusion matrix of logistic regression Table 6.4 Confusion matrix of Naïve Bayes

In addition to the outstanding performance in confusion matrix, Random Forest also has other advantages. For example, it doesn’t need data standardization, both numerical and categorical variable can be well managed. Compared with Decision Tree, Random Forest conquers the problem of outfitting. Therefore, random forest is selected to be the model making the final prediction.

**Section E: Model Implementation**

File 4 is the final trained model. The feature of pdays and poutcome should be dropped before applying the data into the model. When starting prediction, please use “file” node to add the new dataset. Select “load model” to insert the saved model. Using widget to connect both file and load model to “prediction”. Double click “prediction”, the prediction outcome of each instance and the evaluating coefficient such as AUC, CA will be displayed. A “confusion matrix” node can also be connected for further analysis.

**Section F: Business Case Recommendations**

In summary, three recommendations are provided. Firstly, pay close attention to previous campaign outcome. From empirical point of view, the cost of gaining a new customer is more than retaining an existing customer. Thus, it is crucial to maintain good relationship with customers and improve customer satisfaction. Secondly, choose a good timing to launch a campaign. Currently, although the quantity of products being sold in December, March, October and September are less than other moths, their rates of a customer purchasing products are around 50%, which is much higher that other months. Thirdly, the company is recommended to approach customers through additional marketing tools. For example, sending advertisement emails to potential clients.

Reference

Orange Data Mining (2015) *Scoring Methods*. Available at: https://orange3.readthedocs.io/projects/orange-visual-programming/widgets/data/rank.html (Accessed: 15 November 2019)

Provost, F. and Fawcett, T. (2013) *Data Science for Business.* Farnham: O'Reilly Media, Inc.