University of Nottingham Ningbo China

Faculty of Business

Academic Year 2019/20 Autumn Semester

Foundational Business Analytics (BUSI4390)

Lecturer: Dr Jenny Xiaodie PU

Students: Mingyuan Zhu
20219292

Table of contents

Section A: Introduction：

IMBA enterprise decides to sell “IMBA Platinum Deposit” and hope to predict the potential target individuals by using this telemarketing data for almost the same products. This dataset including the personal (age, job, education, etc.) and previous campaign (duration, times, etc.) information.

For telemarketing, it takes many times of cold call to a candidate which wastes time and money. To accurately identify the potential customers, this report is focused on depth analysis of previous campaigns and learning from past mistakes (a large proportion of failure) to determine what is the effective marketing campaign. We hope to convert these data into sales and increase the conversion rates as well as minimize the budget.

Who will buy the term deposit? A Term Deposit is a deposit of financial institution with a fixed higher rate of interest paid (often better than just opening deposit account but than most fixed-rate investments) in which money will be returned at a specific maturity time.
But the customer will consider this investment only for specific conditions:

- Popular for retirees: need to withdraw from their savings to pay for living expenses.
- Extremely safe and risk-free investment: appealing to conservative, low-risk investors
- End-dates to create an investment ladder
- Have enough money for an emergency.

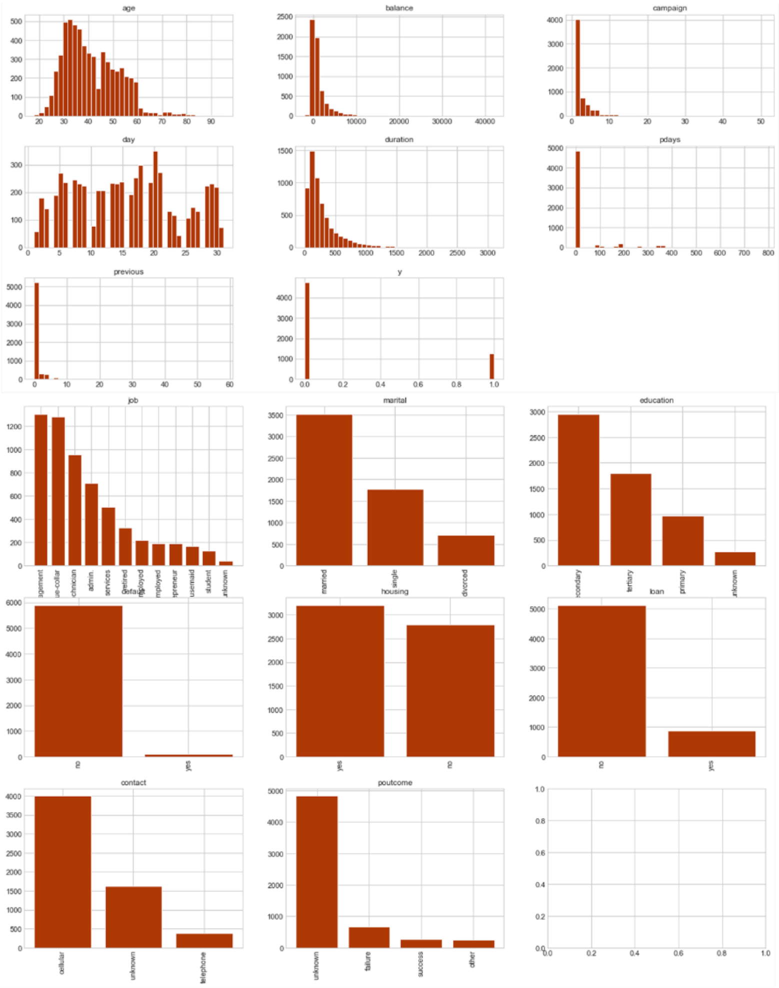
SECTION B. Summarization of telemarketing data

This part will provide summary statistical analysis of the critical points and the relationship of each input and output feature. The input data can be divided into two part: client personal data and campaign. Personal data reflect the original purchase desire of buying, and contact of the campaign show the influence of their mind(Pedregosa , 2012.).

Client personal data	Contact of the current campaign
Age, job, marital, education, default, balance, housing, loan	Contact, day, duration, campaign, pdays, previous, poutcome

Table 1: Divided input data

Here is the summary statistics analysis of bar plot and histogram for all variables.



There exists large variability between variables.

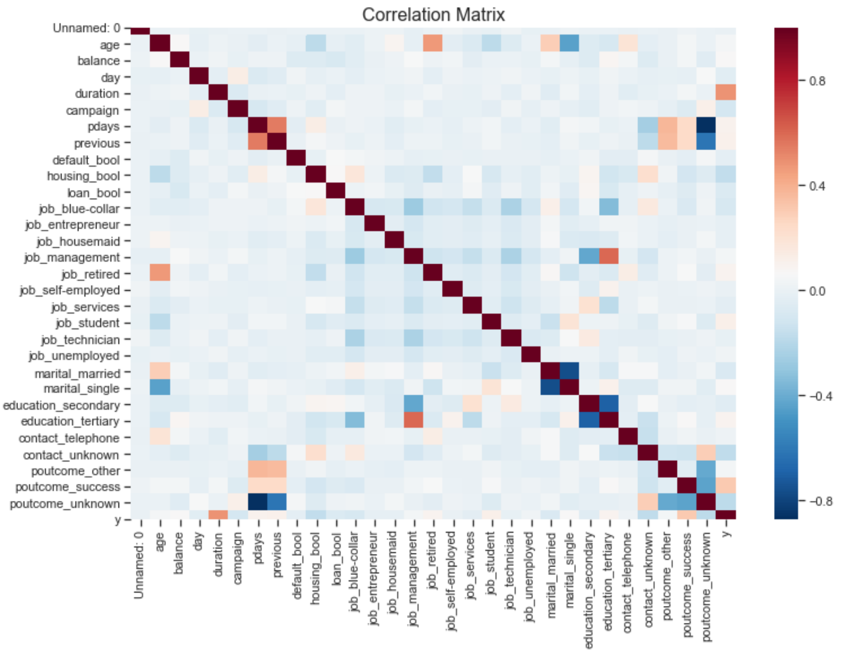
Age vary a lot from 18 to 95, and mean age is 41.

The balance, duration, and pdays are scattered in different areas across the dataset.

Management is the most common job, and most people without default.

Figure 1. Histgram and barchart for each feature

For the "Job" and 'Education' unknown values, data cleaning staps we just drop them. To evaluate the relationship among the numeric and category feature, here we transfer the category data into the format of ‘feature_group ‘+’bool_type’ as a new feature. The correlation matrix is showing the correlation coefficients between the new input and output feature. Apart from same features, absolute value of correlation is coordinate with the linear association, and for the table with the highlight colour, like the deep blue and red, it shows the potential strength and direction of the relationship for two variables.



Weak	Moderate and (0.7 < Absolute r)
Balance: marital_married, marital_single Campaign: y Previous: poutcome_other, poutcome_success default_bool: poutcome_other, poutcome_success job_X: education_X	Previous(pdays, poutcome_unknown default_bool: poutcome_unknown

Figure 2. & Table 2. Heatmap of correlation and selected high score features

The graph shows that most blocks accompanied by light colours and low values indicate those data are not strongly correlated. Table 2 selected Weak(0.3 < Absolute r <0.5) ,Moderate(0.5 < Absolute r < 0.7) and Strong(0.7 < Absolute r) variables from the correlation analysis of the heatmap. The table shows that some input features exist a high correlation between different variable, which cause an error in the model like the logistic regression. Education always plays a role in the job, as the manager always has higher education and lower education levels accompany manual workers. Although the Duration highly affects the output target, there is a weak correlation score with y for it just represents the simple linear relationship. Further analysis will combine the results of this part with prior knowledge and more scientific methods like the decision tree.

Section C: Exploration

This part shows the splitting decided for decision trees to identify the essential variables. We then combine related features to identify the sub-population based on the conclusion of correlation analysis and decision tree.

Here we measure the three different decision trees' methods: information gain, gain ratio and Gini index. Decision trees are the supervised learning white-box models of classification and regression by creating 'if' situation model and studying simple decision rules inferred from features. After computing three indexes, we ranked all feature and picked up the highest score one as our first-cut feature, which indicates the variable contributes mostly for y output.

Information gain	Unnamed: 0, balance,duration,pdays,poutcome
Gain ratio	Poutcome, pdays, Unnamed: 0, balance, duration
Gini index	Age, balance, day, duration, campaign

Table 3: Top 5 higher score for three decision tree indices (rank score form high to low)

This table shows that different methods have different outcomes, but the first top5 variances are all the same. It could be explained by the imbalance data feature- for the success sell makes up a small proportion and imbalance distribution of input variance for the figure 1. The imbalance data will affect the right decision select of the indexes, and decision tree learners create biased trees if some classes dominate. It is therefore recommended to balance the dataset prior to fitting with the imbalance data cross validation. Hence, we choose the optimization Gini index's values to run and visualize the decision trees by scikit-learn. The threes figure shows that "poutcome_success" is the crucial factor for the first cut, which can divide the population into roughly two parts.

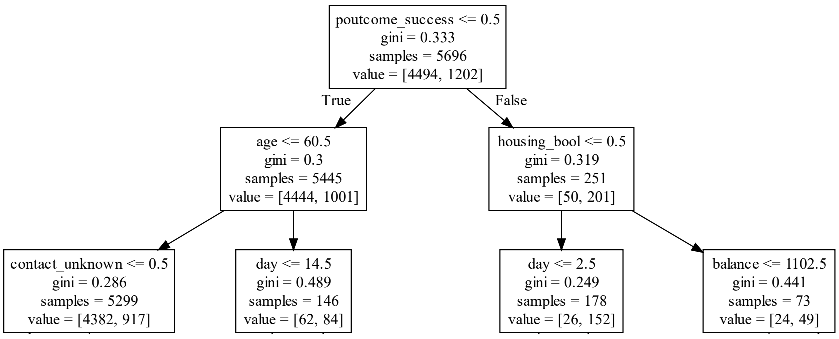


Figure 3: Visualization of CART resulting

But the real sub-group is not divided by isolated decision nor sequence of choices, we must combine and evaluate variable based both on the analysis score and the practical knowledge. We combine some most import features to identify the sub-group with buying decision- that we called customer portrait. Here we combine the variable “poutcome_success” with the “age” and “housing” and identify individuals with feature like outcome unsuccess with age less than 60 (count for 91.76 %), outcome unsuccess without house loan(53.03 %) for a large proportion. It also indicates that those variables have high correlations between each other which will bring multicollinearity during the analysis for methods like logistics regression.

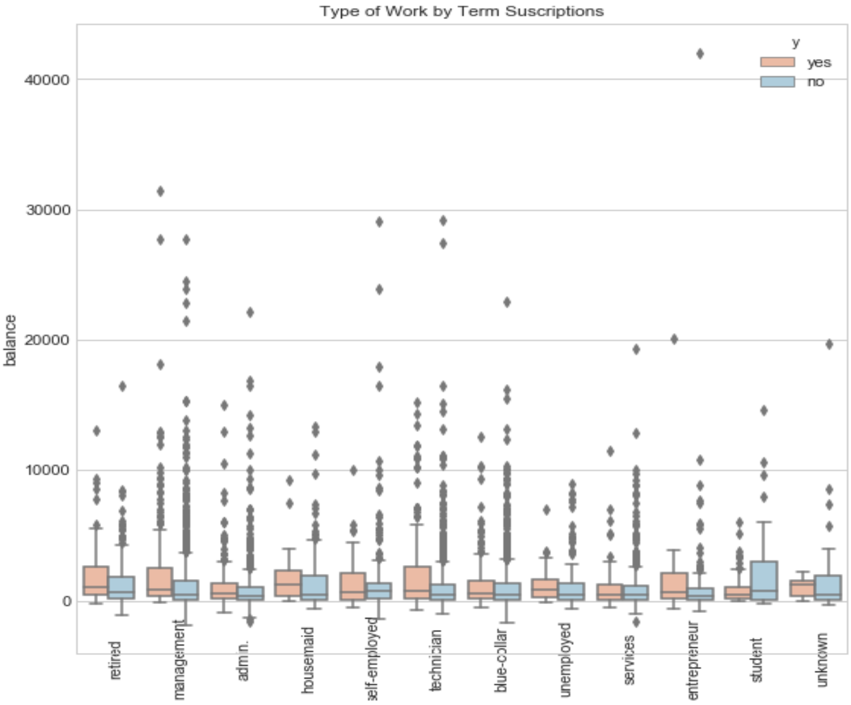


Figure 4: box plot for balance and job (group by y, job)

Some sub-population can be decided by the higher correlation variables, like the job and balance. For deposit focus on the group with a higher balance and lower risk tolerance like the retired. Figure 4 shows that levels of retired balance who choose the deposit are higher compared with the other job. Based on that, we identify the different portrait – retired job(old age), higher balance, willing to accept the deposit.

Section D: Model Evaluation, assessment and Implementation

This section picks up five machine learning models including Logistic Regression, Decision Trees, Random Forests, Naive Bayes Classificatier and nearest neighbors and evaluate it with benchmark. Here we choose the DummyClassifier from the sklearn package that makes predictions using simple rules. We compare the accuracy of classifier with random to measure errors like useless feature, not correct parameters and imbalanced data(Kulkarni, 2012.).

For the input feature, we drop 'Unnamed: 0', 'duration', 'pdays', 'campaign' and 'previous' which is irrelevant or little correlation to they(Section B & C) or would not be used in the next prediction in our training input features. Then computing the Train score, Training time, Cross-validation Mean Scores and Accuracy to evaluate the performance of each model. Train_score focus on accuracy and predict and shows the training performance. The training time will be helpful for some big data, for it calculates the total time we used. Crossval Mean score based on the cross-validation aimed at testing the model's predicting ability on new data that not used in estimating it, to avoiding selection bias or overfitting. Accuracy describes the prediction accuracy of our test data.

Classifier	**Train_score**	**Training_time**	**Crossval Mean Scores**	**Accuracy**
Decision Tree	1.000000	0.093986	0.335537	0.703468
Random Forest	0.989903	0.264965	0.609306	0.806629
Logistic Regression	0.815628	0.099571	0.766110	0.816067
Naive Bayes	0.807507	0.023526	0.276689	0.752195
Nearest neighbors	0.774363	0.774363	0.461709	0.773486
DummyClassifier	0.683333		0.210639	

Table:4 The score of each model

This table shows that the Decision Tree Classifier and Random Forest classifiers are overfitting since they both give us nearly perfect (100% and 98.9%) accuracy scores.
It can also be inferred that the Logistic Regression model is the best for its all good performance in train_score, training_time, Crossval Mean Scores, accuracy and confusion matrices(Figure 5). Above all, the Logistic Regression, as the best classifier, will be selected as our final model.

We further focus on the Recall in Precision(P-R) curve, and as we don't want to lose any potential customer, we must be more biased towards recall. It can be inferred from the Precision and Recall curve (P-R curve), we hope to maximum the recall rate as well as higher accuracy and training score(if possible). Logistic Regression performed well in this indicator. All in all, this classifier is the best model to predict whether a potential client will subscribe to a term deposit with up to 84% accuracy.

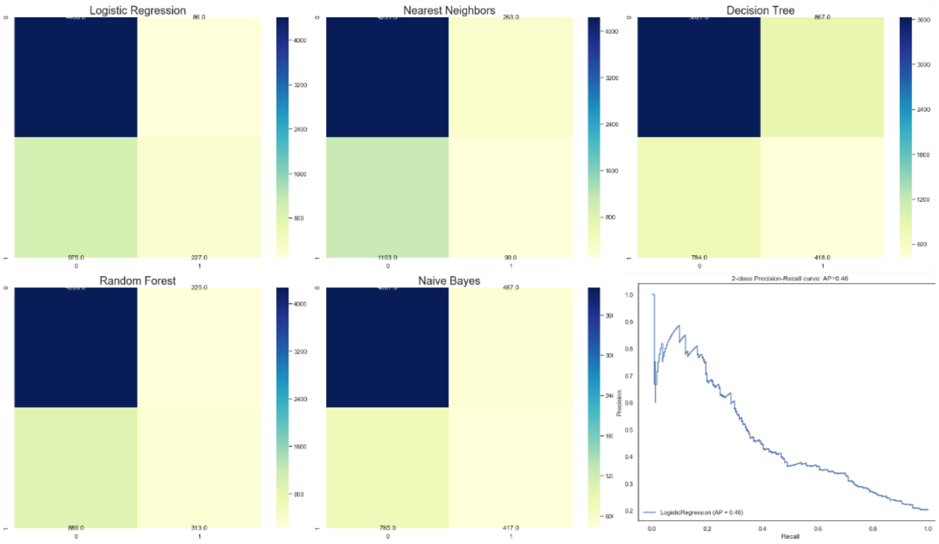


Figure 5: Confusion matrix for five model& Precision and Recall curve(P-R curve) of Logistic Regression

For Logistic Regression model, we select the top3 importance features (poutcome_success 1.540017, job_retired 0.447678 and job_student 0.406556) extract from the model function rank. We notice that for the particular group with the job of retired and students, the success deposit individuals always with more balance as figure 4 shows. It indicates that the retied and students (always related to old and young age) with higher balance would be the target individuals. For the outcomes, we notice that this score performed well in almost every count, like the correlation, decision trees and logistic regression. The results quite same as the previous analysis, for example, the people once bought are more willing to buy the product in the same company again(Atkinson, Jonathan, 2017).

At the end of the Jupiter notebook, here we provide a easy way to predict the results use our previous model, which can be used by changing the path our the file. It will predict the results by using the model that we trained before by this data. If you want to retrain the model by the steps above, you can change the beginning file path and run all the code with the same format.

Section E: Results and Recommendation

The needs and satisfaction of the customer is the key to the selling success, and individual variables will determine the process. What affects the customer to buying decision? We brief conclude it into the 1st and 2nd decision stage model. For the first buying decision, customers only know it once and made decision mostly based on their conditions (like the balance, preference, job). Furthermore, they will be affected by the campaign and advertisement. The process (Begin -> self-condition -> 1st decision -> 2nd decision ->purchase or not) is the key that we must figure out(Mitchell, 1997).

All in all, solutions and conclusion are as follows, and combining strategies and a crucial select feature in the next campaign should address and will be more effective than the current one:

1st stage
- Age: Age is a higher correlation variance. For the young (less than 25) and old (more than 60), they are more likely to subscribe to a term deposit. It's the target potential sub-population
-Occupation: students or retired would be more likely to purchase the deposit for lacking job income
-House loan and balance: They are stable correlation variables for the low balance category were more likely to have a house loan by the bank. So we focus on the group without house load or gave higher balance who have more money subscribing to a term deposit

2nd stage
- Previous contacts: Frequent calls will cause a decline in our product. Less than four times wouldn't affect our customers
- Outcome: the impression is crucial in sales for the previous long-lasting impact of like or dislike with the company. Customers who buy frequently will have a higher success rate next time

All in all, we should combine these strategies and collect more data to portray user portraits, which is crucial for us to find intent groups and business success.

Reference

Atkinson, Jonathan A et al., 2017. Combining semi-automated image analysis techniques with machine learning algorithms to accelerate large-scale genetic studies.

Kulkarni, P., 2012. Reinforcement and systemic machine learning for decision making, Piscataway, NJ : Hoboken, NJ: IEEE Press ; Wiley.

Mitchell, T.M., 1997. Machine learning, New York ; London: McGraw-Hill.

Pedregosa et al., 2012. Scikit-learn: Machine Learning in Python. , pp.Journal of Machine Learning Research (2011).