Prediction of N/LAB Financial Product

Ying Tang [1, a], Xianqi Zhou [2, b], Wenchang Li [3, c]

[1] School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, China

[2] School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, China

[3]School of Economics and Management, Jiangsu University of Science and Technology, Zhenjiang, Jiangsu, China

[a] 544624288@qq.com, [b] 385986706@qq.com, [c] js_liwenchang@126.com

Keywords: Business Analytics Reports; Data Preprocessing; Prediction Models; Confusion Matrix; Marketing Insights

Abstract: Financial product is not only the important financing tools for enterprises who sell this type of product, but also income resource for investors who purchase the product. The N/LAB enterprise aims to market a financial product called N/LAB. There are 5000 customers database who have bought similar financial products. Therefore, the dataset could be used to analyse and predict which type of customers is most possible to buy N/LAB financial products. This report starts from data pre-processing, data exploration, and then the mathematical models of prediction will be justified and selected based on pros and cons of each models. The next stage is model training and an analysis of implication. Final business recommendations and business insights will be given on the basis of previous analysis. In this report, Python and Orange will be used in terms of techniques (McKinney, 2013) [1].

Section A: Data Pre-processing and Data Exploration

Section A1: data pre-processing. Handling missing data performs an important role in data analytics, because missing values may lead to the deviation with exact results (Alvira, 2018) [2]. The first step is to check if there are any missing data and the Fig. 1 shows the number of missing data. It means zero missing values in the dataset.

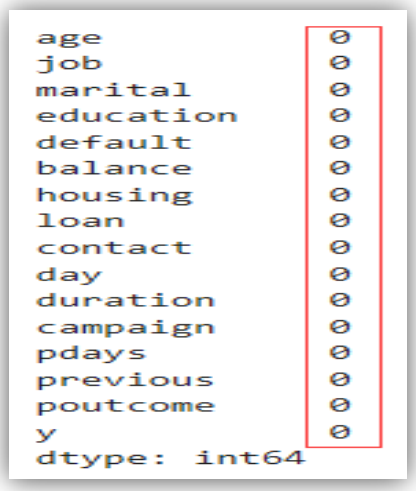


Fig 1. The number of missing data

Section A2: data summarisation. There are 16 features, consisting of 15 input features and one output feature. In terms of 15 input features, some of them are numeric values and some are discrete

values. Age, balance, days for example, are numeric values. Job martial, education are discrete values. More importantly, the last feature, final decision, will become the output feature. Firstly, it is important to visualise the number of customers who purchases the financial product, labelled as 'YES' and vice versa. It is illustrated in Fig. 2. It is obvious that most customers rejected to invest in this product (4405 customers). The number of customers who purchased this product is only 595.

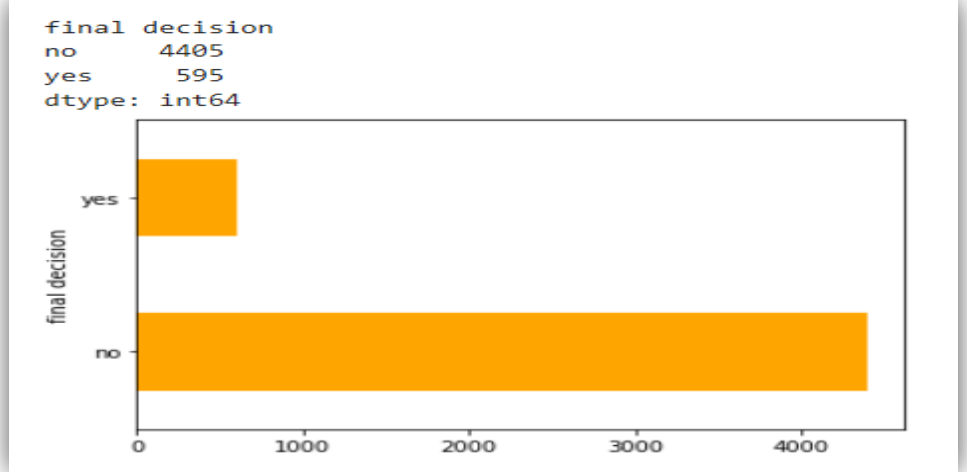


Fig 2. The ratio of final decision

Section A3: feature engineering. Feature engineering, which uses input data to create output, is vital step for machine learning (Emre, 2019) [3]. Specially, it is suggested to examine which types of input features are direct factors that encourage customers to reject the product. The Fig. 3 shows the comparison for two types of final decision in terms of each feature.

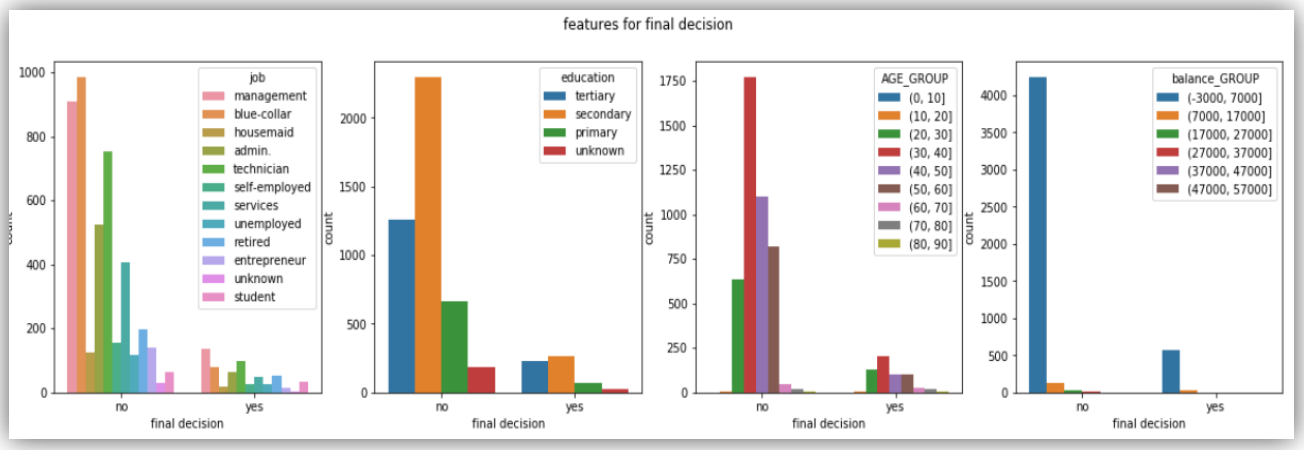
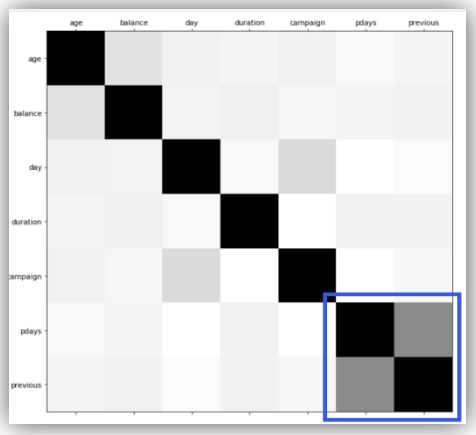


Fig 3. The ratio of each feature for final decision

Firstly, for the job feature, the management, blue-collar and technician has the most probabilities of rejecting to buy the product. Therefore, marketing department should avoid to market these types of customers. For education feature, the tertiary and secondary have the most percentage of customers who are unlikely to buy the product. In terms of age, customers aged 30-40 and 40-50 may be unwilling to invest. For balance feature, most of customers only have -3000 to 7000 deposits regardless of rejection and acceptance of the product. And customers who rejected to purchase the product take account of most part. Therefore, the enterprise should focus on other balance group. Similarly, for default, duration, campaign, pdays, loan feature, the largest percentage in each feature should be not focused. Instead, the rest of part are targeted customers. For instance,

customers who have last contact duration between 500 and 1000 in seconds, and especially for customers who did not hold on housing may be worthy to paid attention. For other features, martial and day feature are similar as job and education. Almost most of customers refuse to invest. Therefore, what marketing department should concentrate on is small part of each feature, such as divorced customers and customers contacted between 24th and 30th.

In addition to data exploration, data correlation is also an essential stage. the correlation graph between each feature is visualised in Fig. 4.



	age	balance	day	duration	campaign	pdays	previous
age	1.000000	0.116662	0.012360	-0.009007	0.019590	-0.046664	-0.012246
balance	0.116662	1.000000	0.003314	0.039377	-0.016964	-0.003679	0.013980
day	0.012360	0.003314	1.000000	-0.037732	0.164897	-0.101644	-0.065875
duration	-0.009007	0.039377	-0.037732	1.000000	-0.097653	0.018035	0.017125
campaign	0.019590	-0.016964	0.164897	-0.097653	1.000000	-0.101282	-0.028284
pdays	-0.046664	-0.003679	-0.101644	0.018035	-0.101282	1.000000	0.488543
previous	-0.012246	0.013980	-0.065875	0.017125	-0.028284	0.488543	1.000000

Fig 4. The correlation between each feature

The pdays and previous have relatively strong correlation compared to other features (0.488543). In fact, pdays and previous feature have similarly descriptive introduction: “pday is the number of days that passed by after the client was last contacted in a previous campaign (numeric; -1 means client was not previously contacted), and previous is the Prior number of contacts performed before this campaign and for this client (numeric)”. Therefore, it is unnecessary to remove any features.

Section B: Model Evaluation

Process of analysis. After data preprocessing and data exploration, model evaluation is the next stage. This step aims to select the most appropriate models for fitting dataset. The key step of model evaluation is to use validation dataset to assess the models (Steve, 2019) [4]. In this section, there are four statistical prediction models that will be evaluated: baseline model, decision tree classifier, logistical regression and K nearest neighborhood. The main technical tool for analyzing the four models is Orange. The Fig. 5 illustrates the entire process of analysis.

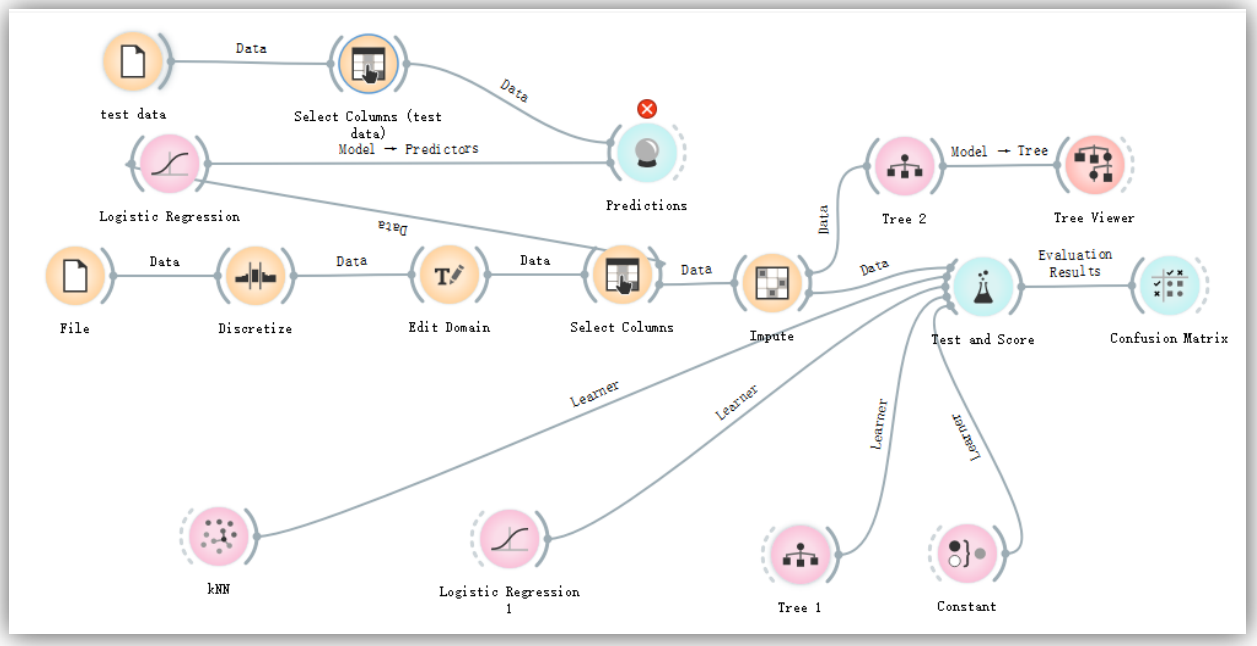


Fig 5. The whole stages of analysis in Orange

File widget is used to load dataset files. Discretize widget aims to adjust the frequency or width of numeric features, such as age, balance. Edit domain widget can change features name and their attributes, such as from numeric values to categorical values. Select columns widget can decide which features are output or input. The function of impute widget is to remove missing data. Importantly, test and score widget could link with four models to evaluate their prediction score. The score is based on the examination for the training dataset and test dataset. In this training process, the cross-validation method is used and the number of folds is 5. Confusion matrix widget is linked with test and score widget. It can measure success for each score. It is obvious that test and score also links with four models. Therefore, test and score can display different scores for different models. After decision on final model, it is available to evaluate new test data. Prediction widget not only connects with logistic regression (final model selected), but also links with new test data for prediction.

After introduction to the entire evaluation process, it is also important to assess the cons and pros of each model. Firstly, baseline model, whose calculation is based on average values, is the object of reference (Brownlee, 2014) [5]. The aim is to evaluate the other models. Additionally, decision tree model is a useful and effective statistical model for prediction. It can list all possible consequences based on given dataset. “Compared to other algorithms decision trees requires less effort for data preparation during pre-processing. A decision tree does not require normalization of data. A decision tree does not require scaling of data as well. Missing values in the data also does not affect the process of building decision tree to any considerable extent. A decision trees model is very intuitive and easy to explain to technical teams as well as stakeholders.” (Dhiraj, 2019) [6]. However, the huge number of possibilities may lead to over-fitting for prediction dataset. Appropriate parameters for decision tree model are the key for fitting prediction dataset. Therefore, for this decision tree model, parameters for minimum number of instances in leaves are only 3. The tree viewer is shown in Fig. 6.

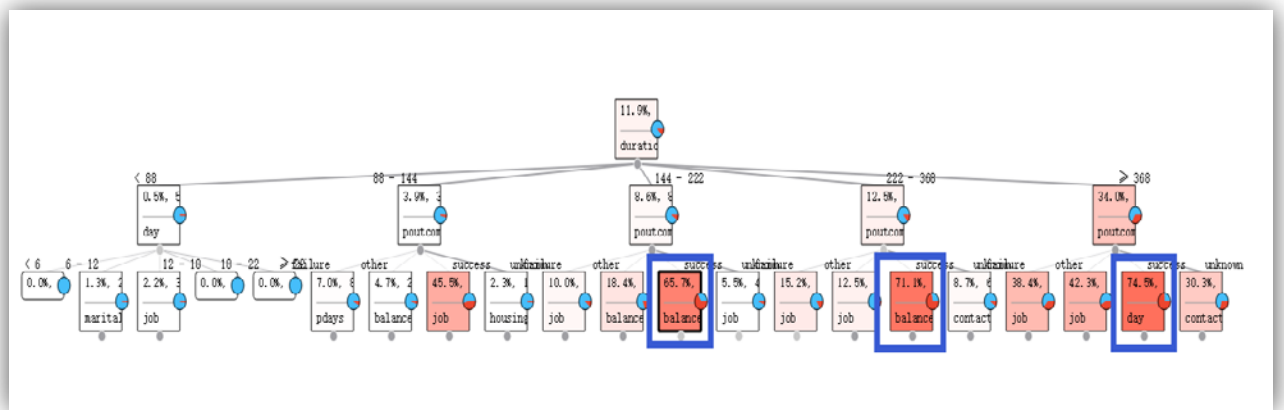


Fig 6. The decision tree viewer

It is clear that when duration is between 144-222 days, 222-368 days, more than 368 days and poutcome is success, it is highly possible to market this financial product successfully. Logistic regression is also a useful model for categorical output feature. The benefit is that "Logistic regression is less prone to over-fitting but it can overfit in high dimensional datasets. You should consider Regularization (L1 and L2) techniques to avoid over-fitting in these scenarios" (Naresh, 2019) [7]. The disadvantages in this report could be avoided, because it is unnecessary to analyse by linear regression. The parameter for the logistic regression is L1 Lasso regularisation, which discourage over-fitting to some extents. The final model is KNN, whose advantages are "no training Period. It does not derive any discriminative function from the training data. This makes the KNN algorithm much faster than other algorithms that require training e.g. SVM, Linear Regression etc" (Naresh, 2019) [8]. There are two parameters for KNN model. The first one is the number of K. in fact, K is difficult to decide precisely. Usually K is 5. Another parameter is metric. Euclidean distance is appropriate formula in two-dimension areas.

The final selection of models not only depends on their pros and cons, but also depends on their score. And score standard is based on what enterprises require. In this report, CEO leaves some take-home messages. The first aim is to attract every possible customer to buy N/LAB product as soon as possible, and not missing any customers. Secondly, enterprise cannot afford the costs for contacting all customers. It means that data analyst should precisely find the potential customers. Finally, it is unnecessary to waste time on contacting customers who are not interest in the N/LAB. To sum up, it is important to consider the costs associated with false alarms and missing prediction equal. Classification accuracy is the suitable scores for selecting models. It is used when correct prediction of all customers is equally important (Provost & Fawcett) [9]. And the Fig. 7 shows the classification accuracy score for each model. The highest score is logistic regression (0.893).

Model	AUC	CA	F1	Precision	Recall
kNN	0.760	0.880	0.272	0.487	0.188
Tree	0.773	0.878	0.327	0.476	0.249
Logistic Regression	0.874	0.893	0.370	0.612	0.266
Constant	0.500	0.881	0.000	0.000	0.000

Fig 7. The scores of four models

		Predicted		
		no	yes	Σ
Actual	no	4305	100	4405
	yes	437	158	595
	Σ	4742	258	5000

Fig 8. Confusion matrix

The Fig 8 shows confusion matrix that measure success and could calculate the score. Classification accuracy is calculated by: (4305+100) / (4305+100+437+158) = 0.893

All steps have been completed, and data analysis team only needs to insert related N/LAB product data into our test data. The prediction result will display.

Section C: Business Recommendations and Insights

In this section, it will be analysed from marketing's and data's perspectives. Marketing department is recommended to pay attention to those potential customers by using marketing techniques in section A. in terms of data analysis, there are more statistical models that can be applied, such as random forest, Naïve bayes and adaboost (Provost & Fawcett) [10]. These models introduced may have higher scores. It means that the target of enterprise can be achieved more successfully.

Reference

[1] McKinny, W., (2013) Python for Data Analysis: *handling missing data* (Chapter 5). United States of America. Published by O'Reilly Media, Inc. CA95472. 2012: First Edition., P142-P146

[2] Alvira, S., (Jan 31, 2018). Towards Data Science: *How to handle missing data* Source from: https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

[3] Emre, R., (April 1, 2019). Towards Data Science: *Fundamental Techniques of Feature Engineering for Machine Learning* https: // towardsdatascience. com/ feature- engineering- for-machine- learning-3a5e293a5114

[4] Steve, M., (April 16, 2019). HeartBeat: *Introduction to Machine Learning Model* Evaluation Source from: https: // heartbeat. fritz.ai/ introduction- to- machine- learning-model- evaluation-fa859e1b2d7f

[5] Brownlee, J., (November 5, 2014). Machine Learning Strategy: *How to get baseline result and why they matter* http: // theprofessionalspoint. blogspot. com/ 2019/ 02/ advantages- and-disadvantages- of-knn.html

[6] Dhiraj, K., (May 27, 2019). MEDIUM: *Top 5advantages and disadvantages of Decision Tree Algorithm*. Source from: https: // medium. com/ @dhiraj8899/ top- 5- advantages- and-disadvantages- of- decision-tree-algorithm-428ebd199d9a

[7] Naresh, K., (March 2, 2019). The Professional Point: *Advantages and disadvantages of logistic regression in machine learning*. Source from: http: // theprofessionalspoint. blogspot. com/ 2019/ 03/ advantages- anddisadvantages- of.html

[8] Naresh, K., (March 2, 2019). The Professional Point: *Advantages and disadvantages of logistic regression in machine learning*. Source from: http: // theprofessionalspoint. blogspot. com/ 2019/ 02/advantages- and-disadvantages-of-knn.html

[9] Provost F & Fawcett T., (2013) Data Science for Business: *Decision Analytics Thinking I: What Is a Good Model?* (Chapter 7). United States of America. Published by O'Reilly Media, Inc. CA95472. First Edition., P187-P208.

[10] Provost F & Fawcett T., (2013) Data Science for Business: *Evidence and Probabilities* (Chapter 9). United States of America. Published by O'Reilly Media, Inc. CA95472. First Edition., P233-P248.