

2204207

by Chenlong Jiang

Submission date: 05-May-2023 02:35PM (UTC+0100)

Submission ID: 204935284

File name: 42003_Chenlong_Jiang_2204207_493224_615937630.docx (279.98K)

Word count: 2310

Character count: 12730

1. Data Set Description

This experiment selected the bank marketing dataset from the UCI machine learning library ([https://archive.ics.uci.edu/ml/datasets/Bank +Marketing](https://archive.ics.uci.edu/ml/datasets/Bank+Marketing)). These data are related to the direct marketing activities of Portuguese banking institutions. These direct marketing activities are based on the telephone. Generally speaking, customer service personnel of banking institutions need to contact customers at least once to know if they will subscribe to bank products (time deposits). Therefore, the task corresponding to this dataset is a classification task, and the classification goal is to predict whether the customer is (yes) or (no) subscribing to a fixed deposit (variable y). This has a huge impact on actual production, and the predicted results can provide a preliminary plan for future work, as well as a reference for whether certain users will subscribe to fixed deposits;

The dataset contains four CSV files:

- 1) Bank additional full. csv: Contains all samples (41188) and all feature inputs (20), sorted by time (from May 2008 to September 2010);
- 2) Bank additional. csv: Randomly select 10% of samples (4119) from 1);
- 3) Bank full. csv: Contains all samples (41188) and 17 feature inputs, sorted by time. (This dataset is an older version with fewer feature inputs);
- 4) Bank. csv: Randomly select 10% of 4119 samples from 3).

The provision of small datasets (bank additive. csv and bank. csv) is to enable rapid testing of computationally expensive machine learning algorithms such as SVM. This experiment will select a relatively new dataset, which includes 1) and 2) of 20 feature quantities.

2. Attribute Descriptions

The input variables of the dataset are 20 characteristic variables, which are divided into numerical variables and categorical variable.

The output variable is y, which means whether the customer has subscribed to a fixed deposit (binary: "yes", "no").

Customer Information:

1. Age: Age
2. Job: Job, job type (classification: "Administrative Management", "Blue Collar", "Entrepreneur", "Maid", "Management", "Retirement", "Self employed", "Service", "Student", "Technician", "Unemployed", "Unknown")
3. Marital: Marriage, marital status (classification: divorced, married, single, unknown) (note: "divorce" refers to divorce or widowhood)
4. Education: Education (Classification: 'Basic. 4y', 'Basy. 6y', 'Basy. 9y', 'Health. School', 'Illiteracy', 'Professional' Courses, 'University Degree', 'Unknown')
5. Default: Default, credit default? (Classification: "No", "Yes", "I don't know")
6. Housing: Housing, do you have a housing loan? (Classification: "No", "Yes", "I don't know")
7. Loan: Is there a personal loan available? (Classification: "No", "Yes", "I don't know")

Other data related to prediction:

8. Contact: Contact method (classification: "mobile phone", "fixed phone")
9. Month: Month, the last contact month (classification: 'MAR',..., 'NOV', 'DEC')
10. Day_ of_ Week: Number of days per week, contact days of the last week (classification): "Mon", "Tee", "We", "TUU", "FRI"
11. Duration: Duration, the last contact duration in seconds
12. Campaign: Number of executives contacted by this client in this campaign
13. Pdays: The number of days since the customer last contacted from the previous event (999 indicates that the customer has not been contacted before)
14. Previous: Number of contacts with this client before this event (number)
15. Proutcome: The result of the previous marketing campaign (classification: failed, non-existent, successful)

Social and economic background attributes

16. EMP.var.rate: Employment Change Rate - Quarterly Indicator (Figures)
17. Cons. price. idx: Consumer Price Index - Monthly Indicator (Number)
18. Cons. conf.idx: Consumer confidence index - monthly index (number)
19. Euribor3m:: Euro Interbank Offered Rate 3-Month Rate - Daily Indicator (Number)
20. Nr. employed: Number of employees - quarterly indicator (number)

3. Data Pre-processing

3.1.First, load the data,

```
In [6]: import pandas as pd
In [7]: input_data_path = "../data/bank-additional/bank-additional-full.csv"
In [8]: data = pd.read_csv(input_data_path, sep=';')
```

Then view the basic information of the dataset. (Using the info() and describe() functions)

```
In [9]: data.info()
<class 'pandas.core.frame.DataFrame'>
Int64Index: 41188 entries, 0 to 41187
Data columns (total 21 columns):
age                41188 non-null int64
job                41188 non-null object
marital            41188 non-null object
education          41188 non-null object
default            41188 non-null object
housing            41188 non-null object
loan               41188 non-null object
contact            41188 non-null object
month              41188 non-null object
day_of_week        41188 non-null object
duration            41188 non-null int64
campaign           41188 non-null int64
pdays             41188 non-null int64
previous           41188 non-null object
poutcome           41188 non-null object
emp.var.rate       41188 non-null float64
cons.price.idx      41188 non-null float64
cons.conf.idx       41188 non-null float64
euribor3m          41188 non-null float64
nr.employed         41188 non-null float64
dtypes: float64(5), int64(5), object(11)
memory usage: 6.9+ MB

In [10]: data.describe()
Out[10]:
```

	age	duration	campaign	pdays	previous	\
count	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000	
mean	40.02406	258.285010	2.567593	962.475454	0.172963	
std	18.42125	259.279249	2.770014	186.910907	0.494901	
min	17.00000	0.000000	1.000000	0.000000	0.000000	
25%	32.00000	102.000000	1.000000	999.000000	0.000000	
50%	38.00000	180.000000	2.000000	999.000000	0.000000	
75%	47.00000	319.000000	3.000000	999.000000	0.000000	
max	98.00000	4918.000000	56.000000	999.000000	7.000000	

	emp.var.rate	cons.price.idx	cons.conf.idx	euribor3m	nr.employed
count	41188.000000	41188.000000	41188.000000	41188.000000	41188.000000
mean	0.081886	93.575664	-40.502600	3.621291	5167.035911
std	1.570960	0.578840	4.628198	1.734447	72.251528
min	-3.400000	92.201000	-50.800000	0.634000	4963.600000
25%	-1.800000	93.075000	-42.700000	1.344000	5099.100000
50%	1.100000	93.749000	-41.800000	4.857000	5191.000000
75%	1.400000	93.994000	-36.400000	4.961000	5228.100000
max	1.400000	94.767000	-26.900000	5.045000	5228.100000

3.2. Missing value processing

From the basic information provided in section 2.2 of the dataset, it can be seen that numerical variables are not missing. Non numerical variables may have unknown values.

Use the following code to view the number of unknown values for character variables. rical variables are not missing. Non numerical variables may have unknown values. Use the following code to view the number of unknown values for character variables. There are usually the following methods for handling missing values:

- ① For variables with fewer unknown values, including job and capital, delete the rows where these variables are missing values;
- ② If it is expected that the variable will not have a significant impact on the learning model's performance, the unknown value can be assigned a mode. It is assumed that all variables have a significant impact on the learning model and this method will not be adopted;
- ③ Complete rows of data can be used as a training set to predict missing values for variables such as housing, loan, education, and default. Since sklearn's model can only deal with numerical variables, it is necessary to first digitize categorical variable and then make predictions. This experiment uses random forest to predict missing values.

3.3 Numerical value of categorical variable

In order to enable categorical variable to participate in model calculation, we need to digitize categorical variable, that is, code them. Categorical variable can be divided into binomial categorical variable, ordered categorical variable and unordered categorical variable. The coding methods of different types of categorical variable are also different.

3.3.1 Code of the second categorical variable

According to the above description of input variables, the variables default, housing, and loan can be considered as two categorical variable, and 0 and 1 codes can be applied to them.

3.3.2 Ordered categorical variable Code

According to the above description of the input variables, we can think that the variable education is an ordered categorical variable. The order of influence is "illustrative", "basic. 4y", "basic. 6y", "basic. 9y", "high. school", "professional. course", "university. grade". The order of variable influence from small to large is 1, 2, 3,...

3.3.3 Unordered categorical variable code

Based on the description of the input variables mentioned above, it can be considered that the variables job, market, contact, month, day_of_Week is an unordered categorical variable. It should be noted that although the variables month and day_of_Week is ordered from a temporal perspective, but unordered for the target variable. For unordered categorical variable, dummy variables can be used for coding. Generally, n-1 dummy variables need to be set for n classifications. For example, the variable marital is divided into divorced, married, and single, encoded using two dummy variables V1 and V2. The Pandas package in Python provides functions for generating dummy variables.

3.4 Numerical feature preprocessing

3.4.1 discretization of continuous features

One advantage of discretization of continuous features is that it can effectively overcome the hidden defect in the data: make the model results more stable.

the variable duration is discretization. Specifically, use the pandas. qcut () function to discretization continuous data. It uses the quantile to divide the data (sub box: binding), and you can get boxes (bin) that are basically equal in size, expressed in interval form. Then use the pandas. factorize () function to convert the interval to a numerical value

3.4.2 Standardization

Due to the frequent use of different units of measurement for different variables, they differ significantly in numerical terms, making distance based learning models more susceptible to the influence of variables with larger values. Data normalization is the process of compressing data into a range, so that the unit effects of all variables are consistent.

3.4.3 Data after persistence preprocessing

Due to the prediction of unknown values, the preprocessing process requires a considerable amount of time, so the preprocessed data will be persistent. Save it to a file, and subsequent learning models will read the file for data import, training, and prediction without further preprocessing.

It should be noted that since the raw data is ordered, after reading it, I need to randomly arrange it into an unordered dataset.

4. Model training and evaluation

4.1 Partition Dataset

Firstly, the processed dataset is divided into three parts, namely the training set, validation set, and testing set. The training set is used to train the model and is used during the training phase. After the model is trained, I don't know how it performs. At this point, you can use the validation set to see how the model performs on new data. At the same time, the model is in the best state by adjusting hyperparameter. Conduct final evaluation through test sets

4.2 Training Set Resampling

By conducting statistics on the imported dataset, it can be found that the number of samples ($y=1$) is less than the number of negative samples ($y=0$), with the former having approximately 8 times the number of the latter.

This is an imbalanced classification problem. In this classification problem, the classification goal is to predict whether customers are willing to purchase bank wealth management products. Obviously, what we are more concerned about is which customers have subscribed to time deposits. In order to mitigate the adverse effects of data imbalance, there are two simpler methods at the data level: oversampling and under sampling.

Over sampling: The most commonly used method for sampling and processing imbalanced data, with the basic idea of eliminating or reducing data imbalance by changing the distribution of training data. The oversampling method improves the classification performance of minorities by adding minority samples. The simplest way is to simply copy minority samples. The disadvantage is that it may lead to overfitting, no new information is added to minorities, and the generalization ability is weak. The improved oversampling method involves adding random Gaussian noise to a few classes or generating new composite samples.

Under sampling: Under sampling methods improve the classification performance of minority classes by reducing the number of majority class samples. The simplest method is to randomly remove some majority class samples to reduce the size of the majority class. The disadvantage is that it will lose some important information of the majority class and cannot fully utilize existing information.

In this experiment, the Smote algorithm [Chawla et al., 2002] was used to add new samples for oversampling; Using the method of randomly removing some majority class samples for under sampling. The basic idea of the Smote algorithm is to calculate the distance from each sample x in the minority class sample set to all samples using Euclidean distance as the standard, and obtain its k -nearest neighbors. Then set a sampling ratio based on the sample imbalance ratio to determine the sampling rate N . For each minority sample x , randomly select several samples from its k -nearest neighbors to construct new samples. For the data of this experiment, in order to prevent excessive noise in the newly generated data, only numerical variables in the new sample are truly newly generated, and other variables are consistent with the original sample.

4.3 Model training and evaluation

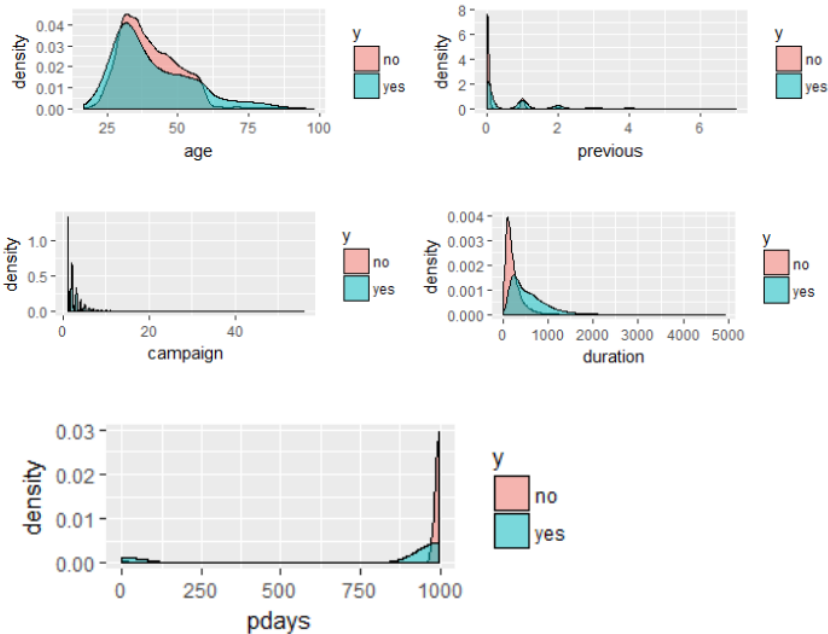
Common classification models include random forest, linear regression, logical regression, decision tree, svm and other algorithms. In this report, logical regression and random forest were selected to train on the training set, and random forest was evaluated on the cross test set. According to the results, I finally chose random forest model to test on the test set.

The evaluation of the model may vary among different tasks. As mentioned above, the dataset is imbalanced, with negative sample values accounting for the majority of the total dataset. If our model "predicts" that all target variable values are 0, then the accuracy should be very high. However, it is evident that this' prediction 'is redundant. Therefore, we prefer to choose models that can predict samples ($y=1$). Therefore, the criterion for evaluating the quality of the model in the experiment is the f1 score of the positive sample.

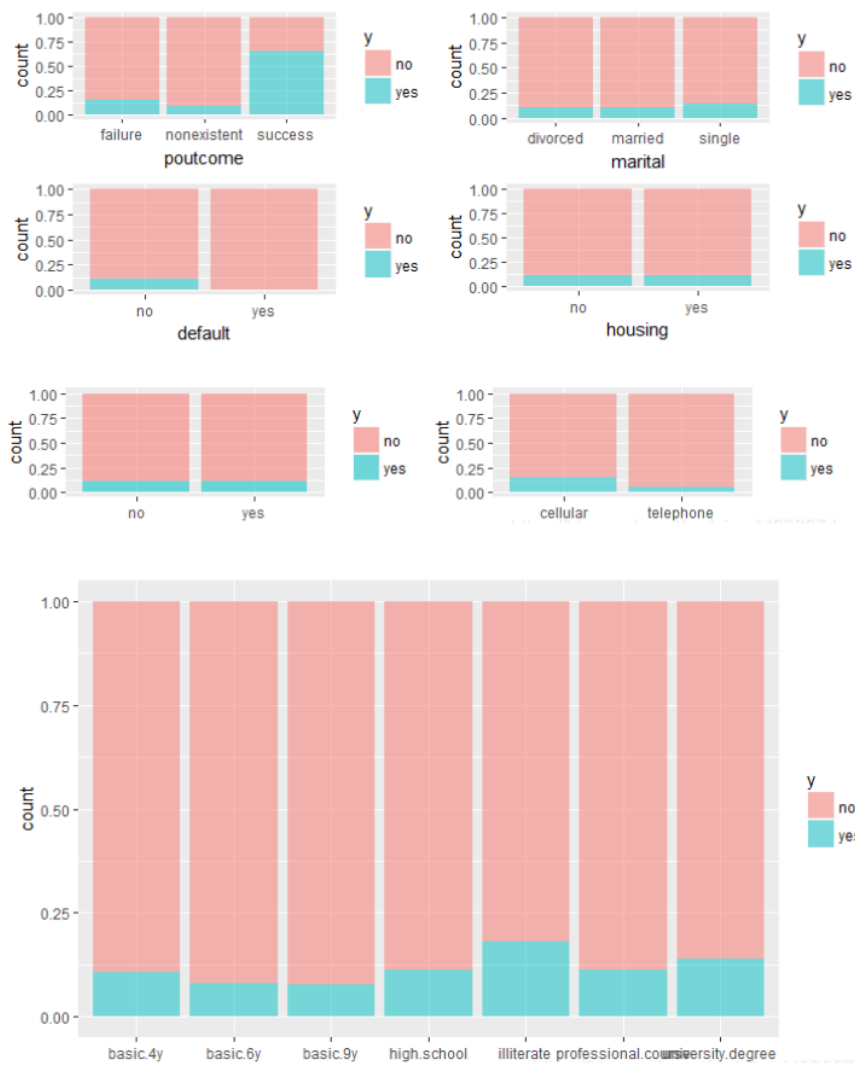
The result of selecting a model depends on the training evaluation function. Logistic regression model and random forest model can be selected for experiment, and their parameter values can be adjusted. Finally, the random forest model is chosen because of the f1 score. Specifically, when `n_When Estimators` is set to 400, 7 times oversampling is applied to positive samples ($n=7$), while negative samples are not applied to negative samples (`frac=1.0`).

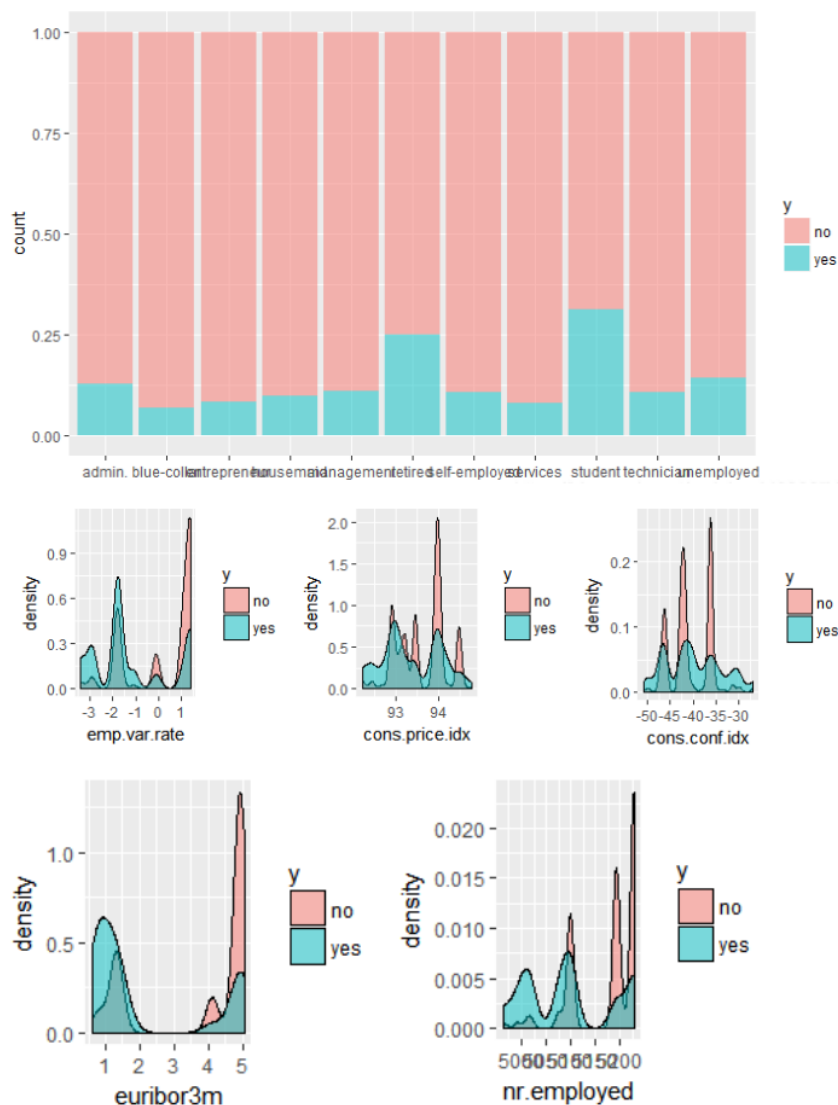
5.Summary

- ① Young customers are more willing to purchase bank fixed deposits,
- ② Customers who have been in contact for too long will not make purchases, and those who have had a large number of previous contacts will also not make purchases.



- ③ Customers with loans are more willing to participate in activities, possibly because customers with loans themselves do not have the ability to consume such a fixed deposit product.
- ④ The probability of regular customers continuing to purchase is much higher. Compared with other wealth management products, the risk of fixed deposits is lower, and most buyers are unwilling to bear the high returns brought by the high risks of other wealth management products. Therefore, there are fewer risk averse customers among regular customers, and the purchase rate will be relatively high.
- ⑤ Single individuals have a higher purchasing rate, possibly due to their greater financial freedom.
- ⑥ The better effect of using a mobile phone for communication may be due to its easier access to customers.
- ⑦ Among different educational backgrounds, illiterate individuals have the highest purchase rate, followed by undergraduate degrees, possibly due to their limited understanding of other investment methods
- ⑧ Among different professions, the highest purchasing rate is in order of students, retirees, and administrative management





It can be understood from the importance of the independent variable that the overall social environment has a significant impact on sales.

The factors related to the contact between sales personnel and customers also have a significant impact. In order to have a good sales effect, the contact duration should reach over 200 hours; At the same time, it is important to choose a fixed executor for communication with customers, as frequent replacement of candidates can lead to a decrease in customer purchase rates.

Secondly, the rate of repeat customers is very high, and banks can focus more on repeat customers; For new clients, they can focus on elderly retirees.

FINAL GRADE

GENERAL COMMENTS

Instructor

50/100

PAGE 1

PAGE 2

PAGE 3

PAGE 4

PAGE 5

PAGE 6

PAGE 7

PAGE 8

GRADING FORM: DATA MINING / VIS A2_2022/3

CHENLONG JIANG

50

KNOWLEDGE AND UNDERSTANDING

Dataset description (15 marks)



Bank marketing dataset used but no clear objectives for undertaking your analysis is provided in submission. Description of attributes are however given.

10

COGNITIVE/ INTELLECTUAL SKILLS

Primary and Secondary data analysis, reflection (35 marks)



Description of performed analysis not given in sufficient detail for ensuring that algorithmic choices made are clear.

20

PRACTICAL APPLICATION SKILLS

Primary and Secondary data processing (20 marks)



Some evidence of data pre-processing evident in submission.

12

TRANSFERABLE SKILLS

Data reporting and report writing (30 marks)



Key headings in submission can be found easily. Images included are not connected to listed themes in assignment.

8