

Bank Deposit

Data Mining

Zixun Zhang, Wenliang Hu, Miaomiao Zhao

Abstract

Due to the increasing volume of data, it is impossible for humans to analyze the data sources and make any predictions in order to bring more benefits to the company, such as bank marketing. Data mining has greatly improved the performance of these activities. This article describes the most critical data mining techniques to analyze data and predict business decisions. Here we implement a decision tree and logistics regression (LR) on the actual data of bank deposit subscriptions. Python is a widely used high-level programming language. It is a general-purpose programming language. As an interpretive language, Python's design philosophy emphasizes the readability and concise syntax of the code (especially using space indentation to separate code blocks rather than curly braces or keywords). The experimental results of performance are represented by three statistics; classification accuracy, sensitivity and specificity. The higher the accuracy, the better the model is. We used python to process the dataset. Finally, we got the result of accuracy of decision tree and logistic regression and compared with them.

Key Words: *Bank Deposit; Exploratory Data Analysis; Decision Tree; Logistics Regression*

I. Introduction

In general, bank tries to search the client who will deposit more money in the next terms and get the profits from it. Whether people are talking about “data analysis” or “big data” or “predictive analysis,” a large number of the results show that data analysis is one of the most important trends facing the bank marketing today. At the same time, because of the continuous development and progress of big data and data analysis, the market competition of banks has become more intense. Then, the higher the consumer's expectations, the more data the bank needs to process and predict. If banks cannot meet the needs and expectations of consumers in time, they will “die” quickly. Therefore, for the bank, they could balance the deposit and loan by predicting the potential clients and invest more in the clients. For the client, they can make a better investment plan or project. The reason that we choose this topic is this is closer to our lives, and make us know more knowledge of investment management.

Our goal is help bank to predict the deposit behavior of the client, according to the predicting, they can make the appropriate investment and financing plan and manage fund flow. We can use bank client data, related to the last contact of the current campaign, and some social and economic context to build a model and analyze the next financing behavior of the clients. We use python to analyze and predict data. Python is a widely used high-level programming language. It is a general-purpose programming language. As an interpretive language, Python's design philosophy emphasizes the readability and concise syntax of the code (especially using space indentation to separate code blocks rather than curly braces or keywords). This final project adopts two methods that are Decision Tree and Logistic Regressive to analyze and predict data. Passing data processing, we can get two accuracies by Decision Tree and Logistic Regressive, and compare with them.

Our Webpages: <https://github.com/huwenliang777/SpaceProject.git>

II. Dataset

This dataset is from the UC Irvine Machine Learning Repository. In this dataset, there are 45211 numbers of samples with 17 attributes (like age, job marital, education, default, housing, loan, contact, etc.) without missing values, and these data is related with direct marketing campaigns of a Portuguese banking institution. Calling is the main way of marketing activities. Staff member often need to contact the same customer multiple times to determine whether a bank's product is subscribed or purchased. The following Table 2.1, showing the features for dataset.

Table 2.1 Features of Dataset

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no

Here is the web link of the dataset:

<http://archive.ics.uci.edu/ml/datasets/Bank+Marketing>

III. Related Work

Bank customers' data plays an important role in all bank marketing campaigns, so that there are many research to attempt to predict whether the client has subscribed a term deposit. Here we present two papers which related to our project.

“Bank Direct Marketing Analysis of Data Mining Techniques” presents a good test accuracy with Multilayer perception neural network (MLPNN), Naïve Bayes Classifier, logistic regression and decision tree model technique. They use three statistical measures to evaluate each classification model, which are classification accuracy, sensitivity, specificity. The logistic regression shows the result that the prediction accuracies are 90.09% and 90.43, and C5.0 with the prediction of 93.23% and 90.09% [1].

In another research report that describes prediction performance, logistic regression models usually perform well when the problem is roughly linearly separable; whereas, the decision tree model does not assume the linear nature of the problem and can therefore handle the more general situation well. Specifically, in this study, the logistic regression model achieves a prediction error rate of 0.1, and the error rate of the decision tree results can be predicted, indicating that the two methods can predict whether customers will purchase fixed deposits [2].

Table 3.1 Compare results for group 9

	Unpruned Tree	Pruned Tree	Bagging	Random Forest
Training Error	0.09986	0.09986	0.1067	0.103
Testing Error	0.1057	0.1057	0.1085973	0.101

IV. Implementation

This section will introduce two methods to help us analysis the model, part of them will be applied to next section.

A. Decision Tree Model

Decision tree algorithm is one of the most popular data mining techniques, which originated in statistics discipline. Decision trees are applied on both classification and regression problems.

Decision tree algorithm separates the data into different subsets in order to make the data more homogeneous than the previous data set. For the binary decision tree, the algorithm partitions the data into two subsets, and splits each subset again until the model reach the homogeneity criterion, which means the criterion is satisfied [1].

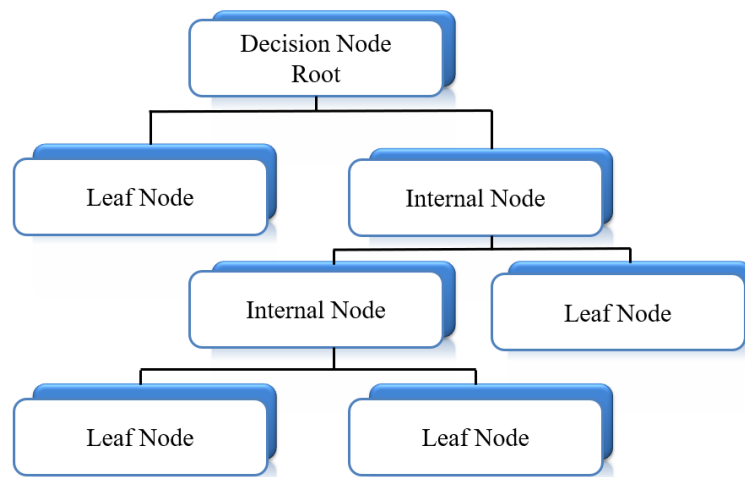


Fig 4.1 Binary Decision Tree

There are two reasons for explaining the popularity of decision trees. First, the process of the decision tree technique is easily understood and explained. Second, the technique solves the number of data complexities, for instance, nonlinearly and interactions, which commonly occur in real big data environment.

There are several types algorithms to build a decision tree model, here we focus on CART (Classification and Regression Trees), which use Gini Index as metrics (explained later), and ID3 (Iterative Dichotomiser 3), which use entropy function and information gain as metrics in details [3].

We have 16 values being categorical and one y value (Yes or No) in our data set, so that this is a binary classification problem.

1. Classification with using the Iterative Dichotomiser 3 (ID3) algorithm

First, we need to determine the attribute that best classifies the training data and use this attribute as the root of the decision tree. The way which could choose the best attribute is to calculate entropy and information gain for all attributes.

Entropy

Entropy $H(s)$ is a measure of uncertainty associated with a random variable. The calculation of the entropy is defined as [4]:

$$\text{entropy}(Y) = - \sum_{i=1}^m p_i \log(p_i), \text{ where } p_i = P(Y = y_i).$$

When entropy value equal to 0, when the set Y is perfectly classified. In ID3, entropy is calculated for each remaining attribute. The factor which has the smallest entropy value is used to split the data set [5].

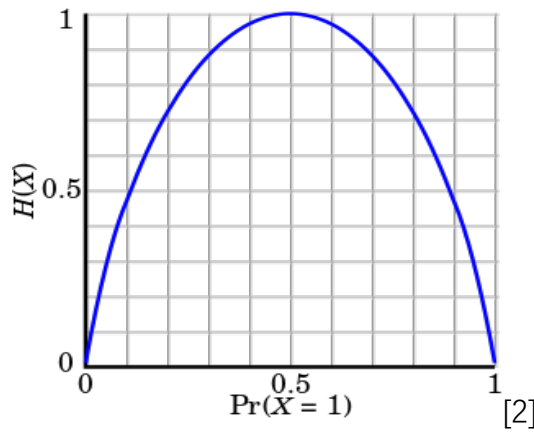


Fig 4.2 The entropy function relative to a binary classification, $X=1$

(in our data set: $y = \text{"yes"}$)

Information Gain

The measure of the effectiveness of an attribute in classifying the training data is called information gain [6].

$$\text{gain}(y, \text{attribute}) = \text{entropy}(y) - \sum_{v \in \text{values}(\text{attribute})} \frac{|y_v|}{|y|} \text{entropy}(y_v)$$

The highest gain attribute which is defined as the best attribute is picked as the node of the decision tree.

Then repeat the same thing as the algorithm for sub-trees till we get the final decision tree.

➤ Classification and Regression Trees (CART) algorithm

In CART model, we use Gini index as cost function used to evaluate splits in the dataset. Since we explained that our dataset is Binary variable which means it take two values [9].

The calculation of Gini Index is defined as:

$$\text{gini}(Y) = 1 - \sum_{i=1}^n p_j^2$$

If a data set Y is split on A into two subsets Y1 and Y2, the Gini index $\text{gini}(Y)$ is defined as:

$$\text{gini}_A(Y) = \frac{|Y_1|}{|Y|} \text{gini}(Y_1) + \frac{|Y_2|}{|Y|} \text{gini}(Y_2)$$

A Gini score gives an idea of how good a split is by how mixed the classes are in the two groups created by the split. A perfect separation results in a Gini score of 0 and the maximum Gini Index value is 0.5.

Then for every attribute, after we calculate Gini Index for all categorical values, we need take average information entropy for the current attribute and calculate the Gini Gain. The calculations are similar to ID3, expect the formula changed.

Finally, we get the best attribute as the root of the decision tree. Then repeat the same thing as the algorithm for sub-trees till we get the final decision tree.

1. Data Clean and transformation

In our dataset, there are attributes with different type of data so that we need to transform the data type.

For example, for the feature “job”, I define this feature as an ordinal feature. I separate 12 jobs into four parts.

```
# high income=3  
(management, admin., entrepreneur )
```

medium income=2
(blue-collar, self- employed, technician)

low income=1
(services, retired, housemaid)


no income=0
(unemployed, student, unknown)

For the binary feature like “loan”, “housing”, I transform the value which is “yes” equal 1, and the other equal 0.

For further analysis, I create new columns to store the numeric data so that the original dataset is unbroken.

After the transformation, a part of the new data set shows below:

	age	job	marital	education	default	balance	housing	loan	contact	day	month	duration	campaign	pdays	previous	poutcome	y
0	58	management	married	tertiary	no	2143	yes	no	unknown	5	may	261	1	-1	0	unknown	no
1	44	technician	single	secondary	no	29	yes	no	unknown	5	may	151	1	-1	0	unknown	no
2	33	entrepreneur	married	secondary	no	2	yes	yes	unknown	5	may	76	1	-1	0	unknown	no
3	47	blue-collar	married	unknown	no	1506	yes	no	unknown	5	may	92	1	-1	0	unknown	no
4	33	unknown	single	unknown	no	1	no	no	unknown	5	may	198	1	-1	0	unknown	no



...	campaign	pdays	previous	poutcome	y	jobnum	maritalnum	educationnum	housingnum	loannum
...	1	-1	0	unknown	no	3	2	3	1	0
...	1	-1	0	unknown	no	2	1	2	1	0
...	1	-1	0	unknown	no	3	2	2	1	1
...	1	-1	0	unknown	no	2	2	0	1	0
...	1	-1	0	unknown	no	0	1	0	0	0

Fig.4.3 Dataset Transformation

2. Exploratory Data Analysis

In order to find the best features and make the prediction better, we try to exploratory data analysis.

In the midterm, we analysis three variables (ages, balance, Campaign) to build decision tree and predict if client will do the term deposit.

```
In [18]: fig=plt.figure(figsize=(5,5))
plt.title("age")
pos_age.hist(alpha=0.7, bins =30, label='yes')
neg_age.hist(alpha=0.7, bins =30, label='no')
plt.legend(loc="upper right")

fig=plt.figure(figsize=(5,5))
plt.title("balance")
pos_balance.hist(alpha=0.7, bins =30, label='yes')
neg_balance.hist(alpha=0.7, bins =30, label='no')
plt.legend(loc="upper right")

fig=plt.figure(figsize=(5,5))
plt.title("campaign")
pos_campaign.hist(alpha=0.7, bins =30, label='yes')
neg_campaign.hist(alpha=0.7, bins =30, label='no')
plt.legend(loc="upper right")
```

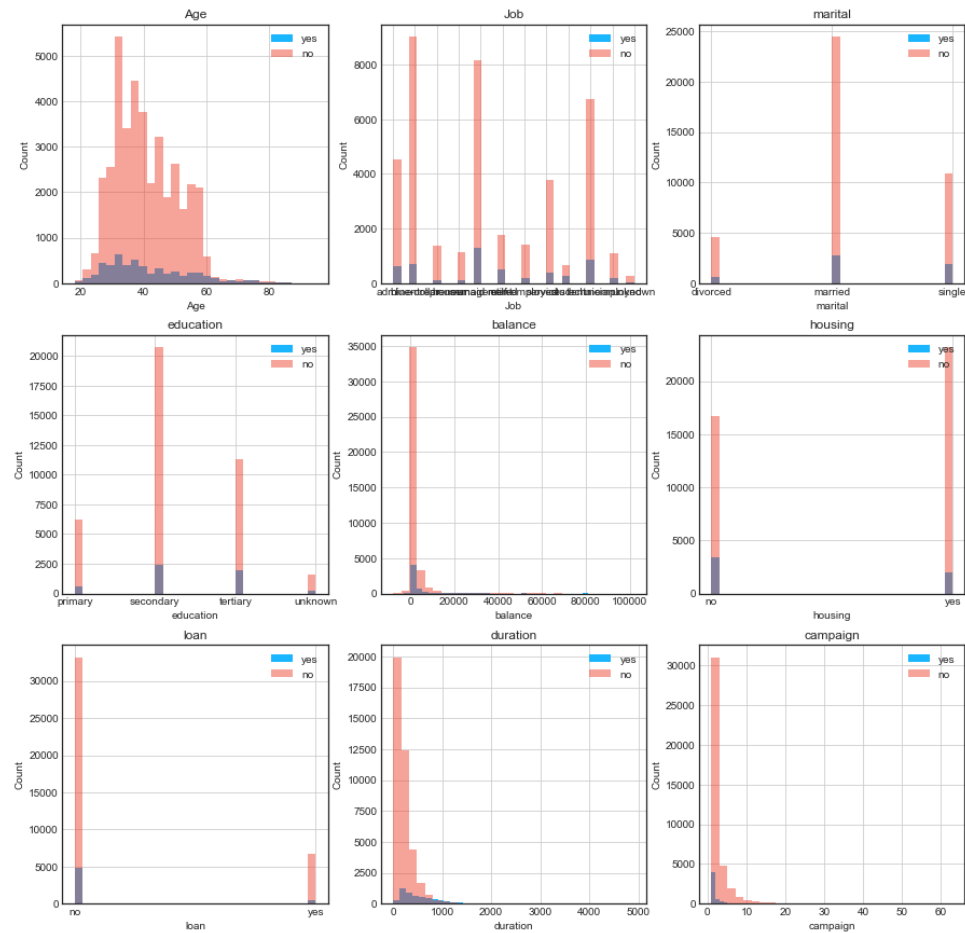


Fig.4.4 Partial code and Histogram all attributes vs Deposit

From the figure 4, we get the base idea is most of clients do not have term deposit, according to histogram, the client who has the high balance and more campaigns will deposit in the bank. The age between 30 to 60, the probability that the client between 30 to 60 deposits is higher than other clients [7].

3. Build CART Model

We try to build decision tree to predicting. The first step is to classifier the data by function `DecisionTreeClassifier()` and combine the three features that we get from the exploratory data analysis. The next step is fitting the model and building the decision tree.

```
: train, test= train_test_split(bank, test_size=0.15)

: print("Training Size: {}; Test Size: {}".format(len(train),len(test)))

Training Size: 38429; Test Size: 6782

: x_train=train[features]
  y_train=train["y"]

  x_test=test[features]
  y_test=test["y"]

: dt=c.fit(x_train, y_train)

: import os

def conda_fix(graph):
    path = os.path.join(sys.base_exec_prefix, "Library", "bin", "graphviz")
    paths = ("dot", "twopi", "neato", "circo", "fdp")
    paths = {p: os.path.join(path, "{}.exe".format(p)) for p in paths}
    graph.set_graphviz_executables(paths)

: def show_tree(tree, features, path):
    f= io.StringIO()
    export_graphviz(tree, out_file=f, feature_names=features)
    pydotplus.graph_from_dot_data(f.getvalue()).write_png(path)
    img= misc.imread(path)
    plt.rcParams["figure.figsize"]=(20,20)
    plt.imshow(img)

    show_tree(dt, features, 'tree.png')
```

Fig. 4.5: Code for building CART model

- Build a Decision Tree classifier using scikit-learn
- Build a Random Forest classifier using scikit-learn
- Build an Artificial Neural Network classifier using Keras

4. The Experimental Results

The performance of the classification model will be evaluated; the estimation methods are classification accuracy, sensitivity and specificity, respectively. Four things will be used here: true positive, true negative, false positive and false negative. [1]

		Predicted	
		Positive (yes)	Negative (no)
Actual	Positive (yes)	TP	FP
	Negative (no)	TN	FN

Fig. 4.6 Confusion Matrix

The actual value of the variable is the percentage of correct/incorrect classification, and the prediction value of the variable is also the percentage of correct/incorrect classification, but they are not the same. True value (TP) refers to the quantity that is predicted by the trueness of an instance. True-negative (TN) is an example (possibly wrong) that appears when both the classifier and the target attribute yield a positive prediction. A false positive (FP) is to suggest that the number of instances (false predictions) is true. Finally, False Negative (FN) is the number of instances (false predictions) that are false. The confusion matrix of the two classifiers is the performance content of Table 4. Classification accuracy is defined as the proportion of correctly classified cases ((TP+TN)/N) [8].

5. Results of Decision Tree

By the CART, we got the final decision tree.

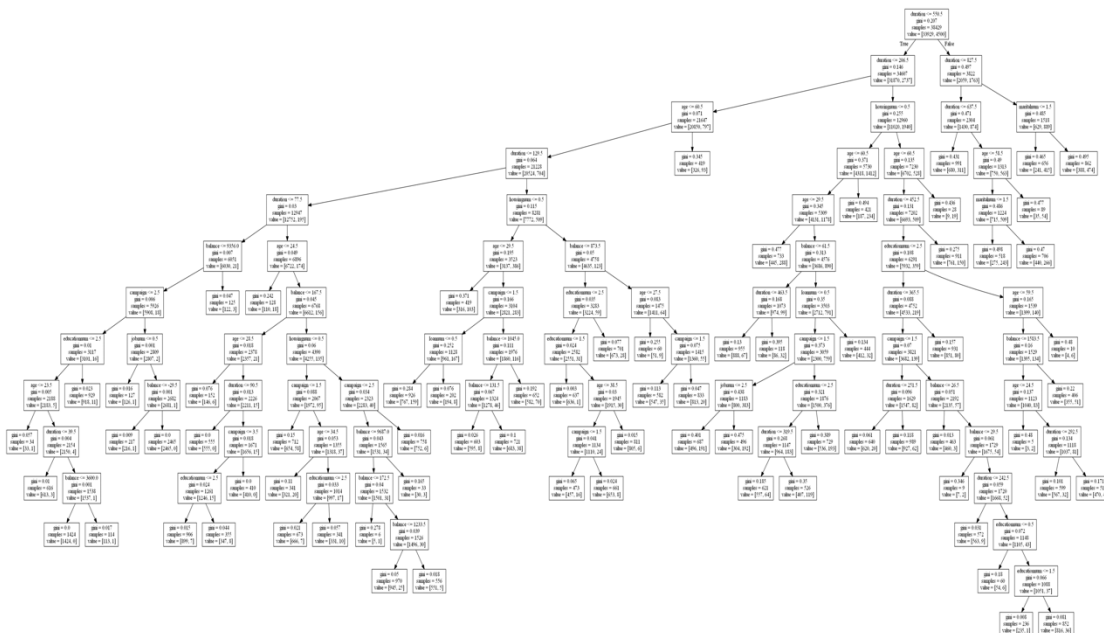


Fig. 4.7: Decision Tree by CART algorithm

Through calculate the gini index, we could find that the attribute of duration has the largest gini index so that duration is defined as the root of the decision tree.

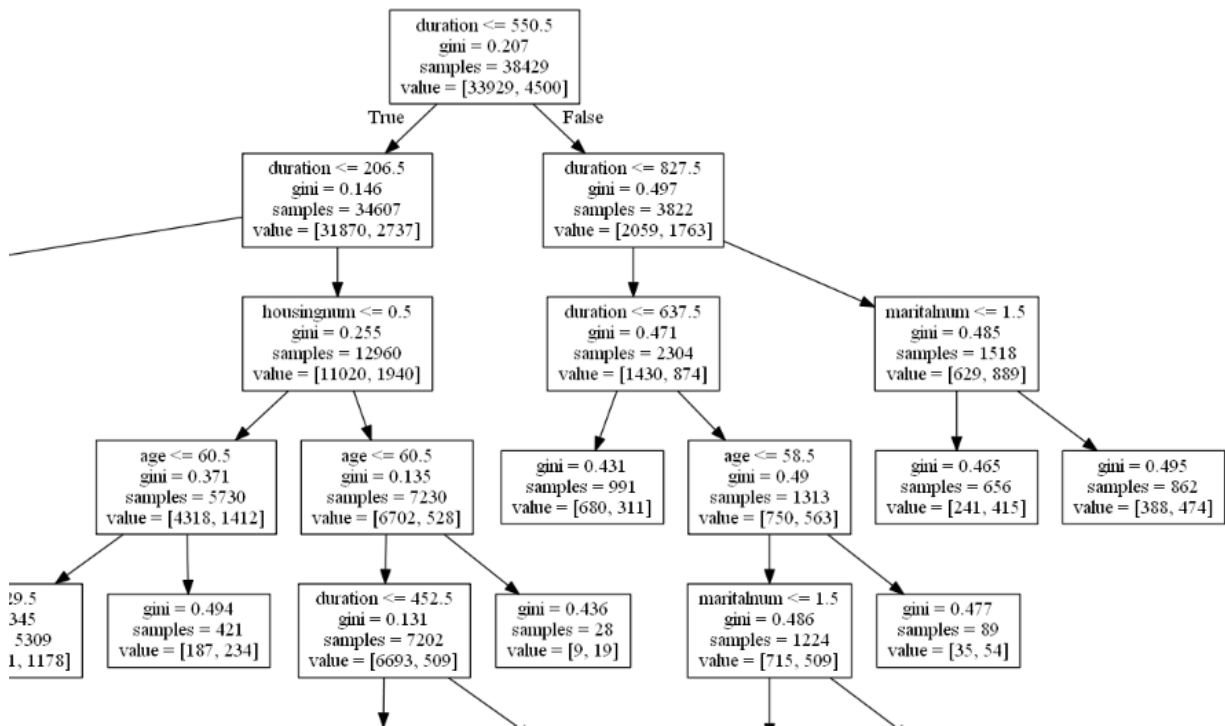


Fig. 4.8: Details of Decision Tree

By the CART model, we got the accuracy is 0.8400, The Confusion_matrix shows that clients truly won't subscribe next term deposit are 9029, and clients will truly subscribe next term deposit are 459.

	precision	recall	f1-score	support
no	0.91	0.90	0.91	9998
yes	0.32	0.35	0.34	1305
avg / total	0.85	0.84	0.84	11303

Fig. 4.9: CART vs. ID3

We also calculate the accuracy of the ID3 Decision Tree model, which is 0.84119. Although the accuracy is higher than CART, all values of confusion matrix are similar to the CART model [9].

<pre>from sklearn.metrics import accuracy_score accuracy_score(y_test,y_predict)</pre> <p>0.8400424666017872</p>	<pre>from sklearn.metrics import accuracy_score accuracy_score(y_test,y_predict)</pre> <p>0.8411926037335221</p>
<pre>from sklearn.metrics import confusion_matrix confusion_matrix(y_test, y_predict)</pre> <p>array([[9036, 962], [846, 459]], dtype=int64)</p>	<pre>from sklearn.metrics import confusion_matrix confusion_matrix(y_test, y_predict)</pre> <p>array([[9029, 969], [826, 479]], dtype=int64)</p>

Fig. 4.10: CART vs. ID3

According to the accuracy of both decision tree models, we need to add more useful which would influence whether the client will subscribe the deposit in order to improve the decision tree model.

B. Logistic Regressive

Logistic regression is a statistical method to analyze a dataset in which there are two or more independent variables. In fact, the regression is to estimate the unknown parameters of the known formula [10]. For instance, we now have a lot of real data (training samples). The regression is to use these data values to automatically estimate. The method of estimation is that, given a training sample point and a known formula, the machine will automatically enumerate all possible values of the parameter for one or more unknown parameters (enumerate their different combinations for multiple parameters) Until you find the parameter (or combination of parameters) that best matches the distribution of the sample points.

1. Data clean and Transformation

In data processing part, we try to use different data transformation method with the Decision Tree model, For the missing data, we use Na and mode to replace because some of data is important to analysis for us; therefore, using mode is a kind of easy and quick method [11].

For transformation, I use “0” and “1” to decide “no” and “yes”. For instance, the feature job, there are about 7 different jobs: housemaid, services, admin, and so on. If a client is housemaid, we will set “1” for housemaid, others will be “0”; The same way, we transform the marital feature and education attribute. For education attribute, there is particular problem, “basic.4y”,

“basic.6y”, and “basic.9y” are the different education status, but in fact, three of them has same class which we will predict. Therefore, we combine three into one education status “basic” which will make our logistic regression model easy to fit. Table 4.1 and table 4.2 show the partial of data processing.

Table 4.1 The Dataset before Transformation

	age	job	marital	education	default	housing	loan	contact	month	day_of_week	.
0	56	housemaid	married	basic.4y	no	no	no	telephone	may	mon	.
1	57	services	married	high.school	unknown	no	no	telephone	may	mon	.
2	37	services	married	high.school	no	yes	no	telephone	may	mon	.
3	40	admin.	married	basic.6y	no	no	no	telephone	may	mon	.
4	56	services	married	high.school	no	no	yes	telephone	may	mon	.
5	45	services	married	basic.9y	unknown	no	no	telephone	may	mon	.

Table 4.2 The Dataset after Transformation

	age	housemaid	services	Admin.	married	divorced	single	basic	high.school
0	56	1	0	0	0	1	0	0	1	0	0
1	57	0	0	1	0	1	0	0	0	0	1
2	37	0	0	1	0	1	0	0	0	0	1
3	40	0	0	0	1	1	0	0	1	0	0
4	56	0	0	1	0	1	0	0	0	0	1
5	45	0	0	1	0	1	0	0	1	0	0

2. Exploration Data Analysis

The objective of this project is to predict if the clients will subscribe deposit in the next term. So, first of all, we count the numbers of no and yes to show the initial image for this dataset.

From the Fig.4.11, it is easy to find most of client will not subscribe, further, the numbers of no are eight times than the numbers of yes.

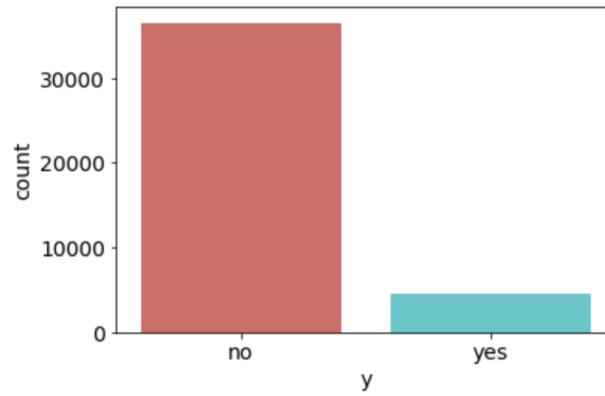


Fig. 4.11 The Count of Class

In this section, we try to find the best predictors which will add into our model, so we plot the histogram of job and marital status.

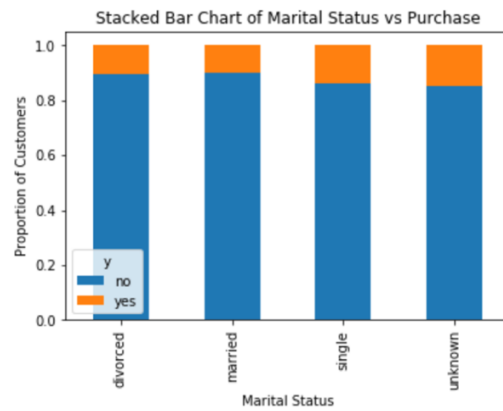


Fig. 4.12 Histogram of Marital Status

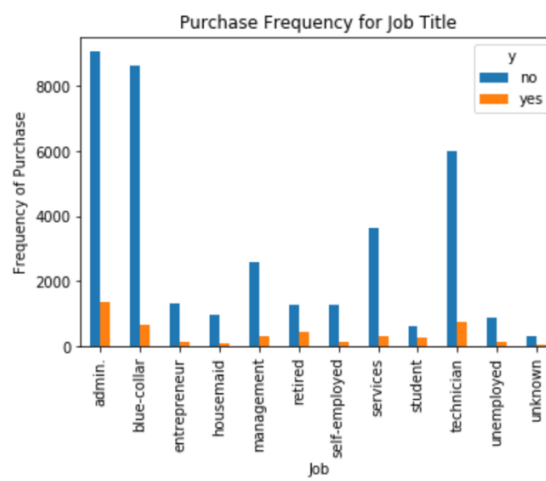


Fig. 4.13 Histogram of Job

There are obvious differences for different job to classify the “yes” and “no”, but for the Marital Status, it is almost same. Therefore, we assume job is a better predictor, but Marital Status is not when we fit features into our logistic regression model.

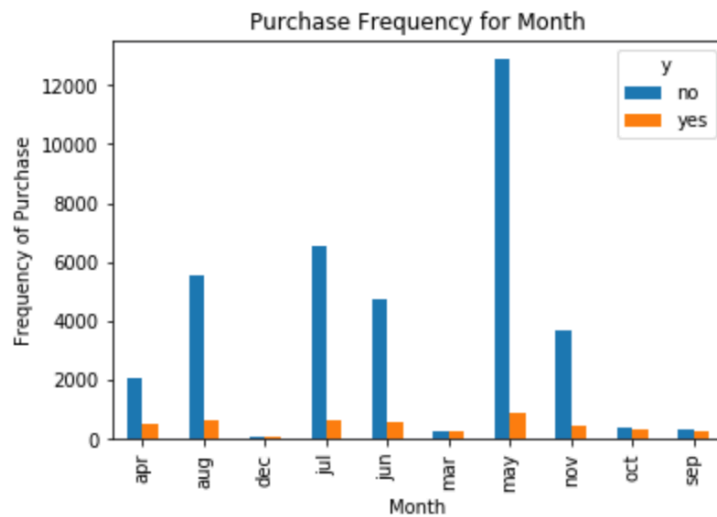


Fig. 4.14 Histogram of Month

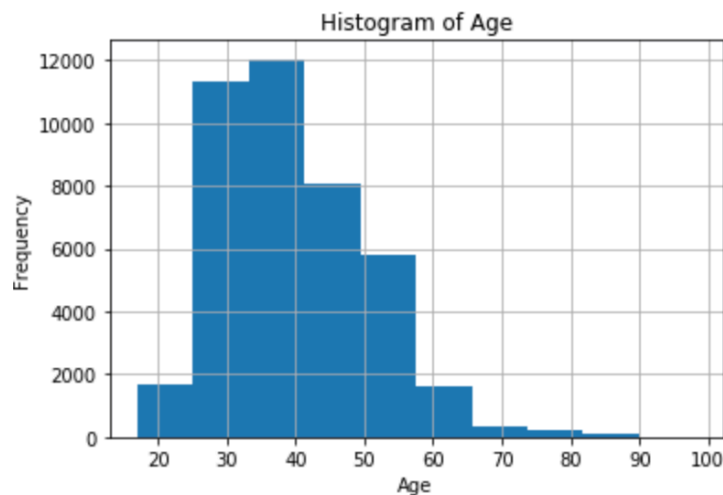


Fig. 4.15 Histogram of Age

When we analyze feature Month, it is obvious to find few clients will subscribe deposit in Dec, the reason could be there are some holidays in this month, so people will spend a lot. On the other hand, most subscribe will happen in May. That will be useful for us when we predict in different month.

In the histogram of age, most age between 30 and 50, so considering this aspect, we could scale Age data. Using this, feature will become more useful.

3. Build Logistic Regression Model

In this section, first of all, I split our dataset into train and test dataset (train: test=80:20), After this, we are using sklearn tools to build model [12].

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, ra
from sklearn.linear_model import LogisticRegression
from sklearn import metrics
logreg = LogisticRegression()
logreg.fit(X_train, y_train)

y_pred = logreg.predict(X_test)
print('Accuracy of logistic regression classifier on test set: {:.2f}'
      .format(logreg.score(X_test, y_test)))
```

Fig. 4.16 Partial Coding

We use function LogisticRegression() to fit our training dataset into a logistic regression model, and function predict() will help us to get the class of test dataset.

4. Results

By the Logistic Regression model, we got the accuracy is 0.898 (10-fold), the Confusion_matrix [13], and some statistics.

	[[7236 83] [750 169]]			
	precision	recall	f1-score	support
no	0.91	0.99	0.95	7319
yes	0.67	0.18	0.29	919
avg / total	0.88	0.90	0.87	8238

Fig. 4.17 Result of Logistic Regression

Accuracy is up to 0.89, we could use add more meaningful feature interactions among features to improve model.

V. Conclusion

In this project, we predicted if the client will subscribe deposit in next term for bank marketing. There are two method which we used: Decision tree and Logistic Regression, we try to compare different accuracy of different algorithms, that is why we choose two to fit the same method. For the Decision Tree, we got the accuracy is 0.81, however Logistic Regression is up to 0.90, so if we just consider accuracy, the Logistic Regression model is better than Decision Tree [14]. But in fact, we need consider a lot when we check which model is better. For example, when we see the recall (Logistic Regression is 0.90, but Decision Tree is 0.84), we think Decision Tree is better (For our dataset, the lower recall, the better model) [15].

By this group project, we learned a lot, not only the professional knowledge, but also the team work. Most of time, how to co-work with other people is important than the knowledge.

In the future, in order to improve our model, we could add more features from the exist features, and perfect our model and algorithms. Finally, we appreciate our supervisor Dr.Praveen Madiraju and all of classmates in Data Mining.

VI. Reference

- [1] Hany A. Elsalamony (2014, January). Bank Direct Marketing Analysis of Data Mining Techniques. Retrieved <http://research.ijcaonline.org/volume85/number7/pxc3893218.pdf>
- [2] Group 9. The Application of Classification Methods on Bank Telemarketing Data.
- [3] Madhu Sanjeevi. Decision Trees Algorithms. Retrieved <https://medium.com/deep-math-machine-learning-ai/chapter-4-decision-trees-algorithms-b93975f7a1f1>
- [4] Praveen Madiraju, "Classification". Retrieved www.mscs.mu.edu/~praveen/Teaching/Sp18/DM/DM-Sp18.html
- [5] Decision Tree-Classification. Retrieved www.saedsayad.com/decision_tree.htm
- [6] Information Gain in decision trees. Retrieved https://en.wikipedia.org/wiki/Information_gain_in_decision_trees
- [7] Wes Doyle. Machine Learning with Python: Decision Tree. Retrieved <https://www.youtube.com/watch?v=XDbj6PxaSf0&t=619s>
- [8] Confusion matrix. Retrieved https://en.wikipedia.org/wiki/Confusion_matrix
- [9] Chinmay Pradhan. What are the differences between ID3, C4.5 and CART? Retrieved <https://prezi.com/lnwwqmsvalun/using-data-mining-for-bank-direct-marketing/>
- [10] StatisticsSolutions, "what is logistic Regression". Retrieved <https://www.statisticssolutions.com/what-is-logistic-regression/>
- [11] MuleSoft, "Data Transformation". Retrieved <https://www.statisticssolutions.com/what-is-logistic-regression/>
- [12] Jason Brownlee, "Logistic Regression for Machine Learning", April 1, 2016. Retrieved <https://machinelearningmastery.com/logistic-regression-for-machine-learning/>
- [13] "Accuracy and Precision". Retrieved https://labwrite.ncsu.edu/Experimental_Design/accuracyprecision.htm
- [14] William Koehrsen, "Beyond Accuracy: Precision and Recall", Mar 2018. Retrieved <https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>
- [15] William Koehrsen, "Beyond Accuracy: Precision and Recall", Mar 2018. Retrieved <https://towardsdatascience.com/beyond-accuracy-precision-and-recall-3da06bea9f6c>