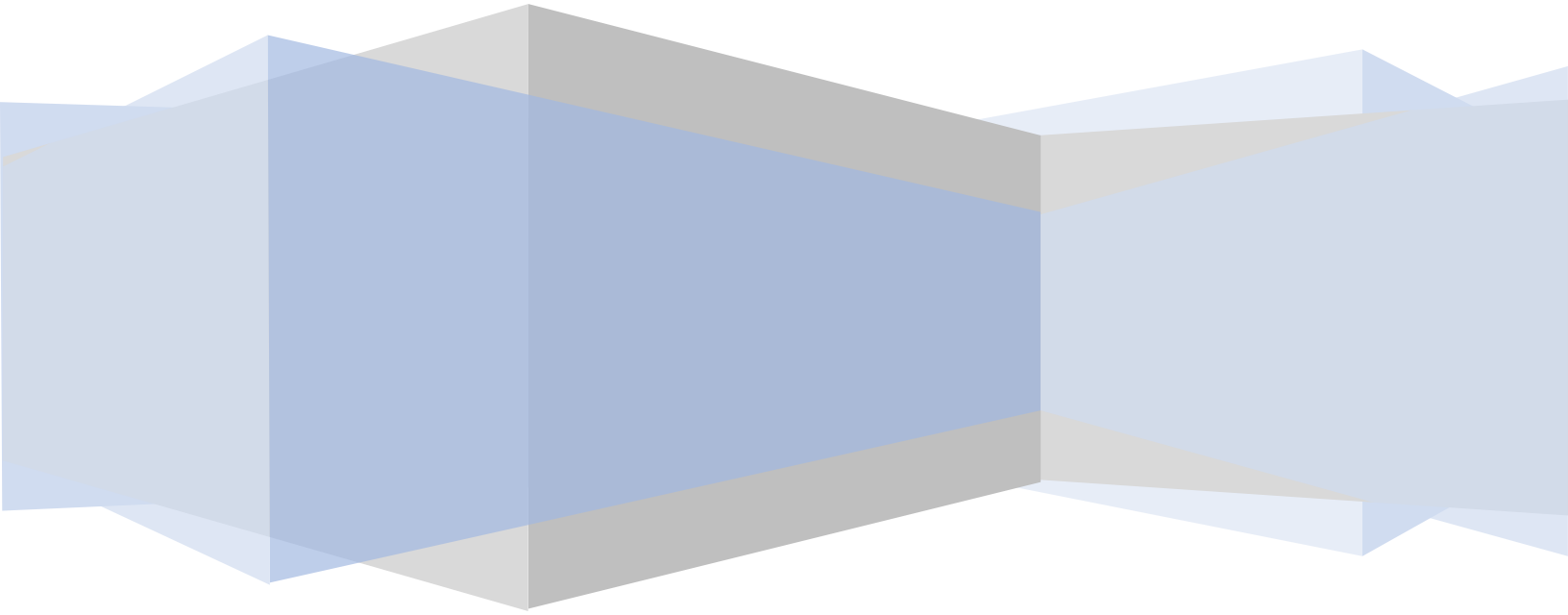
Data Mining

Project Midterm Report

Bank Deposit

Zhang, Zixun
Hu, Wengliang
Zhao, Miaomiao



Abstract

Since the increasing volume of the data, it’s impossible for s human to analysis the data sources and come up with any prediction in order to bring more benefits for the company such as bank marketing. Data mining has extremely improved the performance of these campaigns. This paper introduces the most critical techniques of data mining to analysis data and predict the business decision. Here we implement Decision tree (C5.0), Logistics Regression (LR), Naïve Bayes (TAN) on a real-world data of bank deposit subscription. Python is a widely used high-level programming language. It belongs to the universal programming language. As an interpretive language, Python's design philosophy emphasizes the readability and concise syntax of the code (especially the use of space indentation to divide code blocks instead of braces or keywords). The experimental results of the performances are indicated by three statistical measures; classification accuracy, sensitivity, and specificity. The higher the accuracy is, the better the model is.

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

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. 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.

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

II. DataSet and Features

Here is the web link of the dataset:

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

The data is related with direct marketing campaigns of a Portuguese banking institution. The marketing campaigns were based on phone calls. Often, more than one contact with the same client was required, in order to access if the product (bank term deposit) were (or not) subscribed. The bank direct marketing data set contains 45211 number of samples with 17 attributes without missing values. 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

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 weather 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 prediction accuracies of training and test samples by Naïve Bayes model are 89.16% and 88.75%. 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].

IV. Method

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

- Decision Tree

  Compared to the naïve Bayes algorithm, the decision tree construction process does not rely on domain knowledge. It uses attribute selection metrics to select the best division of tuples into attributes of different classes. The construction of the so-called decision tree is to determine the topological structure between attribute attributes by attribute selection metrics.

  The mean idea of the decision tree algorithm that begin to a feature that cannot be directly classified. The classification result does not get the prefect effect, but the classification can make problem size smaller and easier to classified by this classification. Then repeat this process for the last subset of sorted samples. Finally, after several layers of decision tree, we will get pure subset.

- Naïve Bayes

  The classification principle of the Bayesian classifier is to use the Bayesian formula to calculate the posterior probability through the prior probability of an object, that is, the probability that the object belongs to a certain class and select the class with the maximum posterior probability as the object. The class it belongs to. There are four main types of Bayesian classifiers currently studied: Naive Bayes, TAN, BAN, and GBN. We will use the TAN (naïve Bayes).

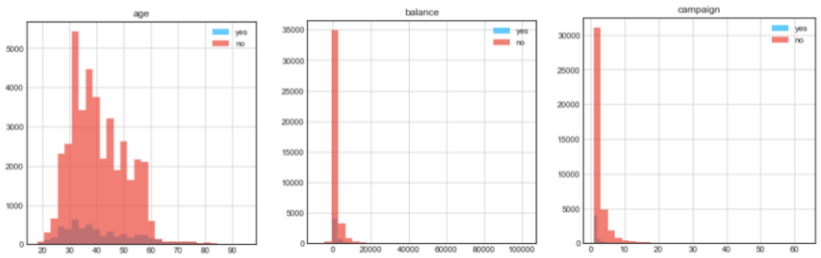
V. Data Analysis and Prediction

- Data Clean and transformation
- 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")
```



Graph 3.1 Histogram Ages/Balance Campaign vs Deposit

From the Table 3.1, 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.

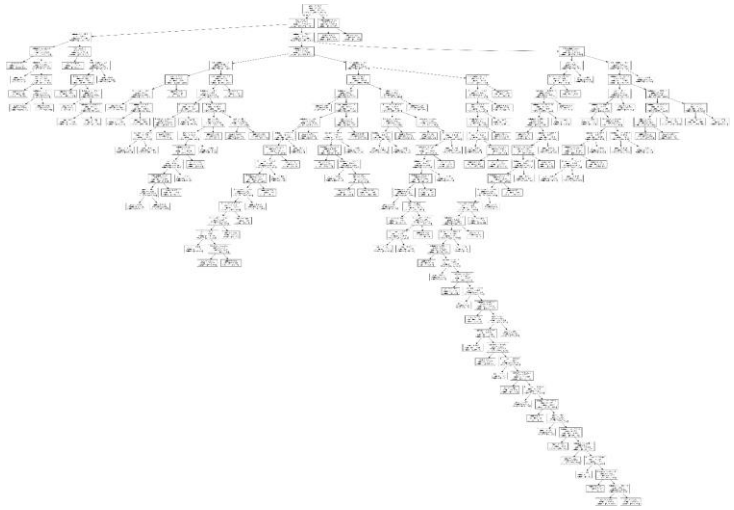
- Building Model

  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.

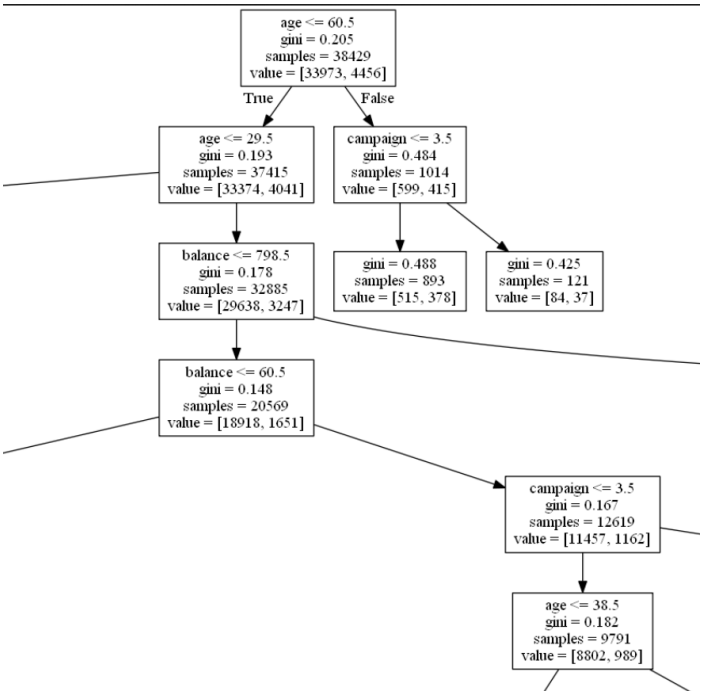
```
In [11]: c=DecisionTreeClassifier(min_samples_split=1000)

In [12]: features=["balance","age","campaign"]

In [13]: x_train=train[features]
         y_train=train["y"]

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

In [14]: dt=c.fit(x_train, y_train)

In [15]: 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)


In [16]: 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')
```

- Result

  Up to now, we have built the decision tree for this bank deposit dataset, the following Graph 3.2 Decision Tree and Graph 3.2 Details of Decision Tree.



Graph 3.2 Decision Tree



Graph 3.2 Details of Decision Tree

VI. Discussion

VII. Conclusion

VIII. Plan for Final Report

In the next step, we will use the decision tree what we get in the midterm to predict. According to result, we will add more related features (housing, load, education, and so on) to make the decision tree better and the prediction precise… We would like to compare different methods which can predict the class, so we want to use Logistics Regression to build model, that would be help us to improve our model.

IX. 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] Sergio Moro and Paul M. S. Laureano, Using Data Mining for Bank direct marketing: an application of the CRISP-DM methodology. Retrieved https://pdfs.semanticscholar.org/a175/aeb08734fd669beaffd3d185a424a6f03b84.pdf

[3] A. Floares., A. Brilutiu. “Decision Tree Models for Developing Molecular Classifiers for Cancer Diagnosis”. WCCI 2012 IEEE World Congress On Computational Intelligence June,10-15, 2012- Bresbance, Australia.

[4] Guoxun Wang, Liang Liu(2010). Predicting Credit Card Holder Churn in Banks of China Using Data Mining and MCDM. Retrieved http://ieeexplore.ieee.org/document/5615798/