An Empirical Investigation of Bank Marketing Data by Using Machine Learning

Submitted by Md. SALIMULLAH (ID:18-37541-1)

COURSE: DATA WAREHOUSING AND DATA MINING SUPERVISED BY **DR. MD. MAHBUB CHOWDHURY MISHU**

Table of Content:

1. Introduction

2. Dataset and Features

3. Methods

4. Results and Analysis

5. Conclusion and Future Work

**Introduction:**A bank is a financial institution that accepts deposits from the public and creates a demand deposit while simultaneously making loans. A deposit is a financial term that means money held at a bank and it can refer to a portion of money used as security or collateral for the delivery of a good. So, how many clients are deposited in a bank daily matter a lot. In a bank, every client has their previous transition record and their bank records. By using this data, we can use machine learning to predict a client will deposit in a bank or not. It will help a lot in a bank to increase their profit. In this paper, we examine best machine learning technique to classify. To get the best result we used a variety of metrics, including classification accuracy and required dataset so that we can get real time predict result correctly. So, I used Decision Tree technique to predict a client is deposit in a bank or not.

**Dataset and Features**:

In this report I have used a data set “Bank Marketing Data - A Decision Tree Approach” from Kaggle database. The aim of this attempt is to predict if the client will subscribe (yes/no) to a term deposit, by building a classification model using Decision Tree. In this data set there are 11163 data points, each containing 17 features. In the dataset, each time-step is labeled with 6\17 different activities that the subjects were engaged in. **The 17 activities are the following: age, job, marital, education, default, balance, housing, loan, contact, day, month, duration, campaign, pdays, pervious, poutcome and deposit**. Summary of data Categorical Variables: 1. job: admin, technician, services, management, retired, blue-collar, unemployed, entrepreneur, housemaid, unknown, self-employed, student 2. marital: married, single, divorced 3. education: secondary, tertiary, primary, unknown 4. default: yes, no 5. housing: yes, no 6. loan: yes, no 7 deposit: yes, no 8. contact: unknown, cellular, telephone 9. month: jan, feb, mar, apr, may, jun, jul, aug, sep, oct, nov, dec 10. poutcome: unknown, other, failure, success Numerical Variables: 1. age 2. balance 3. day 4. duration 5. campaign 6. pdays 7. previous. In this dataset, I don’t do any data preprocessing and cleaning because there was no missing data and extreme outlier. Also, it is balanced dataset because of each class have almost equal samples and normalized. I combined each samples data into a single matrix and divided that into a training and test set. I used 80% of our data for training and 20% for testing. To avoid over-fitting on only a certain subset of classes, I randomly split the data between training and testing. To train our models. I then used this optimized model and analyzed its performance on the test set.

**Methods:**

In this paper, I used Decision Tree for short DT (Weka Tool J48). Decision tree is one of classification technique in data mining that uses branches method to depict each feasible result of a decision making in each possible outcome. Three kinds of node that frame an established a tree which a tree required to have ‘root node’, ‘internal node’, and ‘leaf’ comprises in DT. a set of data into smaller and smaller subsets while at the same time an associated decision tress is incrementally build in decision tree. Root node known as initial attribute or the topmost decision node in a tree which corresponds to the best predicator for a tree to make decision making that have zero incoming and outgoing edges. Both incoming and outgoing edges at least one is allowed in case of internal nodes. Followed by leaf node which has no outgoing edges represents a classification or decision. DT learn from data to approximate a sine curve with a set of IF-THEN rules and used for decision making. The deeper the tree, the more complex DT can be in decision rules and the fitter the model. Also, based on the complexity of a tree will tend to affect the result of accuracy for a tree to do decision making. Ac-cording to DT much more convenient to do classification when it involved decision making, instead of able to compute both categorical and numerical data, it easily accessible and interpreted, involved less calculation, capable to illustrates relationship be-tween dependent and independent variables and computationally low end. For document auto classification, DT is suitable to be applied into a simple framework that setting a set of rules and used for decision making to classify document based on its con-tent into its category. In this study a quick comparison table for all selected classification technique is called followed that discussed previously.

**Results and Analysis:**

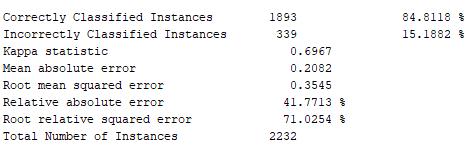


Fig: Accuracy

By using J48 tree classifier 85% accuracy have been found in weka tool.as well as 15% incorrect classified Instances have been found. Mean absolute error is .2082 which measure how close the predictions are to the eventual outcomes. Root mean squared error is .3545 represents the sample standard deviation of the differences between predicted values and observed values.

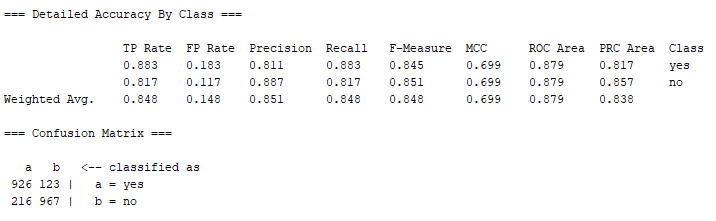


Fig: Accuracy by Class

Confusion matrix is used in supervised learning techniques which is a visualization tool typically. The confusion matrix for decision tree is presented above and also shown the detailed accuracy for each class. One advantage of confusion matrix is that it is easy to see if the system is confusing two classes. The target variable was either "Yes" or "No;" the right side of the matrix tells you that column "a" is yes, and "b" is no. Here value of “a” 926 means which is kind of correct. And value of a-216 and b-128 are kind of error. The value b-967 is actually b.

**Conclusion and Future work:**

For this project, simple supervised learning method (Decision Tree [Weka Tool J48]) was implemented to predict whether a client would deposit or not in a bank using Bank marketing data. From this dataset, the overall performance for decision tree is around 85% accuracy. So, this model accuracy is sufficient to provide useful predictions about a client would deposit or not in a bank.

In future work, there is a scope to collect more data to get more accurate result as well as to apply some complex models.