Mining Bank Marketing Dataset[*]

Trisha P. Malhotra
Rochester Institute of Technology
tpm6421@rit.edu

Vamsi Chandu Mane
Rochester Institute of Technology
vm6528@rit.edu

Rushik Vartak
Rochester Institute of Technology
rv9981@g.rit.edu

ABSTRACT

Big Data Analysis is nothing but finding useful information from given raw data. It helps generate analytical observation from a large amount of data, which aids in making commercial, logical and useful decisions based on the analysis. Our aim is to perform Campaign Analysis on bank-and-customer based dataset. The Portugal Bank has collected data from all its customers regarding their personal information and whether they have subscribed to the banks term deposit facility or not. We perform data analysis to make the dataset apt for a classification. Then we apply various classification algorithms on the target binary attribute determining whether the customer said Yes or No to subscribing to the bank term deposit. This analysis will help us identify traits of a new customer and foresee if it is worth persuading them to enroll to the facility. One of the best accuracies was achieved by Decision Tree and Random Forest classification techniques.

KEYWORDS

Data Mining, Bank Marketing, Big data analysis, Classification

ACM Reference format:
Trisha P. Malhotra, Vamsi Chandu Mane, and Rushik Vartak. 2018. Mining Bank Marketing Dataset. In *Proceedings of NA, NA, 2018,* 5 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Big data analytics is a big part both the commercial and research industries today. Tons of data are being generated and collected by each passing minute. One can make complete sense of the abstract data and deduce information from it. This is used to make commercial, logical, profitable and better decisions by the industry experts. Analyzing a dataset and predicting the target attribute is what this paper aims to accomplish. In this project, we have used a dataset collected from University of California Irvine's data repository [1]. They provide publicly available large datasets. We selected this dataset for our use as it was the most relatable and relevant topic for today's times. Bank Marketing dataset has records dated 2008 to 2010 collected from a Portugal Bank. This set has 41188 records or instances and 21 attributes. The records are nothing but the marketing phone calls made to customers to confirm whether they would subscribe to the product (bank term deposit) or not. This paper describes the steps performed in carrying out in-depth data mining and analysis on the dataset with the final goal of running classification models.

Section 1 presents the design and architecture of the project. Section 3 talks about our implementation in detail. It entails preprocessing, sampling and final classification phases. Section 4 displays our testing results with accuracies achieved for the different classification algorithms applied to the dataset. They are evaluated using Precision, Recall and F-measure. The paper contains future work planned for this project in the section 5. Finally, the paper is concluded in Section 6.

2 ARCHITECTURE

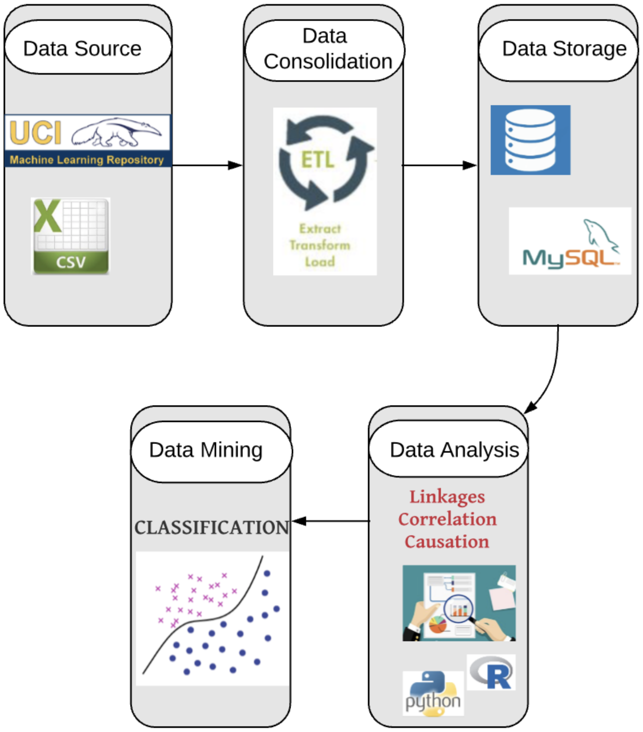


Figure 1: Architecture (PNG).

The architecture of the project is as shown in Figure 1. We obtained the CSV (Comma Separated Values) dataset from the University of California Irvine data repository which provides analysts and learners with many choices of datasets. The next phase is Data Cleaning. Missing values, outliers were removed in the Data Consolidation phase. MySQL was used as the choice of RDBMS to set up the database engine. The original data were divided into

[*] Produces the permission block, and copyright information

normalized data tables, making sure that it follows all rules of Data Integrity and Atomicity.

Consequently, data was visualized using R, Rattle, Python and their libraries, wherein we analyzed data correlations, causations and linkages. We plotted graphs and histograms to demonstrate the distribution of the data attributes across the records. Ultimately, we perform the in-depth campaign analysis and apply classification algorithms to compare their results.

3 IMPLEMENTATION

The project was implemented using Python and R. In the python implementation, the sklearn library was used for implementing the preprocessing functions, PCA, classification models, etc.; the matplotlib library was used to plot the EDA, PCA, ROC plots; and the pandas library was used to read the data file.

3.1 Data cleaning

There are several missing values in some categorical attributes, all coded with the ”unknown” label. There were few possible options to handle them, either treating these missing values as a possible class label which would create unnecessary labels affecting the performance of the model or using deletion or imputation techniques. We chose to delete the records containing such missing and unknown values. After data cleaning, the dataset had around 30,488 records/instances. We would also suggest some ways to reduce the effect of missing values in future work section.

3.2 Exploratory Data Analysis & Preprocessing

Plotting histograms of all attributes gave a sense of the overall skewness and distribution of the data. Figure 2 shows sample distributions of attribute ”age”.

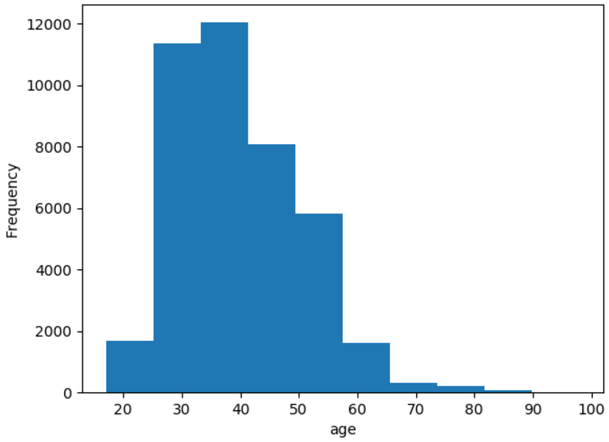
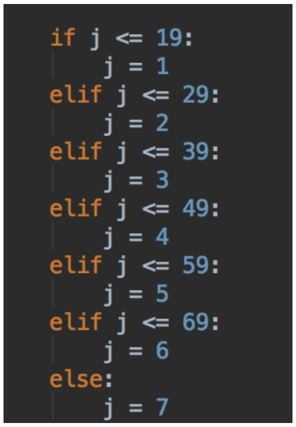


Figure 2: Sample Attribute distribution(age) (PNG).

Similarly, we plotted histograms for all attributes and studied them. The data overall is of both categorical and continuous types. After observing the distributions of the data, we decided to convert some of the continuous data to categorical to reduce the sparsity of the feature matrix. Figure 3 shows an example conversion of continuous data to categorical.



```
if j <= 19:
    j = 1
elif j <= 29:
    j = 2
elif j <= 39:
    j = 3
elif j <= 49:
    j = 4
elif j <= 59:
    j = 5
elif j <= 69:
    j = 6
else:
    j = 7
```

Figure 3: Categorization of Age attribute

The modified attribute was repeatedly adjusted to reach a decent distribution and then the data was normalized using the min-max method. Then we performed PCA on the data to reduce the number of attributes to consider while building a model and found that 99 percent of the variance is distributed between 2 attributes. We figured out that attribute ”duration” has a misleading effect on our model.

For example, the call duration attribute indicates the duration of the call with the customer. The model we try to build would predict whether a person would accept for a term deposit or not. Also, for us to get a value for duration attribute, representative have to complete the call with the customer. The longer the duration of the call, the higher the chance that the customer is interested to make a term deposit. In a way, it would make sense that call duration highly affects the model prediction. But the customer would let representatives know his/her decision by the end of the call i.e we would get value for the duration of the call only after getting the decision from the customer. We decided to drop the attribute ”age” from the model. Next, we performed PCA again on the data and removed unwanted attributes before building the model. And also we decided to take features that contribute to 99.9 percent of the variance. The figures 4, 5, 6 show the cumulative curve of variance ratio and the number of features for few scenarios.

With the help of the correlation matrix and the PCA covariance ratios, we decided on the 15 attributes we would consider for building a classifier model.

3.3 Sampling

The dataset was divided into training and testing data by sampling using a 80:20 split where 80% of the data was used as training data and the rest of the 20% of data was used for testing. As the dataset contains just 4640 instances with positive outcome out of the total 41188 instances, i.e. just 11.26% of the data has a positive outcome; stratified sampling was used to reduce the sampling errors in order to improve the precision.

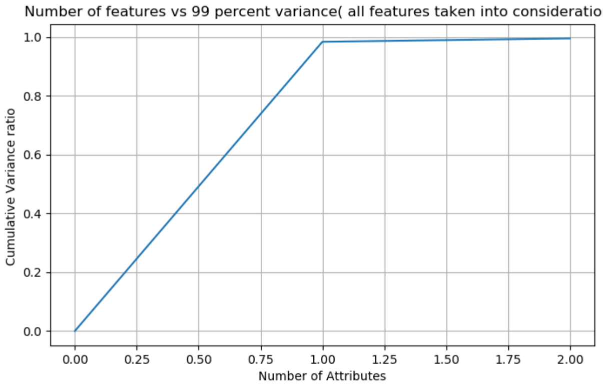


Figure 4: PCA on all the given attributes(99 percent variance)

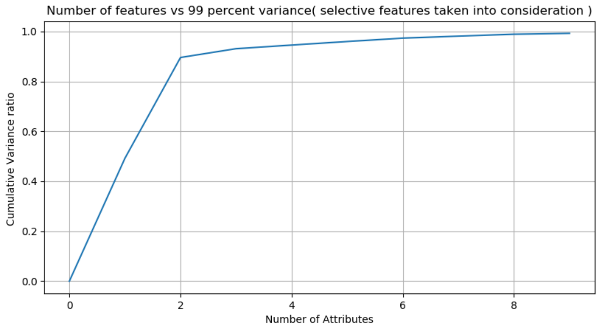


Figure 5: PCA on selective attributes (99 percent variance)

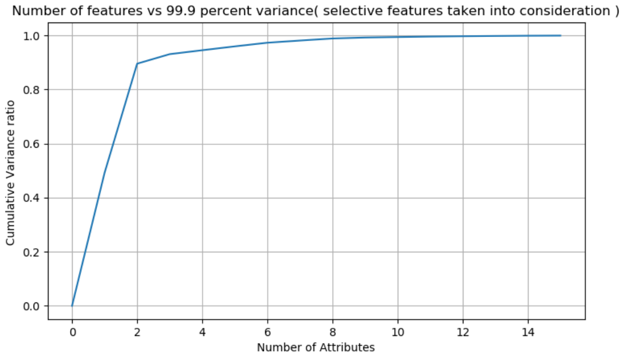


Figure 6: PCA on selective attributes (99.9 percent variance)

3.4 Classification

The classification algorithms used in this project are Nearest Neighbors classifier, Naive Bayes classifier, Decision Tree classifier, Random Forest classifier, Linear SVM classifier, RBF SVM classifier, Neural Network classifier, and AdaBoost.

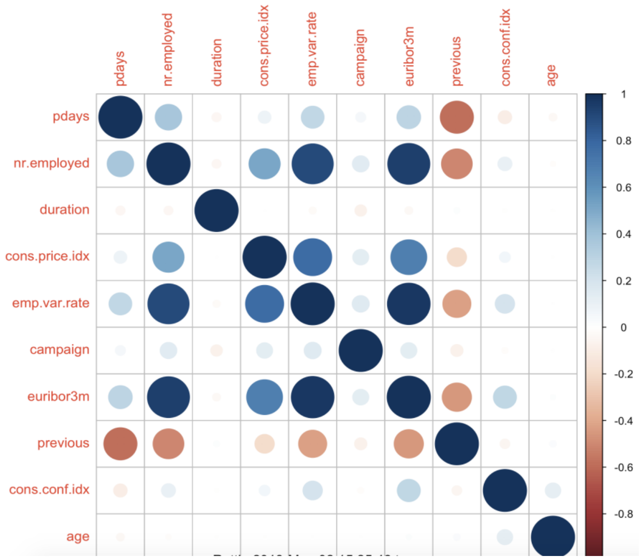


Figure 7: Pearson Correlation (PNG).

3.5 Evaluation Metrics

The metrics used for evaluation of the classification models are Accuracy, Precision, Recall, F-measure, AUC score, and ROC curve. In this project, the Accuracy measure specifies the percentage of clients that were correctly predicted as clients that would buy or not buy the product. The Precision measure specifies the percentage of clients from the predicted buyers that actually bought the product. The Recall measure specifies the percentage of clients from the actual buyers that were predicted as buyers. If the company decides to use a classification algorithm to call only the clients predicted as buyers, then the greater the Precision the less will be the number of clients who denied buying the product when they were called during the current campaign; and the greater the Recall the less will be the number of actual buyers that weren't called during the current campaign.

4 EXPERIMENTATION AND RESULTS

Figure 8 displays the final decision tree that is used for binary classification. The important attributes used are duration of the call, previous outcome, and number of employees.

The first experiment was performed on the preprocessed dataset (having numerical values, but no instances with missing values were removed). Table 1 shows the performance metrics and Fig. 9 shows the ROC curve for all the classification algorithms used in this experiment. In this experiment, the Decision Tree classifier performs the best with an accuracy of 0.91, F-measure of 0.57 and AUC score of 0.73.

In the second experiment, the instances with missing values were removed as there were 10700 out of 41188 instances that had missing values. This left us with 30488 instances which used to evaluate the performance of the selected classification algorithms

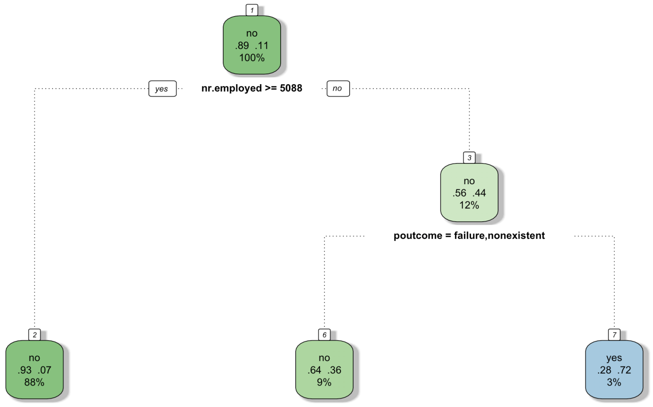


Figure 8: Decision Tree.

Table 1: Performance metric for experiment 1

Algo	Accuracy	Precision	Recall	F-Measure
Nearest Neighbors	0.89	0.54	0.41	0.47
Naive Bayes	0.86	0.42	0.54	0.47
Decision Tree	0.91	0.65	0.50	0.57
Random Forest	0.90	0.70	0.18	0.28
Linear SVM	0.90	0.67	0.19	0.30
RBF SVM	0.91	0.66	0.35	0.46
Neural Net	0.91	0.69	0.32	0.43
AdaBoost	0.91	0.67	0.39	0.49

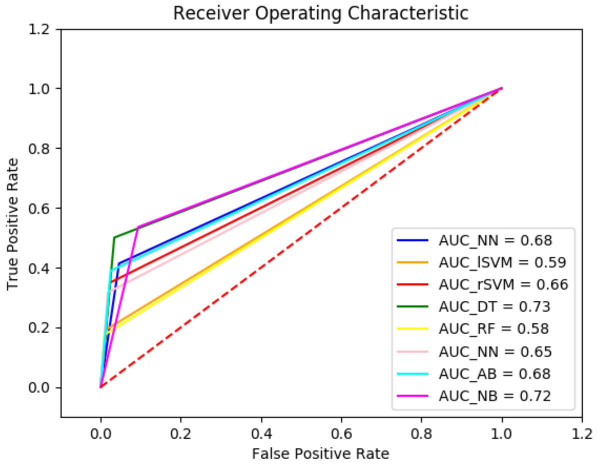


Figure 9: ROC curve for Experiment 1.

in this experiment. Table 2 shows the performance metrics and Fig. 10 shows the ROC curve for all the classification algorithms used in this experiment. Even in this experiment, the Decision Tree classifier performs the best with an accuracy of 0.90, F-measure of 0.56 and AUC score of 0.73.

Table 2: Performance metric for experiment 2

Algo	Accuracy	Precision	Recall	F-Measure
Nearest Neighbors	0.88	0.56	0.43	0.49
Naive Bayes	0.66	0.25	0.81	0.39
Decision Tree	0.90	0.64	0.51	0.56
Random Forest	0.88	0.81	0.13	0.22
Linear SVM	0.88	0.71	0.19	0.30
RBF SVM	0.89	0.67	0.31	0.43
Neural Net	0.89	0.69	0.29	0.41
AdaBoost	0.89	0.65	0.39	0.49

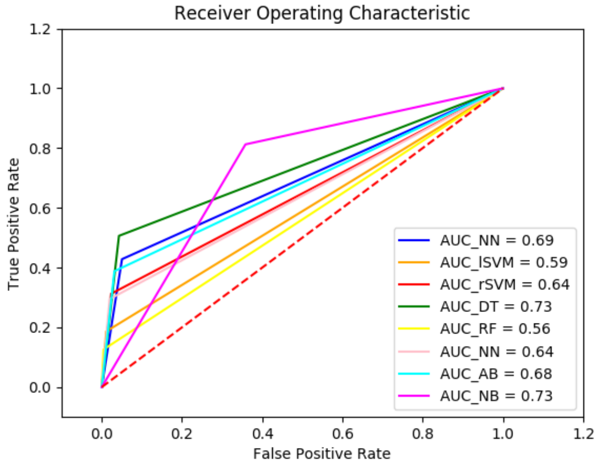


Figure 10: ROC curve for Experiment 2.

In the third experiment, the instances with missing values were removed as well as the attribute d́urationẃas removed. The duration attribute specifies the call duration of the current campaign. The outcome strongly depends on the call duration, as if the client is interested in buying the product he/she tends to be on the call for a longer duration; but the value of this attribute cannot be determined until the call is disconnected, and when the call is disconnected then the outcome is already known. Thus, this attribute was removed in this experiment. Table 3 shows the performance metrics and Fig. 11 shows the ROC curve for all the classification algorithms used in this experiment. In this experiment, the Nearest Neighbor classifier and the Decision Tree classifier perform the best with an accuracy of 0.87 and 0.88 respectively, F-measure of 0.38 and 0.35 respectively, and AUC score of 0.63 and 0.61 respectively.

In last experiment, the instances with missing values were removed as well as the attribute d́urationẃas removed, and in addition stratified sampling was used to split the data into training and testing data. As the dataset contains just 11.26% of the data with a positive outcome; stratified sampling was used. Table 4 shows the performance metrics and Fig. 12 shows the ROC curve for all the classification algorithms used in this experiment. In this experiment, the Nearest Neighbor classifier, the Decision Tree classifier

Table 3: Performance metric for experiment 3

Algo	Accuracy	Precision	Recall	F-Measure
Nearest Neighbors	0.87	0.53	0.30	0.38
Naive Bayes	0.63	0.23	0.77	0.35
Decision Tree	0.88	0.66	0.24	0.35
Random Forest	0.88	0.73	0.11	0.19
Linear SVM	0.87	0.60	0.11	0.18
RBF SVM	0.88	0.63	0.19	0.30
Neural Net	0.88	0.67	0.10	0.18
AdaBoost	0.88	0.66	0.19	0.30

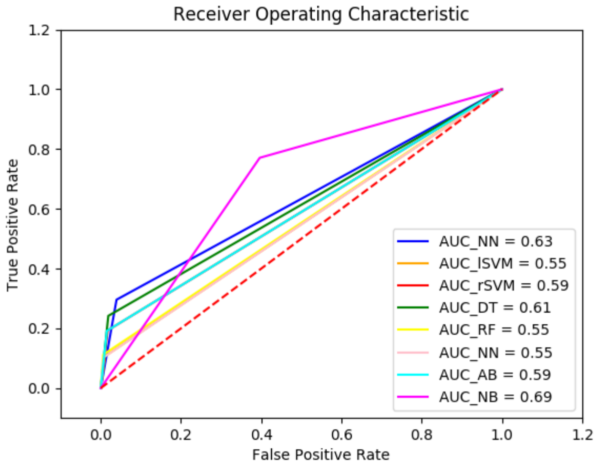


Figure 11: ROC curve for Experiment 3.

Table 4: Performance metric for experiment 4

Algo	Accuracy	Precision	Recall	F-Measure
Nearest Neighbors	0.87	0.49	0.31	0.38
Naive Bayes	0.62	0.22	0.77	0.34
Decision Tree	0.88	0.59	0.25	0.35
Random Forest	0.88	0.64	0.12	0.21
Linear SVM	0.88	0.64	0.06	0.12
RBF SVM	0.88	0.62	0.23	0.34
Neural Net	0.88	0.72	0.13	0.22
AdaBoost	0.89	0.64	0.22	0.32

and the RBF SVM classifier perform the best with an accuracy of 0.87, 0.88 and 0.88 respectively, F-measure of 0.38, 0.34 and 0.35 respectively, and AUC score of 0.63, 0.59 and 0.61 respectively.

In all the above experiments, based on the Accuracy, F-measure and AUC score; Decision Tree classifier performs really well. The Nearest Neighbor and the RBF SVM classifier also perform well when stratified sampling is used and the attribute ´durationís not considered.

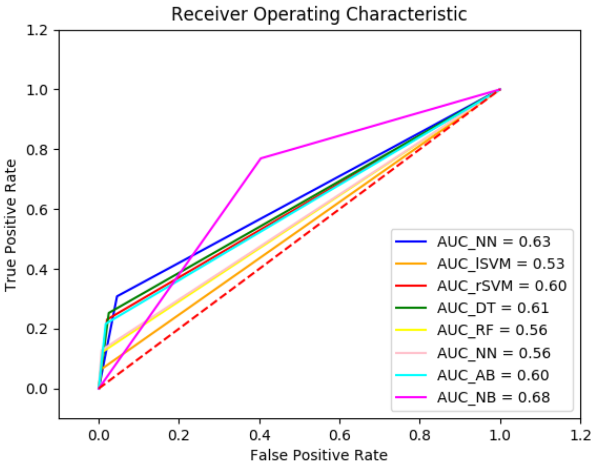


Figure 12: ROC curve for Experiment 4.

In all the experiments, the Naive Bayes classifier gives the highest recall measure; thus, Naive Bayes classifier can also be used when the company does not concern with the number of calls they make but are very concerned about not losing (not calling) clients who would actually buy their product.

5 FUTURE WORK

The tasks that can be carried out as an extension to this project are as follows.

- To impute missing and unknown values in the data, either a regression model in case of continuous type attribute or a classification model in case of categorical attribute can be used to predicts the missing values .
- Discovering potentially useful patterns using Apriori and performing Frequent pattern mining is another future task. This could help the bank to better understand the set of attributes that play a major role in accepting to make a term deposit.

6 CONCLUSION

Based on the experiments performed, the classification algorithms like Decision Tree classifier, Nearest Neighbors classifier and RBF SVM classifier will be the best choices if the company intends to increase the efficiency of the campaign (more number of clients buying the product in less number of calls made during the campaign). If the company intends to just increase the number of clients buying the product irrespective of the number of calls made during the campaign, then the Naive Bayes classifier can be used; which will make sure that most of the clients that would subscribe to the product (here the bank term deposit facility) are targeted and very few of such clients are lost (not called).

REFERENCES

[1] 2012. UCI Bank Marketing Data Set. uci.edu. (July 2012). https://archive.ics.uci.edu/ml/datasets/Bank+Marketing.