Portuguese Bank Marketing Dataset

# Introduction

Marketing campaigns are a common practice within the banking industry. This research uses a dataset provided from the UCI Machine Learning repository [5] about a Portuguese bank. This is basically a binary classification problem. The purpose of this research paper is to provide an effective telemarketing strategy to sell term deposits to the bank’s clients.

Classification models (such as Decision Tree, Naïve Bayes, Random Forest, K-Nearest Neighbour, Support Vector Machine, Neural Network, and Logistic Regression) are used to predict the outcome that a bank customer will accept a term deposit. The performance evaluation of these algorithms will be verified using ROC (Receiver Operating Characteristics) curve and/or AUC (Area Under the Curve).

# Literature Review

We wanted to review literature that showed which models were used to solve a binary classification problem, and in particular relating to bank data if possible. Other secondary questions include how was imbalanced data handled, and what models or processes were used to select features (dimensionality reduction).

In paper [1] Moro et al discussed how to use data mining techniques (the Cross-Industry Standard Process for Data Mining) when formulating models that could explain the success of a contact i.e. if the client subscribes to the deposit. This article is one of the earliest ones on the Portuguese bank marketing dataset. The authors used three different models (Naïve Bayes, Decision Trees and Support Vector Machines). All three models were tested and validated with a holdout split of a training set (2/3) and test set (1/3). They then ran 20 iterations for each model. ROC curves and Lift analysis were performed to compare the models. The top five attributes were given (call ‘Duration’, ‘Month’ of Contact, Nr of ‘Previous’ Contacts, Days Since Last Contact or ‘pdays’, and Last Contact Result or ‘poutcome’). Their results were good, with a minimum AUC of 87%. However, they did not discuss rebalancing, normalizing, discretizing, or what to do with any outliers and “unknown” categories.

In paper [2] Ladyzynski et al tried to identify models for identifying customers interested in credit products for a large bank in Poland. The data was time-series dependent and so they used a 3-month window of data. The three machine learning algorithms they used were: 1) classification trees - CART, 2) Random forests and 3) Deep belief networks – DBN. The authors discussed ‘dimensionality reduction’ and that it can fall into two categories: feature extraction (i.e. Principal Component Analysis) and feature selection - by selecting a subset of the original features (i.e. by using an algorithm called Boratu). They then used the Boratu algorithm and decreased the number of attributes from 616 to 164.

In paper [3] the author discusses four classification models that will be used on the Portuguese bank data to determine which models work well, and what factors are the most important. The models used are multilayer perception neural network (MLPNN), augmented Naïve Bayes (TAN) or Bayesian Networks, logistic regression, and decision tree model C5.0. The author starts with the same dataset we use and seems to leave it as is as there is no more discussion on modifying the data. Further, the author did not discuss rebalancing, normalizing, discretizing, or what to do with any outliers and “unknown” categories. The author then shows a table of attributes vs. model to show what the attributes were rated as, thereby showing the highest-rated attributes. These were Duration (highest in 3 models), Job, Marital, Education, Poutcome, and possibly Age. The author used classification accuracy, sensitivity and specificity to measure the performance of the models. The models performed well, with a minimum accuracy of 89.16% on the training set and 88.75% on the testing set for TAN, and the other 3 models performing even better.

In paper [4], the authors discuss using the same Portuguese bank data with models used in a WEKA package. The two classification models used are Bayes Net algorithm and Naïve Bayes algorithm. The authors briefly discuss the dataset attributes but then move on quickly to the models. They used 10-fold cross validation for the training and test sets. Also, they did not discuss rebalancing, normalizing, discretizing, or what to do with any outliers and “unknown” categories. Further, they did not discuss what attributes worked the best. Although the authors use WEKA to run the models, it was useful to know what models were actually used. The evaluation metrics they used were the Confusion Matrix, Kappa, TP Rate, FP Rate, Recall, Precision, F-Measure and ROC Area. Both models performed well with Bayes Net having 81.5% accuracy and Naïve Bayes with 78% accuracy.

[note: other literature reviews will be included in further versions]

# Dataset

There are two datasets in the University of California at Irvine Machine Learning Repository [5]. The bank dataset was collected by S. Moro et al [1]. The end result is to obtain a model or models that will provide an effective telemarketing strategy to sell long-term deposit accounts. Data from old and current telemarketing campaigns were obtained as part of this dataset. Multiple contacts were often needed to determine whether a customer would subscribe to a long-term deposit account. The data is described below in Table 1. The steps used are best summarized in the diagram (Figure 1) below:

*Figure 1: Data Pre-Processing*

The ‘full’ dataset has 45,211 examples and is ordered by date from May 2008 to November 2010. The second bank dataset has only 10% of the examples (4,521) randomly selected from the full dataset to allow for testing more computationally demanding machine learning algorithms. We will be using the full dataset as much as possible. The dataset attributes are shown below in Table 1.

Table 1: Data List

|  |  |  |
| --- | --- | --- |
| # | Attribute | Purpose |
| 1 | Age | Age of the customer |
| 2 | Job | Type of job |
| 3 | Marital | Marital status |
| 4 | Education | Education of the customer |
| 5 | Default | Shows whether the customer has credit in default or not |
| 6 | Balance | Average yearly balance in Euros |
| 7 | Housing | Shows whether the customer has housing loan or not |
| 8 | Loan | Shows whether the customer has personal loan or not |
|  |  | ***Related to current campaign:*** |
| 9 | Contact | Shows how the last contact for marketing campaign has been made |
| 10 | Day | Shows on which day of the month last time customer was contacted |
| 11 | Month | Shows on which month of the year last time customer was contacted |
| 12 | Duration | Shows the last contact duration in seconds |
| 13 | Campaign | Number of contacts performed during the marketing campaign for this customer |
|  |  | ***Related to previous campaign:*** |
| 14 | Pdays | Number of days that passed by after the client was last contacted from a previous campaign (-1 means client was not previously contacted) |
| 15 | Previous | Number of contacts performed before this campaign for this client |
| 16 | Poutcome | Outcome of the previous marketing campaign |
| 17 | Y | **Class attribute showing whether the client has subscribed a term deposit or not** |

There is a total of 17 attributes with 7 numeric features and 10 nominal / categorical features (of which four are binary – Default, Housing, Loan and the “Y” target). The purpose of each feature is described in Table 1. The numeric features are Age, Balance, Day, Duration, Campaign, Pdays and Previous. The categorical features are Job, Marital, Education, Contact, Month, Poutcome, and the four binary features previously noted. The next two Tables further detail the attributes to get a better understanding of the data itself. There was no missing data, and no obvious incorrect data.

Table 2 shows the numeric data with the minimum value, Q1, median, mean, Q3, the maximum value and the standard deviation. We will show plots, after the tables, that show that all the numeric data is not normal and is right-skewed. The outliers noted will **NOT** be taken out of the dataset. They were mostly on the upper end, and all will be normalized, when required, for the models that need it. Also,

Pdays contains a “-1” and even though it doesn’t make sense in the definition, and “0” doesn’t either, ***all values of “-1” will be replaced with zero*** (which currently doesn’t exist in this attribute) *knowing that 0 means NO CONTACT and not 0 days.* This attribute can then be normalized (0 to max) for models.

Also, when pdays = -1, previous = 0, which makes sense, but this means that the two variables are correlated when they shouldn’t be. We will see the correlation table further down and note if there are any large correlations. ***In ‘previous’, there is one exceptional last value of 275 that will be deleted****.*

Table 2: Numeric Attributes Information

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| # | Attribute | # outliers | Min | Q1 | Median | Mean | Q3 | Max | Std Dev. |
| 1 | Age | 487 | 18 | 33 | 39 | 40.94 | 48 | 95 | 10.6 |
| 6 | Balance | 4729 | -8019 | 72 | 448 | 1362.3 | 1428 | 102127 | 3044 |
| 10 | Day |  | 1 | 8 | 16 | 15.8 | 21 | 31 | 8.3 |
| 12 | Duration | 3235 | 0 | 103 | 180 | 258.2 | 319 | 4918 | 257.5 |
| 13 | Campaign | 3064 | 1 | 1 | 2 | 2.8 | 3 | 63 | 3.1 |
| 14 | Pdays | 8256 | -1 | -1 | -1 | 40.2 | -1 | 871 | 100.1 |
| 15 | Previous | 8256 | 0 | 0 | 0 | 0.6 | 0 | 275 | 2.3 |

Table 3 shows the categorical attributes and it is interesting to note that there is an “unknown” category in four of these attributes (job, education, contact, and poutcome). The ‘job’ category has 12 classes, which is a lot to discern from. The ‘married’ class has only three classes. Education has four, but one of them is ‘unknown’. Contact is either by ‘cellular’ or regular ‘telephone’, and also ‘unknown’. Month has the 12 months of the year. Poutcome has ‘failure’ and ‘success’ but also ‘other’ and ‘unknown’ again. And finally, the binary attrbutes are either ‘yes’ or ‘no’ for Default, Housing, and Loan.

Table 3: Categorical Attributes Information

|  |  |  |  |
| --- | --- | --- | --- |
| # | Attribute | # of classes | Attribute Classes |
| 2 | Job | 12 | ‘admin’, ‘blue-collar’, ‘entrepreneur’, ‘housemaid’, ‘management’, ‘retired’, ‘self-employed’, ‘services’, ‘student’, ‘technician’, ‘unemployed’, ‘unknown’ |
| 3 | Marital | 3 | Marital status – ‘married’, ‘divorced’, ‘single’ (divorced includes widowed) |
| 4 | Education | 4 | ‘primary’, ‘secondary’, ‘tertiary’, ‘unknown’ |
| 5 | Default | 2 | ‘yes’ or ‘no’ |
| 7 | Housing | 2 | ‘yes’ or ‘no’ |
| 8 | Loan | 2 | ‘yes’ or ‘no’ |
| 9 | Contact | 3 | ‘cellular’, ‘telephone’, ‘unknown’ |
| 11 | Month | 12 | ‘jan’, ‘feb’, ‘mar’ …. ’nov’, ‘dec’ |
| 16 | Poutcome | 4 | ‘failure’, ‘success’, ‘other’, ‘unknown’ |
| 17 | Y | 2 | ‘yes’ or ‘no’ |

Finally, it is best to get a sense of which classes within the attributes dominate, if any. For that, Table 4 below shows the breakdown for any category over 20%. We see that there is definite dominance of categories in quite a few of the attributes. We further checked to see what, if any, dominance it would also have for the target answer of ‘yes’ and found that for ‘Default’ it was 99%.

Table 4: Attributes with Dominant Categories contributing over 20% of the outcome

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| # | Attributes | Category | Dominant category | For same dominant category, what % is for “yes” answer only? |
| 1 | Age | 31-40  41-50 | 39%  25% | 34%  19% |
| 2 | Job | ‘blue-collar’  ‘management’ | 22%  21% | 13%  25% |
| 3 | Marital | ‘married’  ‘single’ | 60%  28% | 52%  36% |
| 4 | Education | ‘Secondary’  ‘tertiary’ | 51%  29% | 46%  38% |
| 5 | **Default** | ‘no’ | **98%** | **99%** |
| 6 | balance | -8,019 to 10,000 | 98% | 97% (range is too large; leaving as is) |
| 8 | Loan | ‘no’ | 84% | 91% |
| 9 | Contact | ‘cellular’  ‘unknown’ | 65%  29% | 83%  10% |
| 11 | Month | ‘may’ | 30% | 17% |
| 12 | Duration | 0 to 300 | 73% | 35% |
| 13 | Campaign | 1 to 10 contacts | 97% | 99% (range is too large) |
| 14 | Pdays | -1 -> 0 | 81.7% | 64% |
| 15 | Previous | 0 to 25 | 99.9% | 99.9% (range is too large) |
| 16 | Poutcome | ‘unknown’ | 82% | 64% |
| 17 | Y | **‘yes’** | 89% | n.a. |

Based on the above information, the attribute ‘Default’ will be deleted from the dataset as the category ‘no’ is extremely high at 98% of all instances. ‘Balance’, ‘Campaign’ and ‘Previous’ will not be removed as we realized the range we used was too large. This Table was created using Excel (only one) and so doing this was laborious. We note that the ‘target’ is highly imbalanced, with the “yes” answer only 12% of the responses.

The visualization of the data is next. All tests and plots were performed in R code.

Numeric data

In order to perform numerical analysis, the data is presumed “normal”. This implies that the data has a bell-shaped curve (not skewed), and does not have multicollinearity (the dependent variables are not correlated). In reviewing the data in R, we did normality tests, boxplots, histograms, and a correlation map for the numeric data. Q-Q plots were also run in R (not showed here) and they distinctly show that the data is not normal.

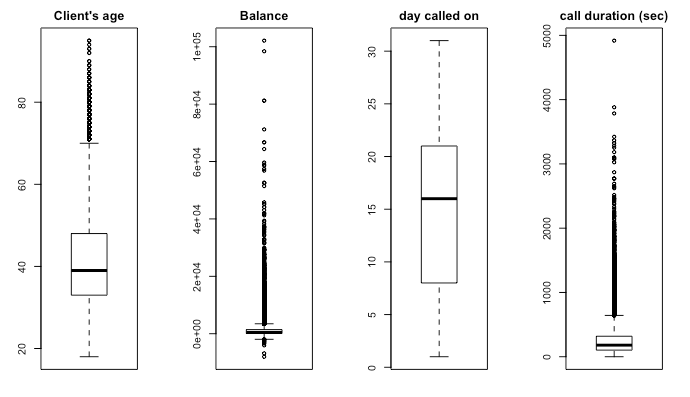
As well, the normality of each numeric attribute was tested using the Shapiro-Wilk test (not shown), and all tests failed, meaning that the numeric attributes were not normal (any value above 0.05 indicates normality). Therefore, the data will be normalized or scaled only for the models that require it. The histograms also show the skewness of all the numeric attributes.

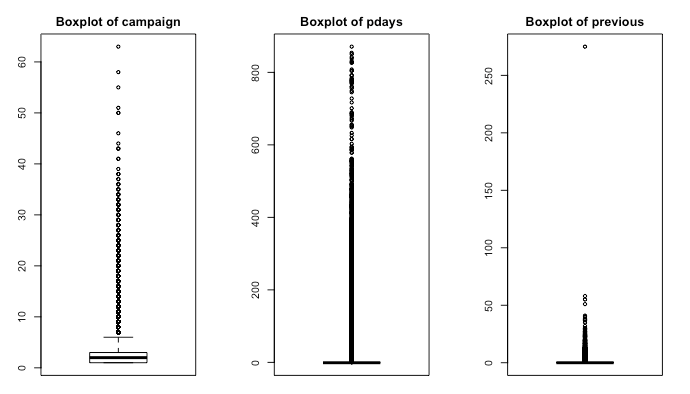
Categorical data

As well, correlations were performed for the categorical data using Pearson’s Chi-squared test and the tests showed that the data was relatively unrelated to each other. The work was done in R and is not shown here.

The boxplots of the numeric data are shown below in Figure 2. From these plots, we can see that the data is not normal, based on the size and location of the boxes. One can see the outliers such as in ‘previous’.

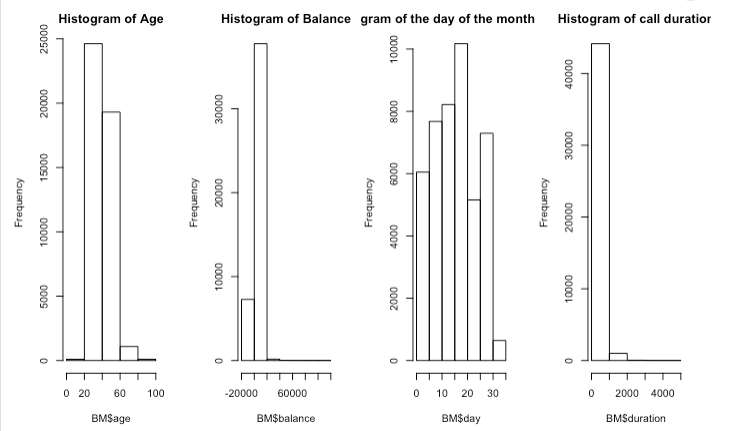
*Figure 2: Boxplots of Numeric data*

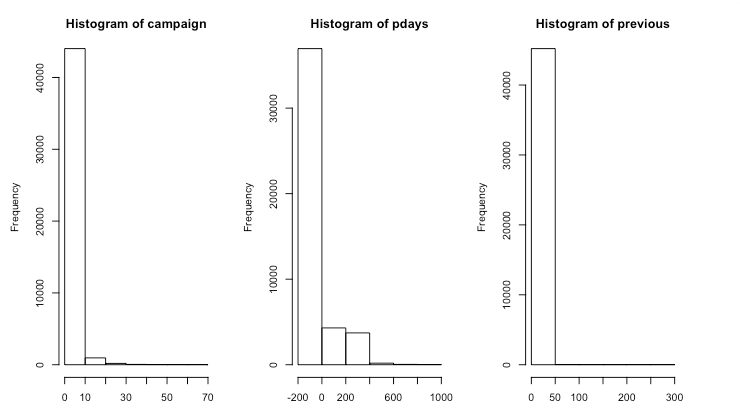




The histograms of the numeric data are shown below in Figure 3. It is obvious from the histograms that all the numeric data is not normal and all were skewed to the right, however this can often happen with real data. It is however required for some models, such as regression, that the data is normal. Hence, the numeric data will be normalized or scaled, where required, for the models that need that (i.e. k-NN or k-means). Notice that for many of the attributes, a large portion of the instances are on the far left.

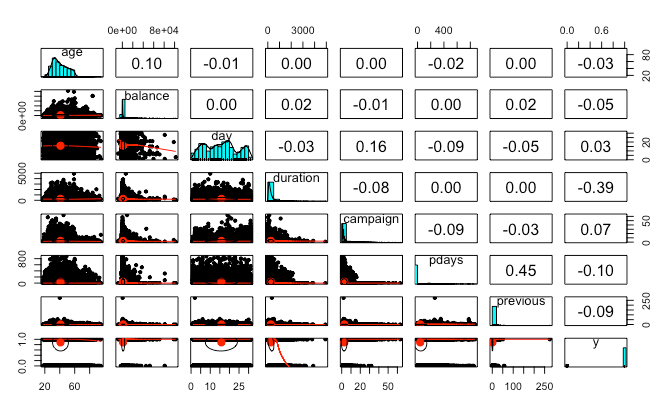
*Figure 3: Histograms of Numeric data*





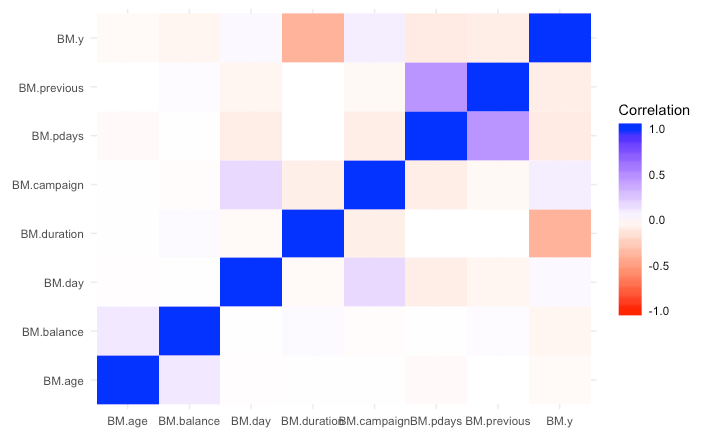
We have a screenshot of a plot from R of the numeric data below in Figure 4. The upper right area shows the correlations, the diagonal shows the histograms, and the bottom left area shows the scatterplots with the circles showing the strength of correlation (circle means a little coorelation, and the oval means a lot). Correlations were performed for the data to ensure there was no multicollinearity – but also to see if there was any correlation between the data and the result, ‘Y’. By reviewing the correlations, we see that the data has low correlations with each other (all under 0.16, one at -0.39) except for pdays to previous at 0.45, which is still an acceptable correlation. Note that the data does not have strong correlations to the “y” answer!

*Figure 4: Correlations/Histograms/Scatterplots of Numeric data*



A correlation heat map of the numeric data was also generated for a visual representation of the correlations (see below). Upon review, the data showed that the correlations were all very low, as above.

*Figure 5: Correlation Heat Map of Numeric data*



To summarize, the value of ‘-1’ has been replaced with ‘0’ in pdays, the value of ‘275’ has been deleted from ‘previous’, and the attribute (#5) ‘default’ has been deleted. The dataset has now been adjusted and will be used ‘as is’ where possible in the models. However, if any conversions need to be made, they are discussed in Step 1 of the next section. The R source code is stored in Github for the Dataset section [see link – <https://github.com/njeanius/bank1/tree/draft> ].

# Approach

The block diagram is shown below. We used a modified version of the CRISP [2,3] methodology. Because the data steps were covered above under Dataset section, they were taken out of this diagram.

The data is a binary classification problem in that (some of) the inputs (both numeric and nominal) should be able to predict the outcome of the class variable (Yes or No) based on the respective models. The steps are broken down as follows.

*Figure 6: Diagram of steps in Modeling Data*

COMPLETED TO HERE….the rest is in draft form, and may be incorrect at this time…

## Step 1: Pre-Processing

Note: not sure yet of the order in which I want to do this as it creates too many datasets….

## Step 1A: Rebalance the data but also keep Unbalanced set

Because the data was imbalanced, with a 89% “no” and only 11% “yes” in the target, rebalancing should be done. The various methods to do this include:

1. Undersampling the majority class
   1. Random
   2. Informative (EasyEnsemble and BalanceCascade)
2. Oversampling the minority class
3. Under- and over-sampling
4. Synthetic Data Generation = ROSE, SMOTE, SMwR? with bootstrapping and k-nearest neighbors
5. Cost Sensitive Learning- choose a classifier with lowest total cost where Total Cost = C(FN)xFN + C(FP)xFP

ROSE (Random Over Sampling Examples) package helps us to generate artificial data based on sampling methods and smoothed bootstrap approach. This package has well defined accuracy functions to do the tasks quickly.

WEKA:

1. the ClassBalancer function, (best in Weka -what is it in R?)
2. SMOTE or Synthetic Minority Oversampling Technique
3. Cost sensitive or “error correct”

Under-sampling reduces the majority class. The sampling can be done 1) randomly or 2) informatively. For the second method, there is EasyEnsemble or Balance Cascade. EXPLAIN.

Over-sampling increases the minority class. The sampling can be done 1) randomly or 2) informatively. With the second method, overfitting can result along with reduced accuracy.

A combination of the two can also be done just by changing a parameter in the R code.

For synthetic data generation, a well-known technique is SMOTE (synthetic minority oversampling technique). Other methods are bootstrapping combined with k-NN, ROSE>>?? We will use SMOTE, ROSE…

For cost-sensitive learning, we don’t create any balance data, but basically “punish” the model using a cost or confusion matrix. We usually punish for getting a FP (?) and this is the worst case because we wouldn’t want to give a client any credit if they end up bankrupt. On the other hand, if we miss giving a client, who would use the money wisely, is not as big a deal. Discuss the process.

We also have to think about the training and test set. We could do a 70/30 split using sampling technique, and verifying the sample is the same percentage for UNBALANCED DATA as the original dataset.

If the sample size is small, 10-fold cross-validation can be used.

However, we will also take a sample of the original data and use that for validation in the evaluation stage. We will ensure that none of the data is replicated in the test, training or validation sets.

## Step 1B: Scaling Numeric data and Normalizing data

There are four methods that can be used to scale numeric data. They are called: 1) the standardized method using z-scores; 2) the mean normalize method; 3) min-max scaling; 4) and unit vector. Method 3 is quite common and we will use that method, where the new values go from zero to 1. Scaling is good when data has ranges that are very different from one another. Scaling is not normalizing data. Normalizing data changes the shape of the data to a bell curve whereas scaling just scales down the numbers so that they are between 0 and 1 (in this case). This version will only be used when the models require a scaled distance. This dataset is called PB\_scale. K-means, K-NN, and PCA require scaling.

min-max scaling is:

As for normalizing, the R command “normalize(dataset, lambda=3)” can be used. This dataset will be used where required. This dataset is called PB\_norm.

## Step 1C: Convert Numeric to Nominal (discretize) and vice versa for a new set of data

Notes here: can we use r code to convert: Numeric to categories: Discretize() ??

Numeric to Nominal - We have added R code to convert the original numeric data to nominal data. Table 5 shows the bins that are used to convert the number of data items into separate bins. For example, the first bin in Age is from 0 to 30 years old, so the number of data items in that group will go into the first new nominal attribute. ---\*\*\*\* there is an alternative that uses quartiles \*\*\*\*.

Table 5: When converting numerical data to Categorical data for models:

|  |  |  |
| --- | --- | --- |
| **Attribute** | **# categories created using binning** | **comment** |
| Age | 6 | [0, 30, 40, 50, 60, 70, 100] |
| Balance | 6 | [-10000, 10000, 30000, 50000, 70000, 90000, 110000] |
| Day | 4 | [1, 8, 15, 22, 31] |
| Duration | 5 | [0, 300, 600, 900, 1200, 5100] |
| Campaign | 5 | [1, 5, 10, 20, 30, 70] |
| Pdays | 6 | [0, 1, 100, 200, 300, 400, 900] |
| previous | 4 | [0, 2, 5, 25,300] |

Nominal to Numeric - We have added R code to convert the original nominal data to numeric data. Table 6 shows the number of new attributes that will be created (n-1) from the number of categories in each original attribute. For example, Job has 12 original categories so 11 new “dummy” (0 or 1) attributes will be created from these.

Table 6: When converting Categorical data to numerical data for models:

|  |  |  |
| --- | --- | --- |
| Attribute | Current # categories | # Attributes created using ‘dummy’ variables (0,1) |
| Job | 12 | 11 |
| Marital | 3 | 2 |
| Education | 4 | 3 |
| Default | 2 | 1 |
| Housing | 2 | 1 |
| Loan | 2 | 1 |
| Contact | 3 | 2 |
| Month | 12 | 11 |
| poutcome | 3 | 2 |

## Step 2: Feature Selection or Dimensionality Reduction (Select Attributes)

Still use train and test set….

Boruta is a feature selection algorithm for finding all relevant variables. The algorithm is designed as a wrapper around a Random Forest classification algorithm.

GainRatioAttributeEval(formula, data, subset, na.action, control = NULL)

InfoGainAttributeEval(formula, data, subset, na.action, control = NULL)

## There are various Feature Selection models that can be used before running any models. These models include:

* Boruta, Lasso regression, Stepwise Forward and backward selection, RFE, Genetic

In WEKA:

**ChiSquaredAttributeEval, InfoGainAttributeEval, CfsSubsetEval**

Algorithms:

Statistical techniques on numeric data include Chi-Square score and Gini index. Meanwhile, most algorithms in this family can only work on discrete data and conventional data discretization techniques are required to preprocess numerical and continuous variables.

Others include Multilayer perceptrons (MLP), Greedy FS, Lasso

## Step 3: Modeling

The models we want to use are the classification models, and with supervised training. The supervised training means that we know the ‘target’ response and we want to train the data to that target so as to obtain an algorithm that works well.

The models used are: a) Naïve Bayes, b) Decision Tree, c) xx, d) xxx…..

Then describe each model

Extra- time permitting-We will also use Unsupervised training, where the target has been stripped out of the data, to get an idea of any relationships in the data (describe these types – Apriori and clustering).

## 3A: Model 1 –

## 3B: Model 2 –

## 3C: Model 3 –

## 3D: Model 4 –

## 3E: Model 5 –

## 3F: Model 6 –

## Step 4: Model Evaluation

## Write details of the steps. If there is any source code that you’d like to share then provide the link of the Github.

## Step 5: Conclusion

## References

1. S. Moro, R. Laureano and P. Cortez. Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology. In P. Novais et al (Eds.), Proceedings of the European Simulation and Modelling Conference – ESM’2011, pp. 117-121, Guimaraes, Portugal, October, 2011.
2. P. Ładyzynski, K. Zbikowski and P. Gawrysiak. Direct Marketing Campaigns in Retail Banking with The Use of Deep Learning and Random Forests. In Expert Systems with Applications, V134, 28-35, 2019.
3. H. Elsalamony. Bank Direct Marketing Analysis of Data Mining Techniques. In Intenational Journal of Computer Applications, V85, 12-22, 2014.
4. M. Purnachary, B. Kumar and H. Shaziya. Performance Analysis of Bayes Classification Algorithms in WEKA Tool using Bank Marketing Dataset. In International Journal of Engineering Research in Computer Science and Engineering (IJERCSE), V5, Issue 2, February 2018.
5. <http://archive.ics.uci.edu/ml> (The UCI Machine Learning Repository collection of datasets).