Design and Implementation of Data Mining Model for Bank Marketing Data set

Deepak Sharma(ds5930@rit.edu), Dipesh Nainani(dsn1945@rit.edu) , Manasi Sunil Bharde(msb4977@rit.edu)

Abstract

The objective of this project is to improve the yield of the direct marketing campaigns, conducted by Portuguese banking institution, by developing a system for data analytics which can predict prospective customers for the Term Deposit service of the banking institution. This system will enable the bank to organize more focused campaigning. In order to develop a system for data analysis, provided dataset was decomposed in relational schemas for addressing the redundancy of data, by removing all partial key dependencies and transitive dependencies present in dataset.

Our designed database relational schema not only enabled faster access to the data but also saved the precious memory of the system. The dataset was analyzed using various statistical and data analytics operations in order to determine appropriate data pre-processing and cleaning. For predicting the target variable, Term Deposit, we constructed different classification models in R namely Random forest, Associative rule-based classification model, Neural network model.

Goals and Related Work

The provided dataset consists of over 20 attributes and 45000 instances. These attributes reflect socioeconomic state of the customer, details of previously subscribed banking products like various type of loans and details of previously made contacts with the customer, for selling the product, like duration of phone call and the number of times customer was offered this product to the prospective customer.

Our goal is to construct a classification model which can predict accurately “Term Deposit” attributes of the above-mentioned dataset using other independent attributes of the dataset. Qeethara Kadhim Al-Shayea [1] have worked on the dataset and they have proposed classification model using Neural Networks. Safia Abbas [3] have proposed relational schema for decomposition of the dataset across various entities.

Analysis , Implementation and Consideration

The dataset was partitioned in the Train and Test datasets. 80% of the data was assigned to training, dataset while the rest 20% was assigned to the Test dataset and for the Test dataset, we are predicting the outcome of target attribute using the constructed classification model.

Preliminary data analysis was performed on data by generating correlation matrix and by plotting graphs for various attributes in order to understand the distribution and correlation among attributes.

During preliminary data analysis, we found the domain distribution of the target variable was skewed as the number of No responses were predominantly higher than Yes responses and Missing values were also observed in multiple attributes.

A robust model cannot be generated using a training data set which contains missing values. The presence of skewness further deteriorates the ability to construct a resilient model. In order to deal with missing values and skewness below data cleaning and transformation operations have been performed.

For addressing the skewness of target variable oversampling and downsampling was performed which balanced the distribution of target variable by generating synthetic samples of the minority class and by removing downsampling the majority class. These operations were performed us SMOTE package.

For treating the null values, first chi square test was applied to check independence between the probability distribution of known and missing values of an attribute and the probability distribution of values in target variable . Based on the results of Chi test, attributes were either assigned values using KNN imputation method or corresponding rows were removed.

By decomposing the Data into various tables storage and retrieval operations can perform efficiently as decomposition promotes atomicity in data and eliminates any redundancy present in data. Hence, After performing cleaning task, Data was decomposed into multiple tables. R doesn't have its own relational data storage system so sqlite was used for storing relational data.

Architecture Design

1. Data integration was performed by joining the CPI data with Bank-Additional dataset.
2. Appropriate Data cleaning was performed as per the observations made while preliminary analysis on training Dataset.
3. For decomposing the data, an Entity-Relationship model as shown in the figure was designed in mySQL workbench. Using ER diagram, corresponding SQL script was generated for implementing the designed entity relationship model. Resultant implemented schema using this script was used for storing the data in decomposition form in SQL.
4.Random Forest, Neural Network, Random forest classification models were implemented in R for predicting the target variable. Associative rule based analysis was also performed on the dataset. Associative rules were extracted using apriori approach and rules are pruned to contain only superset of rules having non-overlapping LHS.

Lessons Learned

Data interpretation is the most important step for making use of the data. After analyzing data, we came to know nature and behavior of the data. Further processing and model creation were based upon this analysis. There is no universal way of cleaning the data. Every model in R requires data in a certain form. To make the model work we need to format the data in a specific way for each model which further requires analysis and data cleaning. Hence this results in putting a lot of efforts to peruse documentation of packages in R. Sometimes it also requires us to visit online forums for posting our queries.

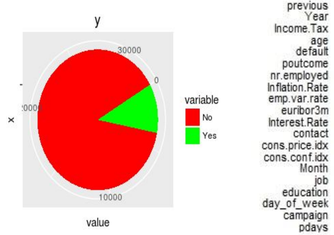

Figure 1. Distribution of Output Variable.

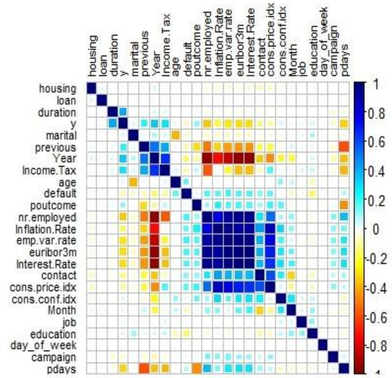

Figure 2. Highly Correlated Attributes.

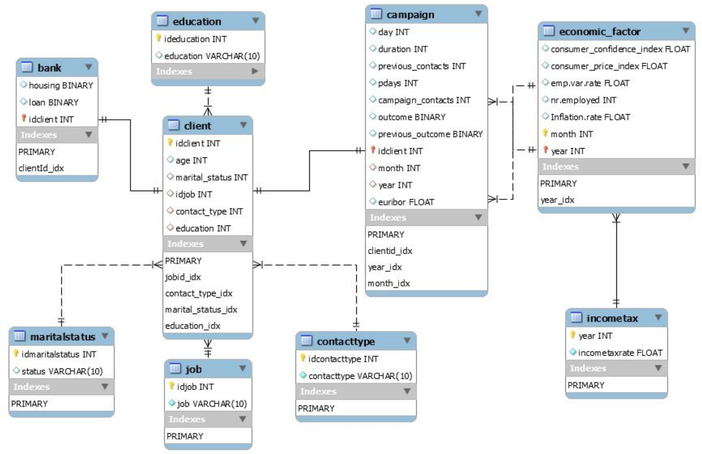

Figure 3 : Entity Relationship Diagram

Results

Random forest model:
Accuracy on training dataset :- 92%.
Accuracy on testing dataset :- 84%.
ROC:- .8845 area under the curve.
Confusion matrix and ROC curves have been shown below
.

Type	No	Yes
No	6395	897
Yes	102	843

Table 1: Confusion Matrix

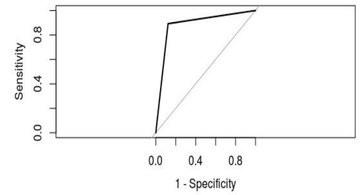

Figure 4 : ROC plot.

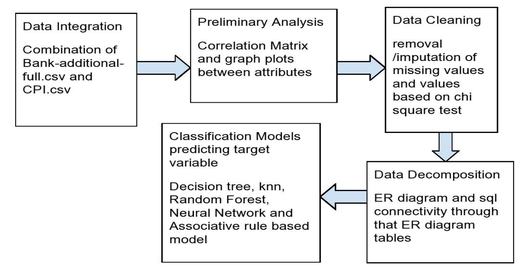

Figure 5 : Architecture Block Diagram

Current status and Future Work

We have developed a successful Random Forest Model which have shown 88% accuracy while predicting the term deposit. We also have developed fancy decision tree model, kNN model, neural network model and associative rule based model for performing the same task to predict the target variable of the dataset.

Neural Network models requires certain formatting for the dataset particularly the size of the column should be equal to build an effective neural network model. We found the center length after data cleaning weren’t equal for all the attributes. As of now, KNN model is not working on testing dataset. Testing dataset requires further processing for treating the missing values. We can neither we can perform imputation nor we can remove rows from testing dataset.

Associative rule based analysis was done for the dataset. For predicting the values of output class using pruned associative rules, for each record of the test data, subset of applicable rules should be fetched from pruned rules based on LHS attributes of the associative rules. Value of output class can be determined using the subset of applicable rules. In order to implement this in R, method should be devised to compare LHS of rules which is in the form of matrix with values of the independent variable in dataset for the test record.

Contact

Deepak Sharma (ds5930@rit.edu)
Dipesh Nainani (dsn1945@rit.edu)(
Manasi Sunil Bharde (msb4977@rit.edu)

References

1) Qeethra Khadim Al-Shayea. Evaluating Marketing Campaigns of Banking
2) Using Neural Networks, July,2013. http://www.iaeng.org/publication/WCE2013/WCE2013_pp759-761.pdf.
2)Safia Abbas. Deposit subscribe Prediction using Data Mining Techniques based Real Marketing Dataset, January, 2015. http://arxiv.org/pdf/1503.04344.pdf