Mining Telemarketing Data

Software Requirements Specification (CSE563), FALL 2015
PROJECT PROPOSAL

Author

Phani Santhosh Vamsi
Deepak, Darbha
1208568487
pdarbha@asu.edu

INTRODUCTION:

In recent times there is a tremendous increase in the number of marketing campaigns targeting different sections of community. As a result the effect of these campaigns on the general public is decreasing day by day. Furthermore, budget constraints and heavy competition from the peers has made marketing managers very selective in choosing the candidates they plan to target for direct campaigns. So making people to subscribe/buy a product or service using these kind of marketing campaigns with limited budgets is becoming quite challenging these days. Different Data mining and Business intelligence techniques can enhance the success ratio of these marketing campaigns.

In this Project real world data [1] from a Portuguese bank's telemarketing campaign related with term deposit subscriptions is taken for analysis. The main goal is to predict the possibility of new user subscribing for a term deposit upon contacting by analyzing users with similar profiles who has been contacted earlier in similar telemarketing campaigns. Such predictions can help marketing managers in increasing the efficiency of marketing campaigns and also analyzing the key factors that can influence users in subscribing for term deposits.

BUSINESS UNDERSTANDING:

Bank Marketing:

Enterprises promote their products and services in two ways. One way is through mass campaigns, targeting general public and the other is direct marketing targeting specific customers. Studies show that mass campaigns success ratio is less than 1% [4] while direct campaigns targeting right customers are proving to be more efficient [5] with success ratio of 10 to 20%.

1. What exactly is the business problem to be solved?

European banks are currently facing severe financial crises, so there is a heavy pressure to increase their financial asset. One of the strategies these banks have adopted to improve financial asset is to increase the number of long term deposit holders by attracting them with good interest rates using telemarketing campaigns. But the success ratio of such telemarketing campaigns are very low (10 to 15%) inferring that a lot of money being wasted by marketing products on wrong people. So if we can analyze and predict the potentially right customers who will subscribe for a term deposit upon marketing, Banks can invest on marketing only those customers which can increase the campaigns success ratio keeping the marketing budget minimum.

In this project marketing campaign data of a Portuguese bank facing similar problems is taken for analysis. The business goal is to build a model on input data which predicts if a customer subscribes to long term deposit or not when contacted through marketing campaign.

DATA UNDERSTANDING:

The data used for this project is taken from UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/datasets/Bank+Marketing) which is donated by [Moro et al., 2014] [1].

Heuristics of Data:

- The Data contains 41188 instances or records where each record corresponds to data of potential users who can be contacted for telemarketing of term deposits.
- There are 17 attributes and a binary target variable (class) which identifies whether a user subscribes or not for a term deposit.
- Of the 41188 users that are contacted only 4640(11.2 percent) has subscribed for a term deposit.

Data Science Solution:

From the data it is clearly evident that this is a binary classification problem with the class label values yes (if contacted person subscribes for long term deposit) or No (otherwise). So Binary classification algorithms like Random Forests, SVM, and Naïve Bayes will be used in this project and a comparative study of algorithms will be made to find the best classifier to model the input dataset. Further comparative study on various attribute values with fixed classifier will be made to find the best possible values of attributes which can improve the success of telemarketing.

As the data is unbalanced(only 11% positive records) different sampling techniques like oversampling, under sampling, cost sensitive sampling and hybrid sampling as discussed in [3] will be employed before using the above said classifiers.

2. Is the data science solution formulated appropriately to solve this business problem?

The proposed data science solution is well formulated to solve the business problem. The more accurately the classifier predicts the target variable, marketing managers will target the right customers who subscribe on telemarketing, there by increasing the success ratio of marketing campaigns.

3. What business entity does an instance/example correspond to?

This entire problem falls under the domain of business intelligence. This problem primarily focuses on making telemarketing campaigns of a Portuguese bank profitable by targeting the right clients who tend to subscribe for term deposits upon contacting.

Each instance of the dataset corresponds to bank client's information like their marital status, gender, education, credit history, other loan information and profession. It also contains information about previous marketing contact history like number of days passed from the day the client has been contacted last, the outcome of previous campaign for this client etc.

4. Is the problem a supervised or unsupervised problem?

The dataset used in this project contains binary target class labeled with values “yes” or “no” which specifies whether the contacted user has been subscribed for term deposit or not. So this problem is a supervised learning problem with class label values specified for training set.

5. Are the attributes defined precisely?

The attributes of the dataset are divided into 3 main categories. Client Personal Data attributes having numerical and categorical values, previous contact information data attributes having numerical values and social-Economic context attributes having numerical values. All the attributes are defined precisely with default values for unknown or missing values.

Attributes and Target Variable:

Attributes:

#Clients Personal Data
Age: (numeric)
Job : (categorical:'admin.','bluecollar','entrepreneur','housemaid','management','retired','self-employed','services','student','technician','unemployed','unknown')
Marital: marital status (categorical: 'divorced', 'married', 'single', 'unknown')
Education:(categorical:'basic.4y','basic.6y','basic.9y','high.school','illiterate','professional.course', 'university.degree','unknown')
Default: has credit in default? (Categorical: 'no', 'yes', 'unknown')
Housing: has housing loan? (Categorical: 'no', 'yes', 'unknown')
Loan: has personal loan? (Categorical: 'no', 'yes', 'unknown')

 # related with the last contact of the current campaign:
Contact: contact communication type (categorical: 'cellular', 'telephone')
Month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec')
Day_of_week: last contact day of the week (categorical: 'mon','tue','wed','thu','fri')
Duration: last contact duration, in seconds (numeric). Important note: this attribute highly affects the output target (e.g., if duration=0 then y='no').

 # other attributes:
Campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
Pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
Previous: number of contacts performed before this campaign and for this client (numeric)
Poutcome: outcome of the previous marketing campaign (categorical: 'failure', 'nonexistent', 'success')

 # Social and economic context attributes
Emp.var.rate: employment variation rate - quarterly indicator (numeric)
Cons.price.idx: consumer price index - monthly indicator (numeric)

Cons.conf.idx: consumer confidence index - monthly indicator (numeric)
Euribor3m: euribor 3 month rate - daily indicator (numeric)
Nr.employed: number of employees - quarterly indicator (numeric)
Output variable (desired target):
Y: has the client subscribed a term deposit? (Binary: 'yes', 'no')

6. Will modeling this target variable improve the stated business problem? An important sub-problem? If the later, is the rest of the business problem addressed?

Modelling this target variable (class variable) predicts whether the client subscribes for term deposit if contacted, which greatly helps marketing managers in targeting the right clients for marketing to increase the number of term deposits while keeping their marketing costs minimum, thereby improving the banking Business.

Also modeling this target variable using individual attributes gives interesting insights for marketing managers on when and how to target the right clients for better subscriptions through marketing.

7. If unsupervised, is there an "exploratory data analysis" path well defined?
The problem addressed in this project is a supervised learning algorithm.

USE CASE SPECIFICATIONS:

Use Case 1: Telemarketing System

Objective: This use case gives the entire flow of system.

Primary Actor: Marketing manager

Dependencies: Marketing Manager and Data analyst inherit Bank Employee (Generalization).Call client include predict client which includes both get client data and build model which are associated with each other.

Trigger: Marketing Manager initiating call.

Secondary Actors: Data Analyst, Client

Preconditions: Existing Client data for training the model.

Post Condition(s): Client subscribes to the term deposit and a new record is added into client data.

MAIN SUCCESS SCENARIO

1. Marketing Manager initiates a call
2. The system then predicts that the client initiated is a potential subscriber from the model built by data analyst on client data.
3. Manager makes the call to the potential client.
4. Client subscribes to the new term deposit.
5. New record will be added to client data with class label **Yes**.

VARIATIONS

Variation ID: *callclient_1*

1. Marketing Manager initiates a call
2. The system then predicts that the client initiated is a potential subscriber from the model built by data analyst on client data.
3. Manager makes the call to the potential client.
4. Client **don't** subscribe to the new term deposit.
5. New record will be added to client data with class label **No**.

FAILURE VARIATIONS

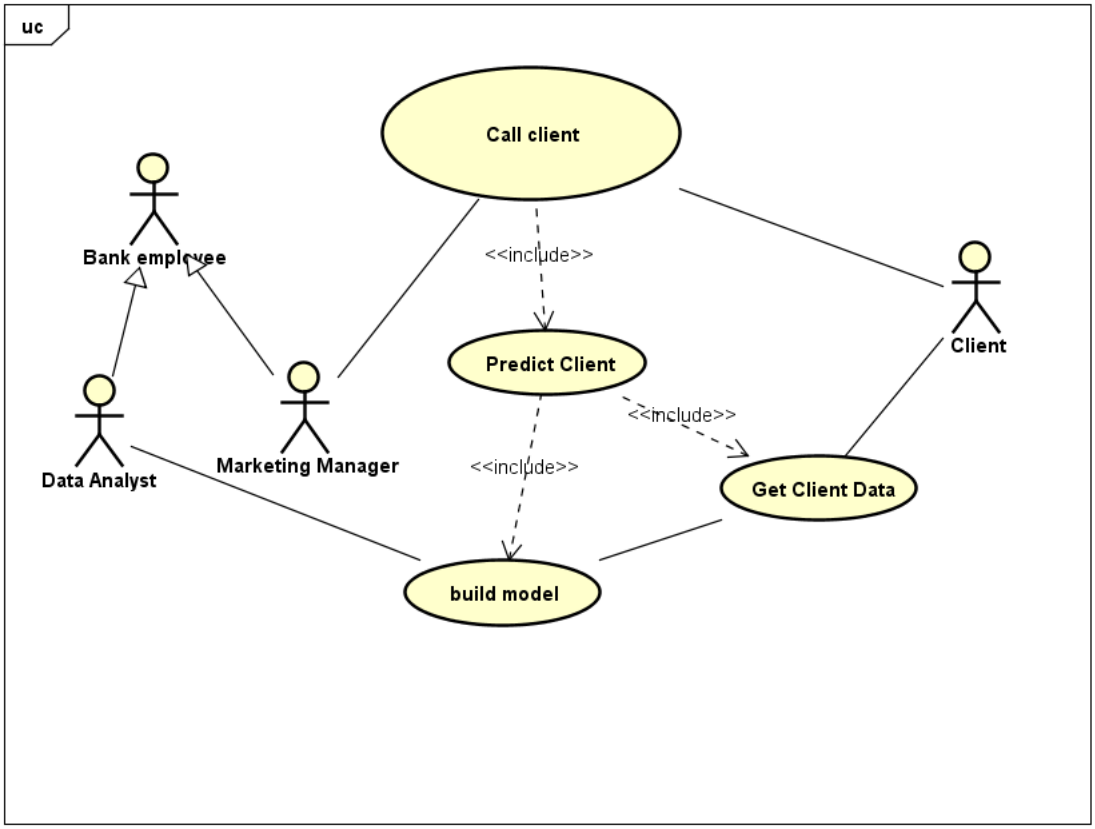
Variation ID: *Callclient_2*

1. Marketing Manager initiates a call
2. The system then predicts that the client initiated is **not** a potential subscriber from the model built by data analyst on client data.
3. Manager drops the call.

BUSINESS RULES

1. Marketing Manager only calls clients those are potential subscribers.
2. Once the action is complete a new record will be added to the client data with Class label Yes or No depending upon the clients subscription to term deposit.

USE CASE DIAGRAM:



DATA PREPARATION:

Will it be practical to get values for attributes and create feature vectors, and put them into a single table?

The data used in this project is collected from a Portuguese bank that used its own contact center to do direct marketing campaigns. Each Campaign was managed in an integrated fashion and the results were outputted together. An attribute of data corresponds to a unique feature of campaign information of a targeted customer or his personal data. So each record in the table corresponds to all campaign related data of a particular customer. The attributes are divided into clients personal information, last contacted details, Social and economic context attributes each impacting customer's willingness to take a long term deposit via tele marketing which is our class label. So it is practical to get values for attributes as they are collected from their own contact center. Also each attribute influences class variable and are independent of one another making it possible to create a feature vector with each attribute as a dimension. The Dataset thus can be populated into a single table where each row corresponding to each feature vector of a customer.

If not, is an alternative data format defined clearly and precisely? Is this taken into account in the later stages of the project? (Many of the later methods/techniques assume the dataset is in feature vector format.)

The Dataset taken in this project is in feature vector format.

If the modeling will be supervised, is the target variable well defined? Is it clear how to get values for the target variable (for training and testing) and put them into the table?

The target variable in this project is a binary variable with values (Yes/No) indicating weather a target customer in marketing campaign has subscribed for a long term deposit or not. All the Data used in this project is historical (collected from May 2008 to Nov 2010) and has got class labels for every record in table. The data will be divided into 70 percent training and 30 percent test. Class labels for 30 percent test set will be then removed. A model is built using 70 percent training data. 30 percent test data would be then inputted to model and classes for test data are predicted using the built model. These class labels will be then compared to original class labels of test set to get test set accuracy. This will be done using confusion matrix feature of R. Training set accuracy is found using 10 fold cross validation feature of R.

How exactly will the values for the target variable be acquired? Are there any costs involved? If so, are the costs taken into account in the proposal?

The values of the target variable are acquired as a part of marketing campaigns conducted by a Portuguese bank using its own contact center. The value of target variable indicates whether a client subscribes to a long term deposit. So the bank is analyzing its own data so there are no costs involved for acquiring values of target variables. This dataset is made available for research for free of cost in UCI Machine Learning Repository with Citation [1]. So there are no costs involved while acquiring data and values of target variable.

Are the data being drawn from the similar population to which the model will be applied? If there are discrepancies, are the selection biases noted clearly? Is there a plan for how to compensate for them?

The Data is drawn from the similar population to which the model will be applied. Both the training and test set is taken from the same data targeting similar customers. The data collected is of potential clients of Portuguese bank from Portuguese population. Most of the attributes are nominal with same set of allowed values in both training and test set. There are a few discrepancies in the clients previous contact data, as half of the clients are contacted for the first time, and so most of the values of those are NULL. In this project those discrepancies are handled using defaults that can minimize the variance. Also each attribute is individually modeled with class variable to find the extent in which it can influence class variable. Also if different values of attribute class have same outcome of class variable those attributes are neglected. In this dataset sex attribute is discarded as it has same rate of success for male and female population [2].Also the attributes that are having a huge variance are normalized before training the model. Attribute call duration has got huge variance ranging from 1 minute to 500 minutes. There are few noise points where in the

call duration is more than 1000 minutes. So normalizing this attribute can bring down the variance and reduce the impact of noise points on the class variable.

Modeling:

Is the Choice of model appropriate for the choice of target variable?

The target variable used in this project is a binary variable with values "yes" or "no" where in "Yes" signifies if a target customer of bank telemarketing campaign when contacted subscribes for a long term deposit. Also the data used in this project is supervised as it is taken from the historical records of telemarketing outcome results. So the problem dealt in this project is a binary classification problem with class labels "yes" or "no". So I have used classification models SVM, Random Forests which best classify binary data in linear time.

Does the model/modeling technique meet the other requirements of the task?

Generalization performance, comprehensibility, speed of learning, speed of application, amount of data required, type of data, missing values?

In this project SVM and Random forest classifiers are used for classifying the data. The choice of classifiers is selected by referring base papers [1], [2] which used similar data and got best accuracies. Other models used are naïve Bayes classifier and Logistic Regression classifier which didn't yield good results. The Data used in this project is biased as the number of positive samples in the data is only 7 percent of the total data. So when the above classifiers are used although the accuracies are good, Precision and Recall values of test set are very less which clearly signify that the models built are unable to classify most of the positive samples correctly. To overcome that the training data is oversampled using ROSS such that there are almost the same number of positive "yes" and Negative "no" Samples in the training data. The models are then built again with the new balanced data and the precision and recall values of positive samples of the test set are computed and their values have significantly improved. Given below are the performance measures.

Model	Accuracy	Precision	Recall
SVM(Original Dataset)	0.737338	0.475	0.082
SVM(Balanced Dataset)	0.7478406	0.518	0.452
Random Forest(Original Dataset)	0.7510797	0.677	0.084
Random Forest (Balanced Dataset)	0.7186887	0.4703	0.641

Is the choice of modeling technique compatible with prior knowledge of problem (e.g., is a linear model being proposed for a definitely nonlinear problem)?

Random Forests can classify both linear and non linear separable data. SVM though the data is not linearly separable it takes attributes to higher dimensions and separates in higher dimension space. So the choice of modeling techniques compatible with prior knowledge of the problem.

Should various models be tried and compared (in evaluation)?
Yes various models can be tried for classifying this binary class data. As the data is collected from external source with no information on the amount of noise that is present, and also given the variations in different attributes of data like Continuous, Categorical and binary attributes, the performance of different classifiers can't be predicted before actually modeling and evaluating as some models can overfit or they can give more priority to columns that are having more defaults. So in this project various models are built on data and prediction accuracies of the models are evaluated and compared and the model with best test set and cross validation accuracies is proposed.

For clustering, is there a similarity metric defined? Does it make sense for the business problem?

All the data used in this project is supervised with class label clearly defined for all the records. So there is no necessity for clustering in this project.

Evaluation and Deployment:

Is there a plan for domain-knowledge validation? Will domain experts or stakeholders want to vet the model before deployment? If so, will the model be in a form they can understand?

The attribute features for this project are chosen using domain knowledge validation. During the modeling phase different categories of features are identified and modeled separately to find out their impact on the predicted class such as bank Client data, previous contact information, and Social and Economic context attributes. Also individual impact of various domain specific features are also found like duration being the dominating feature that can classify the class variable. Although in this project an accuracy of about 75 percent is achieved there is still a lot of scope for improving the accuracy. In these kind of models accuracy mainly depends on the domain specific features that are selected. So there is always a need for domain experts or stakeholders to vet the model before deployment. And for further strengthening their analysis and for giving them better insight on how to proceed further, all the analysis results are presented in the form they can understand(Domain Specific).

Is the evaluation setup and metric appropriate for the business task? Recall the original formulation.

Are business costs and benefits taken into account?

In this project telemarketing data is used for analysis. These kind of business main aim is to target all the customers that can subscribe to their term deposits even at the cost of contacting more irrelevant people (this is so because operating costs are minimum comparative to the profits gained from user's subscriptions to term deposits.) in statistical terms these business main goal is to increase the true positives without having to bother much about true negatives and false negatives. So the deciding factor in these kind of businesses would be sensitivity (in other words recall). To summarize in these business costs almost remain constant with increase in targeted customers while benefits increase.

For classification, how is a classification threshold chosen?

The driving factor for classification accuracy in these kind of businesses is sensitivity. Also the success ratio of these kind of campaigns is very less 7%, so sensitivity or recall value more than 0.5 would be a good threshold for classification.

Are probability estimates used directly?

The Classification models used in this project are SVM and Random Forests which predict the class label directly without probabilities. Though Logistic Regression Model is trained it is not able to classify data as the data contained a lot of missing values substituted with defaults. So there is no scope for using probability estimates in this project.

Is ranking more appropriate (e.g., for a fixed budget)?

Ranking can be appropriate at times when the budget is fixed and only few attribute data can be collected. The attributes which contribute more to the class prediction can be given more precedence compared to attributes which have very less impact on class variable. Here in this project call Duration is given highest priority than any other feature.

For regression, how will you evaluate the quality of numeric predictions? Why is this the right way in the context of the problem?

This project is a classification problem with no scope for regression.

Does the evaluation use holdout data?

Yes, the evaluation of each model in the project is made on holdout data. Training and Test sets are clearly defined and separated with no common instances. The models are built on training set and then test set class variables are predicted using the built models.

Against what baselines will the results be compared?

The results are compared based on the expectant outcome from the analysis. In this project the outcome is predicting the right customers to be targeted for telemarketing campaign for them to subscribe for term deposits. Also the results are compared mainly based on the sensitivity or recall value of the test data.

Why do these make sense in the context of the actual problem to be solved?

Actual problem in here is targeting all the right customers who can subscribe to the tele-marketing data. Though there are not much costs involved in targeting huge section of population, time constraints and availability of customer service representatives can become an overhead. So increasing the recall value or true positives can bring significant improvement in profits and success of such telemarketing campaigns.

Is there a plan to evaluate the baseline methods objectively as well?

Yes, the baseline methods are evaluated across the number of true positive subscribers predicted as this benefits the business more. In this project when Classification models SVM and Random Forests are compared though accuracy and positives predicted values are more for SVM Random forest is finally chosen as it is able to predict more of true subscribers than SVM even though it predicted many false subscribers as true, as when we see objectively in business terms there is literally no or min cost involved in making a phone call when compared to benefit of term subscription of an individual. So the baseline methods are also evaluated objectively while choosing the best model.

For clustering, how will the clustering be understood?

All the data used in this project is supervised with class label clearly defined and there are no missing values in class values. So there is no scope for clustering.

Will deployment as planned actually (best) address the stated business problem?

Deployment using this plan helps in increasing the number of right target customers. The main problem is that only 7 percent of targeted customers subscribe to term deposits. When this solution is deployed only those customers that are predicted right by the classifier can be targeted first so that there is a greater chance to improve the campaigns efficiency by reaching out to more number of right people. The positive target customers can be increased from 7 percent to 26 percent when we deploy this model.

If the project expense has to be justified to stakeholders, what is the plan to measure the final (deployed) business impact?

Before analysis there are only 7 percent of the right target customers i.e. only 7 people subscribe to a term deposit when 100 people are contacted. After deploying this model out of 100 people that are targeted 26 people can be the right customers who can subscribe to the term

deposit(considering 60 percent recall of 44 percent positive predict value of 100 targeted customers who are chosen from the positives predicted by the model). Hence deploying the model can improve the efficiency of marketing campaigns by nearly 400 percent. 400 percent increase in efficiency can bring out more revenue and profits to the bank as a whole.

Even if the model won't predict as expected those 7 percent customers are targeted with a little more population without significant improvement in the campaign cost.

REFERENCES:

[1] S. Moro, P. Cortez and P. Rita. A Data-Driven Approach to Predict the Success of Bank Telemarketing. Decision Support Systems, Elsevier, 62:22-31, June 2014

[2] 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. EUROSIS

[3] Vaishali Ganganwar. An Overview of Classification algorithms for imbalanced datasets. International Journal of Emerging Technology and Advanced Engineering Website: www.ijetae.com (ISSN 2250-2459, Volume 2, Issue 4, April 2012)

[4]Ling, X. and Li, C., 1998. Data Mining for Direct Marketing: Problems and Solutions. In Proceedings of the 4th KDD conference, AAAI Press, 73–79.

[5]Ou, C., Liu, C., Huang, J. and Zhong, N. 2003. "On Data Mining for Direct Marketing". In Proceedings of the 9th RSFDGrC conference, 2639, 491–498.