Module title: [redacted] Machine Learning in Finance

Group assignment title: Bank telemarketing and machine learning

Submission date: Saturday May 7th

Instructions

You should work in a group, of circa 4 students, and each group should nominate an individual to submit a single project report. Maximum word-count for the report is 3,000 words.[1] Front page of submitted assignment should detail module title, membership of the group (i.e., students' names and numbers) as well as the assignment title, and the assignment word count.

Please submit this project by Saturday, May 7 via [redacted]

Assessment and grades

You will be assessed on your ability to respond to questions raised below, i.e., to use machine learning informed alert model algorithms, critically evaluate the performance of these methods, and coherently report your findings. This project counts for 40% of your overall module grade.

Assignment Context

An important source of income at banks is the term deposit, i.e., deposits by customers at a fixed rate for a fixed time. This capital can be used to disburse loans at a higher interest rate. The bank, hence, uses marketing techniques to target customers to save via term deposits. For example: email, advertisement, telephonic and digital marketing. Telephonic marketing (i.e., phone calls) remains an effective way to acquire term deposit customers, especially if enabled with machine learning. Banks can use data and machine learning informed alert models to identify customers who are more likely to save via a term deposit, and to inform a telephonic marketing campaign accordingly.

The dataset, in this assignment, is related to the direct telemarketing campaigns (phone calls) of a European banking institution. You can find the data for the project on Brightspace (MyLearning \ Group Assignment) and variable descriptions below.

The classification goal is to predict if the customer will subscribe to a term deposit (Term

[1] Word count includes an assignment's references section.

Deposit = 1). Tapping into the repertoire of your Machine Learning modelling, evaluation and deployment knowledge, provide recommendations to the bank's Retail Marketing department to achieve its goal.

Questions

Project Tasks

1. (a) Fit a logistic regression model on the dataset. Choose a probability threshold of 10%, 20%, 35% and 50%, to assign an observation to the Term Deposit = 1 class. Compute a confusion matrix for each of the probability thresholds. How do the True Positive and False Positive rates vary over these probability thresholds? Which probability threshold would you choose?

   (b) Divide the dataset into training (70%) and test (30%) sets and repeat the above question and report the performance of these models, across probability thresholds, on the test set.

   (c) Plot the ROC for a logistic model on a graph and compute the AUC. Explain the information conveyed by the ROC and the AUC metrics.

   [10 marks]

2. (a) Fit classification tree, bagging and random forest models on the dataset and comment on the performance of these models. Do you think we are overestimating the performance of these models by fitting them on to the whole dataset? If so, state your reasons.

   (b) Split the dataset in two parts: training (70%) and test sets (30%). Fit the models on the training dataset and evaluate their performance on the test set. Which model would you choose and why?

   (c) For the best model chosen, rank and plot predictors according to their predictive power.

   (d) How do these models perform compared to the model in question 1?

   [10 marks]

3. (a) Standardize your predictors and fit KNN classifier with K equal to 1, 3, 5 and 10, respectively. Evaluate the performance of these models on the test set.

(b) How do these models perform compared to the tree-based models in question 2 and logistic model of question 1?

[10 marks]

4. (a) Fit at least one other binary classifier (e.g. a linear probability model or a Support Vector Machine classifier) to the dataset. Describe its performance relative to the classifiers highlighted above.

(b) Is your training dataset balanced across outcome classes? Comment on the drawbacks of fitting a Machine Learning technique on an unbalanced dataset. Can you identify and deploy a technique to address this concern? If so, why do you think that the method you adopt could work? **Hint:** It is up to each student group to search for a systematic understanding and solution to the phenomenon of imbalanced data across outcome classes.

[10 marks]

Variable Name	Description	Category
Term deposit	Has the client subscribed a term deposit? 1 if yes, 0 if no.	Binary ('1', '0')
Age	Age of Customer in Years	Numeric
Job	Job Status	Categorical
Marital	Marital Status	Categorical
Education	Level of education	Categorical
Default	Default Status – is there a history of default	Categorical
Housing	Has availed Housing Loan?	Categorical
Loan	Has availed Personal Loan?	Categorical
Contact	Contact communication type (landline vs mobile phone)	Categorical
Month	Last contact month of year	Categorical
Day_of_week	Last contact day of the week	Categorical
Duration	Last contact duration, in seconds #	Numeric
Campaign	# of contacts performed during this campaign and for this client (includes last contact)	Numeric
Pdays	# of days that passed by after the client was last contacted from a previous campaign (999 means client was not contacted in a previous campaign)	Numeric
Previous	# of contacts performed before this campaign and for this client	Numeric
Ethnicity	Caucasian is the reference ethnic category: Is the customer of African ethnicity? 1=Yes, 0=No;	Numeric
Poutcome	Outcome of the previous marketing campaign	Categorical
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
Euribor (3m)	Euribor 3 month rate	Numeric
Nr.employed:	# of employees: quarterly indicator	Numeric

Duration: telephone call duration of the Term Deposit outcome call, in seconds (numeric). Important note: this attribute can strongly impact the outcome variable (e.g., if duration=0 then Term Deposit='0'). Thus, this input should only be included for benchmark purposes and should be discarded if the intention is to derive a pragmatic predictive model.

Description of the Dataset