MITx: 15.071x The Analytics Edge

Help

Course > Final Exam > Final Exam > PREDICTING BANK TELEMARKETING SUCCESS

PREDICTING BANK TELEMARKETING SUCCESS

Bookmark this page

PREDICTING BANK TELEMARKETING SUCCESS

The success of marketing campaigns can be highly specific to the product, the target audience, and the campaign methods. In this problem, we examine data from direct marketing campaigns of a Portuguese banking institution between May 2008 and November 2010. The marketing campaigns were based on phone calls. Often, more than one contact to the same client was required, in order to access if the product (bank term deposit) would be or not subscribed.

In this analysis, the goal would be predicting the dependent variable **y**, which takes value 1 if the the client subscribed to a term deposit, and 0 otherwise. The data we will be using bank.csv is a subset of the original data, containing 5000 examples and 20 input variables. The variable information is as follows:

- **age**
- **job** - type of job
- **marital** - marital status
- **education** - Shows the level of education of each customer
- **default** - Whether a customer has credit in default
- **housing** - Does the customer have a housing loan?
- **loan** - Does the customer have a personal loan?
- **contact** - The contact communication type
- **month** - Last contact month of year
- **day_of_week** - Last contact day of Week
- **duration** - Last contact duration in seconds (*Note: this variable is not known before making the call*)
- **campaign** - Number of contact performed for the client during the campaign
- **pdays** - number of days that passed by after the client was last contacted from a previous campaign (value of 999 means the client was not previously contacted)
- **previous** - number of contacts performed before this campaign and for this client

- **poutcome** - outcome of the previous marketing campaign
- **emp.var.rate** - employment variation rate - quarterly indicator
- **cons.price.idx** - consumer price index - monthly indicator
- **cons.conf.idx** - consumer confidence index - monthly indicator
- **euribor3m** - euribor 3 month rate - daily indicator
- **nr.employed** - number of employees - quarterly indicator

Problem 1 - Loading the Data

1.0/1.0 point (graded)

Use the read.csv function to load the contents of bank.csv into a data frame called bank. What is the average age in the data set?

39.5814 ✓**Answer:** 39.5814

Explanation

This can be computed with the mean or summary functions.

Submit You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Problem 2 - Call Durations by Job

2.0/2.0 points (graded)

Build a boxplot that shows the call duration distributions over different jobs. Which three jobs have the longest average call durations? (if it's hard to see from the boxplot, use tapply function.)

□ admin.

□ blue-collar

□ entrepreneur

☑ housemaid ✓

□ management

☑ retired ✓

☑ self-employed ✓

□ unemployed

✓

Explanation

By examining tapply(bank$duration, bank$job, mean), we can see the three jobs with highest mean call durations.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Problem 3 - Multicolinearity

2.0/2.0 points (graded)

As good practice, it is always helpful to first check for multicolinearity before running models, especially since this dataset contains macroeconomic indicators. Examine the correlation between the following variables: emp.var.rate, cons.price.idx, cons.conf.idx, euribor3m, and nr.employed. Which of the following statements are correct (limited to just these selected variables)?

☑ cons.conf.idx does NOT seem to have severe multicolinearity with the other variables. ✓

□ emp.var.rate and nr.employed have the highest correlation between two different variables.

☑ cons.price.idx and cons.conf.idx have the lowest correlation between two different variables. ✓

✓

Explanation

Use cor function to get the correlation matrix and inspect.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Problem 4 - Splitting into a Training and Testing Set

1/1 point (graded)

Obtain a random training/testing set split with:

set.seed(201)

library(caTools)

spl = sample.split(bank$y, 0.7)

Split months into a training data frame called "training" using the observations for which spl is TRUE and a testing data frame called "testing" using the observations for which spl is FALSE.

Explanation

Use the subset function to put the TRUE observations in the training set, and the FALSE observations in the test set.

Why do we use the sample.split() function to split into a training and testing set?

○ It is the most convenient way to randomly split the data

○ It balances the independent variables between the training and testing sets

● It balances the dependent variable between the training and testing sets ✓

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 5 - Training a Logistic Regression Model

2.0/2.0 points (graded)

Train a logistic regression model using independent variables age, job, marital, education, default, housing, loan, contact, month, day_of_week, campaign, pdays, previous, poutcome, emp.var.rate, cons.price.idx, and cons.conf.idx, using the training set to obtain the model. Notice that we have removed duration (since it's not available before the call, so shouldn't be used in a strictly predictive model), euribor3m and nr.employed (due to multicolinearity issue).

Which of the following characteristics are statistically significantly POSITIVELY (at 0.05 level) associated with an increased chance of subscribing to the product?

☑ age ✓

□ default is unknown

□ contact via telephone

☑ month is August ✓

☑ month is March ✓

□ day_of_week is Monday

☑ poutcome is nonexistent ✓

□ emp.var.rate

☑ cons.price.idx ✓

□ cons.conf.idx

✓

Explanation

The model can be trained with the glm function (remember the argument family="binomial") and summarized with the summary function.

Submit You have used 2 of 3 attempts

Answers are displayed within the problem

Problem 6 - Interpreting Model Coefficients

0/1 point (graded)

What is the meaning of the coefficient labeled "monthmar" in the logistic regression summary output?

○ When the month is March, the odds of subscribing to the product are 261.8% higher than an otherwise identical contact. ✓

○ When the month is March, the odds of subscribing to the product are 261.8% higher than the avarage contact.

○ When the month is March, the odds of subscribing to the product are 28.6% higher than an otherwise identical contact.

● When the month is March, the odds of subscribing to the product are 28.6% higher than the avarage contact. ✗

Explanation

The coefficients of the model are the log odds associated with that variable; so we see that the odds of subscribing are exp(1.286)=3.618284 those of an otherwise identical contact. This means the contact is predicted to have 3.618284-1=2.618284 higher odds of subscribing.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 7 - Obtaining Test Set Predictions

0.0/2.0 points (graded)

Using your logistic regression model, obtain predictions on the test set. Then, using a probability threshold of 0.5, create a confusion matrix for the test set.

We would like to compare the predictions obtained by the logistic regression model and those obtained by a naive baseline model. Remember that the naive baseline model we use in this class always predicts the most frequent outcome in the training set for all observations in the test set.

What is the number of test set observations where the prediction from the logistic regression model is different than the prediction from the baseline model?

414 ✗**Answer:** 94

Explanation

Obtain test-set predictions with the predict function, remembering to pass type="response". Using table, you can see that there are 94 test-set predictions with probability less than 0.5.

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Problem 8 - Computing Test-Set AUC

2.0/2.0 points (graded)

What is the test-set AUC of the logistic regression model?

0.7507 ✓**Answer:** 0.7507334

Explanation

The test-set AUC can be obtained by loading the ROCR package, and then using the prediction and performance functions.

Submit You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

Problem 9 - Interpreting AUC

0/1 point (graded)

What is the meaning of the AUC?

○ The proportion of the time the model can differentiate between a randomly selected client who subscribed to a term deposit and a randomly selected client who did not subscribe ✓

● The proportion of the time the model correctly identifies whether or not a client subscribed to a term deposit. ✗

Explanation

The AUC is the proportion of time the model can differentiate between a randomly selected true positive and true negative.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 10 - ROC Curves

1/1 point (graded)

Which logistic regression threshold is associated with the upper-right corner of the ROC plot (true positive rate 1 and false positive rate 1)?

● 0 ✓

○ 0.5

○ 1

Explanation

A model with threshold 0 predicts 1 for all observations, yielding a 100% true positive rate and a 100% false positive rate.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 11 - ROC Curves

1/1 point (graded)

Plot the colorized ROC curve for the logistic regression model's performance on the test set.

At roughly which logistic regression cutoff does the model achieve a true positive rate of 60% and a false positive rate of 25%?

○ 0

● 0.11 ✓

○ 0.29

○ 0.43

○ 0.66

○ 0.87

Explanation

You can plot the colorized curve by using the plot function, and adding the argument colorize=TRUE.

From the colorized curve, we can see that the light turqoise color, corresponding to cutoff 0.11, is associated with a true positive rate of about 0.60 and false positive rate of about 0.25.

Submit You have used 1 of 1 attempt

Answers are displayed within the problem

Problem 12 - Cross-Validation to Select Parameters

0/1 point (graded)

Which of the following best describes how 10-fold cross-validation works when selecting between 4 different parameter values?

● 4 models are trained on subsets of the training set and evaluated on a portion of the training set ✗

○ 10 models are trained on subsets of the training set and evaluated on a portion of the training set

○ 40 models are trained on subsets of the training set and evaluated on a portion of the training set ✓

○ 4 models are trained on subsets of the training set and evaluated on the testing set

○ 10 models are trained on subsets of the training set and evaluated on the testing set

○ 40 models are trained on subsets of the training set and evaluated on the testing set

Explanation

In 10-fold cross validation, the model with each parameter setting will be trained on ten 90% subsets of the training set. Hence, a total of 40 models will be trained. The models are evaluated in each case on the last 10% of the training set (not on the testing set).

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Problem 13 - Cross-Validation for a CART Model

0/2 points (graded)

Set the random seed to 201 (even though you have already done so earlier in the problem). Then use the caret package and the train function to perform 10-fold cross validation with the training data set to select the best cp value for a CART model that predicts the dependent variable y using the same set of independent variables as in the logistic regression (Problem 5). Select the cp value from a grid consisting of the 50 values 0.001, 0.002, ..., 0.05.

What cp value maximizes the cross-validation accuracy?

.008

✗**Answer:** 0.016

Explanation

The cross-validation can be run by first setting the grid of cp values with the expand.grid function and setting the number of folds with the trainControl function. Then you want to use the train function to run the cross-validation.
From the output of the train function, parameter value 0.016 yields the highest cross-validation accuracy.

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

Problem 14 - Train CART Model

1/1 point (graded)

Build and plot the CART model trained with the parameter identified in Problem 13, again predicting the dependent variable using the same set of independent variables. What variable is used as the first (upper-most) split in the tree?

○ cons.conf.idx

○ pdays

○ job

○ day_of_week

● emp.var.rate ✓

○ education

Explanation

The CART model can be trained and plotted by first loading the "rpart" and "rpart.plot" packages, and then using the rpart function to build the model and the prp function to plot the tree.

Submit You have used 1 of 1 attempt

ℹ Answers are displayed within the problem

Problem 15 - Test-Set Accuracy for CART Model

2/2 points (graded)

Using the CART model you created in Problem 14, obtain predictions on the test set (using the parameter type="class" with the predict function). Then, create a confusion matrix for the test set.

What is the accuracy of your CART model?

0.8866667 ✓**Answer:** 0.8866667

Explanation

The test set predictions can be obtained using the predict function. The confusion matrix can be obtained using the table function.
From the table, calculate the accuracy by summing the diagnal and divide by total.

Submit You have used 1 of 2 attempts

ℹ Answers are displayed within the problem

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

English ▼