Course › Predictive Analytics › Homework #5 › Homework 5

Homework 5

The first 8 questions are designed to conduct predictive analysis on the Bank Marketing dataset. The dataset is obtained from the UC Irvine Machine Learning Repository. It contains data related with direct marketing campaigns (phone calls) of a Portuguese banking institution. The outcome is a binary variable with header “Y” indicating whether the customer subscribed a term deposit after the call. For a detailed description of the dataset, please check the source: https://archive.ics.uci.edu/ml/datasets/Bank+Marketing. The exact dataset we use for Part I is a subset of the Bank Marketing dataset. The url that you can use to import the data on AzureML (or download as csv if you prefer that) is:

https://docs.google.com/spreadsheets/d/1a8FyGzrQKTsUGZXEQFC7HuA6XAlJ1gUoQXCcKwA47ik/pub?gid=1921689680&single=true&output=csv.

Question 1

1/1 point (graded)

What is the average conversion rate of this marketing campaign? The conversion rate is the fraction of the outcomes where the outcome (i.e. the dependent variable) is “yes”. You can use whatever method you want to answer this question, such as SQL, a different routine in AzureML, or even downloading the data and examining it on your computer.

○ 15.33%

● 11.52% ✓

○ 3.89%

○ 2.85%

Submit

You have used 2 of 2 attempts

Answers are displayed within the problem

Question 2

2/2 points (graded)

First, assume that we want to use BALANCE and AGE as the explanatory variables, with Y as the dependent variable. Build an experiment, as covered in the lectures, to do this. Use 70% of the data as the training set, with the remaining 30% as the test set and default random seed of 0 as usual. Use the two-class decision forest, and as in class, set the number of decision trees to 1. Set the maximum depth to 3 (please note: this is different from the lectures where we had maximum depth 4). Leave all other parameters in the decision forest at their default values. After you complete building and successfully running your experiment, check the box below. Then, continue to answer the remaining questions using the output of your tree.

☑ I have successfully built and run the experiment.

✓

Submit You have used 2 of 2 attempts

✓ Correct (2/2 points)

Question 3

1/1 point (graded)

Which of the range does the probability of a customer who is 65 years old and has a balance of $698 in his account subscribes a term deposit after the phone call?

● 0 – 30% ✓

○ 30% - 50%

○ 50% - 70%

○ 70% - 100%

Submit You have used 2 of 2 attempts

Answers are displayed within the problem

We will now change the model, to increase its accuracy. Now we will use AGE, BALANCE, DURATION and PREVIOUS as the explanatory variables. Build a decision forest with one decision tree with maximum depth of 4 (instead of 3 earlier) using 70% of the data as the training set and test the model on the remaining 30% of the data. Use your decision tree to answer Questions 4 - 6.

Question 4

1/1 point (graded)

Within which range does the accuracy of the decision tree lie?

○ 0 – 25%

○ 25% - 50%

○ 50% - 75%

● 75% - 100% ✓

Submit   You have used 2 of 2 attempts

Answers are displayed within the problem

Question 5

1/1 point (graded)

Look at the classification matrix and find the number of false positives. In what range does this number lie?

○ 0 – 25

● 26 to 75 ✓

○ 76 to 999

○ 1000 to 1400

Submit   You have used 2 of 2 attempts

✓ Correct (1/1 point)

Question 6

1/1 point (graded)
If the model predicts the outcome will be “yes” for someone, what is the approximate likelihood that the prediction will be correct?

○ 0 – 20%

○ 20% - 40%

● 40% - 70% ✓

○ 70% - 100%

Submit You have used 2 of 2 attempts

✓ Correct (1/1 point)

Question 7

1/1 point (graded)
Which of the following is true if we drop the threshold from 0.5 to 0.1?

○ Since the number of true negative decreases, the accuracy of the model decreases

○ Since the number of true positive increases, the precision of the model increases

○ The area under the ROC curve decreases

● The recall rate increases ✓

Submit You have used 2 of 2 attempts

✓ Correct (1/1 point)

Question 8

1/1 point (graded)

We call the decision tree built using AGE and BALANCE as explanatory variables as tree A, the decision tree built using AGE, BALANCE, DURATION and PREVIOUS as the explanatory variables as tree B. Which of the following comparison between tree A and B is correct?

○ The accuracies of both trees are close to each other, so the predictive power of tree A and B are also close

● Judging by the shape of the ROC curve, tree B is a better predictive model ✓

○ Adding DURATION and PREVIOUS does not make the model better

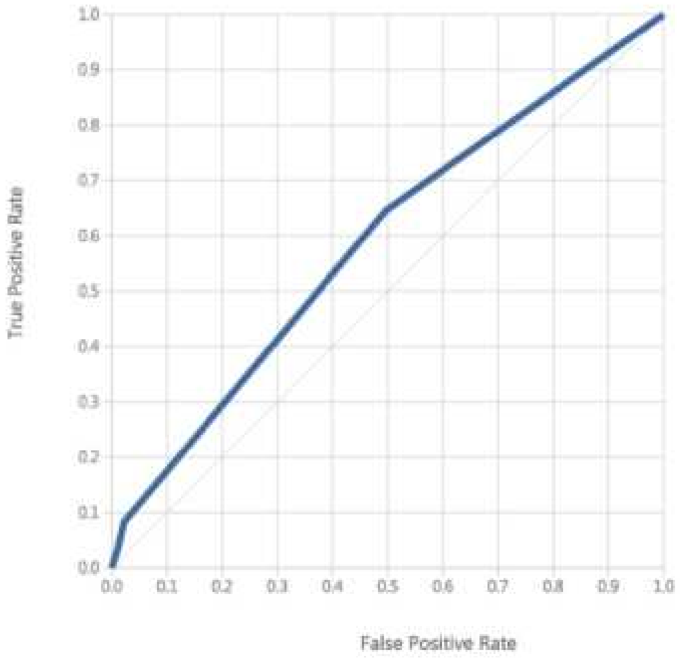
○ There is not enough statistical evidence based on which we can compare tree A and B

Answer

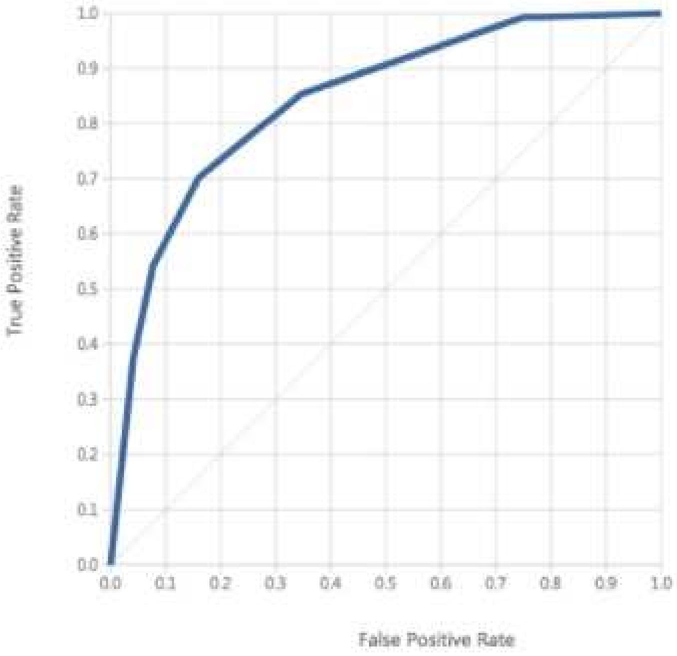
Correct:

The ROC curve for tree A (age and balance only, depth 3) is shown first, followed by the ROC curve for tree B (age, balance, duration and previous, depth 4). As you can see, the ROC curve for tree B has greater area under it, and the curve is more towards the upper left, indicating better predictive power.

Tree A:



Tree B:



Submit You have used 2 of 2 attempts

✓ Correct (1/1 point)

The remaining questions do not require use of the AzureML software, and are purely conceptual.

Question 9

0/1 point (graded)

Which of the following task is not purely predictive analytics?

○ Among several pictures of animals, identify which ones contain a cat

○ Determine the selling price of a new product ✓

○ Estimate the conversion rate of a marketing campaign

● Forecast the winners of the Academy Awards ✗

Submit You have used 2 of 2 attempts

ℹ Answers are displayed within the problem

Question 10

1/1 point (graded)

Linear regression may not be appropriate for which of the following situations?

● Predict the winner of a game using the past number of winnings of the two teams playing in the current season ✓

○ Predict the download rate of a mobile app using the depth of price discount

○ Predicting the stock price of a firm based on its revenue reported in its recent financial report

○ Predicting the number of sales of a product using its price

Answer
Correct:
Answer explanation: Linear regression is most appropriate when the dependent variable (the quantity being predicted) is on a numerical scale. Download rates, stock prices, and number of sales are such quantities. Predicting the winner of a game is a binary outcome (one of two teams will win), so linear regression is not appropriate, although other machine learning methods may be appropriate.

Submit You have used 2 of 2 attempts

✓ Correct (1/1 point)

Question 11

1/1 point (graded)

Which of the following statement about decision forest is correct?

○ For the same set of data, increasing the depth of the decision forest will always improve the prediction accuracy

○ Increasing the number of explanatory variables will always improve the prediction accuracy of the decision tree

● Compared to linear regression, one of the advantage of decision forest is that it can capture non-linear relationships between explanatory and dependent variables ✓

○ In order to improve the prediction accuracy, the training set should be as large as possible

Answer
Correct:
Answer explanation: Although making the tree deeper or adding more variables (options a and b) may improve the model, there is no guarantee that this will happen.

Submit You have used 2 of 2 attempts

✓ Correct (1/1 point)

Question 12

1/1 point (graded)

The confusion matrix of a decision forest is given as follows:

True Positive	False Negative	False Positive	True Negative
254	90	53	863

What is the value of precision?

○ 88.6%

● 82.7% ✓

○ 73.8%

○ 78.0%

Answer

Correct: $Precision = \frac{TP}{TP+FP} = \frac{254}{254+53} = 82.7\%$

Submit

You have used 2 of 2 attempts

✓ Correct (1/1 point)

Question 13

1/1 point (graded)

Suppose you use linear regression to predict the price of houses in your local area, using their size as the explanatory variable. The equation you build using linear regression is Price=120+0.12Size, where Size is measured in square feet. What is the predicted price for a house whose Size is 1500 square feet?

○ 120.00

○ 120.12

○ 180.00

● 300.00 ✓

Submit

You have used 2 of 2 attempts

✓ Correct (1/1 point)

Question 14

1/1 point (graded)

When using linear regression to provide forecasts, why is it more useful to provide an interval forecast, instead of just a point forecast?

● The point forecast is almost certainly not going to be correct, so a narrow interval forecast is more useful. ✓

○ Interval forecasts are easier to construct compared to point forecasts.

○ Interval forecasts are usually too wide to be useful, so it is actually not true that they are better. Point forecasts are generally more useful.

○ Interval forecasts always take into account more explanatory variables compared to point forecasts.

Submit You have used 2 of 2 attempts

✓ Correct (1/1 point)