James Taylor

Assignment 3

University of Maryland University College

DATA630-9041 Spring 2020

Dr. Gates

james.taylor.ps3@gmail.com

**Introduction**

Banks have multiple ways of making money, but one of their most effective ways are loans. Access to loans and interest rates are a crucial driver of the world economy’s growth and business investment. Banks need money before they get to loan it out of course. One way they raise funds is by offering certificates of deposits (CD), which promises someone a return on money they give to the bank, but cannot get back until an agreed upon future date. Banks profit off this money when the return on a loan the bank issues is greater than the return the bank guaranteed to the depositor.

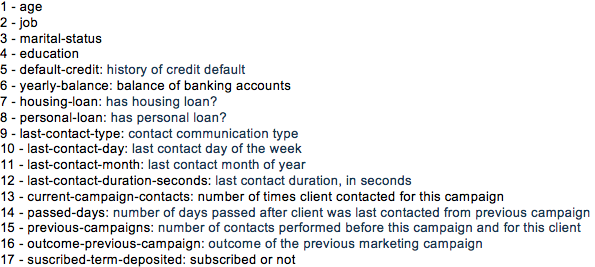
Just like any other business, banks want more customers to drive profit. Advertising and marketing is one way to attract new clients. Money spent marketing to potential clients that are not interested in CDs is waste of resources. Is there a way to get more out of that marketing dollar? The easy answer is targeted marketing, a common marketing approach were dollars are spent only on people who the product is built for. A bulldozer company would not advertise to elderly grandmother for example, and would be smart not to.

The purpose of this paper is to build a decision tree model meant to predict which individuals a bank should target for getting CDs from. This will help banks learn what kinds of people participate on CDs so as to not spend money and tie on others who are not likely to get a CD. An analogy would be like gambling, the bank bets marketing/advertising money on an individual getting a CD. If they end up getting a CD the bank wins, but if they don’t the bank lost those dollars. The model is meant to show which people the bank to focus on so that banks gamble wins more often.

**Analysis and Model Development**

About Dataset

The dataset is from a Portuguese banking institution’s direct marketing campaign from May 2008 until November 2010. The information was collected from calls, the purpose of which was to get bank customers to participate in a bank term deposit. There were 4,521 observations on 17 variables. The dependent (target) variable is whether the individual subscribed to the term deposit or not. Below are the variables and their meanings.



Preprocessing

There were no missing values in the dataset. The yearly balance was forced to be an integer. The ‘passed-days’ variables had a ‘?’ in observations where that client had not been contacted before. These were changed to the numeral value zero so the model recognizes the variable on a continuum, and not categorical. If this was not done, the model would treat 24 and 25 as two mutually-exclusive categories instead of values being one unit apart. The model can use nodes that split the data along its values, like greater than or less than a particular number. The variables last contact type, day, month and duration were all removed because for our model’s prediction, these variables will be in the future. This is because the last contacts’ variable’s value was the last call when the customer either joined or declined to participate in the offer. Imagine a scenario where the client has been contacted already but has yet to confirm if they want to participate, and the model is being used to predict if they will subscribe. The ‘last-contact-duration-seconds’ variable is in the future from this perspective because the client has yet to give a yes/no answer (which is also when the last contact type, day and month is taken down).

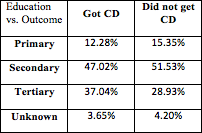
The data was partitioned into training and test data. The training set was 70% of the original dataset and 3,181 observations, leaving 1,340 observations left to test the model with 13 variables.

Exploratory Data Analysis

People of different ages have different priorities with their money. To put it simply, people are either saving money for retirement or spending their retirement. That can be seen here. Retirees have access to lots money but it has to last and keep up with inflation. Investing it in the stock market is risky for the elderly because if they lose money, they do not have the time to make-up for the mistake (whereas the young have maybe 40 years for value to increase). A CD has a guaranteed return and no market risk, making it an enticing investment for retirees. Working-age people priorities higher returns, which CDs don’t often have.



Education level and getting a CD does not appear independent. People with tertiary education are more likely to get a CD than not. Unlike the tertiary level, primary, secondary and unknown education levels are more likely to not get a CD. Education is usually correlated with higher incomes, so tertiary educated may have more money to invest with.



Having money in a bank account doesn’t correlate to more CDs. The rate is nearly uniform across all account balance levels. People likely have money in investment accounts and no in savings accounts that normally have little return.



The success rate appears to be linked to the success of previous campaigns. Individuals who signed up on previous campaigns were far more likely to commit to this CD campaign, at a rate of about 65%. There may be some exterior influences appearing here. The campaigners may know (intuitively) clients that signed up before are likely to agree to this campaign, so they apply extra pressure or work harder to sell them the CD. This same campaigner, if they knew a particular client denied every campaign they’ve been offered, may look at this client as a lost-cause and use little effort to sell the idea. This would be introducing bias from prior-knowledge.



Model

The model will be a decision tree (ctree() function from R package ‘party’) built on the training subset of the dataset. This model is specifically a conditional inference tree, where a model subsets data on conditions to maximize the ‘purity’ of terminal sets. Using the tree analogy, these are the leaves, and the purity is the model’s ability to accurately predict on each leaf. In our example, a pure terminal set would have a model predict either all ‘yes’ or ‘no’ for observations that meet the criteria on that leaf. If over-fitting is a concern the decision tree can be ‘trimmed’, removing branches that over-fit on the training data.

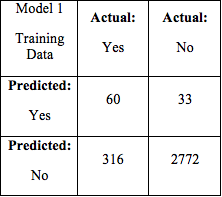
**Results and Model Evaluation**

****

The figure above shows the decision tree on the training data with specifications on number of nodes or branches. Given the correlations in the training data, the model has previous campaign outcome as the first node. If the previous outcome was a success, the model has a terminal node predicting that observation to have gotten the CD. Since the model has a terminal node on just one criteria, that previous outcome variable is a significant predictor of CD subscriptions. The terminal set is not very pure, reporting a 35/65 percentage of ‘no’ and ‘yes’ respectively. If the model was able to be more confident with more nodes on that branch it would have.

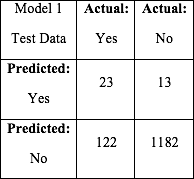
If an observation was not successfully participating in a previous campaign, the node that spits them next is on the ‘passed days’ at 222. Observations that had greater than that ended at a terminal node predicting the observation to not get the CD. The rest of the observations, less that 222 passed days, are then split on their house loan status. Observations that did not have a mortgage were split again based on their job. Both terminal nodes from the job splitting predicted that observation to not get the CD, although purity of the sets were significantly different.

The below is the confusion matrix for the model where ‘yes’ is getting the CD. The model was 89% accurate on the test data. Off the two types of errors, type II error was by far the most numerous. The model only predicted 16% of the customers to get CDs when they in fact did, or sensitivity.



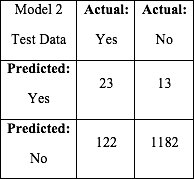
Model Accuracy on Test Data

The model preformed just as well on the training than the test data, returning a 90% accuracy. This is evidence that the model is not over-fitted to the training data because the accuracy metrics are approximately the same. The model again committed many type II errors, or false negatives. This value is crucial for the purposes of our model because it is meant to maximize efficiency of resources spent on getting customers to participate in a CD, but it is also leaving behind customers that could have been customers. These are customers that would have signed up but were not contacted because the model recommended that it wasn’t worth the campaigner’s time.



New Model to Reduce Type II Errors

In an attempt to reduce the type II errors, another model was built with less variables. These were age, education, job and previous campaign outcome. These variables were chosen because the exploratory analysis showed the elderly, educated and previous campaign participants are significant participants in the campaign. This is the theory by which these variables were chosen. The model and confusion matrix is below.



This model has the same confusion matrix as the other model with all variables included. The code was checked multiple times to confirm the accuracy of the confusion matrix. While double-checking, it was discovered if the decision tree was only allowed to predict on the previous outcome variable, it still had the same confusion matrix as the first model.

Analysis

The models only used the previous campaign outcome variable in predicting which observations are going to sign up for the CD. All other nodes did not distinguish between ‘yes’ and ‘no’ on its predications, just different kinds of ‘no’s. It is possible another model could have been built that included a terminal node which would predict a ‘yes’, but this would have concerns with overfitting and data mining (getting a model to display a biased, preconceived notion by manipulating it until you see it).

This is one of the downsides of predicting a Boolean variable with a decision tree model instead of a logistic regression. The decision tree model fits the observation into a True or False value depending on where if fall in the tree, whereas a logistic regression produces a likelihood value that can then be rounded into the True/False predictions. There is no way to adjust the recommendation on the back-end of a decision tree, but a logistic model can have different rounding thresholds.

The models were better at accurately predicting which clients won’t get a CD than which will. Of test data clients that actually got a CD, the had a sensitivity of 16%, compared to a specificity of 99%. Either this sample of data doesn’t have a significant relationship for getting a CD, or those who get a CD have similar attributes to those that don’t get a CD.

**Conclusion**

The created models only used the outcome of previous campaigns to predict which clients were going to get a CD, successful previous outcomes being predicted to get a CD. They were both about 90% accurate but they marked many clients as unlikely to get a CD but in fact did (false negative).

For the purpose of these models, those would-be customers were not contacted because the model predicted them not to get a CD. Those are lost potential customers. More data is needed to get better predictions and find more correlations to getting a CD. A logistic regression may be a good option for this prediction due to the ability to round up and specific thresholds. That would allow for lower levels of false negatives, but more false positives as well.

The models may still be useful however. They accurately predicted many clients to not get a CD when they did not. This could save lots of time/resources spent on calls by limiting or removing these people from the list of potential CD participants.

References

B Fryar CD, Chen T-C, Li X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999–2010. NCHS data brief, no. 103. Hyattsville, MD: National Center for Health Statistics; 2012. Retrieved from <https://www.cdc.gov/nchs/data/databriefs/db103.pdf>

Coronary Heart Disease. (n.d.). National Heart, Lung and Blood Institute. Retrieved from <https://www.nhlbi.nih.gov/health-topics/coronary-heart-disease>.

How Heart Attacks became Less Deadly. (n.d.). Harvard Health Publishing. Retrieved from <https://www.health.harvard.edu/healthbeat/how-heart-attacks-became-less-deadly>

Decline in Deaths from Heart Disease and Stroke -- United States, 1900-1999. (n.d.). Center For Disease Control. Retrieved from <https://www.cdc.gov/mmwr/preview/mmwrhtml/mm4830a1.htm>