1. **Before you use the response for anything**, look at the distributions of all of your predictors. Are there any with a large proportion of missing values? Ignore these variables. Are there any that have a very narrow distribution (e.g., almost entirely 0s or entirely 1s)? Consider ignoring these or transforming/combining them in some sensible way if you can think of one. Feel free to examine any crosstabulation tables between some sets of two predictors as well to see if anything jumps out.
   1. Missing: (ACCTAGE – 546, PHONE – 1075, POS – 1075, POSAMT – 1075, INV – 1075, INVBAL – 1075, CC – 1075, CCBAL – 1075, CCPURC – 1075, INCOME – 1537, HMOWN – 1463, LORES – 1537, HMVAL – 1537, AGE – 1702, CRSCORE – 195)
   2. 0’s and 1’s: (NSF, IRA, LOC, ILS, MTG, SDB, MOVED, CDBAL, DEPAMT, IRABAL, MTGBAL­, NSFAMT, CASHBK, DIRDEP, CD, CDBAL, LOCBAL, ILSBAL, MM, MMBAL, MMCRED, INAREA)
2. Which of your predictors (continuous and categorical) do you think might be important to your problem? Why? This can be based on subject knowledge, literature, test results, or whatever you feel might be important. Fit a logistic regression model with these variables. (If you have no idea and are only going off test results to decide what goes into your model, that’s fine.) Give an interpretation (including the confidence interval) of the odds ratio for the predictor with the largest estimate (in magnitude).
   1. Potential customers with savings and checking accounts and a large amount of funds in their savings and checking accounts would show responsibility and a trend towards financial management. Also, given that many of the variables are zero-inflated or contain missing observations and have removed from the dataset, the few variables leftover should be given a chance to explain variation.
   2. ins = dda + ddabal + dep + checks + teller + sav + savbal + atm + atmamt + branch
   3. The variable with the highest magnitude estimate was BRANCHB14 at -.992. The odds ratio for BRANCHB14 is 0.371 which means customers who bank at branch B14 are 0.371 times more likely to get insurance than customers who do not bank at branch B14 with a 95% confidence interval of (0.254, 0.532).
3. Think of an interesting comparison involving multiple predictors. Compute and interpret the odds ratio for these two subjects.
   1. It would be interesting to compare whether individuals who have a checking or a savings account would be more likely to get insurance. The odds ratios foring savings and checking accounts are 1.779 and 0.379, respectively. This would suggest that customers who have a savings account 1.779 more than those without a savings account to get insurance while customers with a checking account are only 0.379 more likely to get a checking account than those without a checking account. So if all you knew about a potential customer if if they have a savings account or checking account you could guess customers with savings account are more likely to get insurance.
4. The dataset has several variables that might have redundant information (e.g., money market account and money market balance) or might be indicative of the same underlying phenomenon (e.g., teller visits and phone number banking could represent something like actual human contact with the bank). Is anything like this in your model? If so, why do you feel like you need to keep both? (There’s no right or wrong answer.)
   1. We have three sets of variables that might appear to be possibly redundant: savings accounts and savings account balance, checking account and checking account balance, and ATM and ATM withdrawal amount. For both savings and checking it would seem logical to remove both in the model because someone may have an account but have very little to nothing in it giving them a small balance. There is most likely a cutoff value that could be determined in the balance column that would negate the need for the existence of account column. ATM would follow a similar logic because a very large withdrawal from the ATM might indicate instability with finances. We kept all of them in the model because the full model had an AIC of 9835.5 while the reduced model had an AIC of 10107. So, based on preliminary research the full model appears to be better.
5. How many of your predictors have missing values? Earlier, you ignored predictors with a large number of missing values, which is a perfectly valid thing to do—the idea being that they might be likely to be missing in the future as well and thus may not be useful for the application of your model.1 How many observations have missing values? You should keep in mind and make a note of how much of your sample is being discarded when we only do a complete case analysis. Dealing with missing values is challenging to do accurately and beyond the scope of this class, so for now we won’t worry about it aside from noting it here.
   1. 15 of the predictors had missing values.
   2. In total 3034 observations had a missing value.
   3. We removed variables with missing values rather than rows which kept the same total number of observations in our dataset, but information could have been lost in the dropped variables. Looking at the dropped variables’ histograms there are some variables that would interesting to including such as income, age, age of oldest account, and credit score. Credit score in particular would be worth keeping in the model considering it only has 195 missing observations. Testing this returns a model with an AIC of 9857.6 which is worse than our previous best model and the CRSCORE estimate is also insignificant as alpha = 0.5.