OUTLINE

Method

1. Cleaning Data
   1. Deleting default data [(my source: https://towardsdatascience.com/machine-learning-case-study-a-data-driven-approach-to-predict-the-success-of-bank-telemarketing-20e37d46c31c)](https://towardsdatascience.com/machine-learning-case-study-a-data-driven-approach-to-predict-the-success-of-bank-telemarketing-20e37d46c31c)
   2. Code for Duplicate Column
   3. table(is.na(bank)) False if no NA
   4. Deleting unknowns shrinks observations from  41k->38k
2. Make Predictive Classification Models using different Algorithms
   1. Regression tree
   2. Classification w/ Lasso Penalty AIC(logit model with lasso regularization)
      1. Done simple
   3. Classification w/ CV
      1. Done Simple and uses same code as AIC
   4. Random Forest Tree
      1. Done- analysing to make sure it makes sense
3. Compare each model:

1. OOS prediction error/fit using CV or AIC Vs Is Pred
   1. Roc curves tell us how they compare
   2. Expectation: similar but IS may be better
2. Choice of low threshold b/c we care more about getting potential clients to say yes
3. Find In Sample Deviance for each model
4. ROC Curves - We care about the OOS ROC
   1. Important Definitions
      1. Specificity: True negative
      2. Sensitivity= True Positives
      3. 1- Sensitivity= False negatives
      4. 1- specificity= False positives
      5. {pred <- predict(model1$gamlr, X, type ="response")
      6. pred <- drop(pred) ## Remove the sparse Matrix formatting
      7. # false positive rate
      8. sum((pred <0.2)[y ==0]/sum(pred>0.2)) # p=0.2
      9. # false negative rate
      10. sum((pred <0.2)[y ==1]/sum(pred<0.2))
      11. # sensetivity
      12. mean((pred>1/5)[y ==1]) # p =1/5
      13. # specificity
      14. mean((pred<1/5)[y ==0])
      15. }

1. Metrics for choosing best classification models
   1. Accuracy: Fraction of predictions correct:P+TN/TP+TN+FP+RN
      1. Expect to be high b/c bad with class-imbalanced data set( see analysis of class variables in data clean section)
   2. Precision&Recall- Affected by threshhold
      1. Precision: proposition of positive correct=TP/TP+FP up to threshhold
         1. Precision in our model  % of clients predicted to subscribed correctly identified
      2. Recall= proposition of actual positives correctly identified TP/TP+FN
         1. Recall in our model= % of positive subscribers identified
   3. ROC shows true positive rate(sensivitivy rate) vs False positive rate( 1- specificity): TPR(TP/TP+FN) & FPR(FP/FP+TN)
      1. AUC= area under - not affected by threshold
         1. Probability model ranks positive example higher than a negative example
         2. Measures quality of prediction regardless of threshold- pro
            1. Con- sometimes threshold matters i.e. we want to minimize false positives even if it increases false negatives. In oir case, doesnt matter
         3. No scale - pro and con
2. Determining casualty aka making an affect estimation model aka looking for bias: Adding Controls and looking at our models chosen coefficients/interactions
   1. Regressors Used in our top choice of model( aka X or variables kept in model)
      1. Compare coefficients
      2. Interaction terms/powers
   2. Some analytic methods
      1. Plot mosaic like Lec 4 1.1 or in taddy’s credit code
         1. We see some selection bias
         2. Do it for highest coefs/betas
   3. Intuition based sources of Bias
      1. Selection Bias-
         1. Job/anything that would make caller choose a client to class- CONTROL for this and see outcome
      2. Previous Contact what type of bias

Results

1. Univariate/Mulvariate Analysis of independent variables
   1. Class distribution of Y - will affect metric used to evaluate of classification models to achieve desired goal
2. OOS Pred at X threshold, OOS ROC AUC for each model
   1. AIC Lasso(classification)
   2. CV Lasso(classification)
   3. Random Forest(CAT)
   4. Casual Tree(CART)
   5. Add graph of all ROC curves layered on top of each other
3. Limitations/Issues with our Analysis
   1. Issues with the predictive model
   2. Causality: applying the predictive model to population/future
      1. Observational data so we have to make sure independent variable interactions aren’t in error term i.e. where could there be omitted variable bias or selection bias
         1. Potential Selection Bias
            1. Contac

This of course creates a counterfactual situation where future sets are now affected as we can see previous contacts has a high estimation

* + - * 1. Job

Callers choosing clients with high paying jobs

* + - * 1. Default