**SparkCognition Data Science Contest**

**Model created:** Marketing Response Model

**Language of Choice: R (packages used: randomForest,** **caret,** **dplyr,** **Hmisc,** **pROC,ROSE)**

Q1. Describe your model and why did you choose this model over other

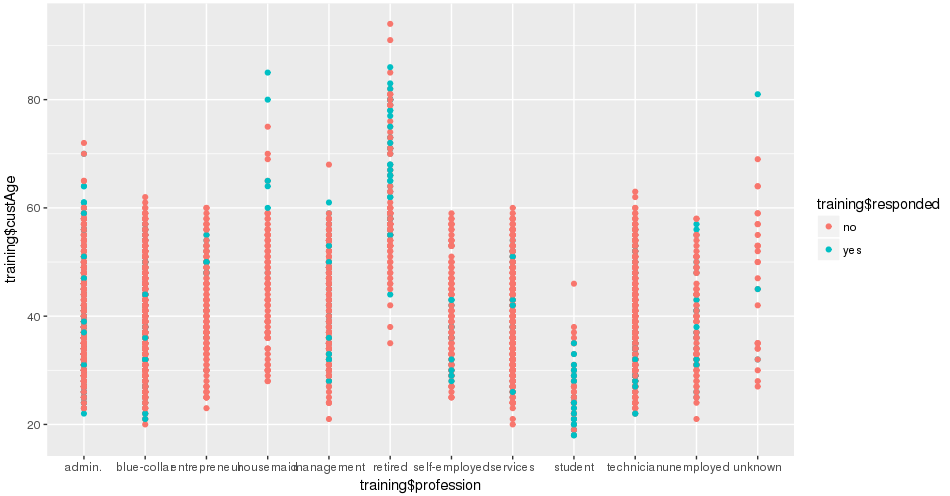
type of models?

A1. Description of model: The data represents whether or not a marketing campaign sent to a customer got any response or not. The data contains a lot of categorical variables and some decision rules can be created to divide the data into responded=yes and responded=no buckets. So necessarily it is a classification problem with some decision rules. Random forest which represents bootstrap aggregating of multiple decision trees is chosen as model of choice after comparison with generalized linear model, support vector machine and gradient boost model.

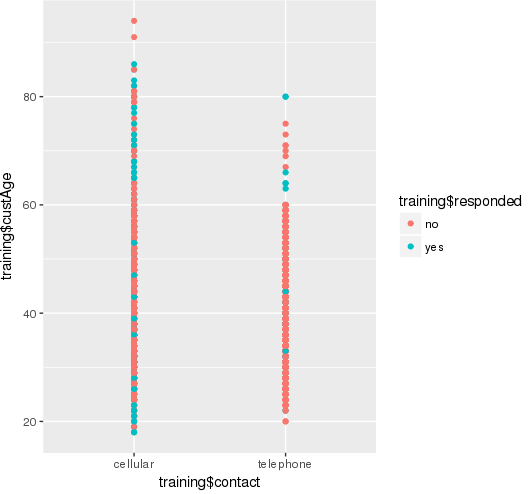
Some examples of decision rules which can be deduced from the data:

1. Student who have age <=30 have 36% response rate as compared to 11% response rate overall
2. Retired customers who have age>=60 have 42% response rate as compared to 11% response rate overall

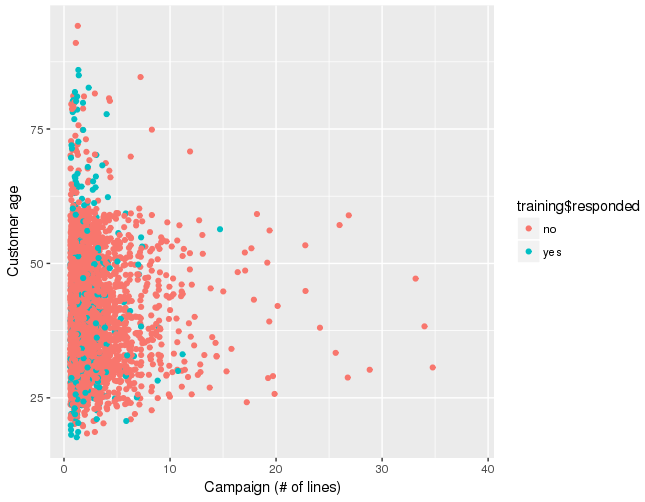
Please see the graph below for evidence on above two claims



1. Contacting potential customers on cellphone as compared to telephone has a better response rate. Please see the graph below.



1. Campaigns with < 18 lines have a much better response rate as compared to campaigns with lengthy text or voice content.



Q2. Describe any other models you have tried and why do you think this

model preforms better?

A2. I tried generalized linear model(glm) which uses logistic regression for classification and assumes that a straight line or plane could be fitted through the data. However our data is much more non-linear and any plane fitted on the data will be less accurate than the creating multiple decision trees and aggregating the result.

Q3. How did you handle missing data?

A3. On checking the summary of the data it could be found that 24.3% of the data has customer age missing. It is a very important variable and has the highest feature importance among all the covariates.

I used predictive mean mapping (pmm) from Hmisc package using aregImpute() function to fix the problem of missing data. PMM produces imputed values that are much more like real values. If the original variable is skewed, the imputed values will also be skewed

How predictive mean mapping works? Suppose variable x has missing values and rest of the variables are Z. We estimate a linear regression of x on Z, producing a set of coefficients C. Now we transform C into C’ by making random draws from posterior predictive distribution of C. Using C we predict all the values of x, even for the cases where the data is not missing. For each of the case with missing x,we identify a set of cases with observed x whose predicted values are close to the predicted value for the case with missing data. Now we calculate mean of the close cases or we can randomly draw one of the close case.

Q4. How did you handle categorical (string) data

A4. Since I am using random Forest model. I don’t need to change the categorical variables. Categorical variables are handled by this model. However , when using logistic regression for classification we need to create multiple dummy variables for each of the categorical variable.

Q5. How did you handle unbalanced data?

A5. 88.7% of the data belongs to a single class. The majority class is responded(=”no”) since most customers will choose not to respond. Hence the data is unbalanced. We can either over sample the minority class responded (=”yes”) or under sample the majority class responded(=”no”). I have used a mixed approach wherein we oversample the minority class and under sample the majority class so that majority and minority class are in 0.7 ratio. 0.7 was selected by using multiple value and checking the performance on the validation set. I have used ovun.sample() function from ROSE package for this.

Q6. How did you test your model?

A6.I used AUC as the performance measure of classification. AUC is the ratio of true positive rate (TPR) and false positive rate (FPR). The focus was on maximizing the AUC while keeping good track of Sensitivity (TPR). TPR is very important metric for us since our model will be used to decide whether or not to contact a customer. We might want to reach out to most of the potential customers even if we have to send out campaigns to a larger group of people. My model has AUC~71.86 , TPR~71.82 and FPR~71.90. We can maximize AUC at the cost of TPR , which I think is not worthwhile.