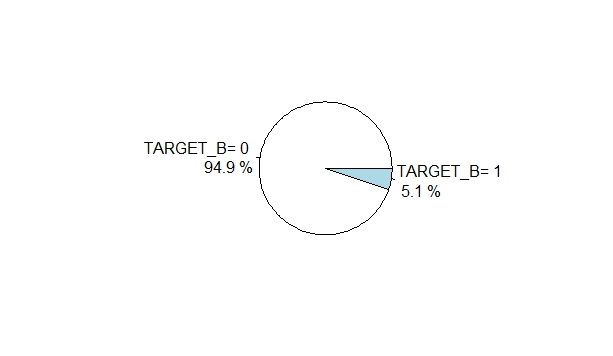
# KDD-project

Customer Response Prediction and Profit Optimization

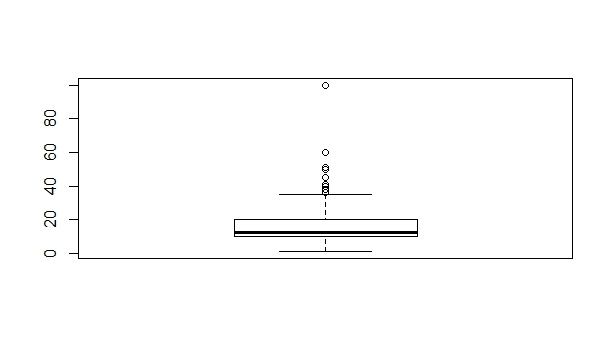
Research question: Estimate the return from direct mailing in order to maximize donation profits. For the purpose of doing feature selection in order to train a model I am using the notion of entropy. Entropy is a measure of disorder, features with low entropy express high variability and are better candidates. I have selected 30 features that have met this criteria. There are 2 target variables in the dataset TRGET\_B=binary variable indicating if participants responded to the mail or not and TARGET\_D= donation amount in $.

 Distribution of response (TARGET\_B) is shown in the pie. As we can see 94% of participants showed 0 response to the direct mail and 5.1% people were the ones who responded. In order to find out how many respondents made positive donation amounts we ran a code where TARGET\_D>0.

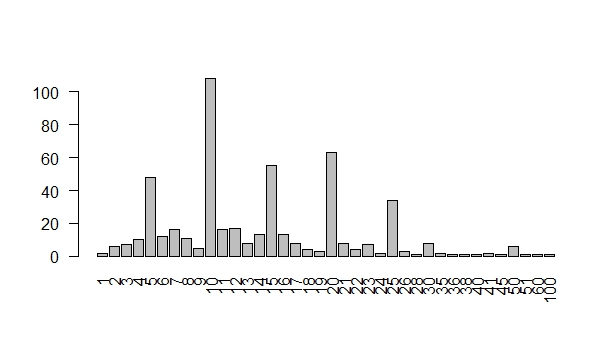
Min. 1st Qu. Median Mean 3rd Qu. Max.

1.00 10.00 12.00 14.58 20.00 100.00

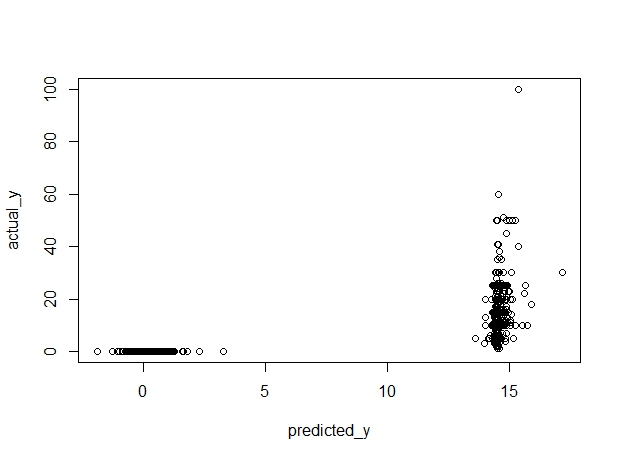
From here we can see that the minimum donation amount was $1, maximum was $100,and $14.58 was the average amount donated by the participants who responded to the mail. The box plot bellow shows the donation $ distribution, the median amount for donations is less than $20, donations range from $1 to $200, data is skewed left, and the inter-quartile range falls between $15-20.



By using a bar-plot, it is also possible to check the number of positive donations. The plot shows that most donations are no more than $25 and are multiples of $5.



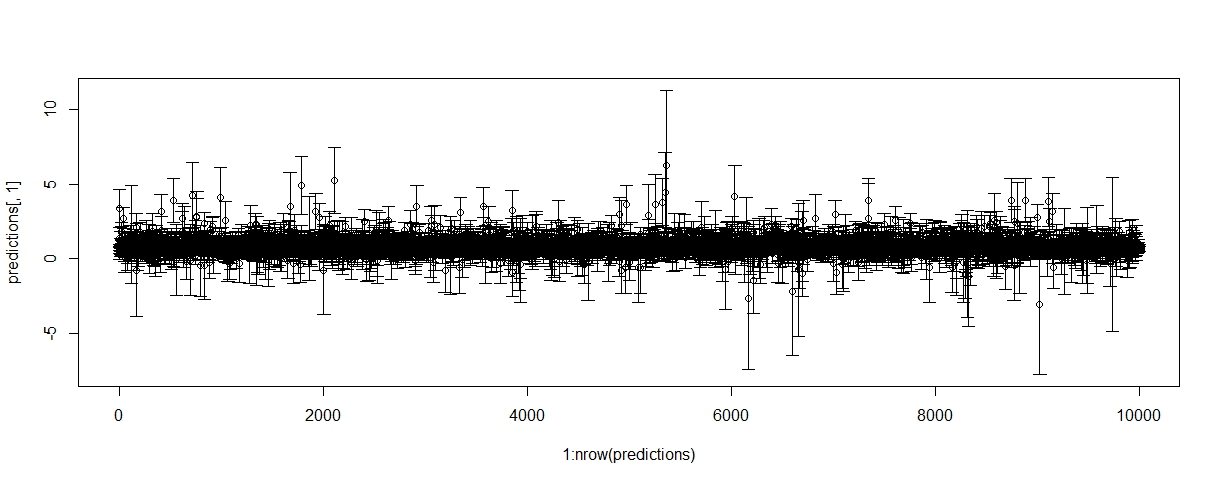
In order to run a T-Test on my dataset I first decided to use regression and create a predicted model for my dataset and only then use T test to compare two groups (the actual and the predicted). Linear method of regression was used and TARGET\_D was the response variable and the rest 30 features were the predicted variables.



The results of my predicted model(predicted\_y) in regards to actual\_y showed that because of high variability of TARGET\_D variable the plots were scattered into two groups; people who made $0 contribution and people whose donation was between $12 to $14- very similar to what the box plot had also predicted. With T tests I will try to measure the accuracy of my model in respect to the actual one.

Ho= true difference in means is equal to zero, Halt= true difference in means is not equal to zero. After the prediction I ran a T test to compare my two groups and here are the results: t=0, df=19374.43, p-value=1, since our P value is greater than zero we keep our hypothesis meaning the predicted model is quite accurate compared with the actual one. We are 95% confident that the error lies between -0.09818937& 0. 09818937 intervals. Our regression and T test showed us that instead of trying to estimate the return with the given dataset, it will be more accurate if try to compute if people will donate or not instead. Since the donations are only forming 5% of entire data, it can be more meaningful to work with the 95% trying to predict whether they will donate or not.

With the help of Confidence Intervals we can measure the error of our predicted model meaning what is the numerical range that our error falls in between. If we assume that our error in the [linear regression model](http://www.r-tutor.com/node/91) is independent of *x*, and is [normally distributed](http://www.r-tutor.com/node/58), with zero [mean](http://www.r-tutor.com/node/35) and constant [variance](http://www.r-tutor.com/node/42), we should be able to construct a Confidence Interval- for a given value of X the interval estimate for the mean of the dependent variable, Y. By using the function predict.lm we are able to construct CIs for all of our rows(observations) in order to be able to compare the means of our actual and predicted models. Since I am moring with more than 2 rows, the display of CI was better usning a plot.Bellow are all the fitted values(X) and their confidence intervals plotted.



I also used the pwr package  to estimate the power of a t-test with unequal variances and with Cohen's d and got 0.0004649306.