Findings

We need a good sensitivity score since we want the targeted audience to respond.

1. Using Classic logistic regression using GLM which takes care of the class imbalance by itself,

    a. The final features are - ['job_retired','job_student','month_mar','month_may','day_of_week_mon','previous_Never_contacted','poutcome_success','cons.price.idx','euribor3m']

    b. the sensitivity is - 97.8%

    c. We prefer to take the cut off probability to 0.4 as opposed to 0.1 as shown in the sensitivity, specificity and accuracy intersection graph since we are interested in a high sensitivity score.

2. Using PCA and Logistic regression.

    a. We are using 22 principal components which explains over 95% data.

    b. The sensitivity is not good - 64%. (This is the most optimistic and highest sensitivity achieved.)

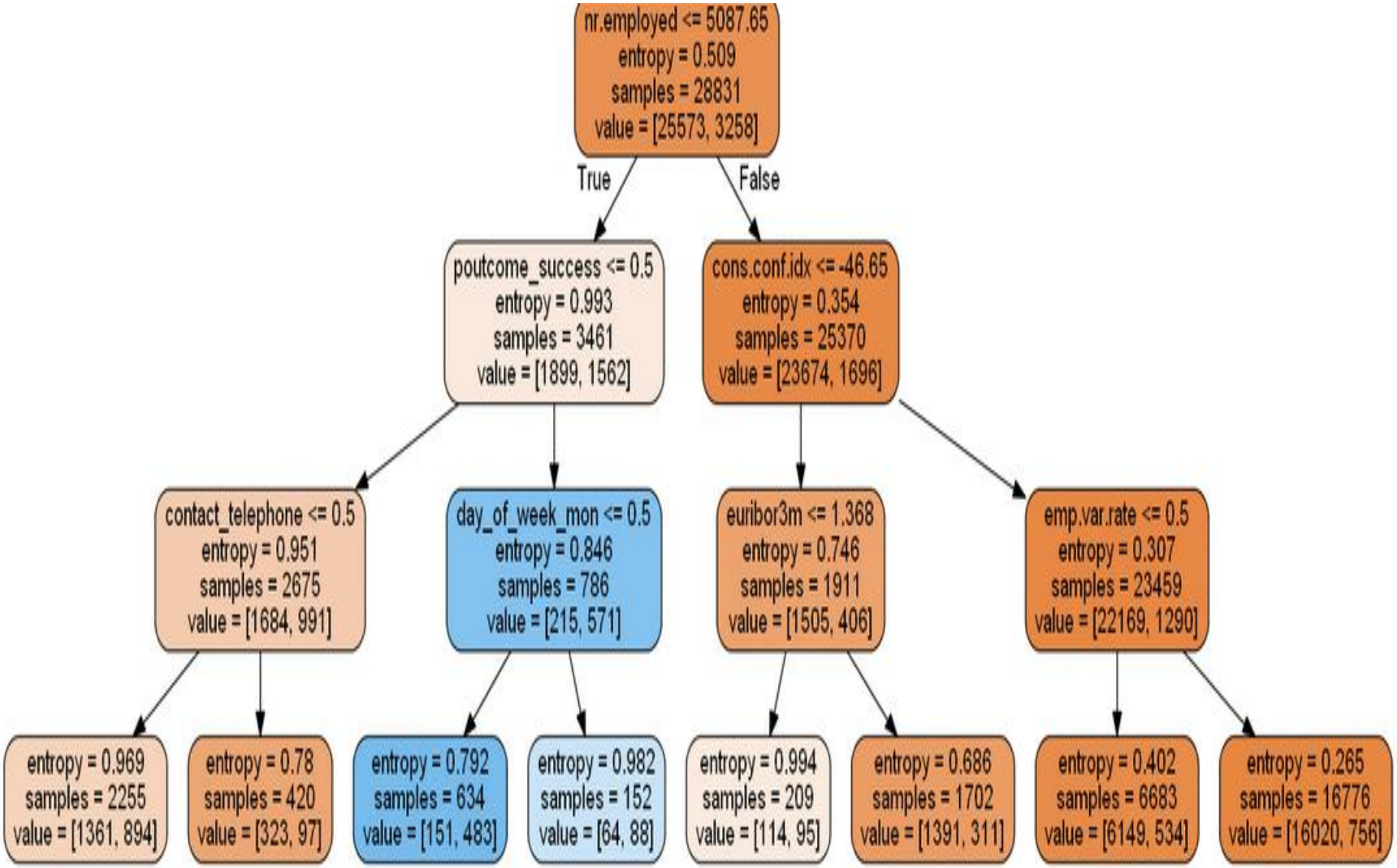
Findings

Using Decision trees on the available training dataset

1. On all features
    a. using, the below values for hyperparameters - we get a score of 89.8%
        min_samples_leaf = 100
        min_samples_split = 50
        max_depth = 5
        criterion = entropy
    b. nr.employed <= 5087.65 (1562,1899)(no. of responders, no.of prospects)
    nr.employed > 5087.65 and cons.conf.idx <= -46.65 (406,1505)
    nr.employed > 5087.65 and cons.conf.idx > -46.65 and emp.var.rate <= 0.5 (534 ,6149)
    9553 prospects(33.13%), 2502 responders (76.79%)

2. On all features selected by logistic regression using rfe
    a. using, the below values for hyperparameters - we get a score of 89.9%
        criterion = "gini",
        max_depth =5,
        min_samples_leaf=50,
        min_samples_split=100
    b. euribor3m <= 1.152 (1374, 1592)
    euribor3m > 1.152 and euribor3m <= 3.24 (963, 5545)
    euribor3m > 1.152 and euribor3m > 3.24 and month_may > 0.5 (181, 5276)
    12413 prospects(43.05%), 2518 responders (77.28%)

Findings – Decision Tree



Findings

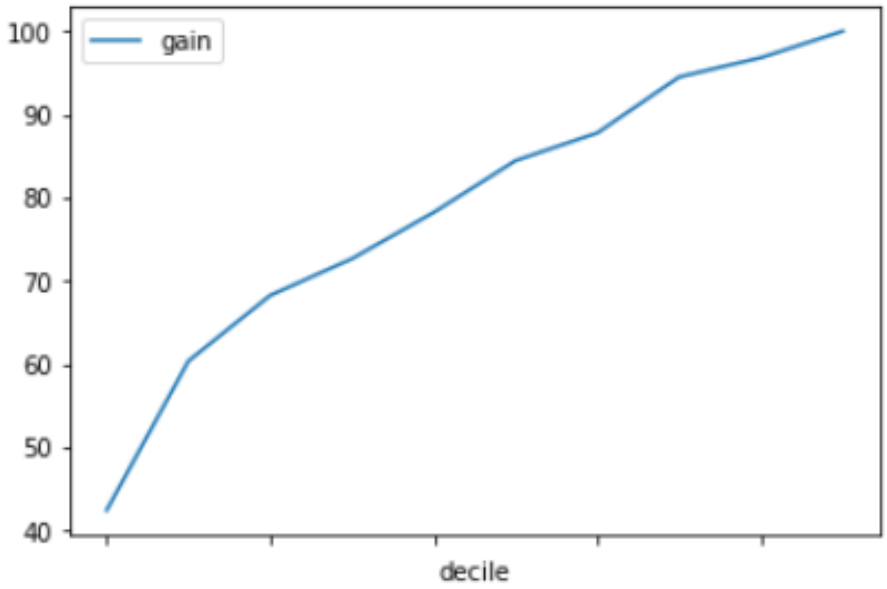
- The decision made on the NON RFE feature narrow down features for logistic regression (i.e. all features) actually gives us a better response by targeting lesser audience.
- This also tells us that features that were actually used for targeting our end audience is not the ones that help us to predict the response of a prospect.

- Ensemble techniques using Random forest on both datasets using all features or using the features selected by logistic regression using rfe, both have a score of 89%

- Using the Decile on the Test set -
- - When we target the 6 Decile, we get a response of 84%
- - No. of prospects - 7410
- - No. of responders - 1167

Findings – Lift Gain Chart

The Gain Chart here shows that there is a steep increase until 60 % (which tells us that most responders will be targeted) and then onwards is gradually sloping which tells you that with increase in population, there will be a gain in responders.



Findings – Cumulative lift

The Lift chart tells us that at the 6th Decile, we get a lift of about 1.4 cumulative lift which means that our response rate is multiplied by 1.4 by targeting 60% of out prospects