ANALYTICS FOR FINANCIAL SERVICES

Project Report and Summary

2018

Scenario A: Technical : Build a model for predictions in financial analytics

Final Project

Name: AHSAN AFTAB
Master of Science in Business Analytics and Big Data
Course: ANALYTICS FOR FINANCIAL SERVICES
Submitted to: **PROFESSOR ANTONIO PITA LOZANO**
Date: 10-06-2018

ahsan.aftab@student.ie.edu

IE Business School, Spain

VERSION-FINAL

ANALYTICS FOR FINANCIAL SERVICES

06-06-2018

A Brief Introduction

The problem provided for this project is that of a binary classification problem.

The problem can also be treated as Logistic Regression problem provided the variable treatment is carried out as such.

We have to predict target variable that is dichotomous in nature.

In finance, such logistic or decision problem are common and require analysis of various variables pertaining to business, customer and time.

We expect the logistic regression and decision trees to perform better over such a problem. Although random forest and tree classifiers can also provide very similar results,

Tools Used for design and implementation

The following tools have been used to analyze the given dataset:

1- **R and Rstudio**

2- **Dataiku, Weka and IBM SPSS** for Analytics and validation

3- **Java Spring with Weka java libraries** for machine learning algorithms.

The results have been shared from various tools however the code provided is the final and working model implemented in R. Accuracy achieved through this model is approx. 70%, though future work might improve further the performance of the model.

The missing values have been treated in different way and are explained in the R code provided and in the text below.

Inside This Report

Data Summary

Training dataset = 522939 instances with class named as 'target'

Test dataset = 174313 instances with class not available and need to ne predicted.

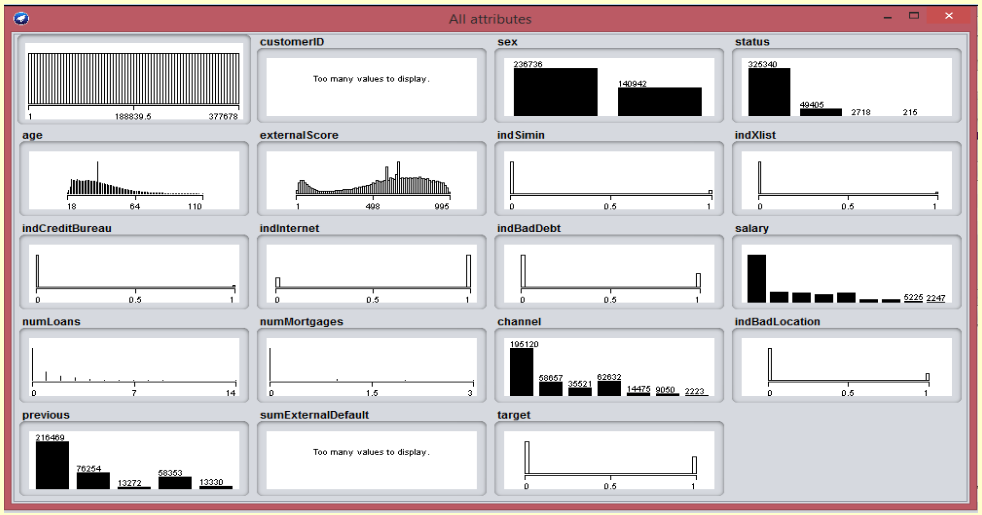
Variables are of type: nominal, numerical, boolean

THE MOST RELEVANT VARIABLES SHOW A CUMULATIVE PERCENTAGE OF MORE THAN 90% IN THE MODELLING OF PREDICTIONS OF TARGET VARIABLE

Summary of results

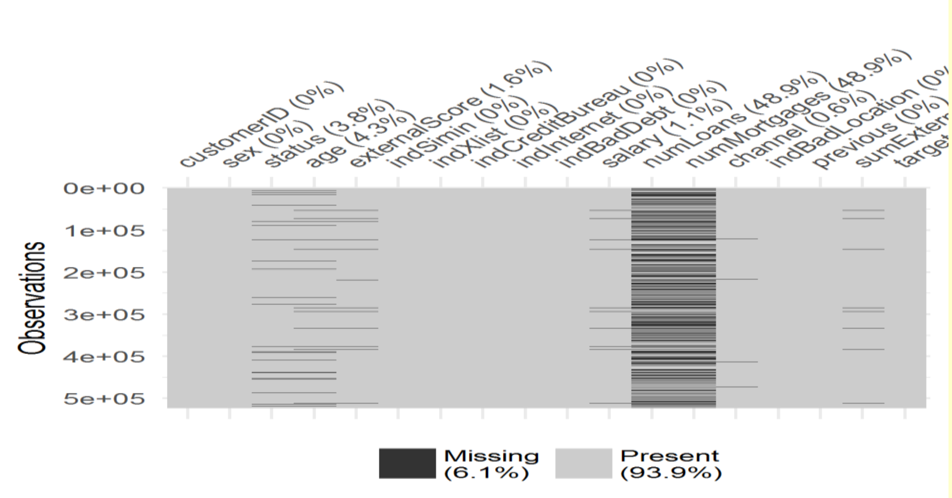
THE DISTRIBUTION OF VARIOUS VARIABLES

The distribution of various variables can be seen in the figure below:



Most of the distributions are skewed to the right.

THE MISSING VALUES IN THE DATA



Missing Values Treatment:

The columns numLoans and numMortgages have more than 50% data missing. We can replace the missing values with 0s.

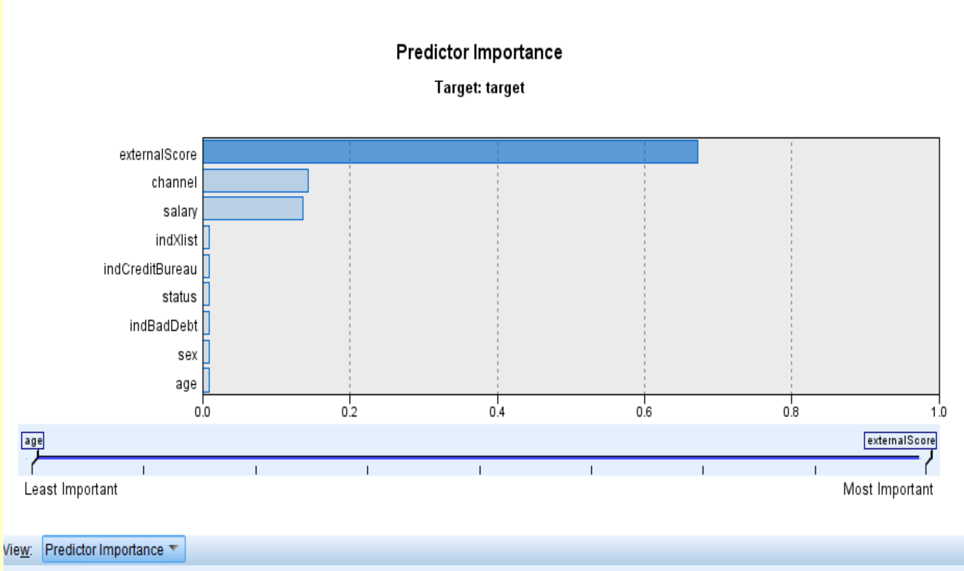
The age variable can be replaced by median since distribution is skewed to the right.

The other missing values can be replaced by 75th quarantile majority value in order to improve accuracy of the model.

The final dataset over which predictions are to be made, also has missing values and will be treated in similar way to training dataset.

THE MOST RELEVANT VARIABLES AS PREDICTORS

The most relevant variables in making predictions are: externalScore, channel, and salary.



Algorithms Used for Modeling

-Binary Tree Classification

-Logistic Regression

-Neural Network

-Random Forest

-Bayesian Network Model

-Decision Tree Model

COMPARISON OF MODELS HAVE BEEN MADE FOR:

The comparison of results for the following algorithms follows:

-Binary Tree Classification

-Logistic Regression

-Neural Network

-Random Forest

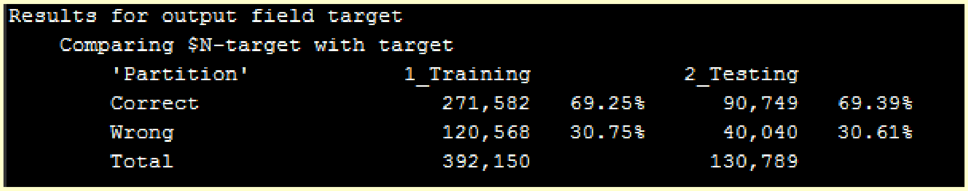
-Bayesian Network Model

-Decision Tree Model

The results showed that the Logistic Regression performed better in terms of resource requirements and results achieved. Accuracy achieved = 69.40%

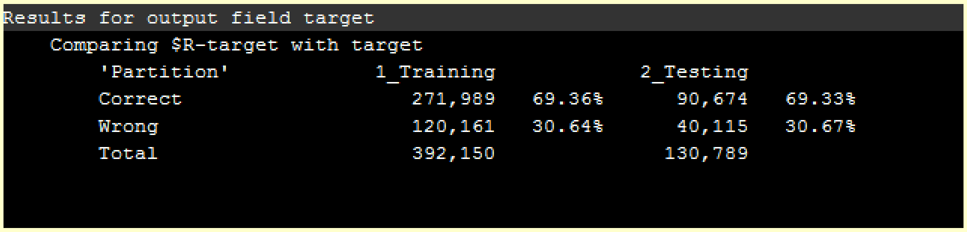
Comparison of Results from different models

Neural network



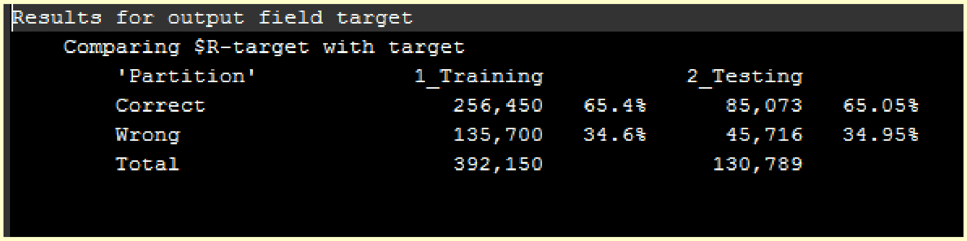
```
Results for output field target
    Comparing $N-target with target
        'Partition'             1_Training             2_Testing
        Correct                    271,582    69.25%      90,749    69.39%
        Wrong                      120,568    30.75%      40,040    30.61%
        Total                      392,150               130,789
```

Decision Tree



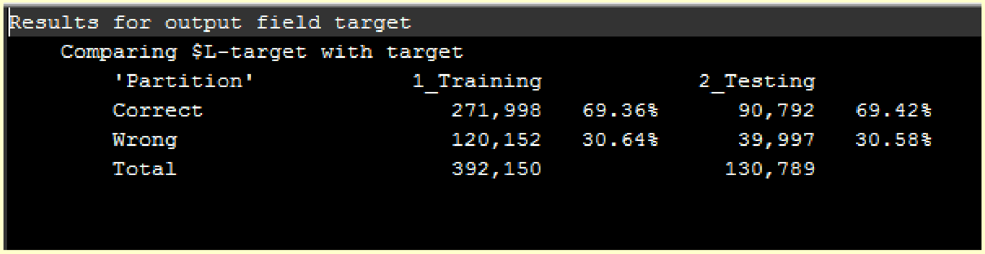
```
Results for output field target
    Comparing $R-target with target
        'Partition'             1_Training             2_Testing
        Correct                    271,989    69.36%      90,674    69.33%
        Wrong                      120,161    30.64%      40,115    30.67%
        Total                      392,150               130,789
```

Random Forest



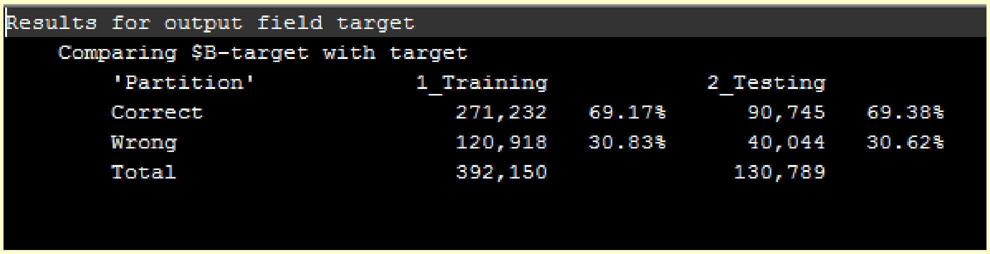
```
Results for output field target
    Comparing $R-target with target
        'Partition'             1_Training             2_Testing
        Correct                    256,450    65.4%       85,073    65.05%
        Wrong                      135,700    34.6%       45,716    34.95%
        Total                      392,150               130,789
```

Logistic Regression



```
Results for output field target
    Comparing $L-target with target
        'Partition'             1_Training             2_Testing
        Correct                    271,998    69.36%      90,792    69.42%
        Wrong                      120,152    30.64%      39,997    30.58%
        Total                      392,150               130,789
```

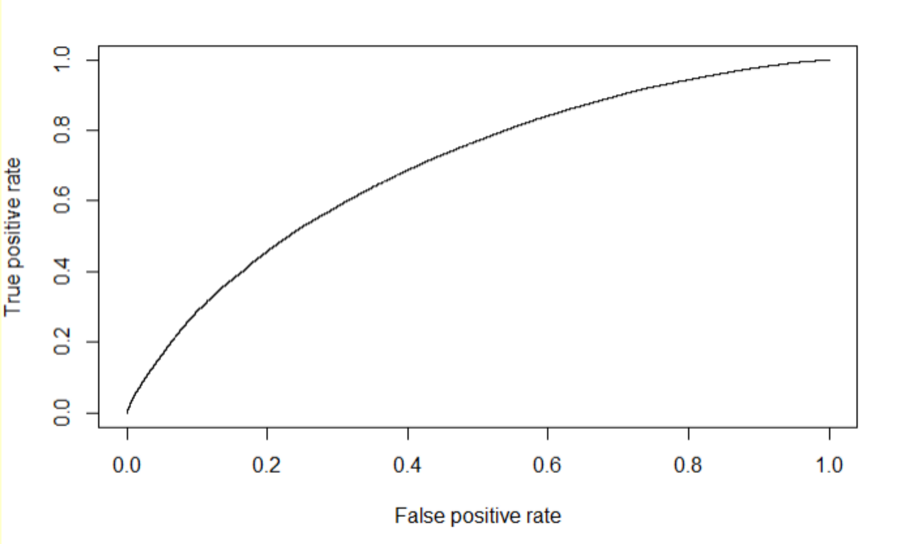
Bayesian network



```
Results for output field target
    Comparing $B-target with target
        'Partition'             1_Training             2_Testing
        Correct                    271,232    69.17%      90,745    69.38%
        Wrong                      120,918    30.83%      40,044    30.62%
        Total                      392,150               130,789
```

Conclusions

- The final code for the project has been implemented in R notebook and has been provided along with the code with description.
- The split used between test and train dataset during validation phase: 75%-25%
- The final accuracy achieved at the end of the project for the given dataset during validation phase : 69.9%
- The ROC curve obtained after validation phase:
- The final output file: output.csv gives the predicted values with customer Ids. To be evaluated by the teacher with actual predictions.



Advice from financial and business perspective

- The prediction are to be made on dataset with no missing values and customer care can take action on relevant predictions.

- Even though system trained on missing data, the model when applied to cases with no missing value will yield better results.

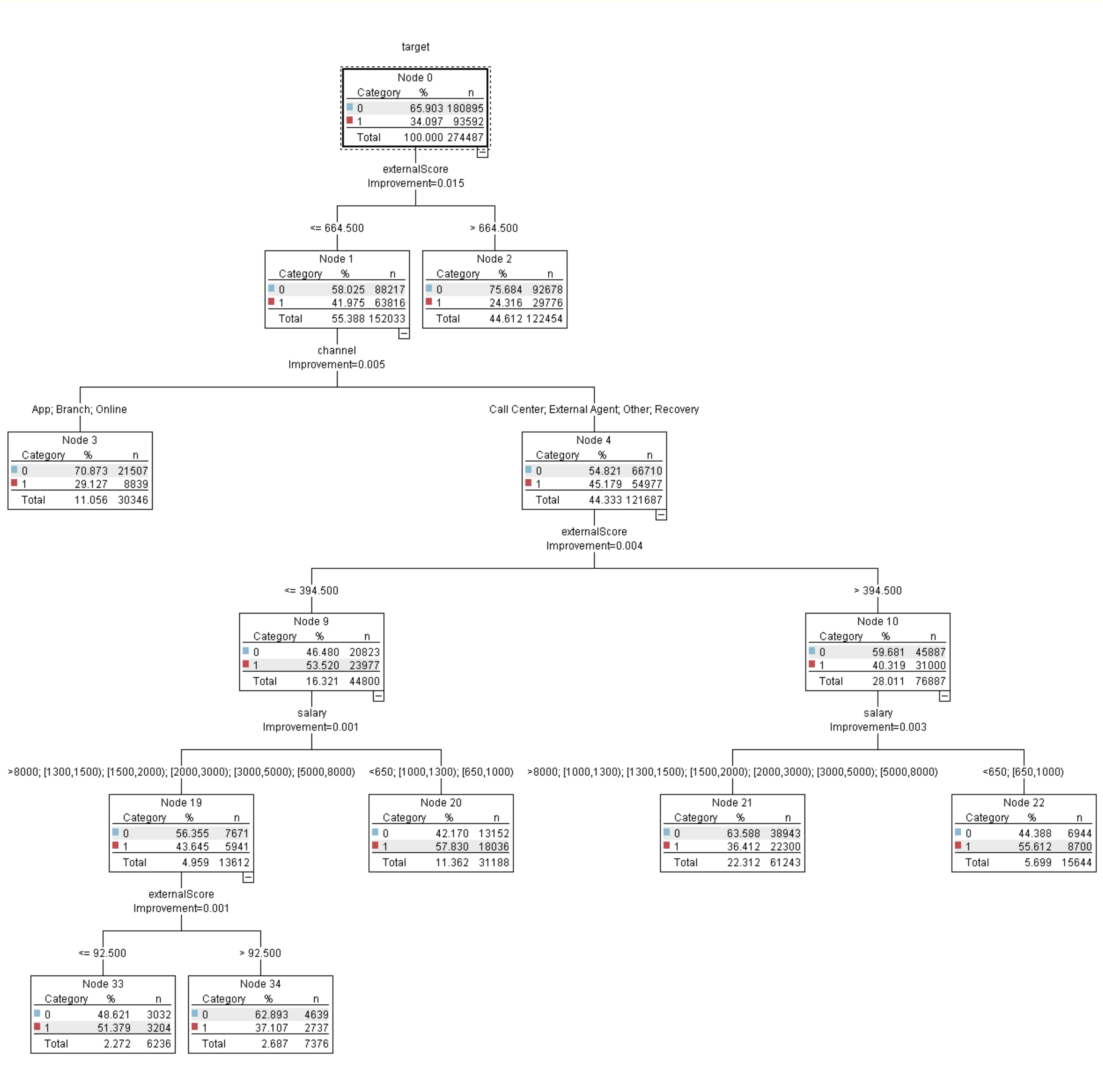
- The externalscore, channel and salary are the major players to decide the prediction values to be 1 or 0 and rightly so from the banking perspective.

- The designed model can be applied to real-time system after further optimization.

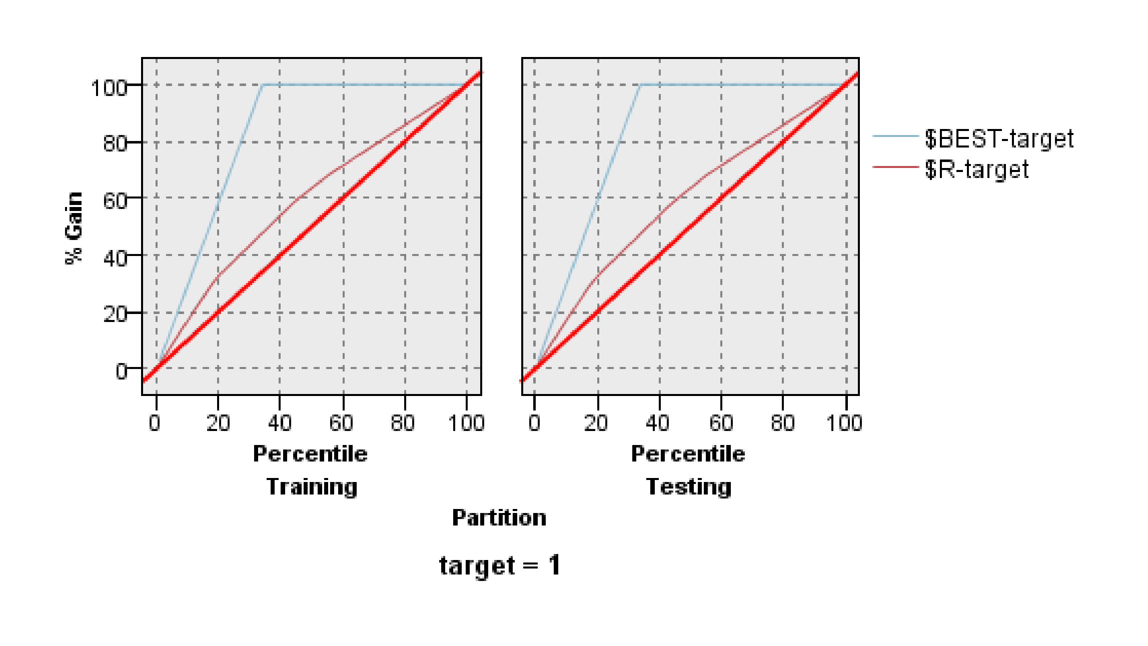
- The model's accuracy can be improved by integrating it with time series data.

Appendix

The decision tree diagram for the model:



The ROC Curve Ideal Behavior:


% Gain
Percentile
Training
Percentile
Testing
Partition
target = 1
$BEST-target
$R-target