Introduction

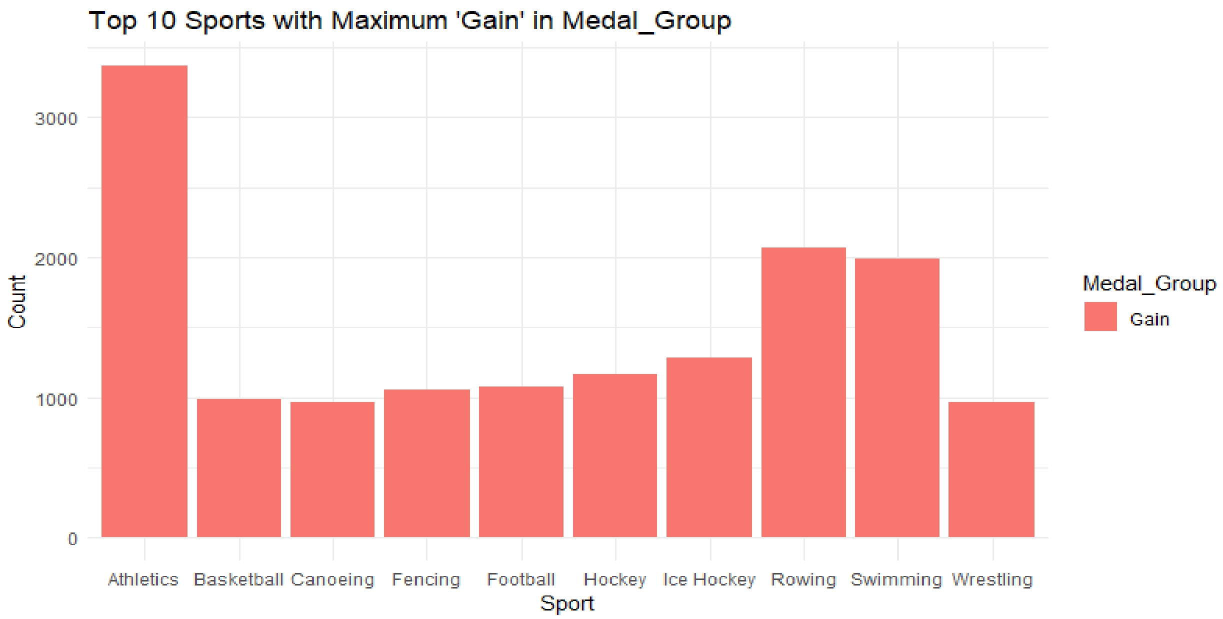
The Olympics, a global celebration of athletic prowess and human achievement, has been a stage for extraordinary feats and triumphs for over a century. Athletes from around the world converge every four years to compete at the highest level. Beyond the spectacle and glory, the Olympic Games provide a unique opportunity to unravel patterns and insights into the factors that contribute to winning a medal.

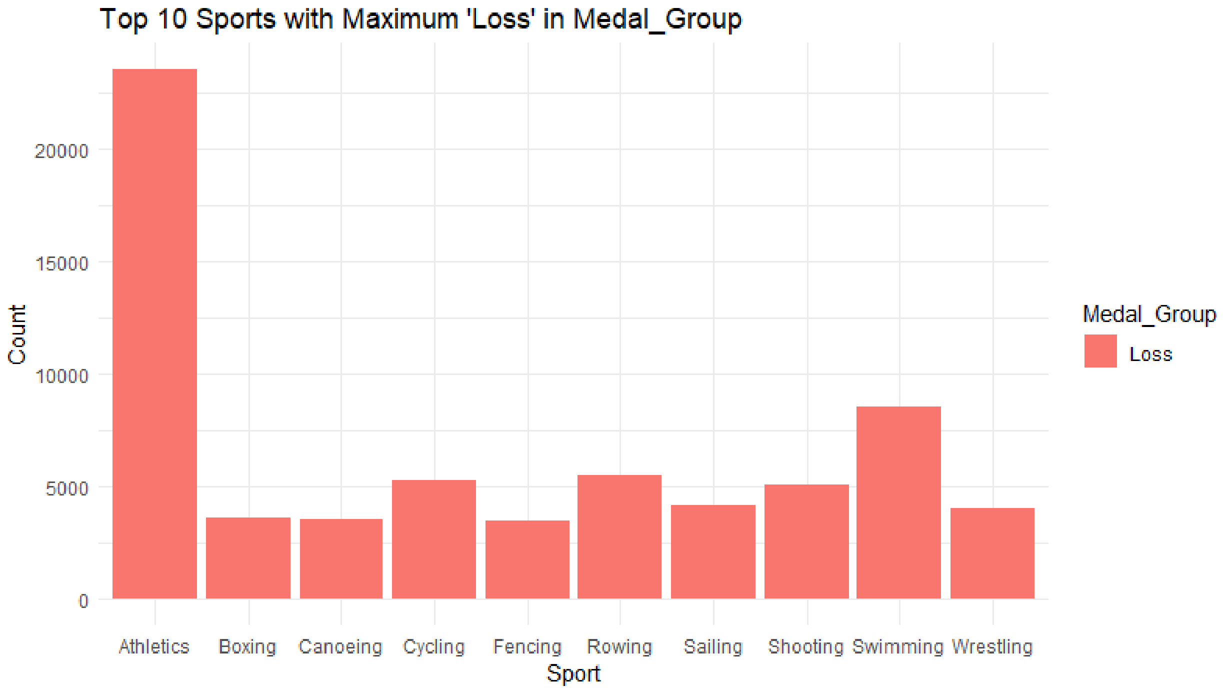
Understanding the dynamics behind Olympic success is not merely a pursuit of statistical curiosity; it holds significant implications for athletes, coaches, and sports analysts alike. In this report, we delve into the intricate world of Olympic data, employing classification techniques to learn about the possible factors for winning a medal. By deciphering the patterns and characteristics common among medalists, we aim to shed light on the multifaceted nature of athletic achievement.

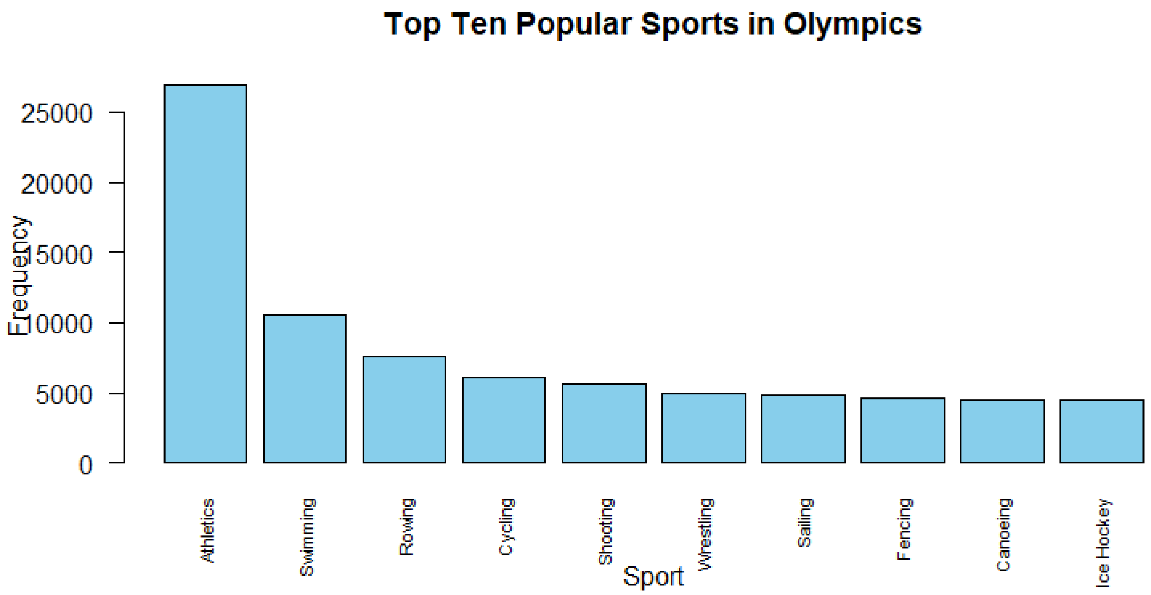
Data Description

Through exploratory data analysis, I found that the dataset contains 4 numeric fields – ID, Age, Height, and Weight. There are 10 categorical fields – Name, Sex, Team, NOC, Games, Season, City, Sport, Event, etc. Since my study focuses on studying the possibility of winning a medal, I created a new column Medal Group where I created two classes - Loss (No Medal) and Gain (if the Medal was Gold, Silver or Bronze).

There were quite interesting insights in the dataset like the most popular sport between both men and women is Athletics. It is also the sport with the maximum number of Gains and Losses.







Preprocessing

By observing the dataset with 15 predictors, I decided to drop ID, Name, Team, City, Games, Year and Event columns. In my classification model, I was trying to study the possibility of winning or losing a medal and ID, Name, and City could not contribute to it. The ‘Games’ column has information about the Year in which the Olympics was held and the season. I decided to drop the Year and Games column because for team sports, it might introduce data leakage issues in the model.

I first replaced the values as NA in the Medal column with None in the dataset. I clubbed the values in the Medal column in 2 groups - Gain and Loss. I stored these values in a new column -

Medal_Group where Gain = Gold, Silver or Bronze, and Loss = None. I then dropped the unwanted columns so that the duplicate values or missing value counts from those columns would not affect the dataset. There were missing values in the Age, Height, and Weight columns. I decided to remove those rows since the number of missing rows was relatively smaller than the size of the dataset and they would not have impacted the result.

Model Selection & Methodology

Since I attempted a classification problem, I decided to go with the DecisionTree Classifier. I chose Decision Tree because it provides a visual way of interpreting the data.

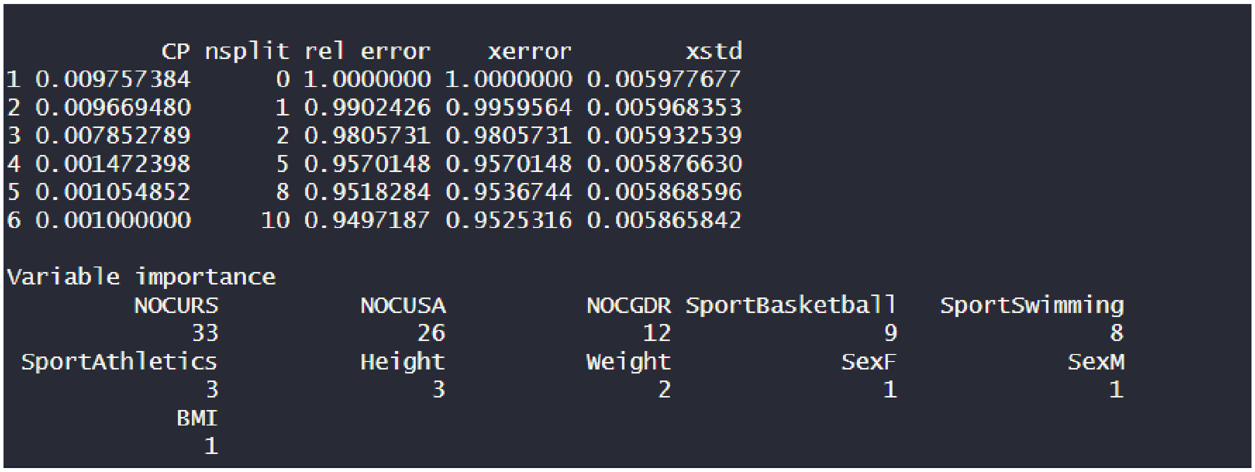
Rationale for Choosing Predictors:

Categorical Predictors: Sex of the player can have an impact on the probability of winning or losing. Team and NOC columns convey the same information and since the number of unique values is less in NOC, I decided to choose it. It would help us reduce the number of dummy variables. Sport and Event also have similar information. The data in the Event column is already provided by the Sport and Sex columns, so I decided to drop it.

Numerical Predictors: Age, Height, and Weight are the 3 numerical predictors I chose for the model. It is so because these features can highly impact the selection of players for that sport and also the chances of winning a medal.

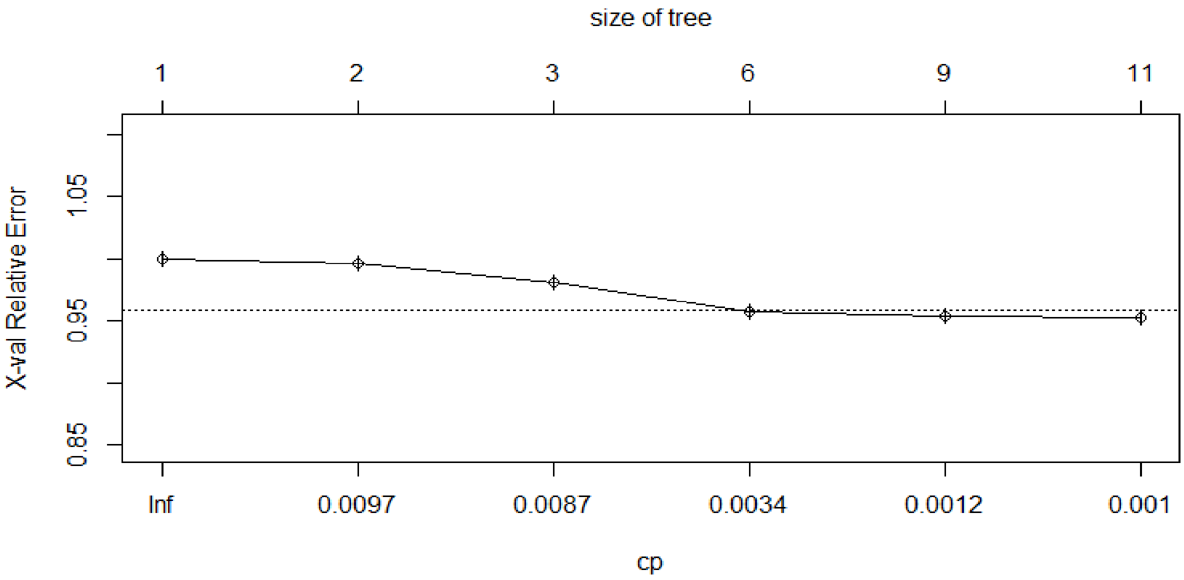
Feature Engineering: I created a column BMI based on the Height and Weight column.

I divided the dataset into Train (80%) and Test (20%). I used the training dataset to first build an overfitted tree with cp = 0.001 and different combinations of minimum split and minimum bucket which is the number of minimum records in the terminal node. Based on the below reading, the optimal cp is where the relative error is the least. It is least at 0.001. The best tree with cp = 0.001 was then tested on the 20% test dataset.



```
          CP nsplit rel error    xerror        xstd
1 0.009757384      0 1.0000000 1.0000000 0.005977677
2 0.009669480      1 0.9902426 0.9959564 0.005968353
3 0.007852789      2 0.9805731 0.9805731 0.005932539
4 0.001472398      5 0.9570148 0.9570148 0.005876630
5 0.001054852      8 0.9518284 0.9536744 0.005868596
6 0.001000000     10 0.9497187 0.9525316 0.005865842

Variable importance
         NOCURS          NOCUSA          NOCGDR SportBasketball   SportSwimming 
             33              26              12               9               8 
 SportAthletics          Height          Weight            SexF            SexM 
              3               3               2               1               1 
            BMI 
              1 
```



I shortlisted these two models:

<u>Model 1:</u> (Figure 3)The first model had predictors - Sex, Age, Height, Weight, NOC, Season, BMI, and Sport. The tree performed well and had considered almost all the predictors during splitting.

Model 2:(Figure 1) The second model had predictors - Sex, Age, Height, Weight, NOC, BMI, and Sport. I selected this model because the Season variable does not add any importance to the model. Sports are by default specific to Summer and Winter Games and since we are interested in the likelihood of winning a medal, the Sport column alone suffices that.

Results

The final model had an accuracy of 82.13%. The accuracy of classifying the majority class (Loss in the Medal_Group) in the dataset, is ~78%. If we use this as a baseline model, we can say that our model is performing better.

The final model has a Precision of 82.68% which means model is classifying Gain labels correctly for 82% of the time. Recall of the model is 98.70% meaning, F1 score of 89.98% and AUC is 54%. (Figure 2)

Conclusions

Based on the final model outcome, the following conclusions can be drawn:

1. Players from Russia, who are not athletes but weigh 75Kg+ we can say with a 34% likelihood that the player will win the medal. The dataset has 1% of such records.
2. However, male players from Russia, who are not athletes and weigh less than 75 kg can say with 53% likelihood that the player will not win the medal.
3. Female players from Russia, who are not athletes and weigh less than 75Kgs will likely win the medal. We can say this with a likelihood of 41%.
4. Russia will not gain medals in Athletics. We can say this with a likelihood of 69%.

5. If the USA is participating in Basketball, the likelihood that it will win the medal ~0%
6. USA has a 33% likelihood of winning a medal in Swimming and a 69% likelihood of not winning one.
7. A female player from Germany has a 44% likelihood of winning a medal.
8. A male player from Germany with a height < 188 cm has a 64% likelihood of not winning a medal.
9. A male player from Germany with height more than or equal to 188 CMS has 39% likelihood of winning a medal.
10. A player not from Germany has 84% likelihood of not winning a medal.

Appendices

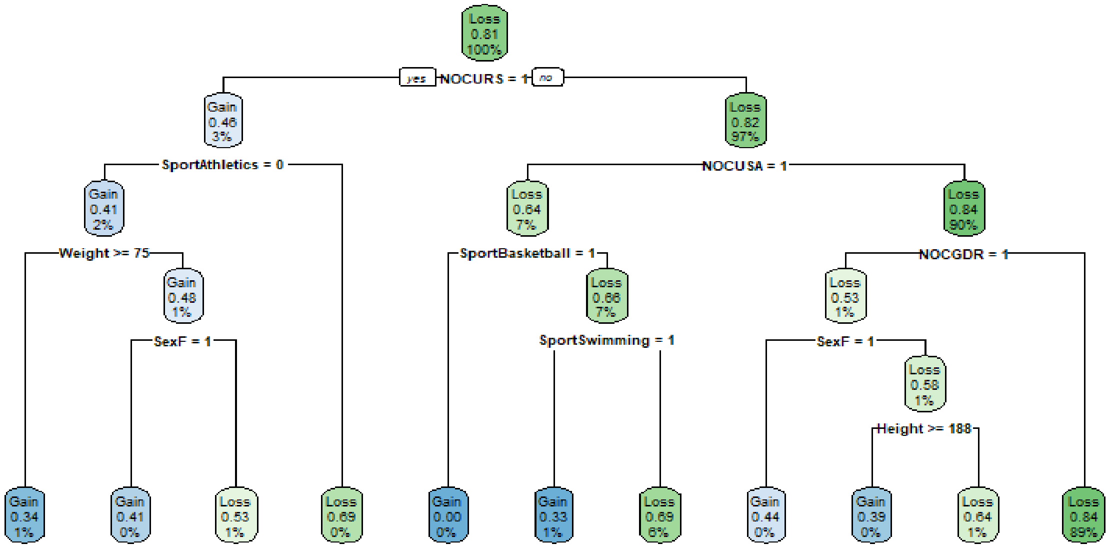


Figure 1

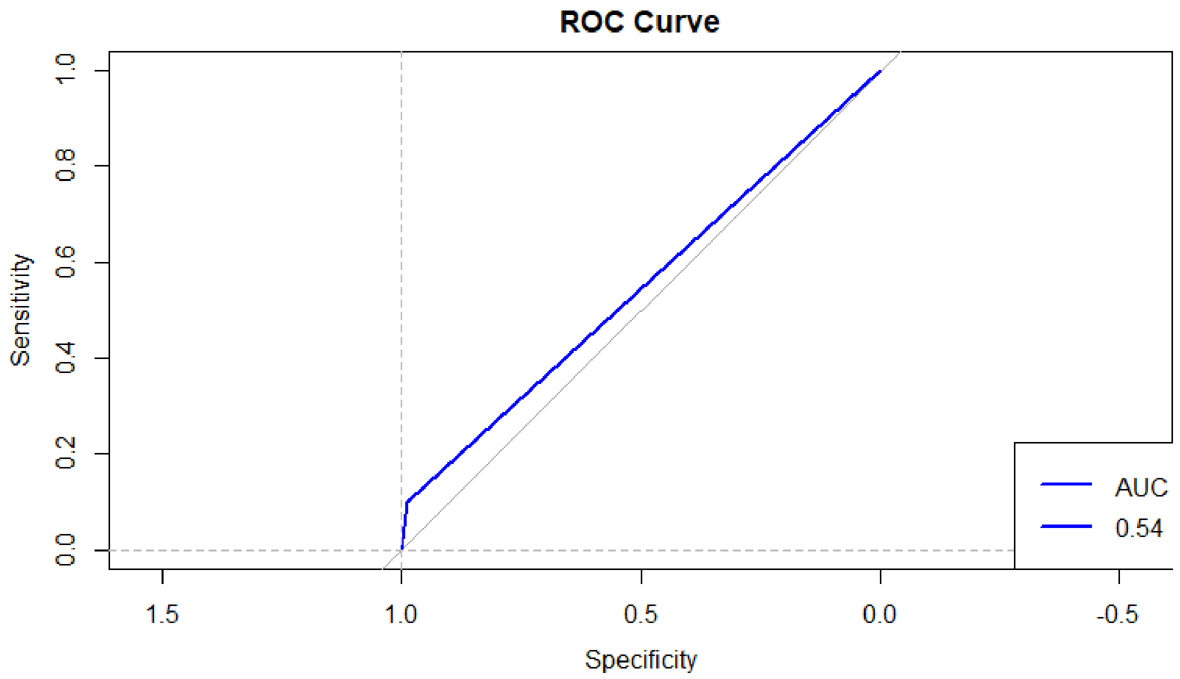

ROC Curve
Sensitivity
0.0
0.2
0.4
0.6
0.8
1.0
Specificity
1.5
1.0
0.5
0.0
-0.5
AUC
0.54

Figure 2

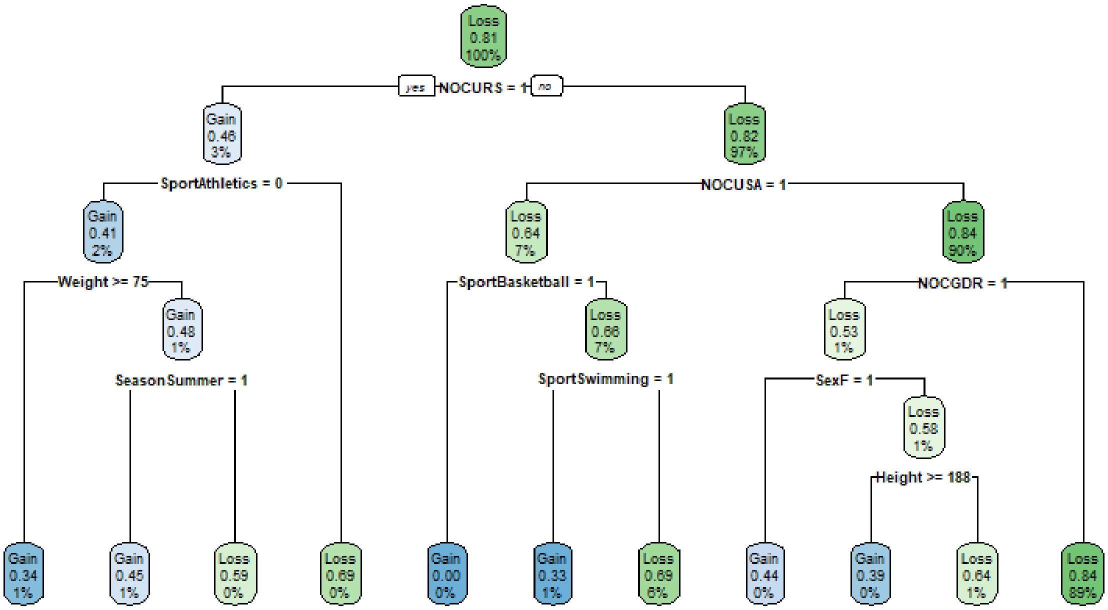

Loss
0.81
100%
yes
NOCURS = 1
no
Gain
0.46
3%
Loss
0.82
97%
SportAthletics = 0
NOCUSA = 1
Gain
0.41
2%
Loss
0.64
7%
Loss
0.84
90%
Weight >= 75
Gain
0.48
1%
SportBasketball = 1
Loss
0.66
7%
NOCGDR = 1
Loss
0.53
1%
SeasonSummer = 1
SportSwimming = 1
SexF = 1
Loss
0.58
1%
Height >= 188
Gain
0.34
1%
Gain
0.45
1%
Loss
0.59
0%
Loss
0.69
0%
Gain
0.00
0%
Gain
0.33
1%
Loss
0.69
6%
Gain
0.44
0%
Gain
0.39
0%
Loss
0.64
1%
Loss
0.84
89%

Figure 3