October 25, 2024

Lalit Boyapati & Augustus Simanson

Predicting NBA All-Stars

1. Background
2. Preprocessing
3. Attribute Selection
4. Classification
5. Results
6. Analysis

October 25, 2024

Background

Named NBA Rookie of the Year
MVP Chicago Bulls
Air Jordan
Dedication
For the Love of the Game
NBA All-Defensive First Team
Named to the All-NBA First Team
BULLS

October 25, 2024

What is our project?

NBA All-Star Selection

- Every February, 24 players are chosen for the NBA All-Star Game.
- Players are divided into Eastern and Western Conference.
- Selection is based on voting: 50% fan vote, 25% current player vote, and 25% by NBA representatives

Research Question & Method

- Can we predict which players will be named all-stars before the voting process?
- Focus: Analyze data from the 2010-2023 seasons.
- Use player statistics to build a predictive model for all-star designation.

Goals & Impact

- Determine the most influential stats for all-star selection.
- Help fans make more informed votes.
- Provide players with insights into the stats they need to excel in for potential selection.

October 25, 2024

Description of Dataset

We gathered data from stats.nba.com and basketball-reference.com. Our dataset includes player statistics from 2010-2023 and identifies whether a player was named an All-Star. This data was collected by web scraping using Python to retrieve 68 attributes from each player, including points, rebounds, assists, and other performance metrics. One limitation of our dataset we would like to point out is that not many players are all-stars in the nba causing our dataset to be heavily skewed.

68

Number of Attributes

7190

Total Instances

538

Number of players who were all-stars

Notable Initial Attributes

GP: Games Played
W: Wins
L: Losses
MIN: Average minutes played a game
PTS: Average points a game
FGM: Field Goals Made on average– Any shot or tap in besides a free throw.
FGA: Field Goals Attempted - the number of field goals attempted by the player on average.
FG_PCT: Field Goal Percentage - the percentage of field goals made by the player on average.
FG3M: Three Pointers Made - the number of three pointers made by the player on average.
FG3_PCT: Three Pointer Percentage - the percentage of three pointers made by the player on average.
FT_PCT: Free Throw Percentage - the percentage of free throws made by the player on average .
OREB: Offensive Rebounds per game - the number of offensive rebounds grabbed by the player on average .
DREB: Defensive Rebounds per game - the number of defensive rebounds grabbed by the player on average .
AST: Assists - the number of assists passed by the player on average .
TOV: Turnovers - the number of turnovers caused by the player on average .
STL: Steals - the number of steals forced by the player on average .
BLK: Blocks - the number of shots blocked by the player on average .
PF: Personal Fouls - the number of personal fouls committed by the player on average .
NBA_FANTASY_POINTS: The number of fantasy points generated by the player on average.
DD2: Double-Doubles - the number of games in which a player achieves double digits in two statistical categories
TD3: Triple-Doubles - the number of games in which a player achieves double digits in three statistical categories
PLUS_MINUS: The point differential when a player is on the court.

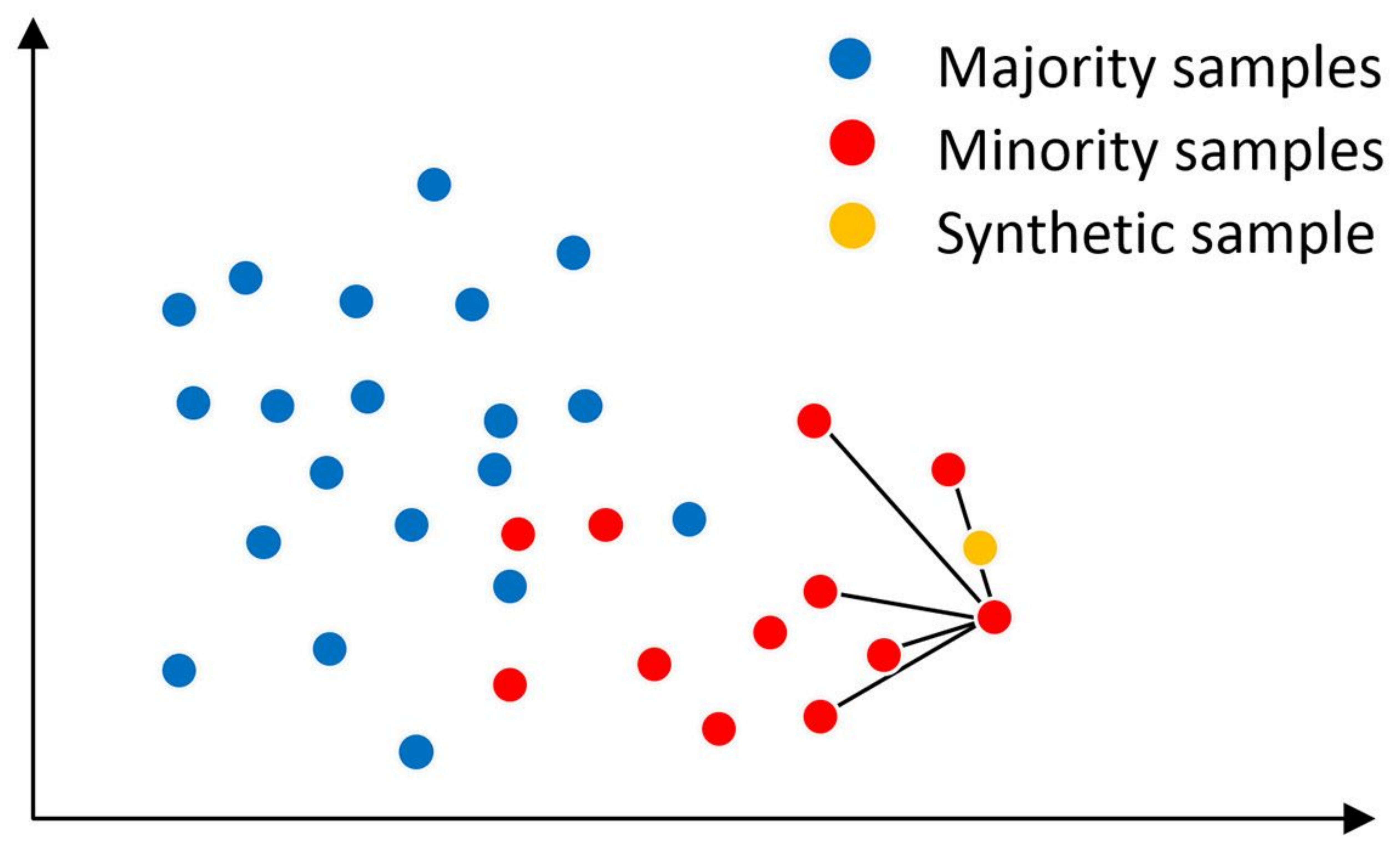
And our class is:
All_Star_Selection: Whether or not player was named All-Star

October 25, 2024

Preprocessing

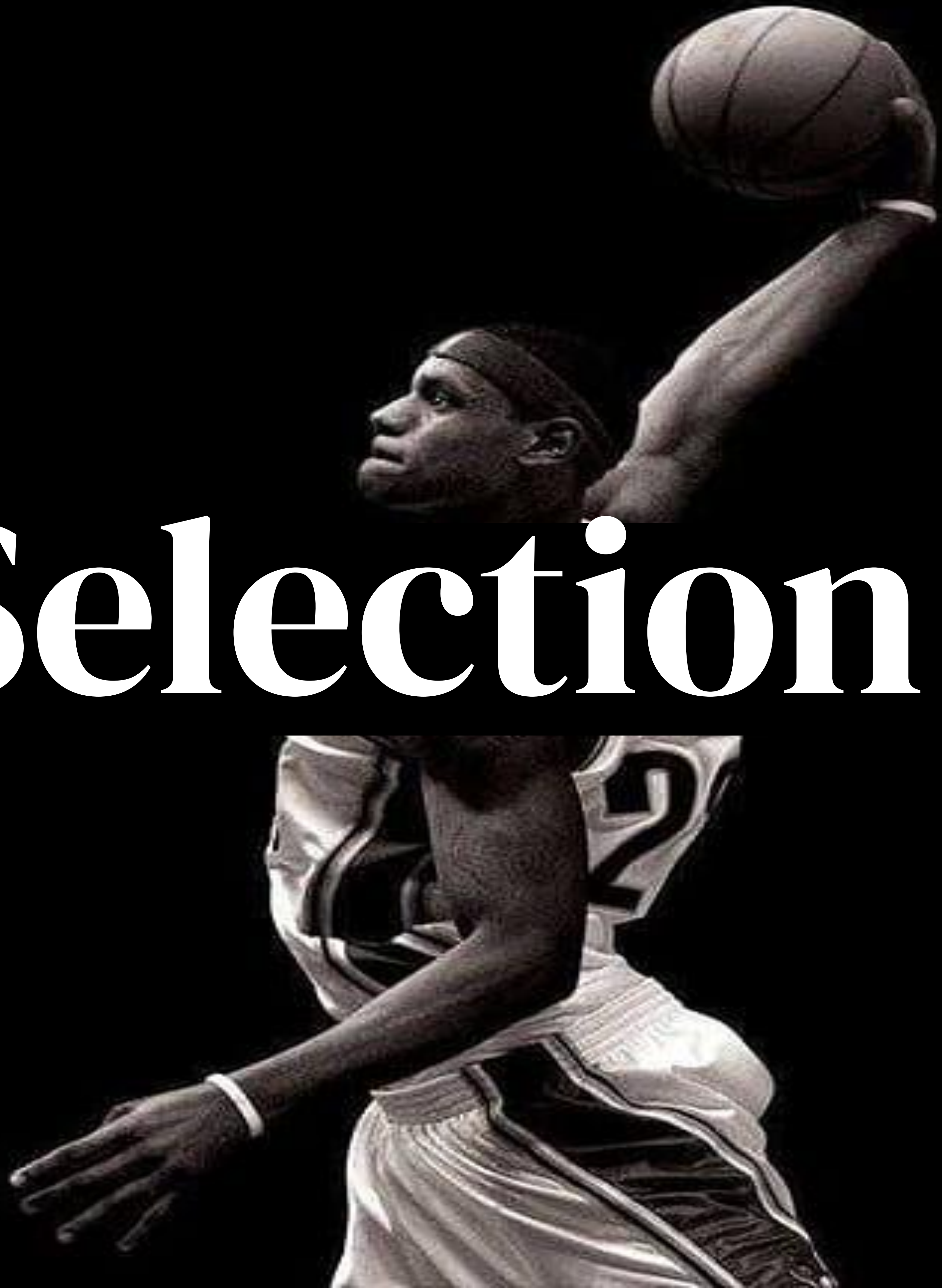
Preprocessing

- Missing Attributes
  - No missing values; comprehensive data source.
- Unrelated Attributes
  - We removed irrelevant attributes such as Player ID, Player Name, Nickname, Fantasy Rank, etc.
- Derived Attributes
  - We removed calculated attributes such as Win PCT, FT PCT, etc.
- SMOTE
  - We addressed the class imbalance mentioned in a previous slide by using SMOTE (Synthetic Minority Oversampling Technique)
    - Achieved a ratio of 2152 all-stars to 8804 instances (24%).


Majority samples
Minority samples
Synthetic sample

October 25, 2024

Attribute Selection



October 25, 2024

Attribute Selection Methods

CoorelationAttributeEval

This calculates the Pearson correlation coefficient between each feature and the class.

Set chosen by us

Using our own prior knowledge we select attributes we believe to be the most correlated with all-star selection

CfsSubsetEval

CfsSubsetEval works by evaluating the degree of redundancy among features associated with the class. It selects features that are highly correlated with the target variable but not correlated with each other.

OneRAttributeEval

OneRAttributeEval in WEKA evaluates attributes by creating one-rule classifiers based on each attribute and measuring their classification error rates.

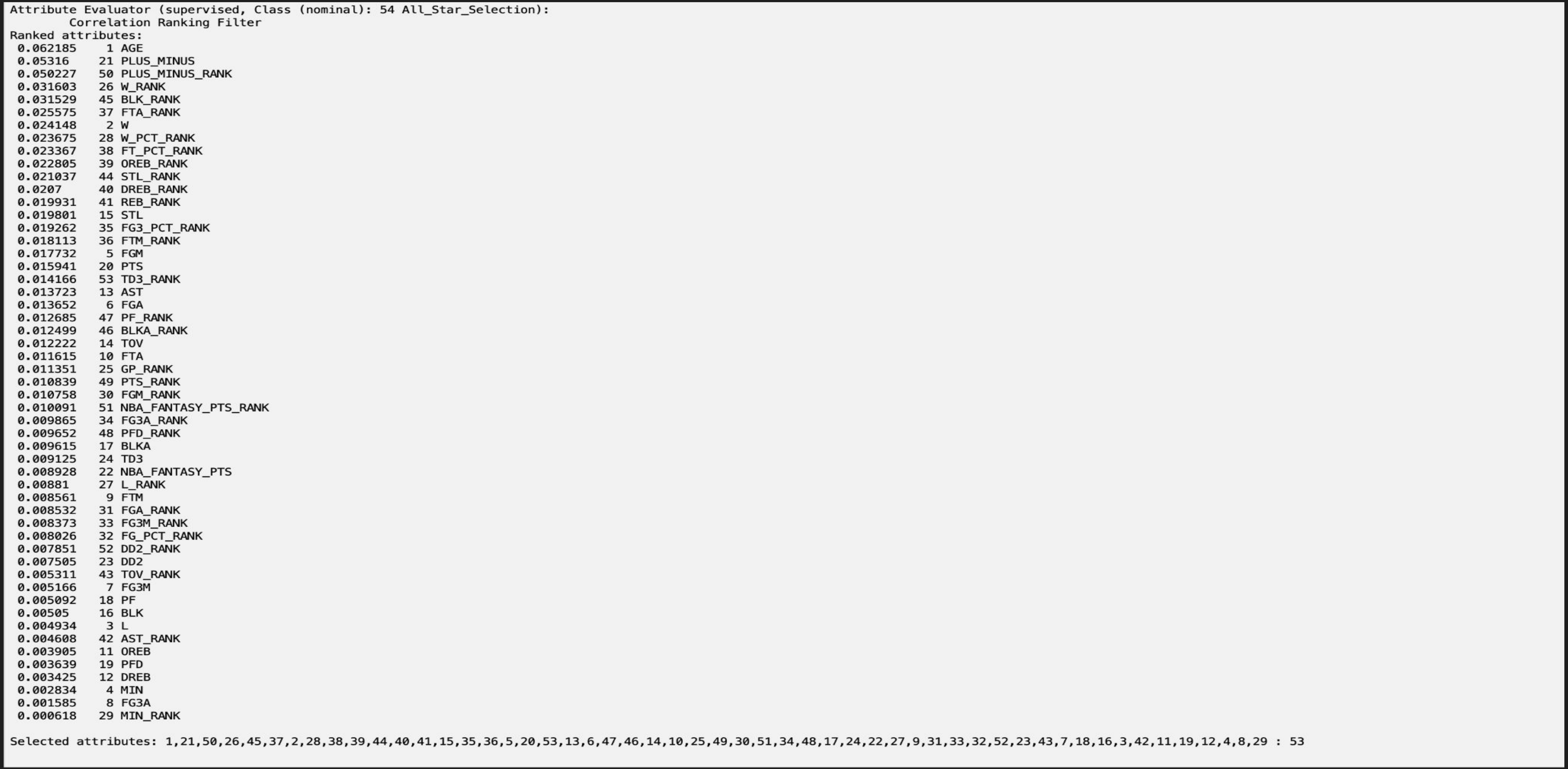
InfoGainAttributeEval

InfoGainAttributeEval determines how well a given attribute separates the training examples according to their class labels. The higher the ranking means the attribute is more informative for classification.

October 25, 2024

CorrelationAttributeEval

Cutoff Value of 0.02, 12 attributes

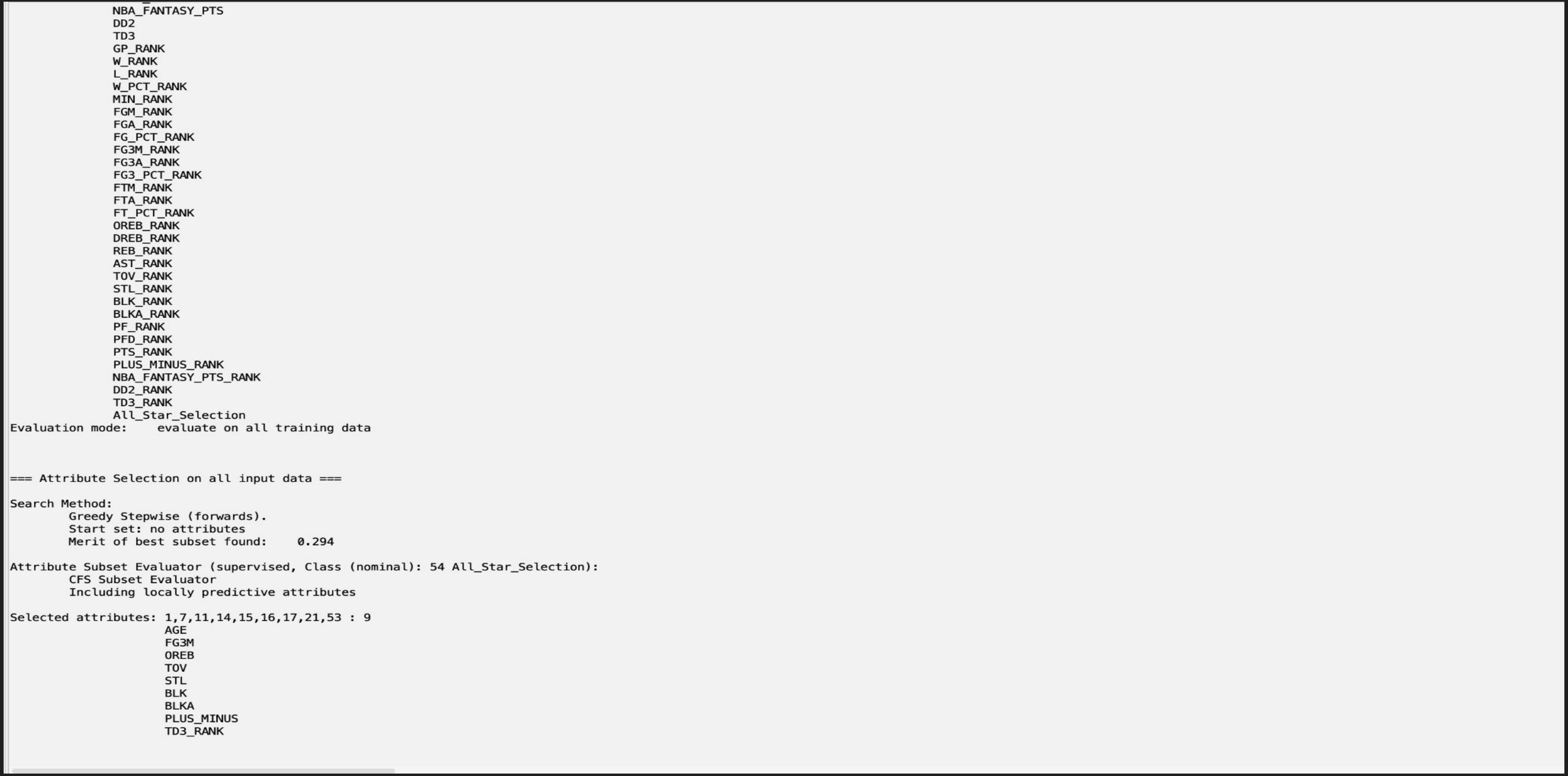


```
Attribute Evaluator (supervised, Class (nominal): 54 All_Star_Selection):
        Correlation Ranking Filter
Ranked attributes:
 0.062185    1 AGE
 0.05316    21 PLUS_MINUS
 0.050227   50 PLUS_MINUS_RANK
 0.031603   26 W_RANK
 0.031529   45 BLK_RANK
 0.025575   37 FTA_RANK
 0.024148    2 W
 0.023675   28 W_PCT_RANK
 0.023367   38 FT_PCT_RANK
 0.022805   39 OREB_RANK
 0.021037   44 STL_RANK
 0.0207     40 DREB_RANK
 0.019931   41 REB_RANK
 0.019801   15 STL
 0.019262   35 FG3_PCT_RANK
 0.018113   36 FTM_RANK
 0.017732    5 FGM
 0.015941   20 PTS
 0.014166   53 TD3_RANK
 0.013723   13 AST
 0.013652    6 FGA
 0.012685   47 PF_RANK
 0.012499   46 BLKA_RANK
 0.012222   14 TOV
 0.011615   10 FTA
 0.011351   25 GP_RANK
 0.010839   49 PTS_RANK
 0.010758   30 FGM_RANK
 0.010091   51 NBA_FANTASY_PTS_RANK
 0.009865   34 FG3A_RANK
 0.009652   48 PFD_RANK
 0.009615   17 BLKA
 0.009125   24 TD3
 0.008928   22 NBA_FANTASY_PTS
 0.00881    27 L_RANK
 0.008561    9 FTM
 0.008532   31 FGA_RANK
 0.008373   33 FG3M_RANK
 0.008026   32 FG_PCT_RANK
 0.007851   52 DD2_RANK
 0.007505   23 DD2
 0.005311   43 TOV_RANK
 0.005166    7 FG3M
 0.005092   18 PF
 0.00505    16 BLK
 0.004934    3 L
 0.004608   42 AST_RANK
 0.003905   11 OREB
 0.003639   19 PFD
 0.003425   12 DREB
 0.002834    4 MIN
 0.001585    8 FG3A
 0.000618   29 MIN_RANK

Selected attributes: 1,21,50,26,45,37,2,28,38,39,44,40,41,15,35,36,5,20,53,13,6,47,46,14,10,25,49,30,51,34,48,17,24,22,27,9,31,33,32,52,23,43,7,18,16,3,42,11,19,12,4,8,29 : 53
```

October 25, 2024

CfsSubsetEval

9 Attributes

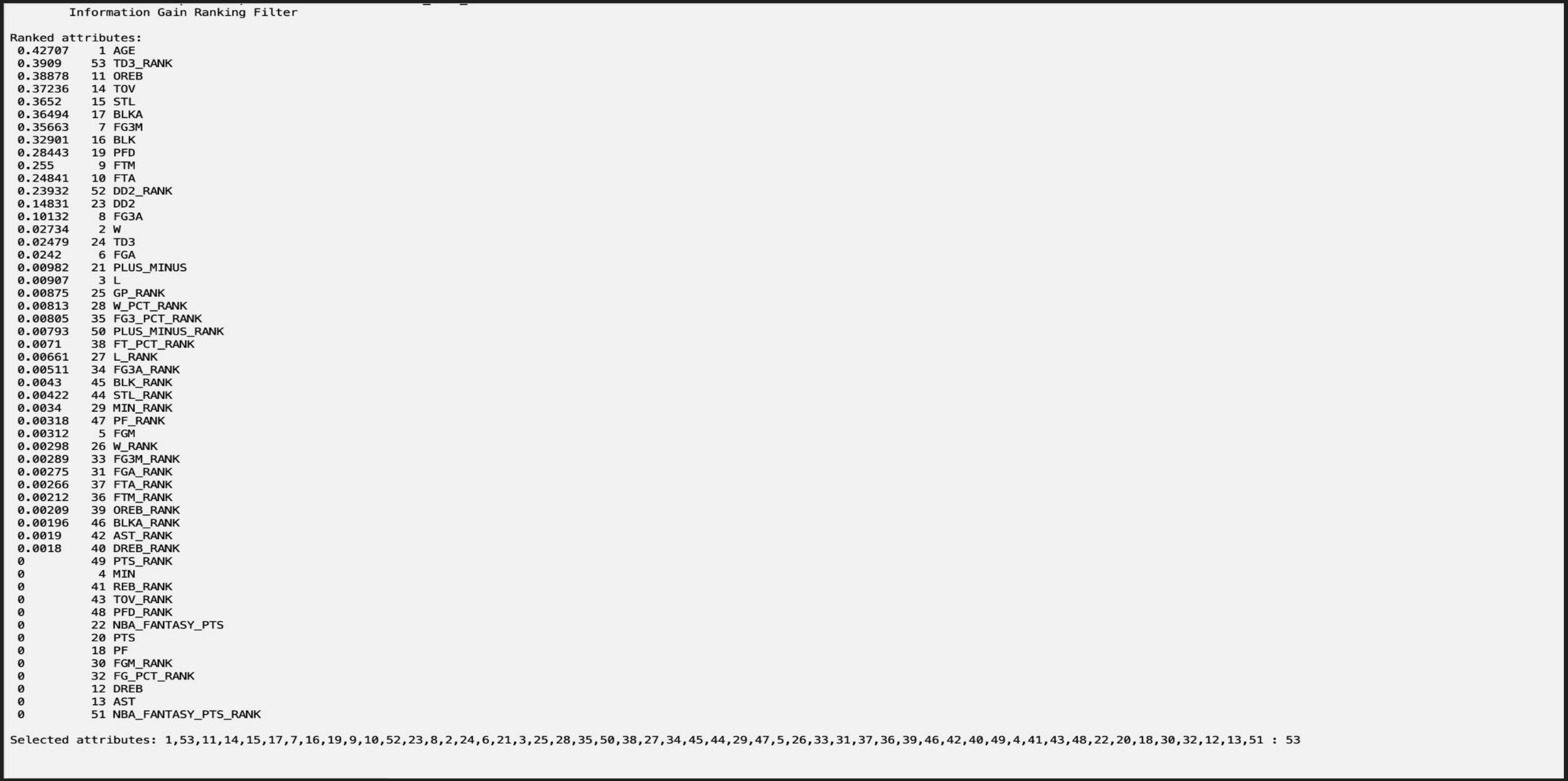


```
              NBA_FANTASY_PTS
              DD2
              TD3
              GP_RANK
              W_RANK
              L_RANK
              W_PCT_RANK
              MIN_RANK
              FGM_RANK
              FGA_RANK
              FG_PCT_RANK
              FG3M_RANK
              FG3A_RANK
              FG3_PCT_RANK
              FTM_RANK
              FTA_RANK
              FT_PCT_RANK
              OREB_RANK
              DREB_RANK
              REB_RANK
              AST_RANK
              TOV_RANK
              STL_RANK
              BLK_RANK
              BLKA_RANK
              PF_RANK
              PFD_RANK
              PTS_RANK
              PLUS_MINUS_RANK
              NBA_FANTASY_PTS_RANK
              DD2_RANK
              TD3_RANK
              All_Star_Selection
Evaluation mode:    evaluate on all training data



=== Attribute Selection on all input data ===

Search Method:
        Greedy Stepwise (forwards).
        Start set: no attributes
        Merit of best subset found:    0.294

Attribute Subset Evaluator (supervised, Class (nominal): 54 All_Star_Selection):
        CFS Subset Evaluator
        Including locally predictive attributes

Selected attributes: 1,7,11,14,15,16,17,21,53 : 9
                     AGE
                     FG3M
                     OREB
                     TOV
                     STL
                     BLK
                     BLKA
                     PLUS_MINUS
                     TD3_RANK
```

October 25, 2024

InfoGainAttributeEval

Cutoff Value of 0.2, 12 attributes

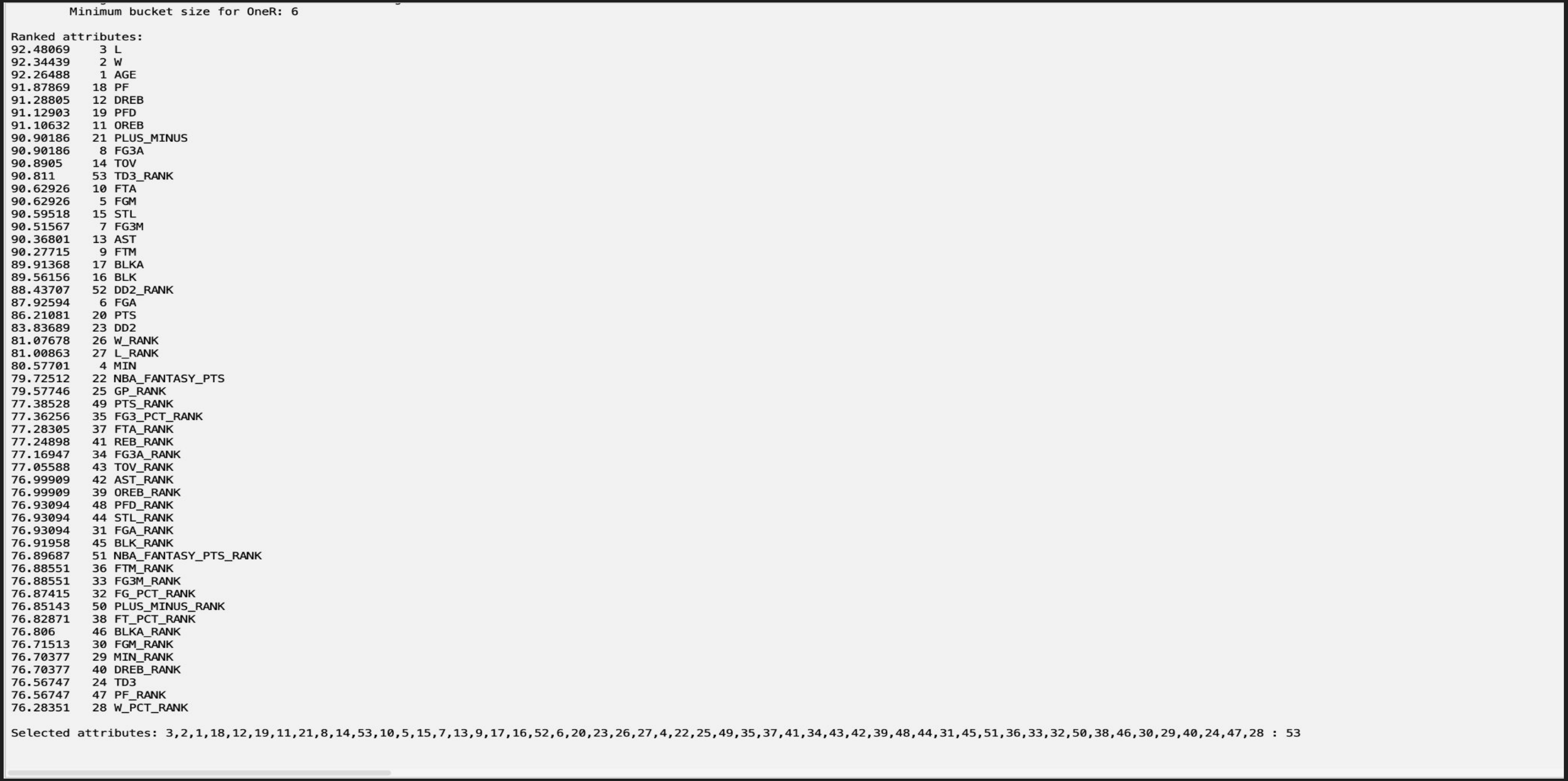


```
        Information Gain Ranking Filter

Ranked attributes:
 0.42707    1 AGE
 0.3909    53 TD3_RANK
 0.38878   11 OREB
 0.37236   14 TOV
 0.3652    15 STL
 0.36494   17 BLKA
 0.35663    7 FG3M
 0.32901   16 BLK
 0.28443   19 PFD
 0.255      9 FTM
 0.24841   10 FTA
 0.23932   52 DD2_RANK
 0.14831   23 DD2
 0.10132    8 FG3A
 0.02734    2 W
 0.02479   24 TD3
 0.0242     6 FGA
 0.00982   21 PLUS_MINUS
 0.00907    3 L
 0.00875   25 GP_RANK
 0.00813   28 W_PCT_RANK
 0.00805   35 FG3_PCT_RANK
 0.00793   50 PLUS_MINUS_RANK
 0.0071    38 FT_PCT_RANK
 0.00661   27 L_RANK
 0.00511   34 FG3A_RANK
 0.0043    45 BLK_RANK
 0.00422   44 STL_RANK
 0.0034    29 MIN_RANK
 0.00318   47 PF_RANK
 0.00312    5 FGM
 0.00298   26 W_RANK
 0.00289   33 FG3M_RANK
 0.00275   31 FGA_RANK
 0.00266   37 FTA_RANK
 0.00212   36 FTM_RANK
 0.00209   39 OREB_RANK
 0.00196   46 BLKA_RANK
 0.0019    42 AST_RANK
 0.0018    40 DREB_RANK
 0         49 PTS_RANK
 0          4 MIN
 0         41 REB_RANK
 0         43 TOV_RANK
 0         48 PFD_RANK
 0         22 NBA_FANTASY_PTS
 0         20 PTS
 0         18 PF
 0         30 FGM_RANK
 0         32 FG_PCT_RANK
 0         12 DREB
 0         13 AST
 0         51 NBA_FANTASY_PTS_RANK

Selected attributes: 1,53,11,14,15,17,7,16,19,9,10,52,23,8,2,24,6,21,3,25,28,35,50,38,27,34,45,44,29,47,5,26,33,31,37,36,39,46,42,40,49,4,41,43,48,22,20,18,30,32,12,13,51 : 53
```

October 25, 2024

OneRAttributeEval

Cutoff Value of 90.8, 11 attributes



```
        Minimum bucket size for OneR: 6

Ranked attributes:
92.48069    3 L
92.34439    2 W
92.26488    1 AGE
91.87869   18 PF
91.28805   12 DREB
91.12903   19 PFD
91.10632   11 OREB
90.90186   21 PLUS_MINUS
90.90186    8 FG3A
90.8905    14 TOV
90.811     53 TD3_RANK
90.62926   10 FTA
90.62926    5 FGM
90.59518   15 STL
90.51567    7 FG3M
90.36801   13 AST
90.27715    9 FTM
89.91368   17 BLKA
89.56156   16 BLK
88.43707   52 DD2_RANK
87.92594    6 FGA
86.21081   20 PTS
83.83689   23 DD2
81.07678   26 W_RANK
81.00863   27 L_RANK
80.57701    4 MIN
79.72512   22 NBA_FANTASY_PTS
79.57746   25 GP_RANK
77.38528   49 PTS_RANK
77.36256   35 FG3_PCT_RANK
77.28305   37 FTA_RANK
77.24898   41 REB_RANK
77.16947   34 FG3A_RANK
77.05588   43 TOV_RANK
76.99909   42 AST_RANK
76.99909   39 OREB_RANK
76.93094   48 PFD_RANK
76.93094   44 STL_RANK
76.93094   31 FGA_RANK
76.91958   45 BLK_RANK
76.89687   51 NBA_FANTASY_PTS_RANK
76.88551   36 FTM_RANK
76.88551   33 FG3M_RANK
76.87415   32 FG_PCT_RANK
76.85143   50 PLUS_MINUS_RANK
76.82871   38 FT_PCT_RANK
76.806     46 BLKA_RANK
76.71513   30 FGM_RANK
76.70377   29 MIN_RANK
76.70377   40 DREB_RANK
76.56747   24 TD3
76.56747   47 PF_RANK
76.28351   28 W_PCT_RANK

Selected attributes: 3,2,1,18,12,19,11,21,8,14,53,10,5,15,7,13,9,17,16,52,6,20,23,26,27,4,22,25,49,35,37,41,34,43,42,39,48,44,31,45,51,36,33,32,50,38,46,30,29,40,24,47,28 : 53
```

October 25, 2024

Set chosen by us

- Based on our NBA knowledge and player performance insight we chose:
    - W
    - FGM
    - PLUS_MINUS
    - PLUS_MINUS_RANK
    - BLK
    - STL
    - AST
    - FG3M
    - MIN

October 25, 2024

Classification

October 25, 2024

Different Classifiers Used

Naive Bayes

A probabilistic classifier that assumes the presence of each feature is independent of the other features.

J48

A decision tree algorithm that splits the data based on attribute values. It creates tree-structures that can handle both categorical and continuous data.

OneR

A simple, rule-based classifier that generates one rule for each predictor and selects the one that performs the best.

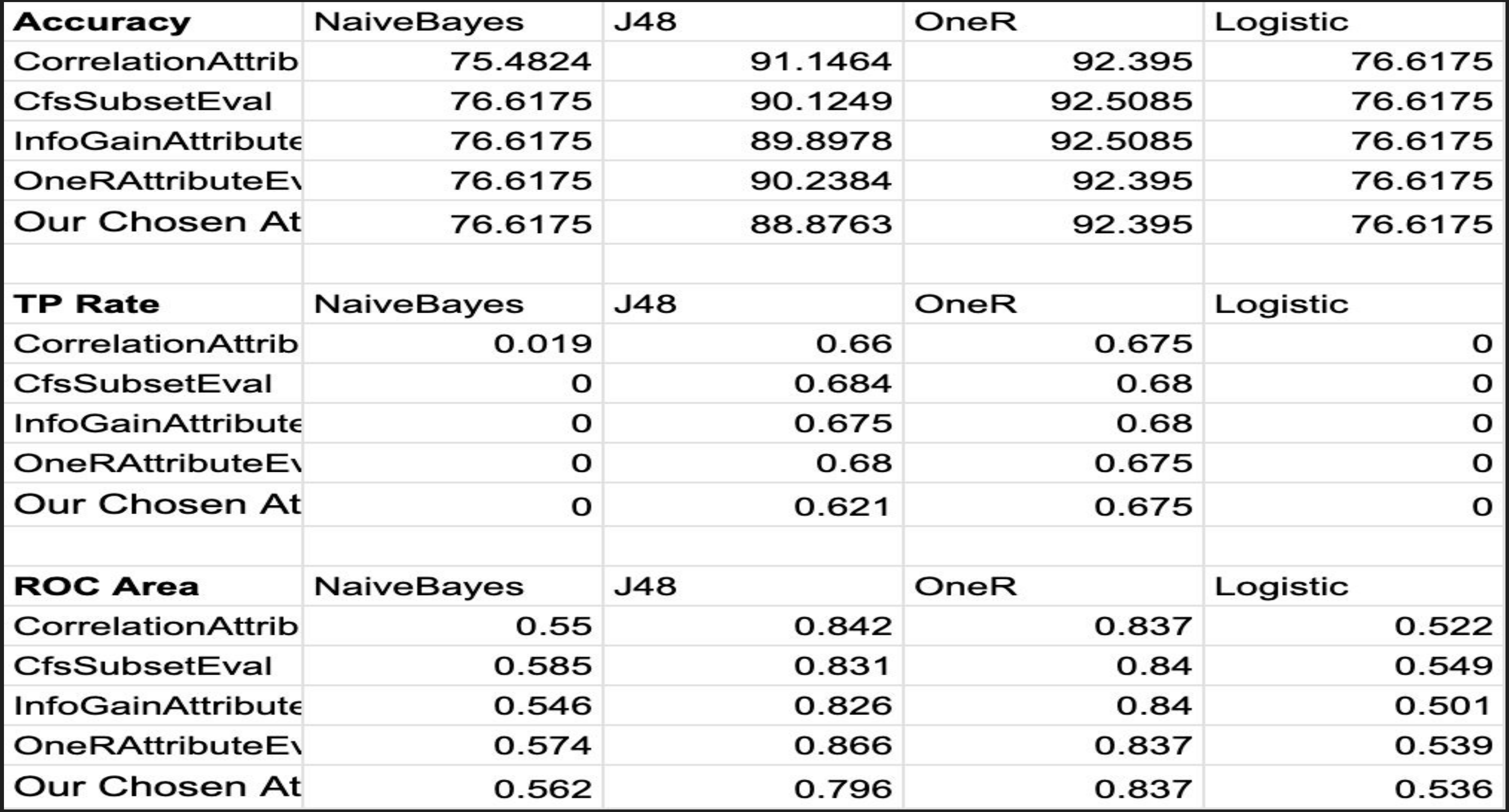
Logistic

A statistical model used for binary classification. It assumes a linear relationship between the independent variable and the log odds of the dependent variable.

October 25, 2024

Results

Results



Accuracy	NaiveBayes	J48	OneR	Logistic
CorrelationAttrib	75.4824	91.1464	92.395	76.6175
CfsSubsetEval	76.6175	90.1249	92.5085	76.6175
InfoGainAttribute	76.6175	89.8978	92.5085	76.6175
OneRAttributeEv	76.6175	90.2384	92.395	76.6175
Our Chosen At	76.6175	88.8763	92.395	76.6175

TP Rate	NaiveBayes	J48	OneR	Logistic
CorrelationAttrib	0.019	0.66	0.675	0
CfsSubsetEval	0	0.684	0.68	0
InfoGainAttribute	0	0.675	0.68	0
OneRAttributeEv	0	0.68	0.675	0
Our Chosen At	0	0.621	0.675	0

ROC Area	NaiveBayes	J48	OneR	Logistic
CorrelationAttrib	0.55	0.842	0.837	0.522
CfsSubsetEval	0.585	0.831	0.84	0.549
InfoGainAttribute	0.546	0.826	0.84	0.501
OneRAttributeEv	0.574	0.866	0.837	0.539
Our Chosen At	0.562	0.796	0.837	0.536

Accuracy

Five Highest Accuracies:

1.CfsSubsetEval with OneR Classification - 92.5085%
1.InfoGainAttributeEval with OneR Classification - 92.5085%
2.CorrelationAttributeEval with OneR Classification - 92.395%
2.OneRAttributeEval with OneR Classification - 92.395%
2.Our Chosen Attributes with OneR Classification - 92.395%

TP Rate*

Five Highest True Positive Rates:

1.CfsSubsetEval with J48 Classification - 0.684

2.CfsSubsetEval with OneR Classification - 0.68

2.InfoGainAttributeEval with OneR Classification - 0.68

2.OneRAttributeEval with J48 Classification - 0.68

5. InfoGainAttributeEval with J48 Classification - 0.675

ROC Area

Five Highest ROC Areas:

1.OneRAttributeEval with J48 Classification - 0.866
2.CorrelationAttributeEval with J48 Classification - 0.842
3.CfsSubsetEval with OneR Classification - 0.84
3.InfoGainAttributeEval with OneR Classification - 0.84
5.Our Chosen Attributes with One R Classification - 0.837

Best Performing Model:

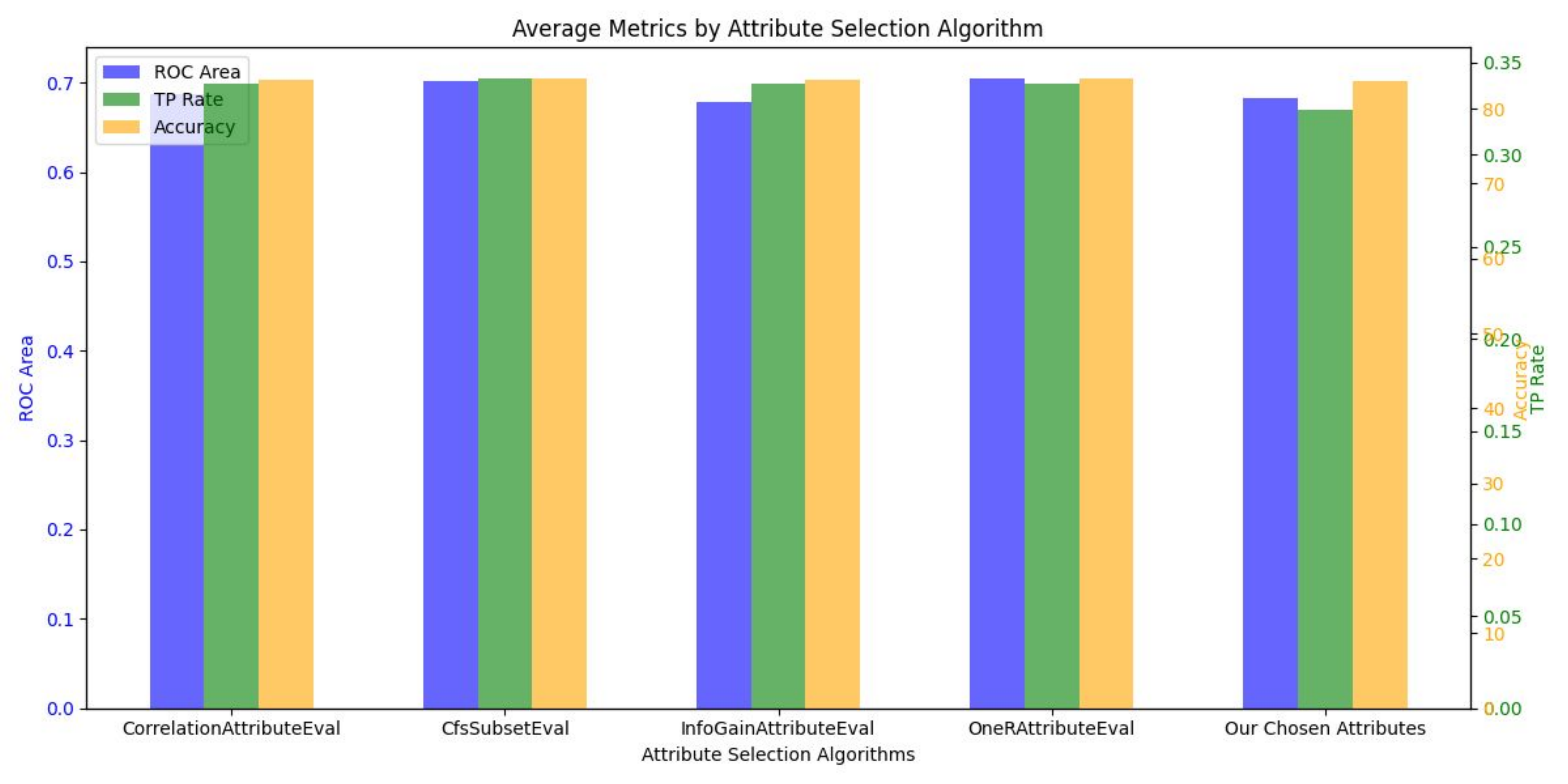
CfsSubsetEval with J48 Classification

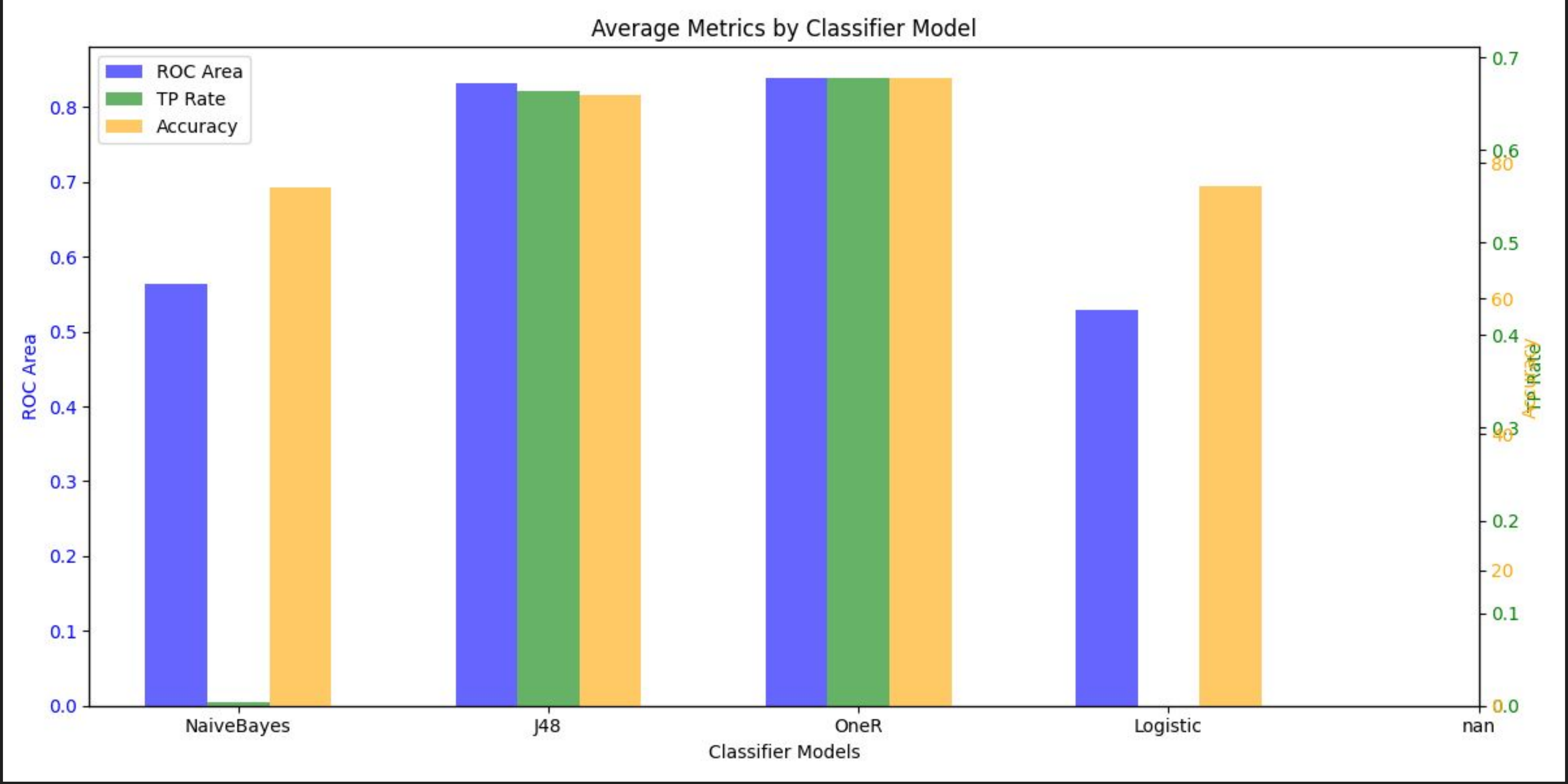
- Highest TP Rate
- Strong Accuracy and ROC Area

Confusion Matrix

	Classified	
Actual	653	22
	65	141

Analysis


Average Metrics by Attribute Selection Algorithm
ROC Area
TP Rate
Accuracy
ROC Area
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.35
TP Rate
0.00
10
20
30
40
50
60
70
80
Accuracy
CorrelationAttributeEval
CfsSubsetEval
InfoGainAttributeEval
OneRAttributeEval
Our Chosen Attributes
Attribute Selection Algorithms


Average Metrics by Classifier Model
ROC Area
TP Rate
Accuracy
ROC Area
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
TP Rate
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
Accuracy
0
20
40
60
80
NaiveBayes
J48
OneR
Logistic
nan
Classifier Models

Key Findings

- High performance of OneR
  - The OneR classification model consistently performs well across all attribute evaluation methods
- Poor performance of NaiveBayes and Logistic Regression
  - both are bad with imbalanced data (not to mention feature independence)
- J48's TP Rate
  - J48 shows superior performance in TP rate despite not leading in accuracy
- ROC Area Consistency
  - Models like OneR and J48 maintain a high ROC Area showing they are quite effective despite accuracy or TP rates

Key Findings

- Attributes selected by OneR
    - OneR picked either Age or Wins for the rule, showing that these are the two most applicable attributes for attribute selection

Why?

Future Outlook

- Look at advanced performance metrics, not just “traditional” statistics
- Gather data prior to 2010
- Use other attribute selection algorithms and classifier models (maybe even a deep learning algorithm)
- Possibly expand to MVP selection, All NBA teams, ROTY, DPOY, OPOY
- Explain why wins are so important

October 25, 2024

Thank you