Checkmate Predictions

By Maeva Assi

Contents

- Introduction:
  - Chess Basics
  - Dataset
  - Goal
- Methods:
  - Data wrangling
  - Data visualization
  - Modeling: random forest model
- Key Findings.
- References

Introduction

Chess Basics

White moves first, after which players alternate turns.

Figure 1: A chess board.

16 pieces each:

- 1 King, 1 Queen;
- 2 Rooks, 2 Knights, 2 Bishops;
- 8 Pawns.

Goal: to **checkmate** the opponent's King

Checkmate means **attacking the King so that it cannot escape capture**, thus ending the game.

The King is never actually captured – a player loses as soon as their King is checkmated.

Dataset:

Chess Game Dataset

Set of **20,058 games** collected via Kaggle by **Mitchell J.** from the free chess server **Lichess.org**.

Over 16 variables, will focus on the 4 following:

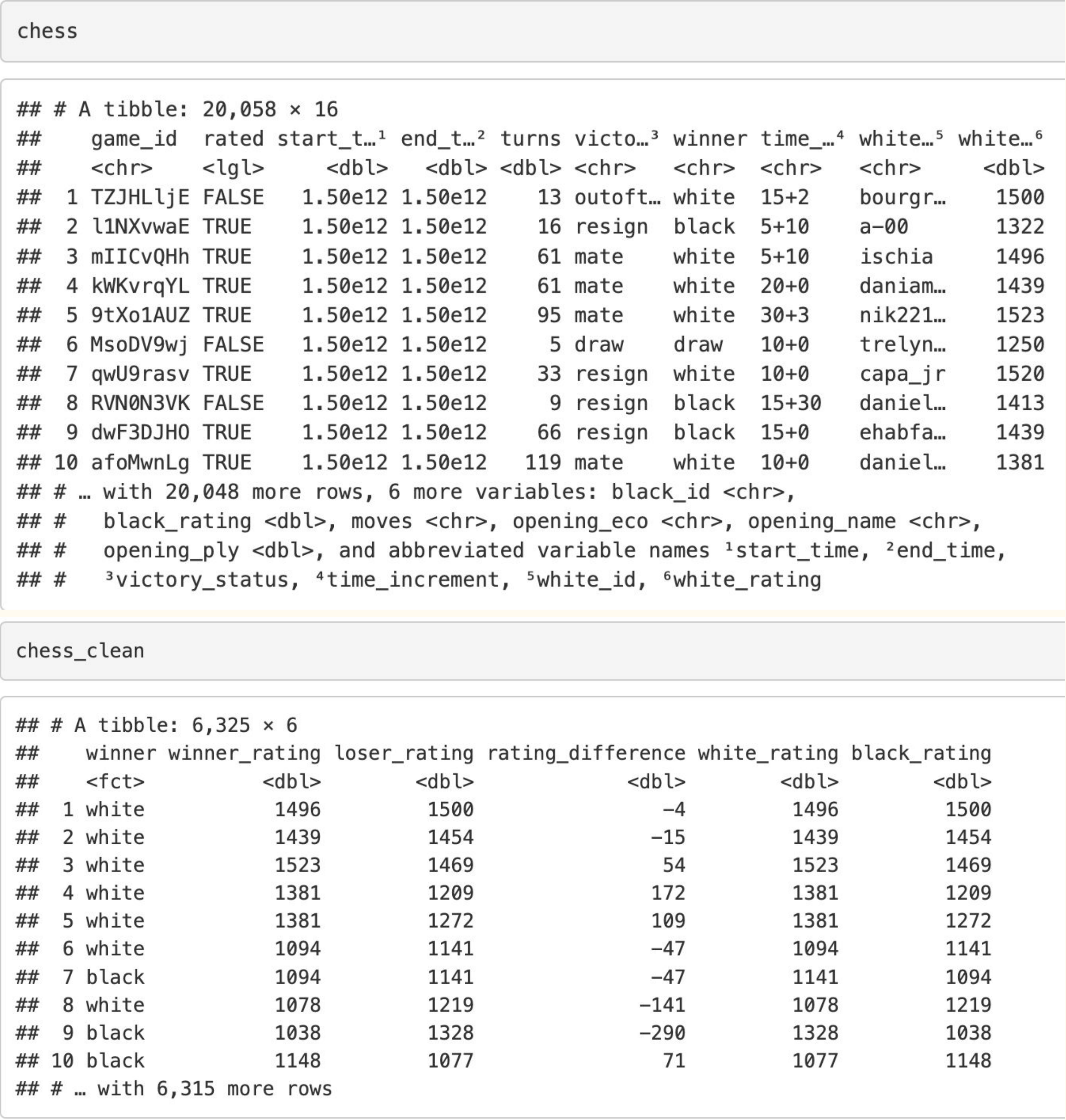
- Game status (mate, resign, draw, outoftime)
- Winner (black, white, draw)
- White player rating
- Black player rating

Is there a relationship between player level & victory? Can we predict if a game resulted in a black or white winner ?

The Methods

Data import & wrangling

- Imported straight from Github
- Filtered for the relevant victory status
- Selected variables of interest
- Factored winner variable
- Created 3 new variables using mutate & if_else() that report:
    - winner rating
    - loser rating
    - the difference between winner & loser rating



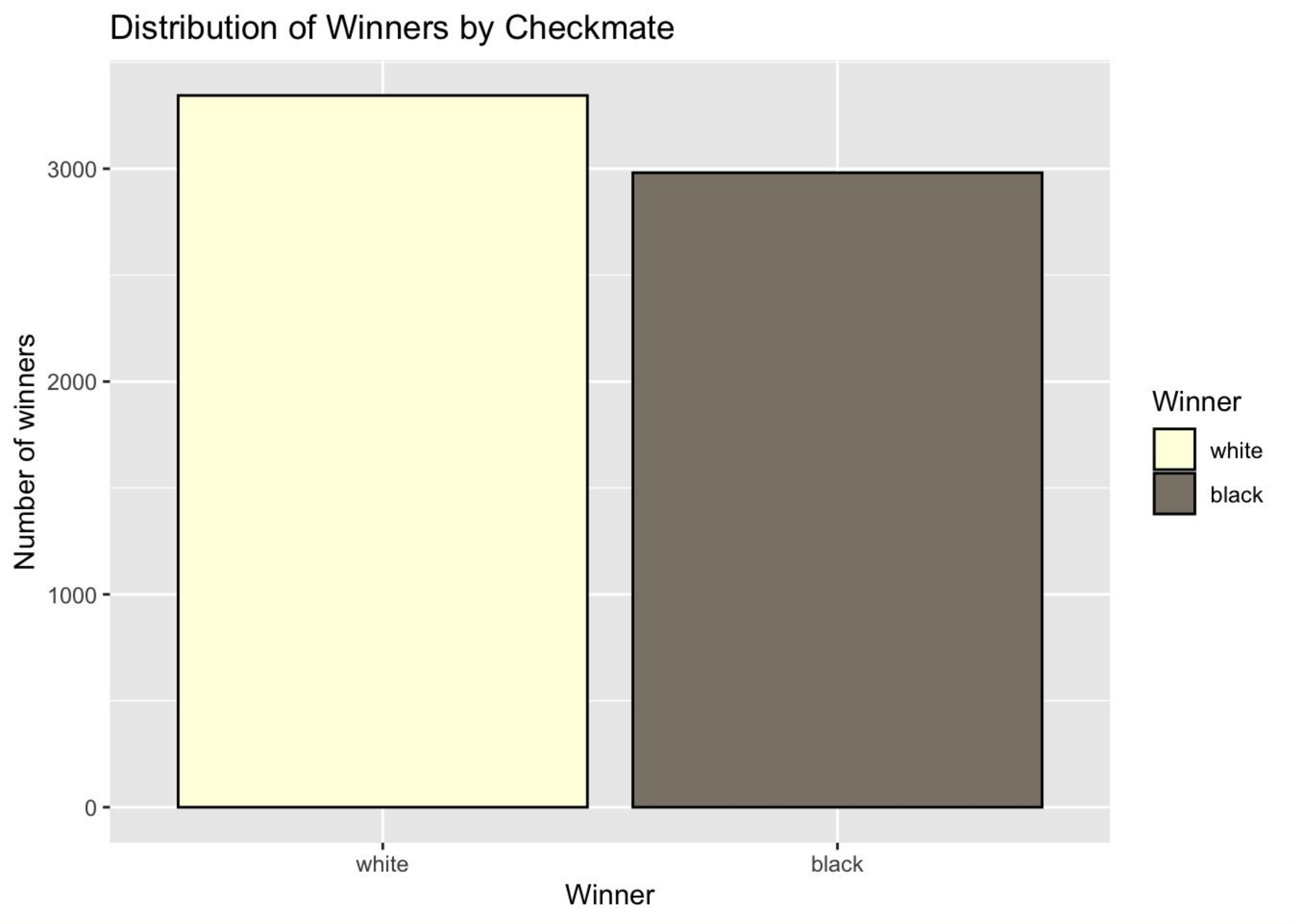
```
chess
```

```
## # A tibble: 20,058 × 16
##    game_id  rated start_t…¹ end_t…² turns victo…³ winner time_…⁴ white…⁵ white…⁶
##    <chr>    <lgl>     <dbl>   <dbl> <dbl> <chr>   <chr>  <chr>   <chr>     <dbl>
##  1 TZJHLljE FALSE   1.50e12 1.50e12    13 outoft… white  15+2    bourgr…    1500
##  2 l1NXvwaE TRUE    1.50e12 1.50e12    16 resign  black  5+10    a-00       1322
##  3 mIICvQHh TRUE    1.50e12 1.50e12    61 mate    white  5+10    ischia     1496
##  4 kWKvrqYL TRUE    1.50e12 1.50e12    61 mate    white  20+0    daniam…    1439
##  5 9tXo1AUZ TRUE    1.50e12 1.50e12    95 mate    white  30+3    nik221…    1523
##  6 MsoDV9wj FALSE   1.50e12 1.50e12     5 draw    draw   10+0    trelyn…    1250
##  7 qwU9rasv TRUE    1.50e12 1.50e12    33 resign  white  10+0    capa_jr    1520
##  8 RVN0N3VK FALSE   1.50e12 1.50e12     9 resign  black  15+30   daniel…    1413
##  9 dwF3DJHO TRUE    1.50e12 1.50e12    66 resign  black  15+0    ehabfa…    1439
## 10 afoMwnLg TRUE    1.50e12 1.50e12   119 mate    white  10+0    daniel…    1381
## # … with 20,048 more rows, 6 more variables: black_id <chr>,
## #   black_rating <dbl>, moves <chr>, opening_eco <chr>, opening_name <chr>,
## #   opening_ply <dbl>, and abbreviated variable names ¹start_time, ²end_time,
## #   ³victory_status, ⁴time_increment, ⁵white_id, ⁶white_rating
```

```
chess_clean
```

```
## # A tibble: 6,325 × 6
##    winner winner_rating loser_rating rating_difference white_rating black_rating
##    <fct>          <dbl>        <dbl>             <dbl>        <dbl>        <dbl>
##  1 white           1496         1500                -4         1496         1500
##  2 white           1439         1454               -15         1439         1454
##  3 white           1523         1469                54         1523         1469
##  4 white           1381         1209               172         1381         1209
##  5 white           1381         1272               109         1381         1272
##  6 white           1094         1141               -47         1094         1141
##  7 black           1094         1141               -47         1141         1094
##  8 white           1078         1219              -141         1078         1219
##  9 black           1038         1328              -290         1328         1038
## 10 black           1148         1077                71         1077         1148
## # … with 6,315 more rows
```

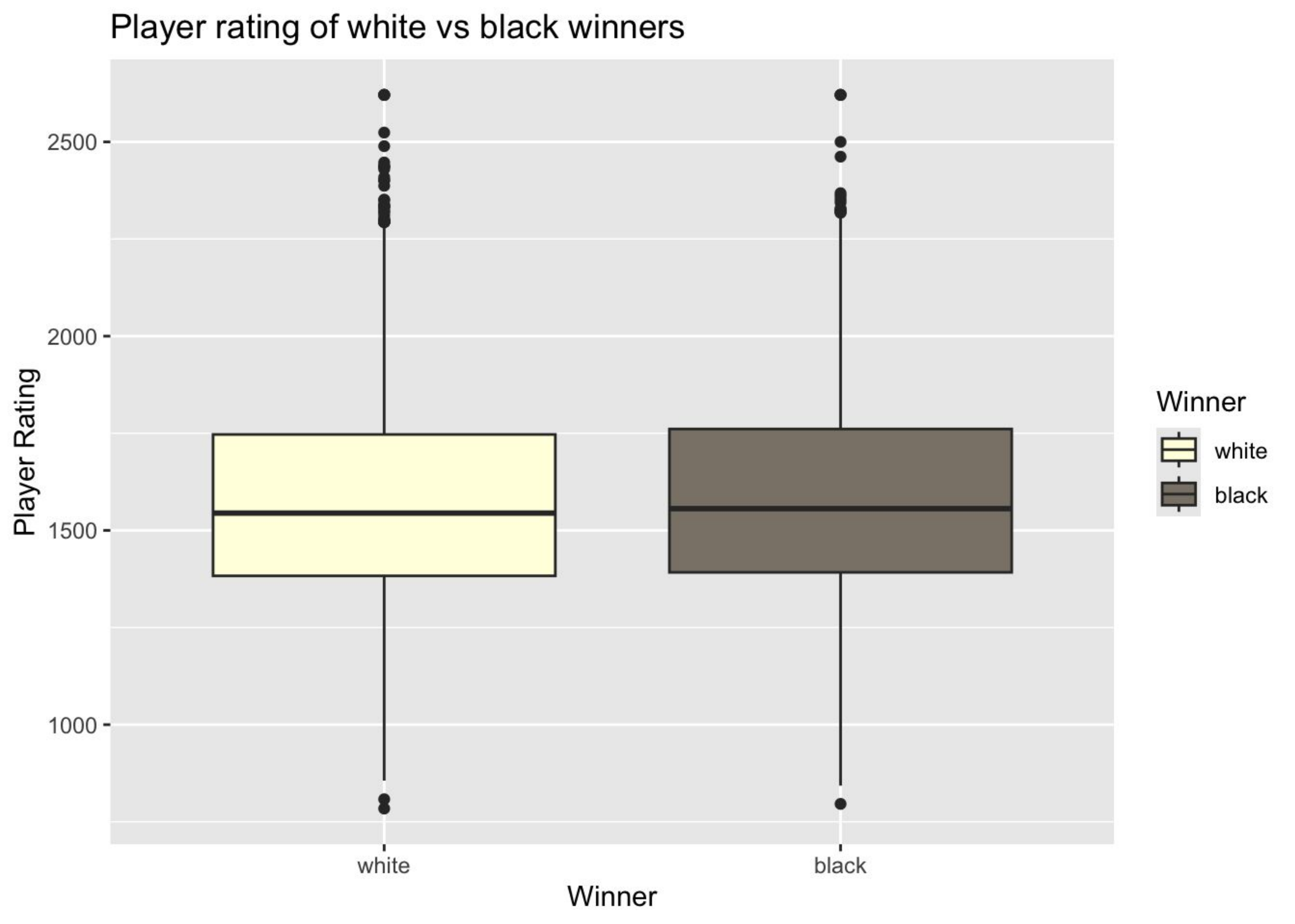
Data visualization

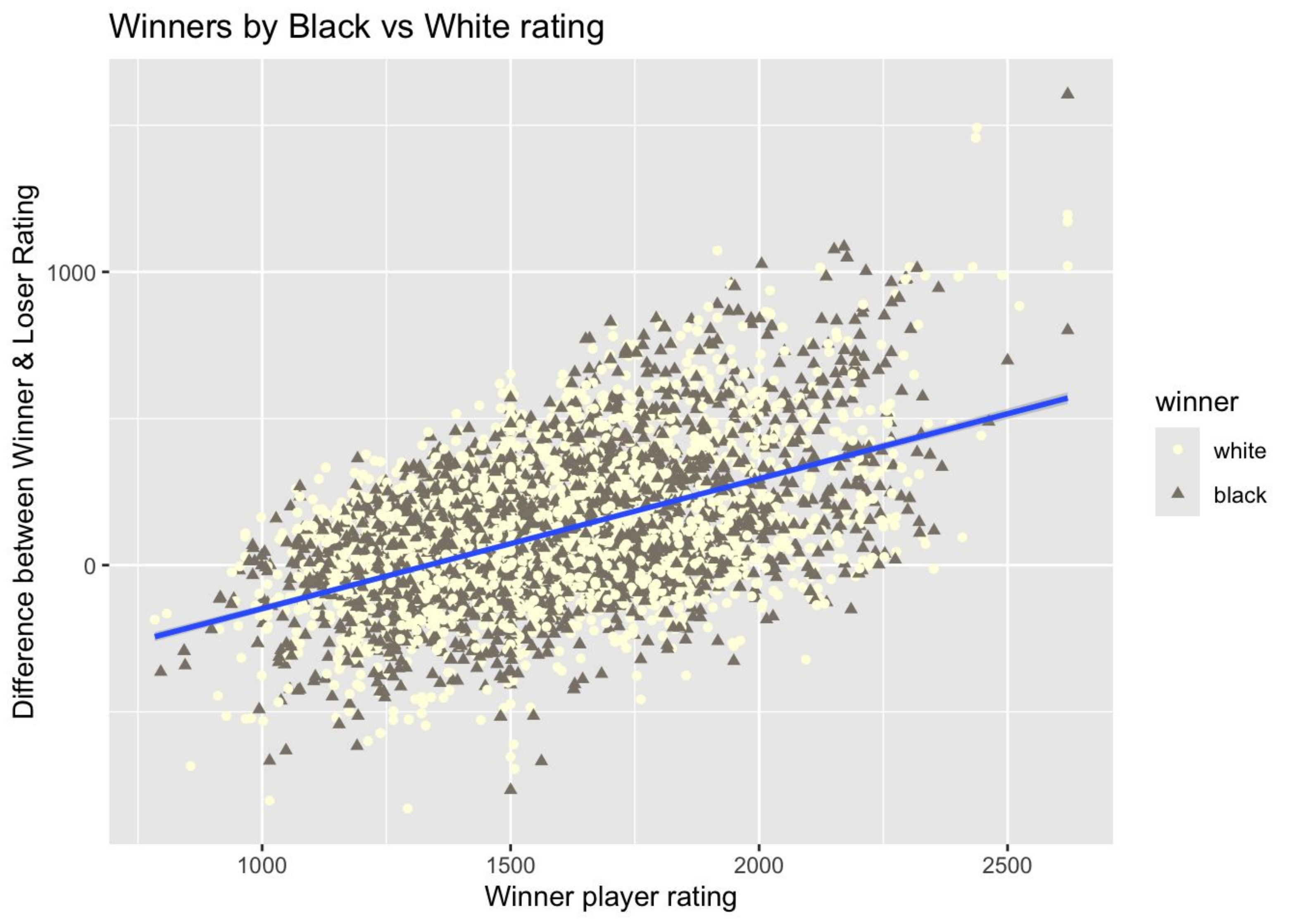


Most checkmates were delivered by white players (52.87% out of 6,325 games).

Data visualization

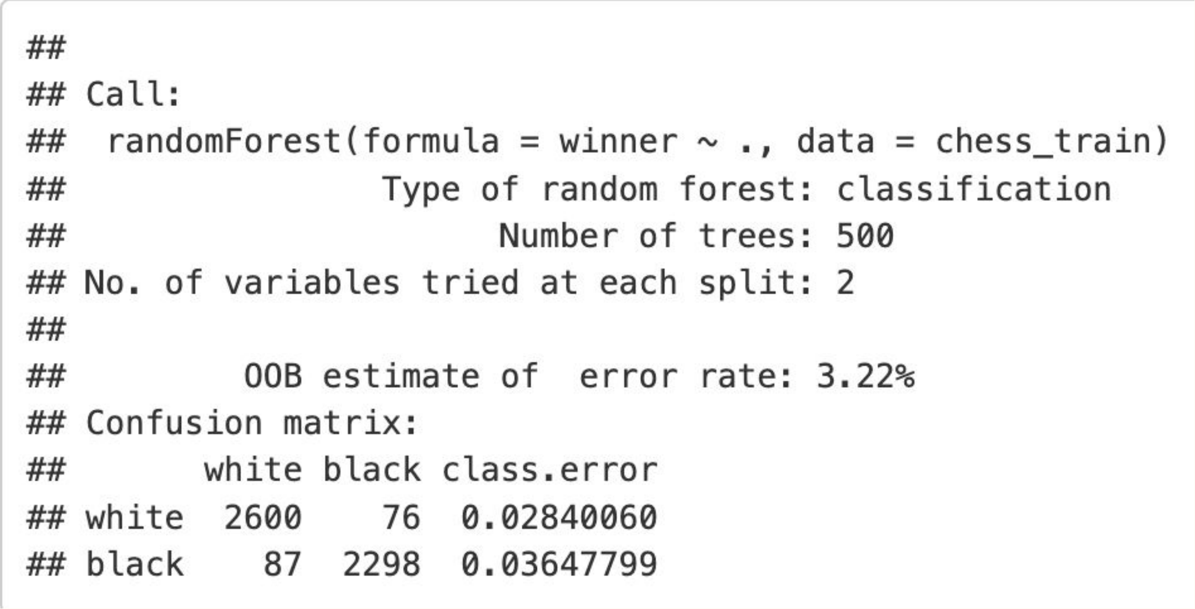
- Black winners have a slightly higher median player rating (1,556 vs 1,544.5)
- High outliers are caused by the same players, though white winners have a higher number of high outliers
- 6 of the top 10 rated players won as white players





There is a positive moderate linear relationship between the winner’s rating and how overleveled or underleveled they were compared to their opponent.

Modeling: Random Forest Model



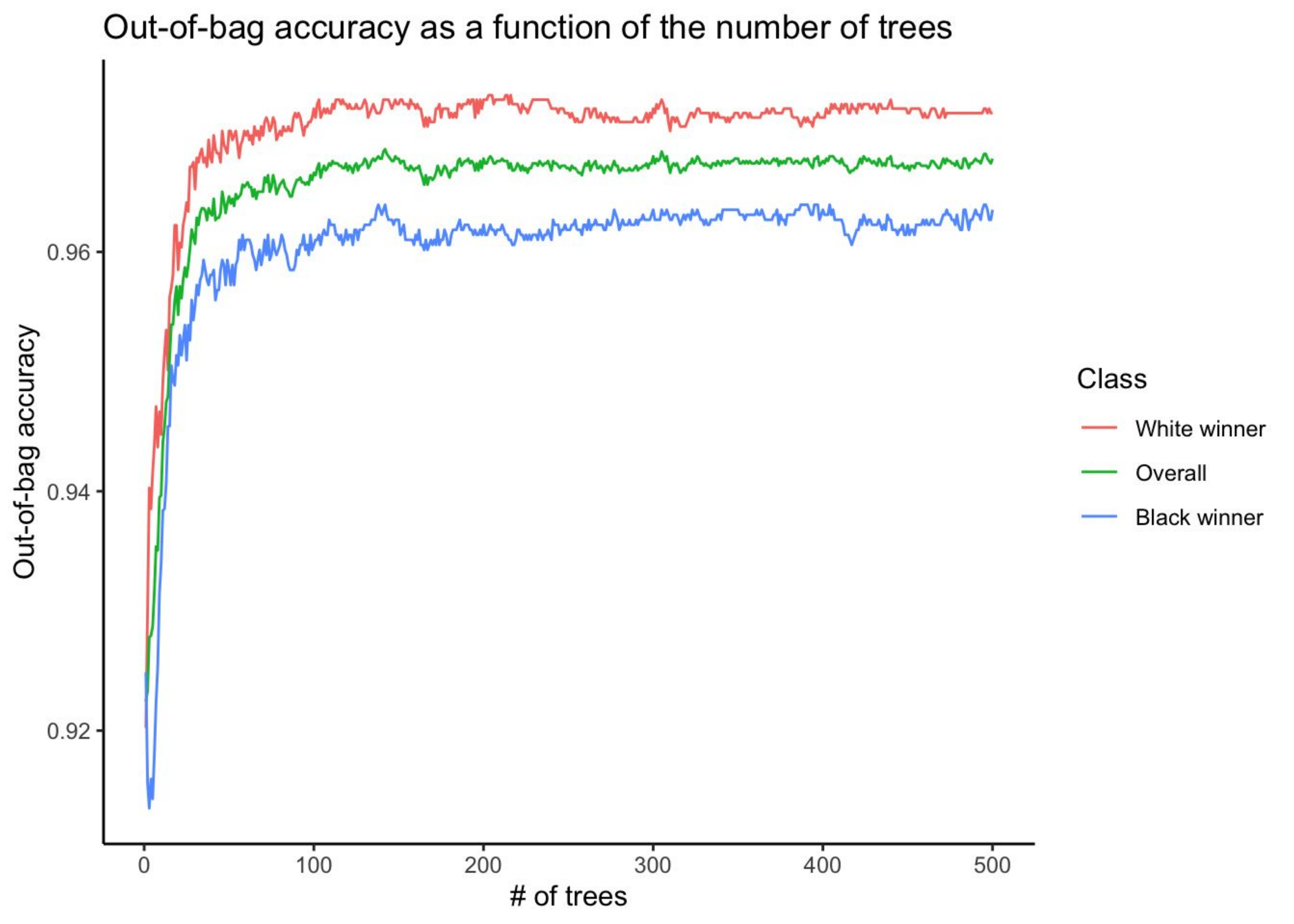
```
##
## Call:
##  randomForest(formula = winner ~ ., data = chess_train)
##                Type of random forest: classification
##                      Number of trees: 500
## No. of variables tried at each split: 2
##
##         OOB estimate of  error rate: 3.22%
## Confusion matrix:
##       white black class.error
## white  2600    76  0.02840060
## black    87  2298  0.03647799
```

Split my data into 80% training and 20% testing data.

- Out-of-bag error estimate was 3.22%.
- 96.78% of the out-of-bag observations were classified correctly.

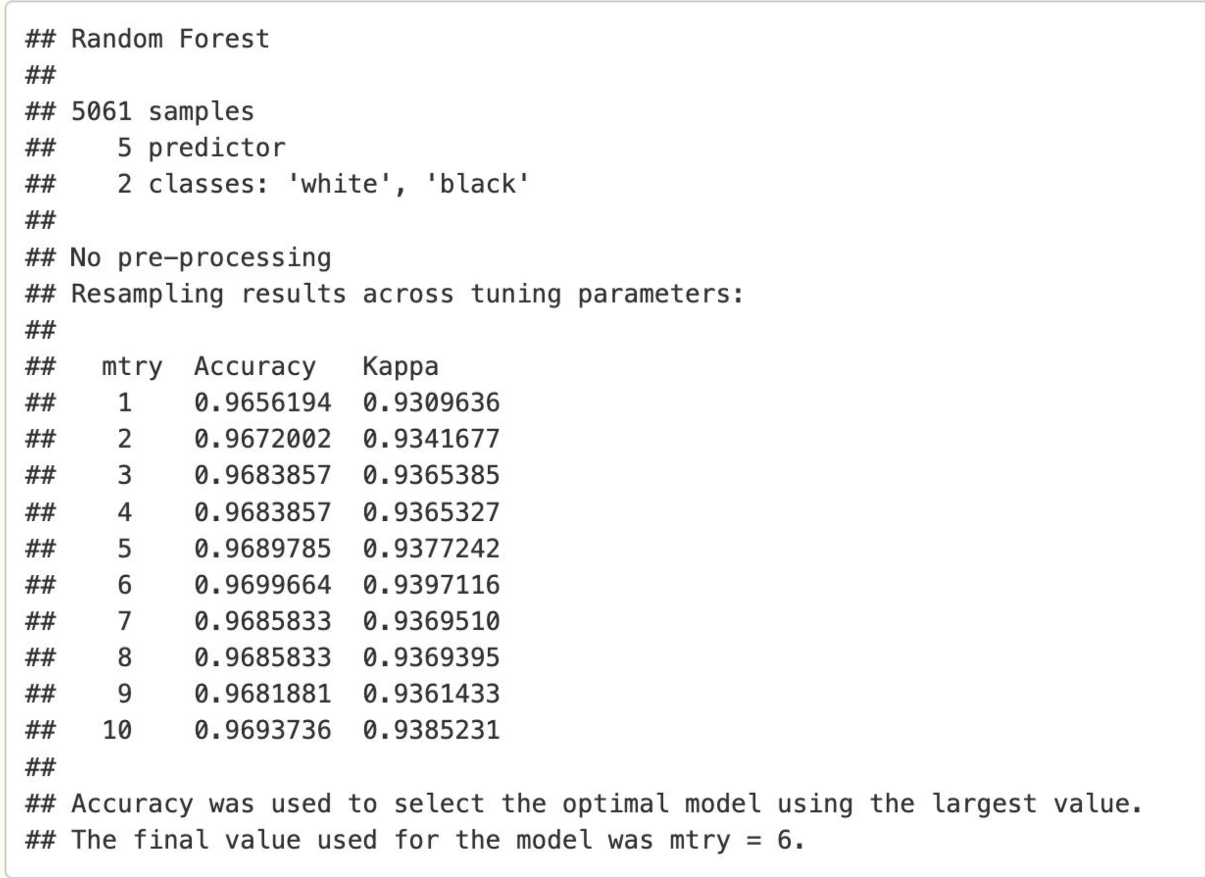
- 2,600 white winners were correctly labeled as white (These are true negatives.)
- 76 white winners were incorrectly labeled as black (This is a false positive.)
- 87 black winners were incorrect labeled as white. (This is a false negative.)
- 2,298 black winners were correctly labeled as black (These are true positives.)

Modeling: Random Forest Model

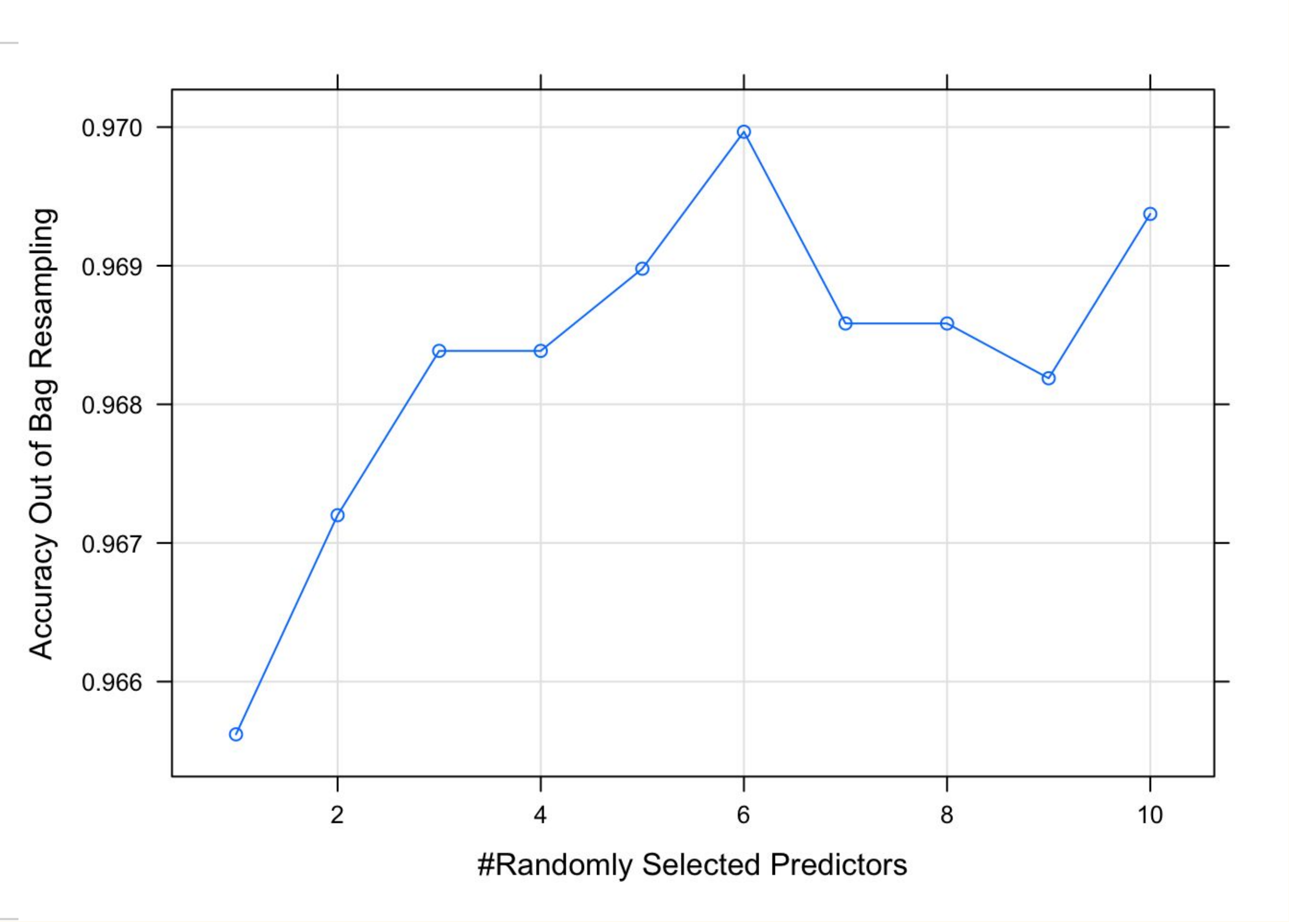


The error rate sort of stabilizes at around 170 trees.

Modeling: Random Forest Model



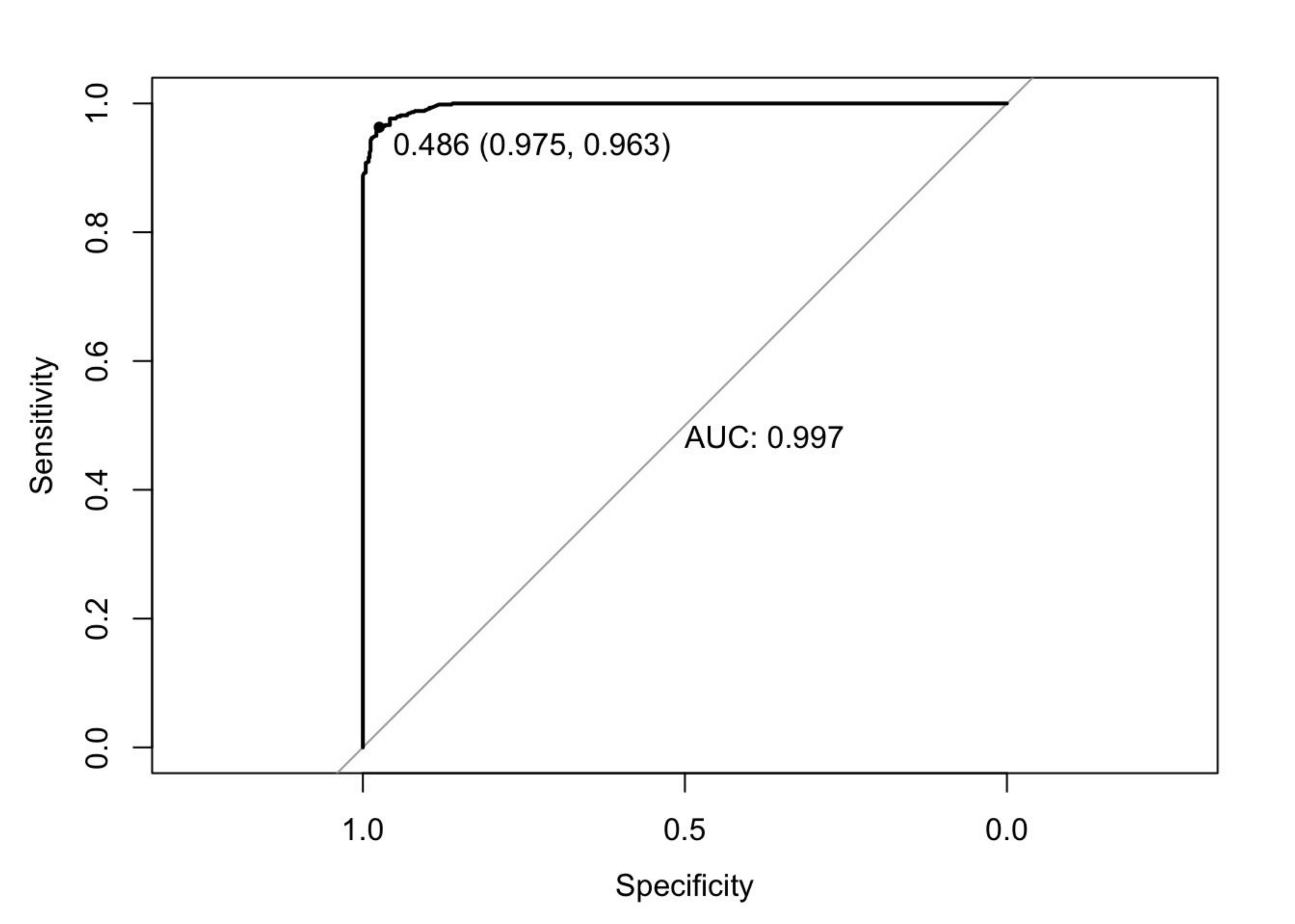
```
## Random Forest
##
## 5061 samples
##    5 predictor
##    2 classes: 'white', 'black'
##
## No pre-processing
## Resampling results across tuning parameters:
##
##   mtry  Accuracy   Kappa
##    1    0.9656194  0.9309636
##    2    0.9672002  0.9341677
##    3    0.9683857  0.9365385
##    4    0.9683857  0.9365327
##    5    0.9689785  0.9377242
##    6    0.9699664  0.9397116
##    7    0.9685833  0.9369510
##    8    0.9685833  0.9369395
##    9    0.9681881  0.9361433
##   10    0.9693736  0.9385231
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 6.
```



Set the number of trees to 250.

Modeling: Random Forest Model

The random forest has a high accuracy with an area under the curve of **0.997**.



Key Findings

Conclusion

There is a **positive moderate linear relationship** between player rating and winning by checkmate, and using **6 features at each split of trees** in a random forest model gives the best out-of-bag accuracy.

Possible next steps

Try to predict if a game resulted in a black or white winner depending on

- the opening move
- whether the game was a rated game or a casual game.

References

Sources:

- Chess Game Dataset:
  - via GitHub: https://github.com/rfordatascience/tidytuesday/blob/main/data/2024/2024-10-01/readme.md
  - via Kaggle by Mitchell J: https://www.kaggle.com/datasets/datasnaek/chess/data
- Wikipedia. "*Checkmate*", *from* https://en.wikipedia.org/wiki/Checkmate

Pictures:

- Title slide picture, from https://www.tapsmart.com/wp-content/uploads/2020/12/chess-header.jpg
- Chess board picture, from https://chessbazaar.gumlet.io/media/catalog/product/y/y/yy.jpg

Thank you!