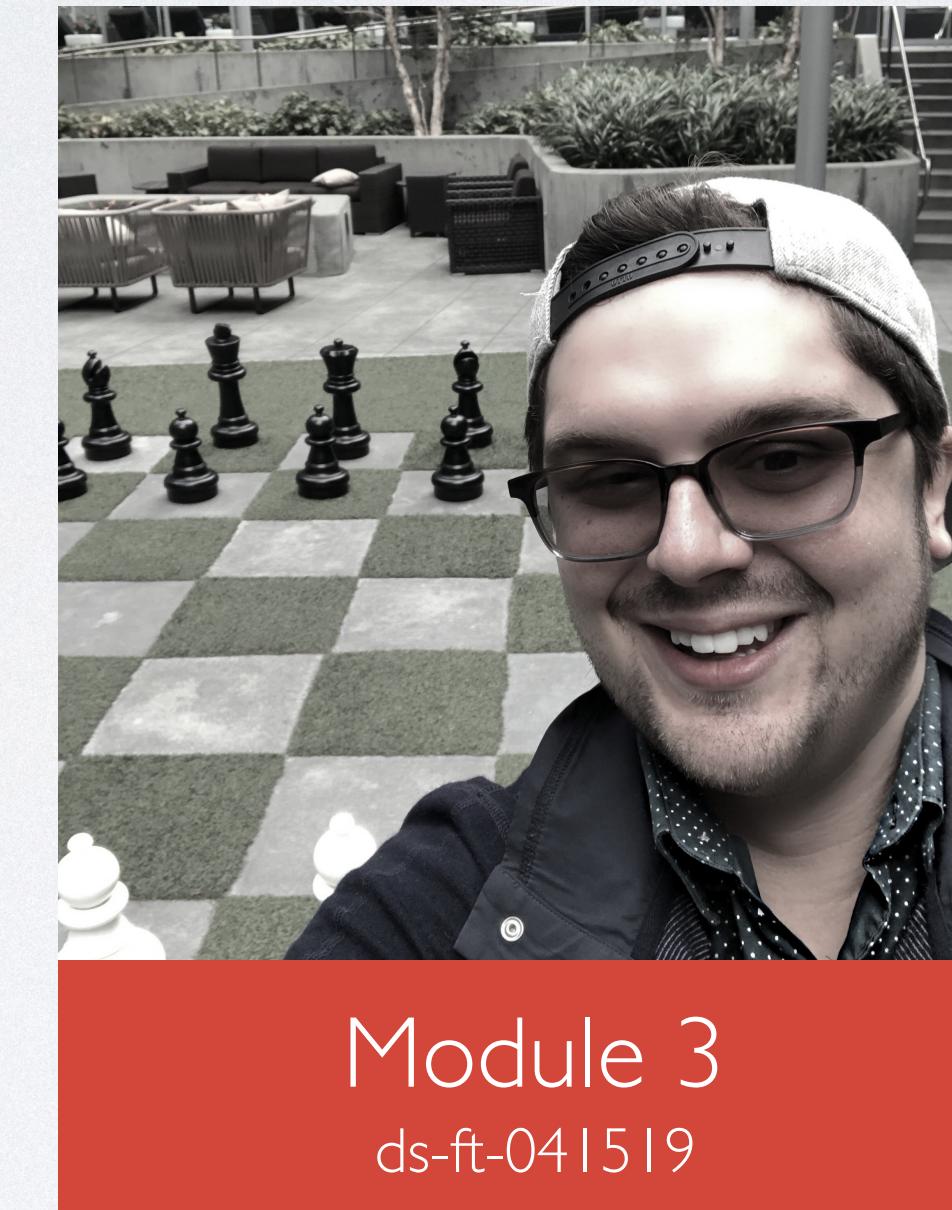


LICHESSE CHESS DATASET ANALYSIS



Module 3
ds-ft-041519

Paul Woody

CLASSIFICATION: PREDICTING THE WINNER OF CHESS GAMES

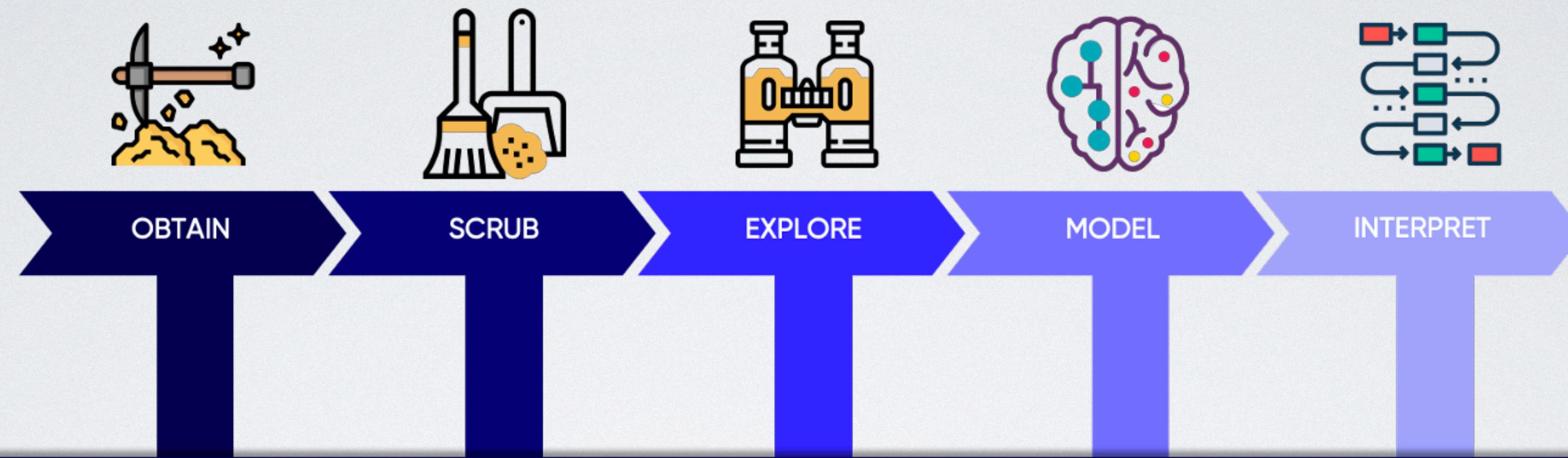
Methodology

Background

Feature Engineering

Insights

METHODOLOGY: OSEMN



O

Gather data from relevant sources

S

Clean data to formats that machine understands

E

Find significant patterns and trends using statistical methods

M

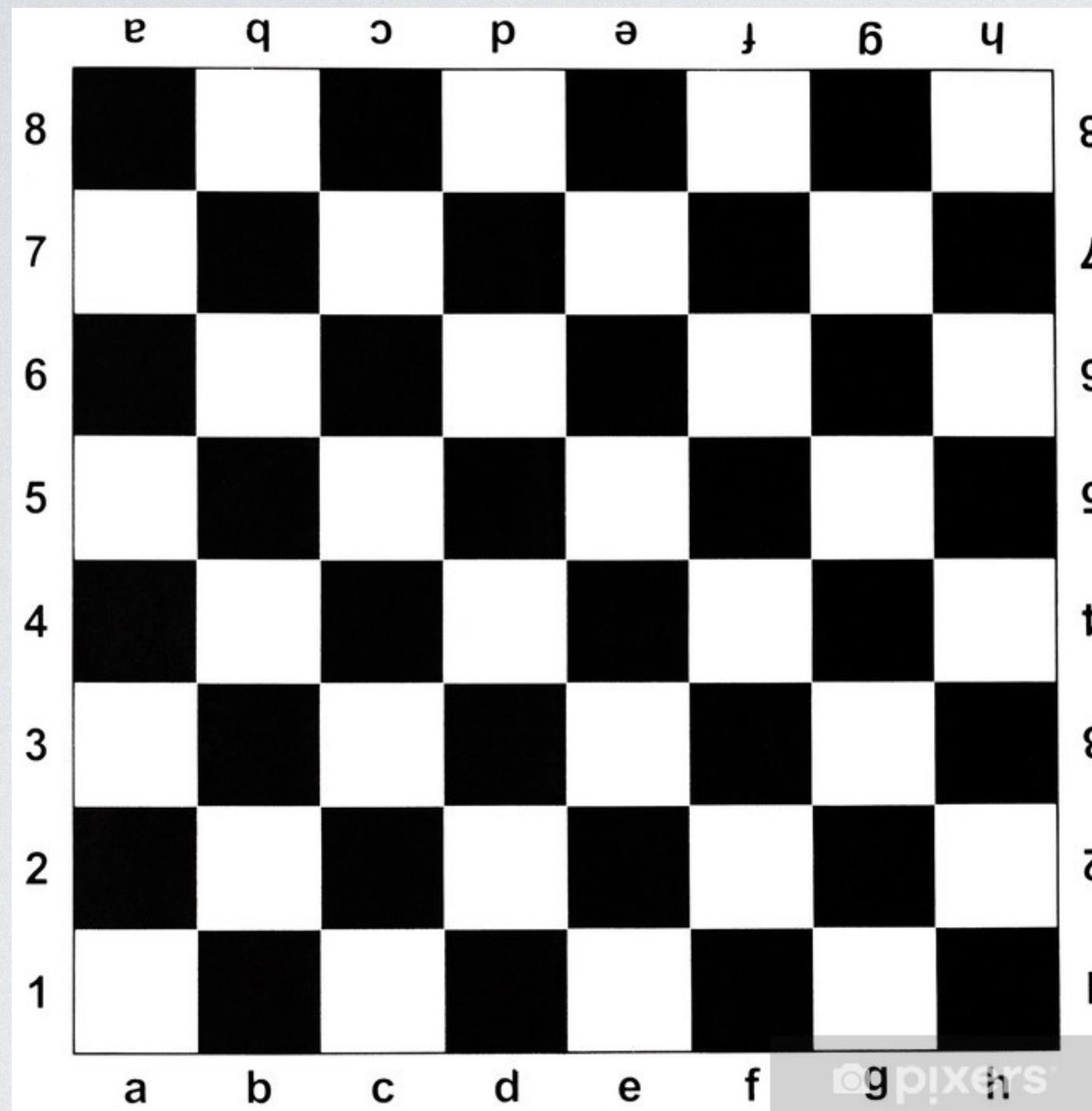
Construct models to predict and forecast

N

Put the results into good use

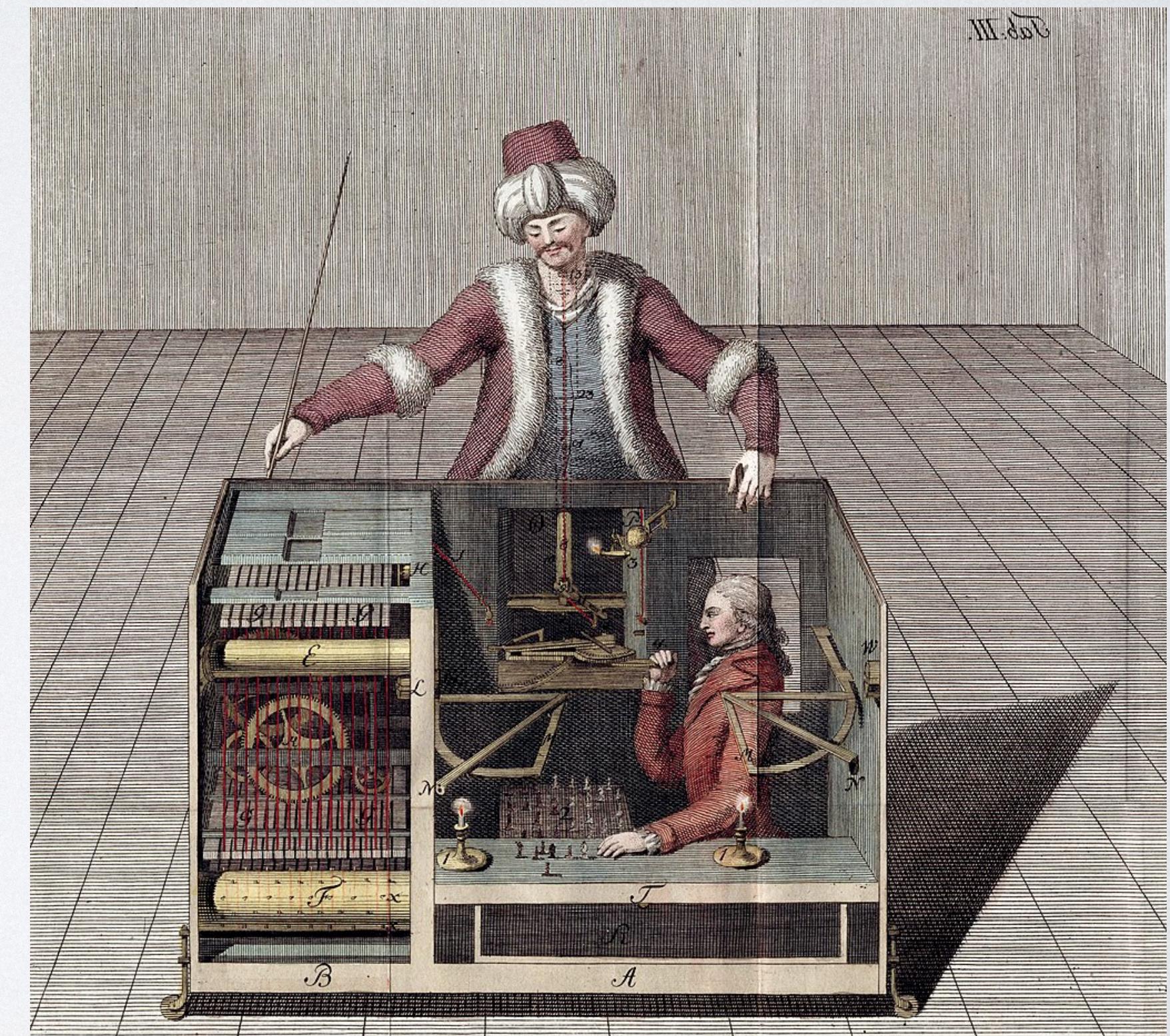
Perform exploratory data analysis to identify meaningful trends.
Use models and appropriate hypothesis tests to evaluate significance.

BACKGROUND



Interest in automation in chess has long captivated audiences, as evidenced by the debunked **Mechanical Turk** of the 18th century.

There are more possible positions in chess than atoms in the observable universe, as characterized by the “Shannon number” - a value which approximates chess game-tree complexity.



BACKGROUND



A chess player's **Elo score** offers an approximation of skill level and fluctuates as games between ranked players are won or lost.

Performance in a game depends on a number of factors such as the first move played and number of opponent pieces captured from the board.



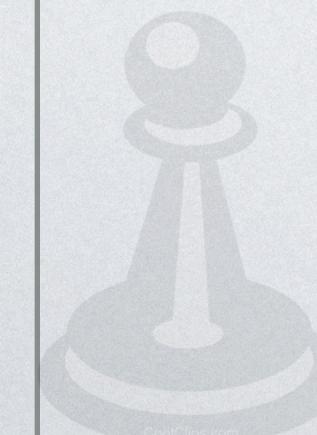
FEATURE ENGINEERING

Without an understanding of Chess, we'd still be able to fit the data above into a model and do an 'okay' job with the features provided by LiChess.

However, applying what we know about Chess, we're able to manipulate the existing features of the dataset to produce new input variables that will hopefully improve the predictive accuracy of our model(s). This process, known as **Feature Engineering** is crucial to the implementation and improvement of classification models.

```
# Feature Importance
features = pd.DataFrame(columns = ['feature',
                                      'importance'])
features['feature'] = x.columns
features['importance'] = rf.feature_importances_
features.sort_values('importance', ascending=False)
[0:15]
```

Without After applying our knowledge of Chess to create new features, we found that of the top 15 'best' predictors, **12/15 were created as a result of feature engineering.**

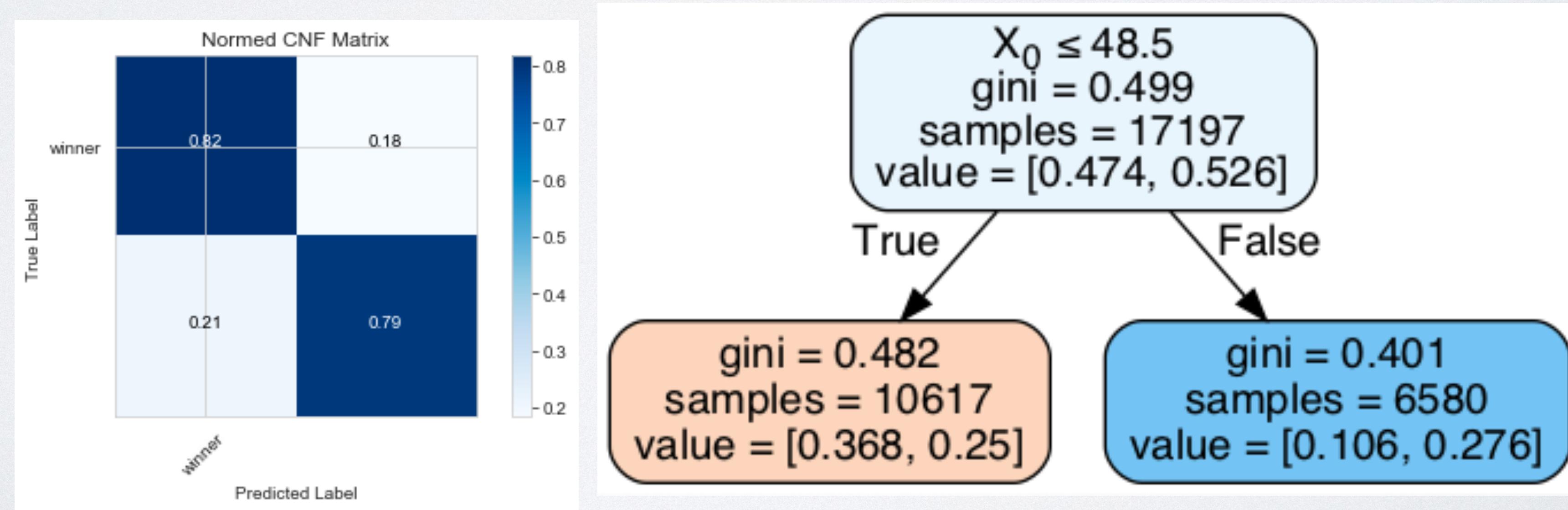


MODEL INSIGHTS

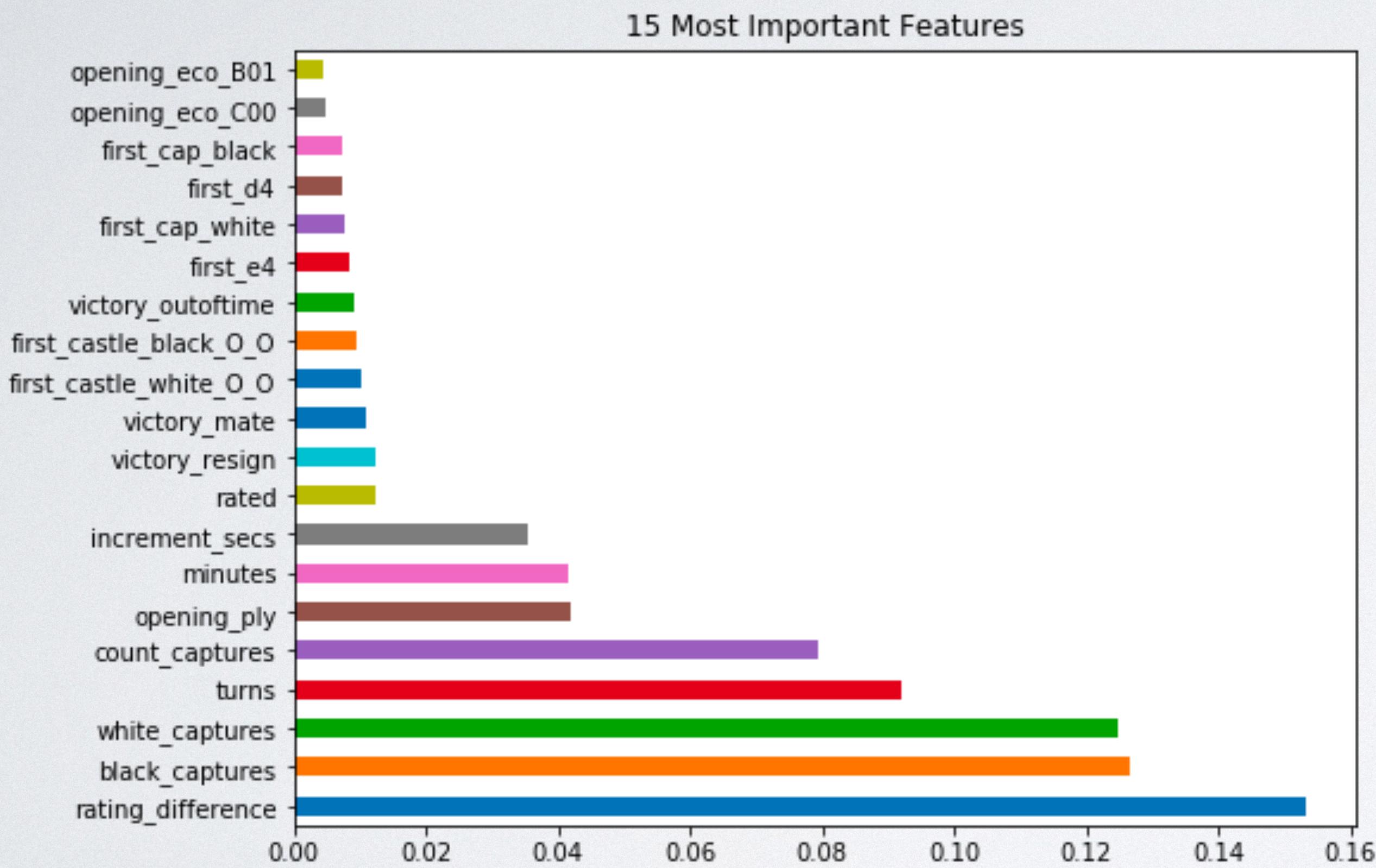
Model Performance:

Logistic Regression mean score: 0.8248381410406866
Random Forest mean score: 0.8219059307271458
Bagged Tree mean score: 0.8057360826397231
Adaboost Tree mean score: 0.8204943718042659

Adaboosting: A Closer Look at Model Performance



CHESS INSIGHTS



- Rating Difference is the strongest predictor
- Maintaining the material balance is vital — capturing pieces and taking favorable trades produces advantages
- Choice in opening moves is important, based on historical data playing 'e4' or 'd4' maximizes chance for victory.
- Castling is important, but perhaps not as crucial as choice in opening and material balance

FUTURE WORK

In this experiment, we put two of the 4 opening principles of Chess to the test in order to determine their importance in deciding the outcome of a Chess game.

1. Control the center
2. Develop your pieces
3. ~~Protect the King and castle early~~
4. ~~Maintain the material balance~~

In a future experiment, additional feature engineering could be implemented in order to seek to quantify the value of controlling the 4 central squares of the board, as well as the effect produced by varying rates of major piece development by each player.

THANK YOU
FOR YOUR
TIME