SPRINT 3

Predicting Tennis Outcomes

TABLE OF CONTENTS

01 Overview of problem statement

02 Dataset and preprocessing procedures

03 EDA Findings

04 Baseline model and evaluation metrics

05 Next steps

Can tennis outcomes be predicted?

Problem

- Debate and general uncertainty around who will win professional tennis matches

Solution

- Build a predictive model to see not only who will win, but what features are important

Impact

- Fans
- Analytics
- Betting

Preprocessing

Filtering dataset down

- 2000s onwards (modern tennis)
- Highest level professional tournaments
  - Grand Slams and Masters 1000s
- Rows where NAN values
- Raw: 73247 entries   Filtered: 25585 entries

Feature Engineering

- Surface win percentage for both winner and loser prior to match
  - Cumulative and chronological

EDA

checking for class imbalance in target variable and simple model

- 67% for higher ranked

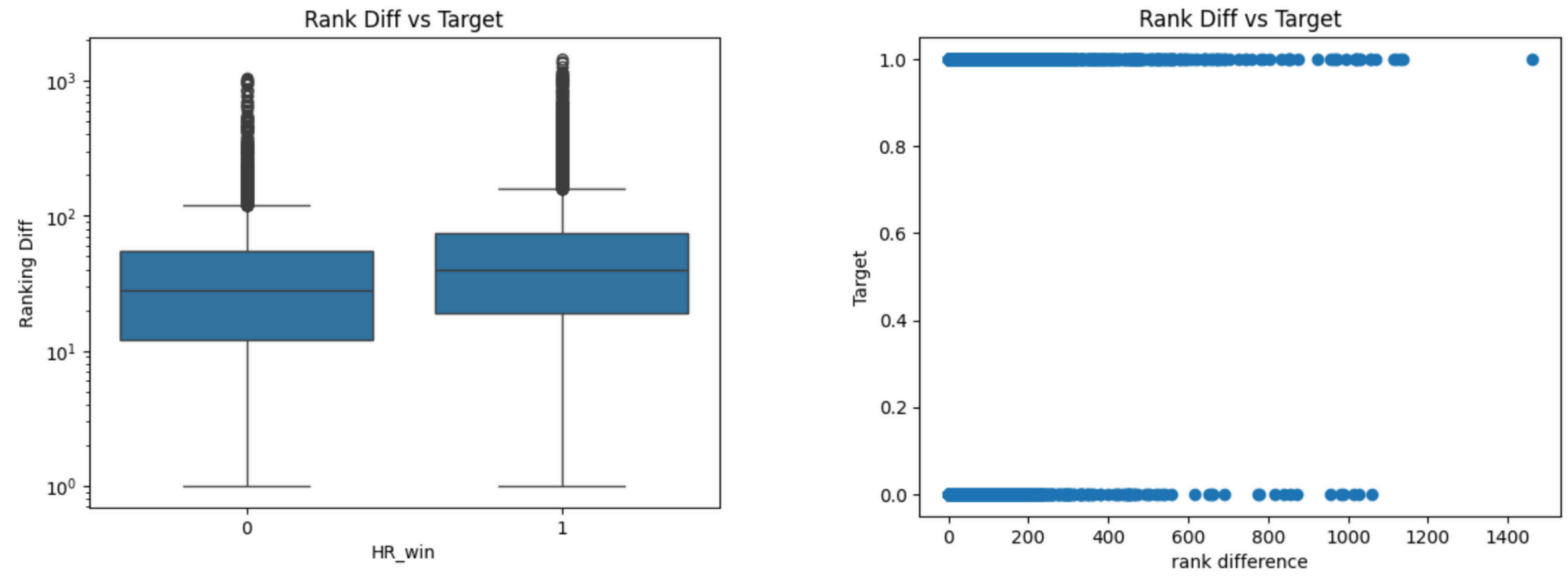
Non-linearity of some features and no real demarcation

- Age
- Height

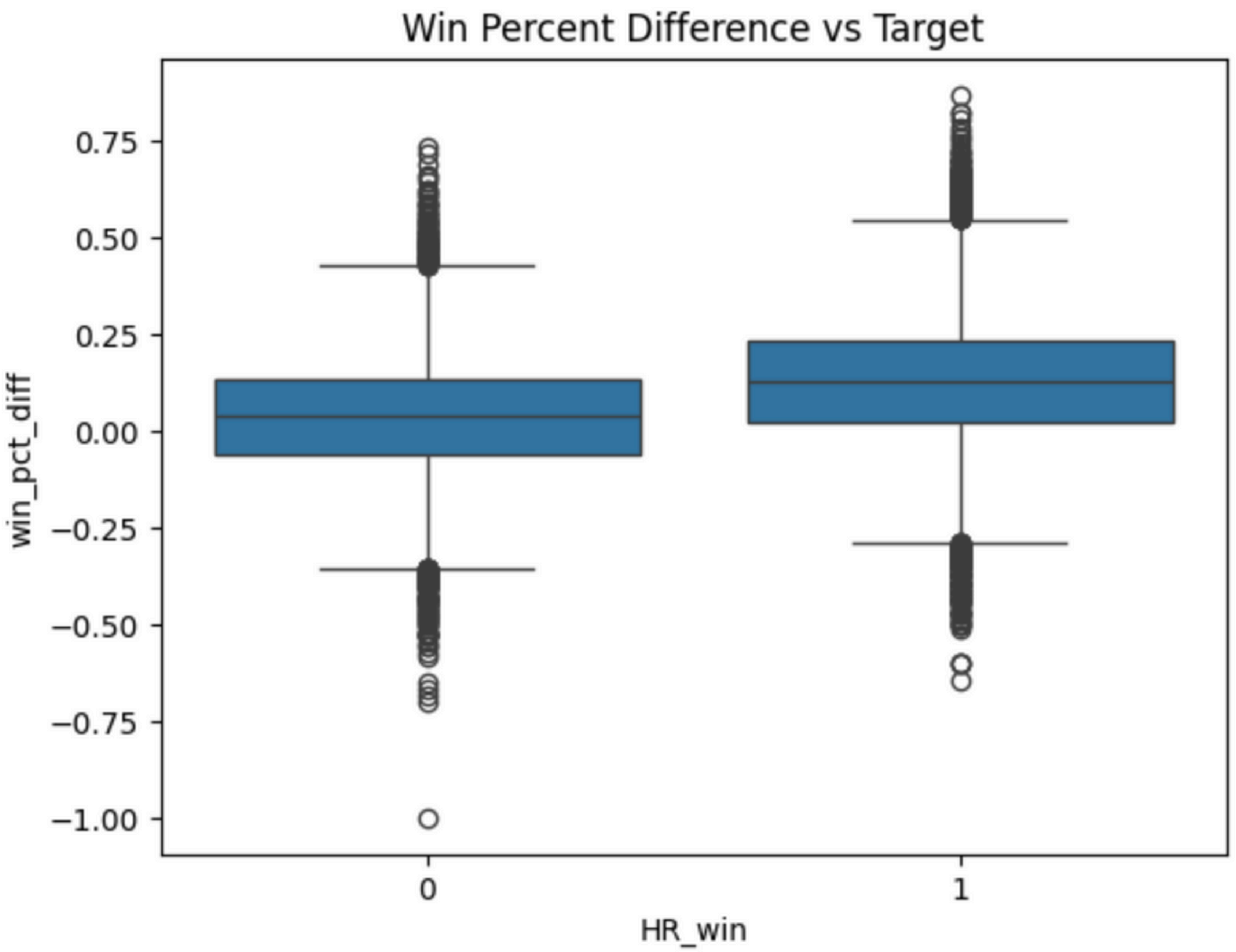
Two features show (marginal) promise!

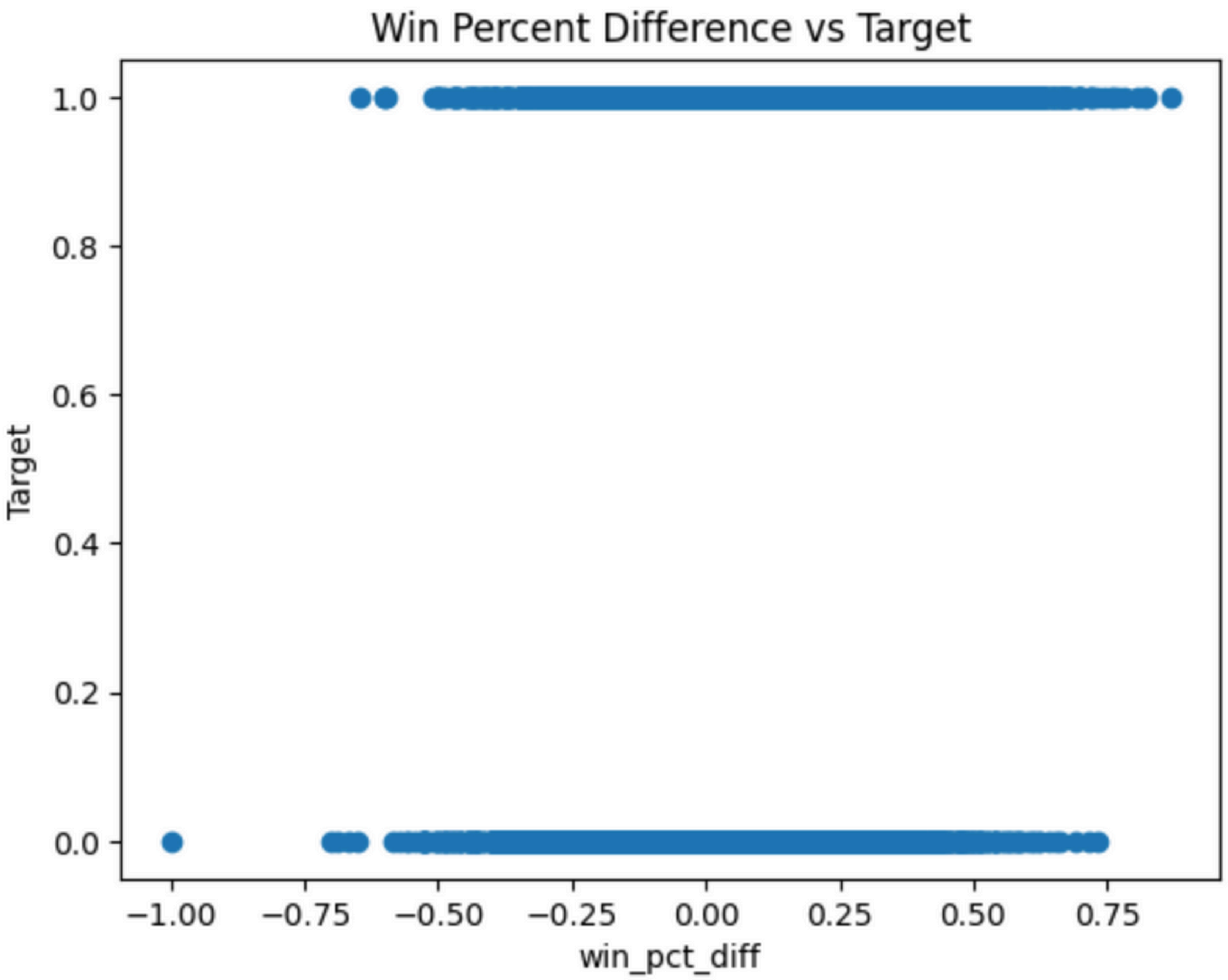
- Ranking difference
- Surface win percentage

Ranking Difference



Surface Win Percent Difference

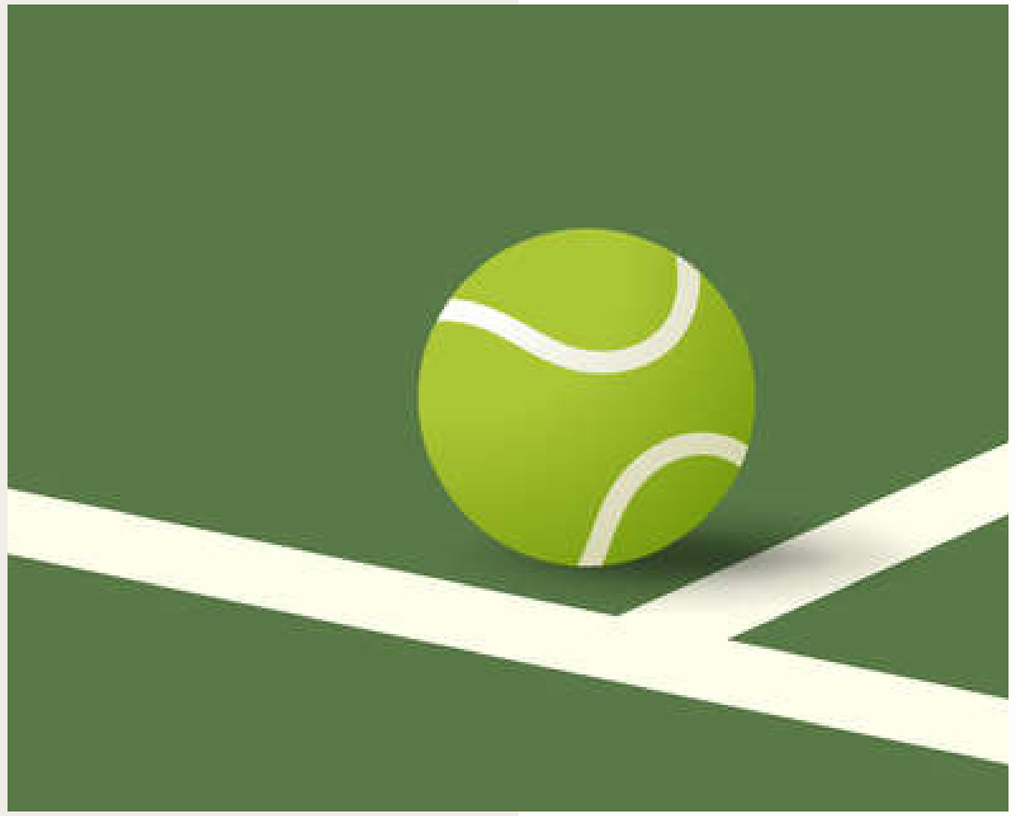




MODELING

Outcome of logistic regression and random forest models

- ~69% accuracy for train and test score
- 2% bump in accuracy
- Minimal overfitting (test score occasionally better)



MODELING RESULTS

Model Type	Train Score	Test Score
Baseline	67.3%	67.3%
Logistic Regression	68.4%	69.6%
Random Forest	69.4%	69.1%

NEXT STEPS & CONCLUSION

Areas for improvement

- Adding additional features
  - Recent form (win percentage for past 6 months)
- One hot encoding court surface
- Gradient Boosting model

Conclusion

- ATP ranking system is strong predictor!
  - point system works quite well
- Surface win percentage valuable analytical feature