

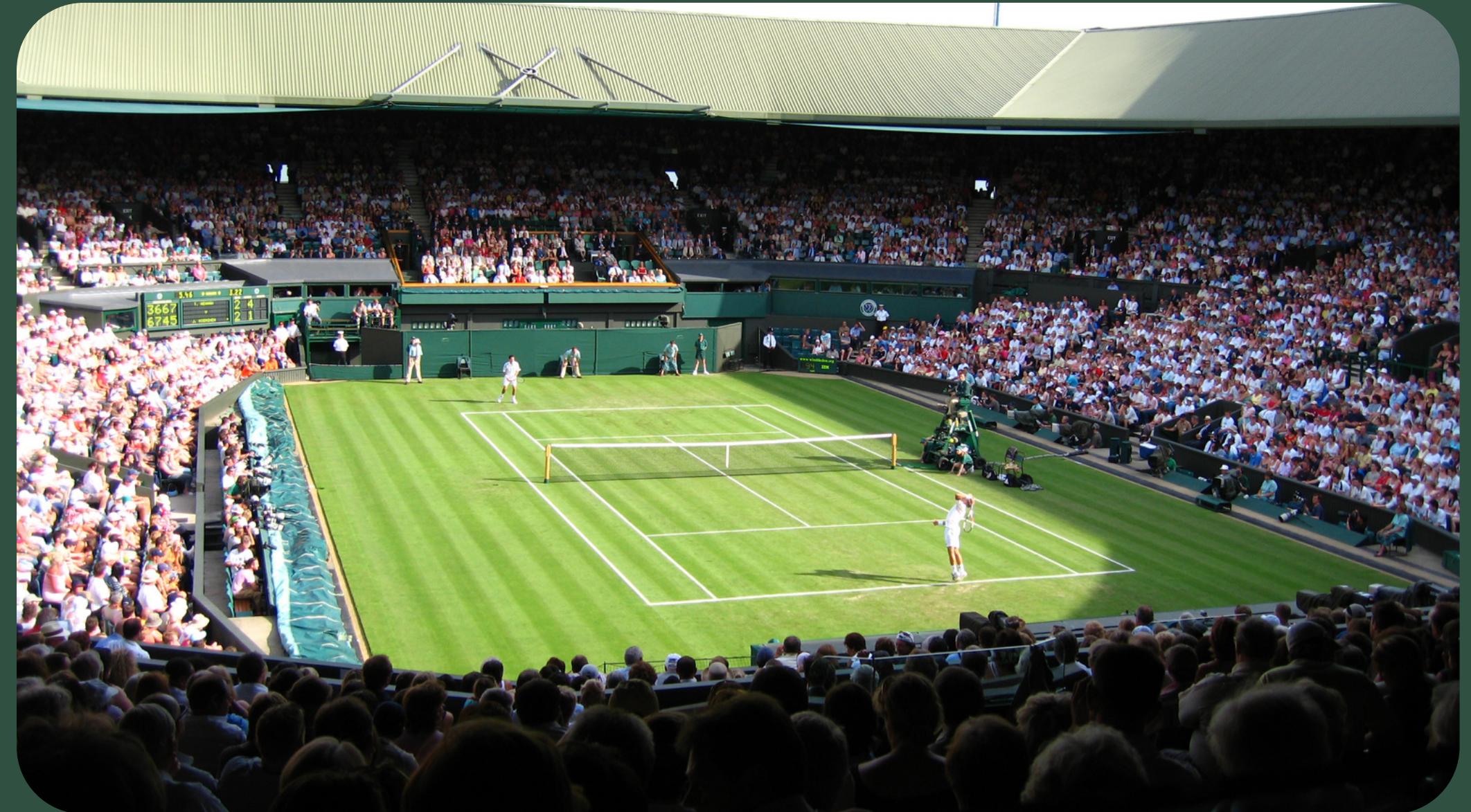
DAVIDE MATTÀ

CAN STATISTICAL LEARNING OUTPERFORM BOOKMAKERS?

Data Science for Economics – University of Milan

CONTENTS

- Research Question
- Features
- Unsupervised Models
- Supervised Models
- Test Results
- Other Players
- New Data
- Conclusions



RESEARCH QUESTION

- We might be interested in predicting the winner of a tennis match
- Odds are a good predictor, but they are often wrong
- Can we use statistical learning instead?

OUR APPROACH

- Choose a reference player first and then replicate the analysis with other players
- Standard data analysis and unsupervised learning for analyzing the data
- Build supervised learning models to predict the winner of a match
- Compare their performance with the odds

OUR PLAYERS



NOVAK DJOKOVIC

ATP #1



DANIIL MEDVEDEV

ATP #3



CASPER RUUD

ATP #4

ILLUSTRATED ANALYSIS IS ABOUT DJOKOVIC

THE FEATURES

RANKINGS

- Player's ranking (log)
- Opponent's Ranking (log)

TYPE OF COURT AND MATCH

- Court (Hard, Clay, Grass, Indoor)
- Maximum number of sets (3, 5)
- Match Importance score

OPPONENT CHARACTERISTICS

- H2Hs
- Percentage of opponent wins with this court and number of sets

FORM

- Recent player performances
- Recent opponent performances

STATISTICAL LEARNING METHODS

UNSUPERVISED TECHNIQUES

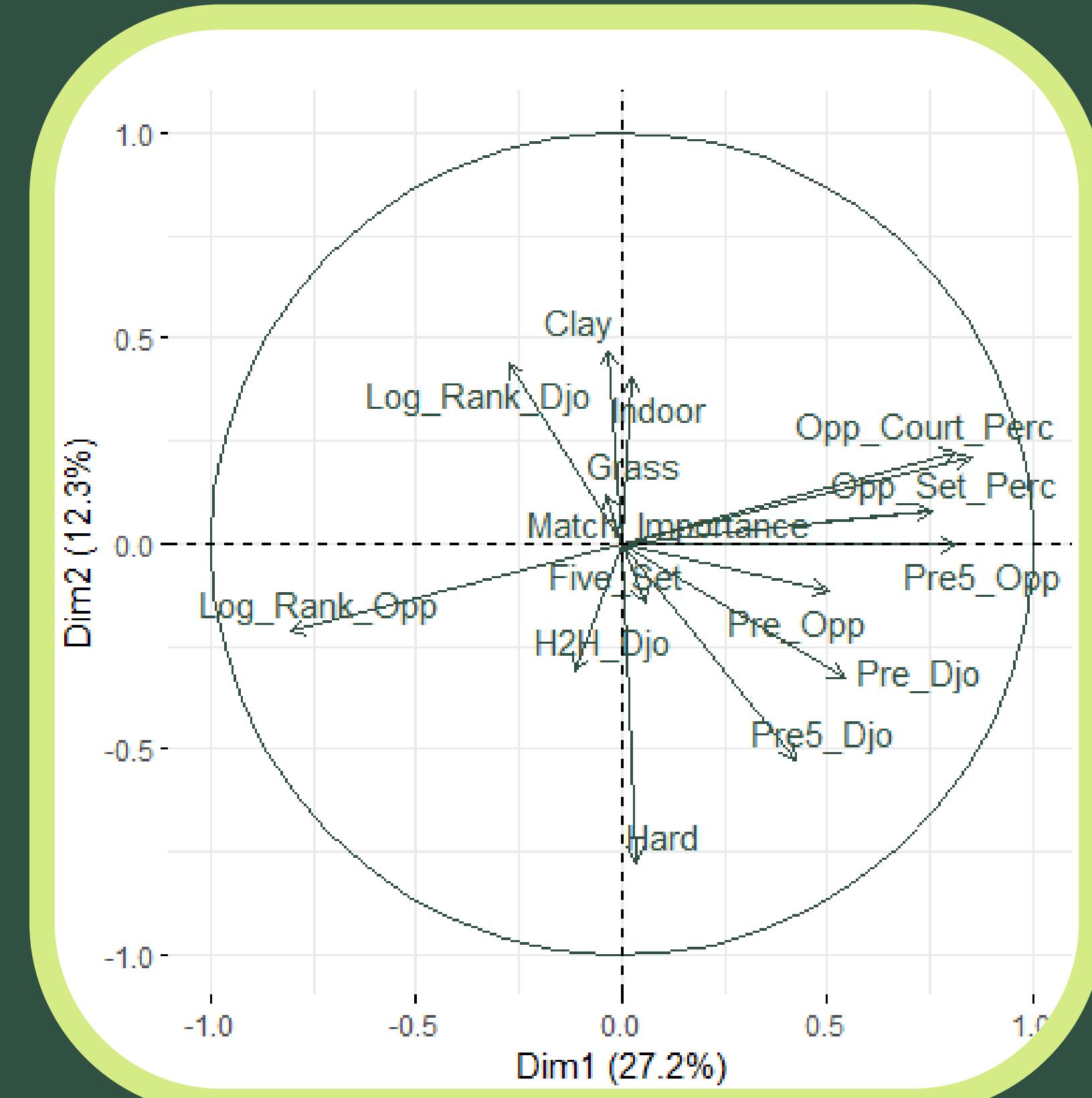
- Principal Component Analysis
- 2-Means Clustering

SUPERVISED TECHNIQUES

- Standard Linear Regression
- Robust Linear Regression
- Logistic Regression

PRINCIPAL COMPONENT ANALYSIS

Can data be represented
in two dimensions?
Yes . . . but no
First two components
explain less than 40%
of the variance



2-MEANS CLUSTERING

Clusters partly resemble
in some way the Win -
Loss partition.

If we use Cl. 1 = Win,
there is a 71%
correspondence.

This is encouraging,
but . . . has no meaning.



Algorithm performed on the whole dataset,
only two variables are shown for simplicity

SUPERVISED LEARNING

- We divide the 1126 observations into training (70%) and test (30%) set
- Odds test accuracy for Djokovic's matches is 86.35%
- Can we do better?



STANDARD LINEAR REGRESSION

We perform best subset selection and keep seven variables. The probability to win a match depends on:

POSITIVELY ↑

- Opponent's Ranking
- Number of Sets
- If the opponent won his last match

NEGATIVELY ↓

- Djokovic's Ranking
- Clay court
- Opponent's Win ratio over his last 5 matches
- Opponent's results in that surface

This model reached a 84.56% accuracy

SOME OLS ASSUMPTIONS

PT. 1

- SENSITIVITY TO OUTLIERS => ROBUST M ESTIMATORS

- We cleaned the dataset, so all data are correct, but it is still worth seeing what happens if we use a robust method
- The robust OLS always predict a Win!
- The only non-zero coefficient is the intercept, which is equal to 1
- We cannot rely on this model

SOME OLS ASSUMPTIONS PT. 2

• LINEARITY => LOGISTIC REGRESSION

- When the response variable is a probability, a logistic model is usually more appropriate
- The logit model with 14 features (there is no multicollinearity) is accurate at 86.94%
- That is enough to outperform the odds!



LOGIT MODEL'S QUALITIES

- It is the most suited model according to theory and it complies with all assumptions
- It is the model with higher accuracy
- The other models beat the odds only when the latter predict a loss; the logit is able to outperform them in both directions

DANIL MEDVEDEV

- The best model is a 5-variables logit with Log_Rank_Med, Log_Rank_Opp, Clay, Opp_Court_Perc and Pre_Med
- Odds accuracy on Medvedev is 73.91%
- Unfortunately our model stops at 73.04%

CASPER RUUD

- The best model is a 6-variables logit with Log_Rank_Ruud, Log_Rank_Opp, Clay, Indoor, Opp_Court_Perc and H2H_Ruud
- Odds accuracy on Ruud is 70.00% and our matches exactly that percentage!

2023 – ROLAND GARROS

- Fresh new data available to test the models again
- Our models only got wrong one match and beat the odds (2 wrong predictions)



ALCARAZ (AND THE ODDS) VS DJOKOVIC (AND THE LOGIT)



Odds only gave Djokovic a 36% chance

Our logit correctly predicted his win (54%)



CONCLUSIONS

- We obtained models that work well and satisfy their assumptions
- Some of them are more accurate predictors than the odds, or at least as good as
- Good predictive power in both train - test partition and new data.
- Some player are harder to predict than others (for both us and the odds)

REFERENCES

Data source: <https://www.kaggle.com/datasets/dissfya/atp-tennis-2000-2023daily-pull>

Code and full report: <https://github.com/DavideMatta/TennisticalLearning>

Pictures in the slides

https://commons.wikimedia.org/wiki/File:Centre_Court_Wimbledon_%282%29.jpg

<https://www.flickr.com/photos/markusunger/25330109179>

https://commons.wikimedia.org/wiki/File:Australian_Open_2020_%2849836756233%29.jpg

https://commons.wikimedia.org/wiki/File:Casper_Ruud_%28NOR%29_%2821311773908%29.jpg

https://snl.no/Novak_Djokovic

<https://commons.wikimedia.org/wiki/File:RolandGarrosCentral.jpg>

https://commons.wikimedia.org/wiki/File:Carlos_Alcaraz.jpg

https://commons.wikimedia.org/wiki/File:Djokovic_MCM22_%2852035377907%29_%28edited%29.jpg