NCAA BASKETBALL

PREDICTING THE OUTCOMES OF BASKETBALL GAMES

CONTENTS

- Introduction
- The Data
- Cleaning the Data
- Analysis
- Machine Learning
- Conclusion

INTRODUCTION

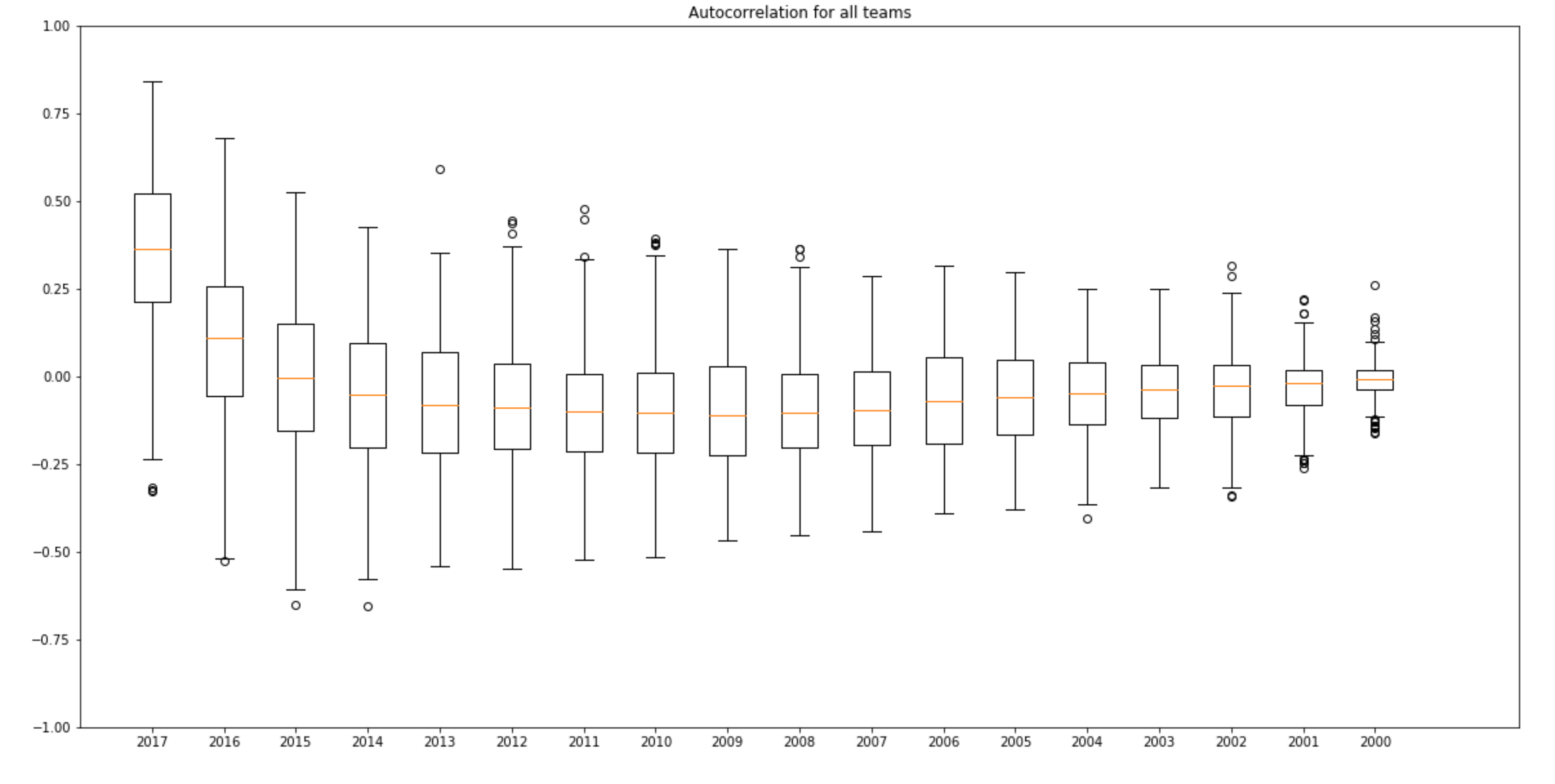
- What variables are correlated with whether or not a team wins a game?
- Why is this important?
- How can we predict which team will win a game?

THE DATA

- Multiple csv files with team stats (from Kaggle)
- Overall ratings for teams (Sports-reference.com)
- Merging multiple files, web scraping
- Data from past 18 years

CLEANING THE DATA

- Tools: pandas, numpy, matplotlib
- Extracting the data needed
- View autocorrelation to determine which data to use
- Preparing the data for merging



For the previous two years, there is a correlation, but after that the correlation is almost zero. This indicates that the data from the previous two years may be useful for predicting the outcomes for the current year.

CREATING FEATURES

- Raw data contains numbers, but we want rates
- Example: Field goal rate instead of number of field goals made
- Created function to calculate rates for each team for previous years
- Also calculated average number of steals, rebounds, blocks, etc.
- Average difference in scores for each game

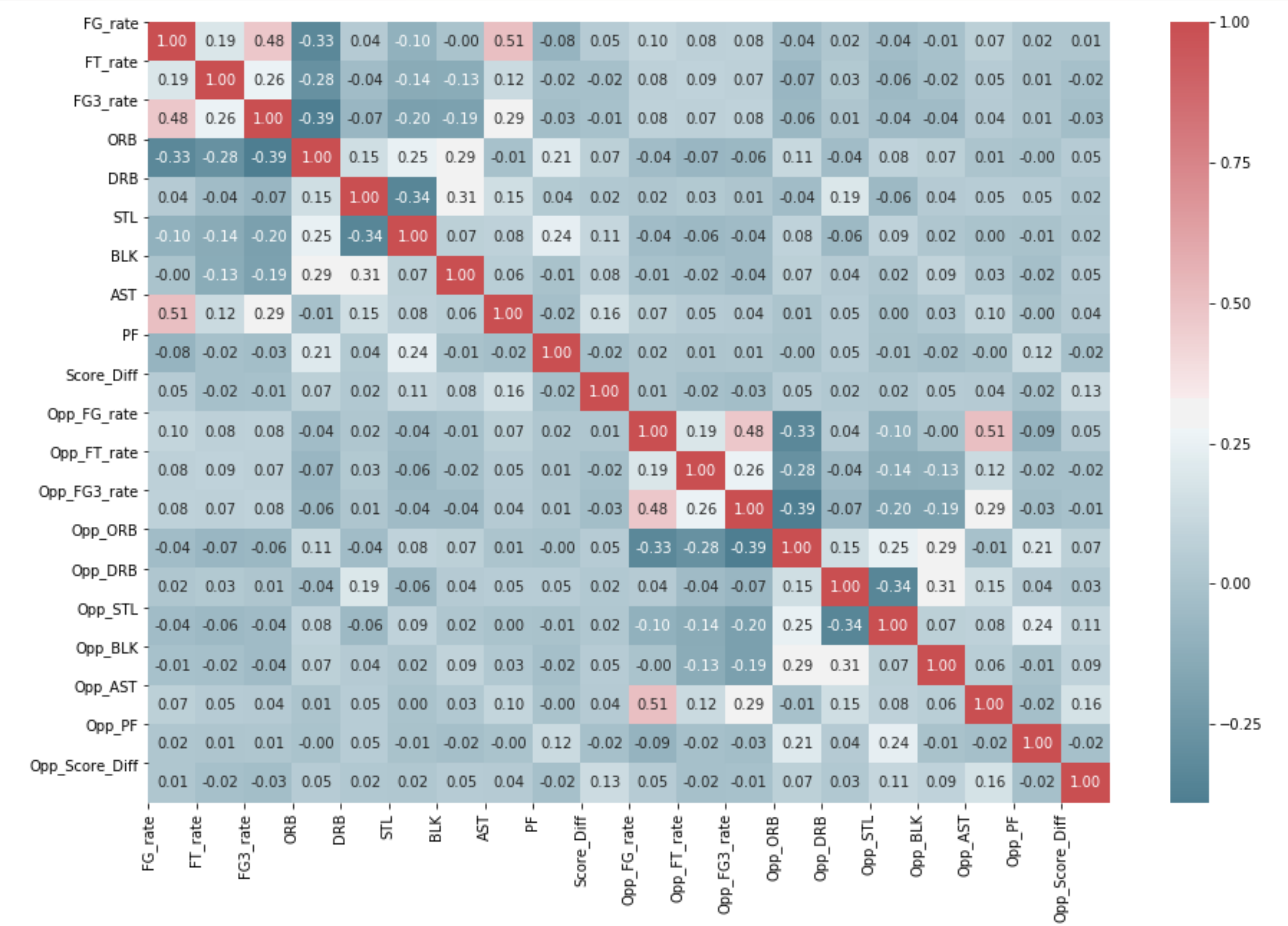
PREPARING THE DATA FOR ANALYSIS

- Merged the dataframes
- Created a column of ones and zeros, indicating whether the team won or lost
- Exported cleaned data as csv file

ANALYSIS

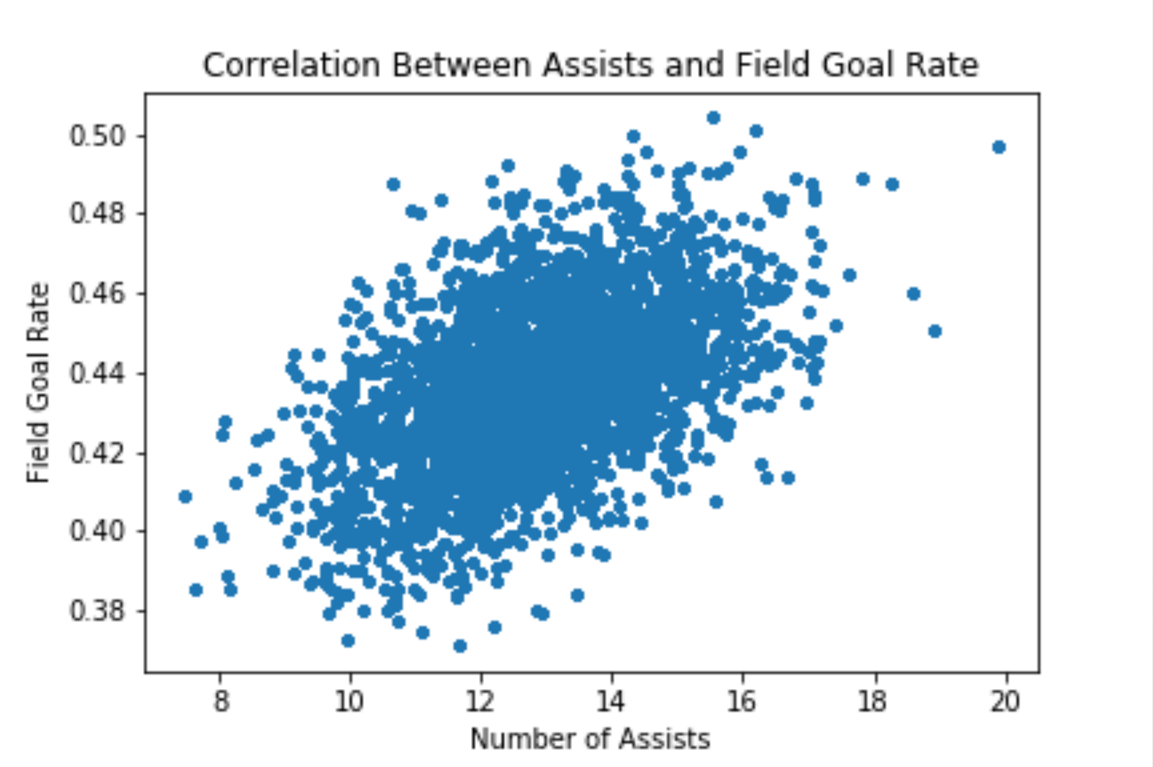
- Which features are correlated with each other?
- Which features are useful for predicting the outcome of a game?
- Testing the difference in a feature for winners versus losers.

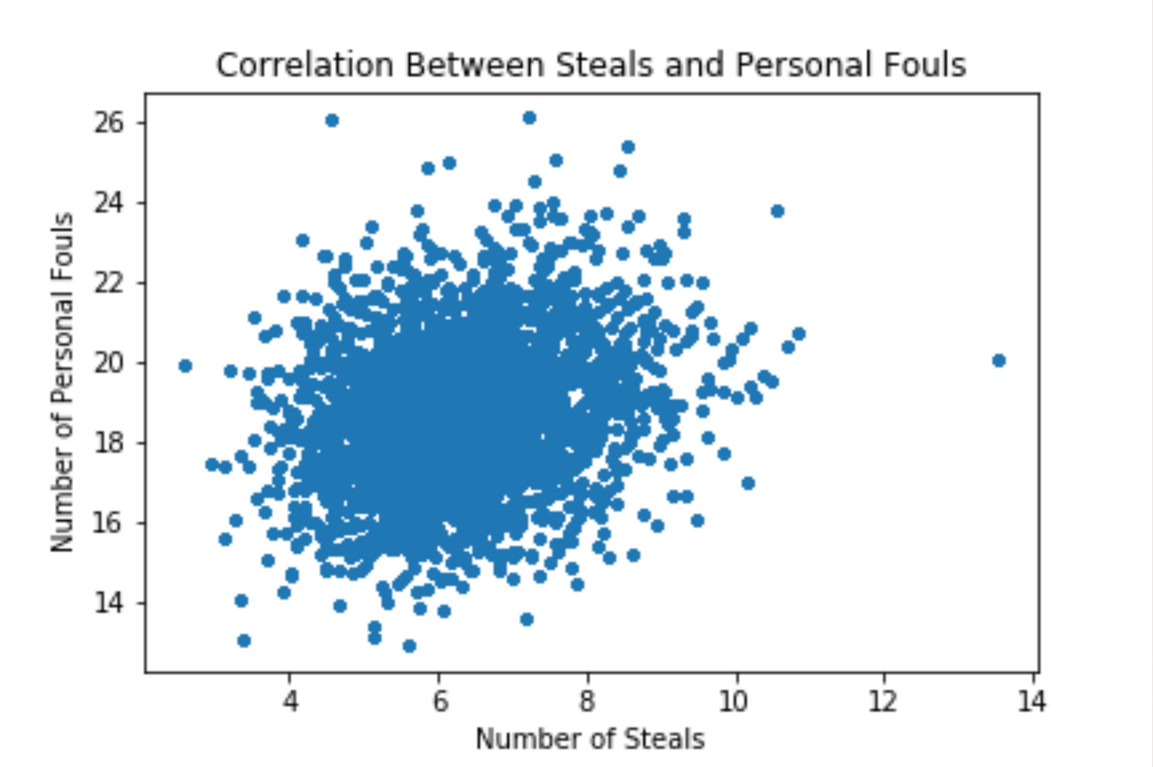
Heat map of the correlation between features



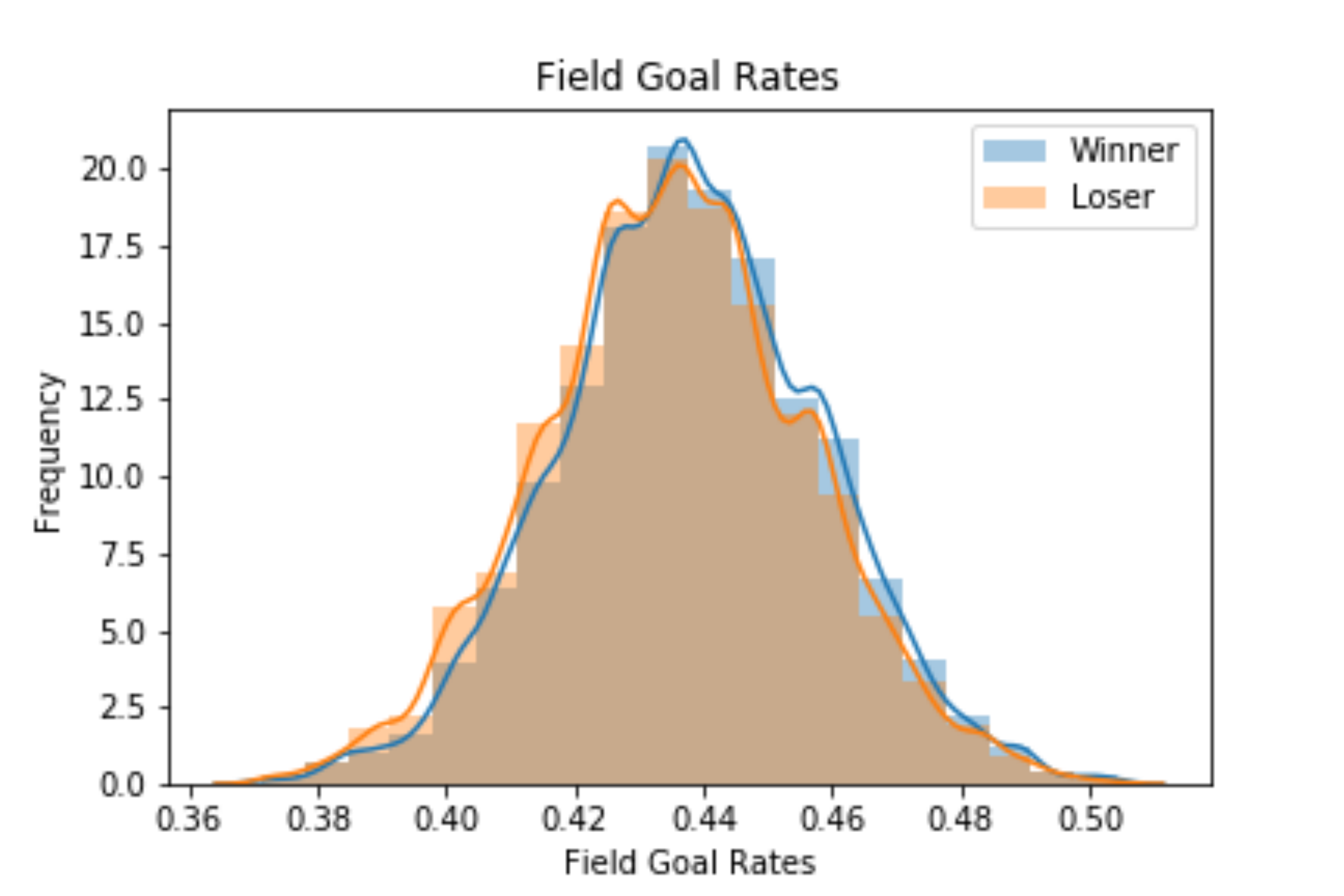
	FG_rate	FT_rate	FG3_rate	ORB	DRB	STL	BLK	AST	PF	Score_Diff	Opp_FG_rate	Opp_FT_rate	Opp_FG3_rate	Opp_ORB	Opp_DRB	Opp_STL	Opp_BLK	Opp_AST	Opp_PF	Opp_Score_Diff
FG_rate	1.00	0.19	0.48	-0.33	0.04	-0.10	-0.00	0.51	-0.08	0.05	0.10	0.08	0.08	-0.04	0.02	-0.04	-0.01	0.07	0.02	0.01
FT_rate	0.19	1.00	0.26	-0.28	-0.04	-0.14	-0.13	0.12	-0.02	-0.02	0.08	0.09	0.07	-0.07	0.03	-0.06	-0.02	0.05	0.01	-0.02
FG3_rate	0.48	0.26	1.00	-0.39	-0.07	-0.20	-0.19	0.29	-0.03	-0.01	0.08	0.07	0.08	-0.06	0.01	-0.04	-0.04	0.04	0.01	-0.03
ORB	-0.33	-0.28	-0.39	1.00	0.15	0.25	0.29	-0.01	0.21	0.07	-0.04	-0.07	-0.06	0.11	-0.04	0.08	0.07	0.01	-0.00	0.05
DRB	0.04	-0.04	-0.07	0.15	1.00	-0.34	0.31	0.15	0.04	0.02	0.02	0.03	0.01	-0.04	0.19	-0.06	0.04	0.05	0.05	0.02
STL	-0.10	-0.14	-0.20	0.25	-0.34	1.00	0.07	0.08	0.24	0.11	-0.04	-0.06	-0.04	0.08	-0.06	0.09	0.02	0.00	-0.01	0.02
BLK	-0.00	-0.13	-0.19	0.29	0.31	0.07	1.00	0.06	-0.01	0.08	-0.01	-0.02	-0.04	0.07	0.04	0.02	0.09	0.03	-0.02	0.05
AST	0.51	0.12	0.29	-0.01	0.15	0.08	0.06	1.00	-0.02	0.16	0.07	0.05	0.04	0.01	0.05	0.00	0.03	0.10	-0.00	0.04
PF	-0.08	-0.02	-0.03	0.21	0.04	0.24	-0.01	-0.02	1.00	-0.02	0.02	0.01	0.01	-0.00	0.05	-0.01	-0.02	-0.00	0.12	-0.02
Score_Diff	0.05	-0.02	-0.01	0.07	0.02	0.11	0.08	0.16	-0.02	1.00	0.01	-0.02	-0.03	0.05	0.02	0.02	0.05	0.04	-0.02	0.13
Opp_FG_rate	0.10	0.08	0.08	-0.04	0.02	-0.04	-0.01	0.07	0.02	0.01	1.00	0.19	0.48	-0.33	0.04	-0.10	-0.00	0.51	-0.09	0.05
Opp_FT_rate	0.08	0.09	0.07	-0.07	0.03	-0.06	-0.02	0.05	0.01	-0.02	0.19	1.00	0.26	-0.28	-0.04	-0.14	-0.13	0.12	-0.02	-0.02
Opp_FG3_rate	0.08	0.07	0.08	-0.06	0.01	-0.04	-0.04	0.04	0.01	-0.03	0.48	0.26	1.00	-0.39	-0.07	-0.20	-0.19	0.29	-0.03	-0.01
Opp_ORB	-0.04	-0.07	-0.06	0.11	-0.04	0.08	0.07	0.01	-0.00	0.05	-0.33	-0.28	-0.39	1.00	0.15	0.25	0.29	-0.01	0.21	0.07
Opp_DRB	0.02	0.03	0.01	-0.04	0.19	-0.06	0.04	0.05	0.05	0.02	0.04	-0.04	-0.07	0.15	1.00	-0.34	0.31	0.15	0.04	0.03
Opp_STL	-0.04	-0.06	-0.04	0.08	-0.06	0.09	0.02	0.00	-0.01	0.02	-0.10	-0.14	-0.20	0.25	-0.34	1.00	0.07	0.08	0.24	0.11
Opp_BLK	-0.01	-0.02	-0.04	0.07	0.04	0.02	0.09	0.03	-0.02	0.05	-0.00	-0.13	-0.19	0.29	0.31	0.07	1.00	0.06	-0.01	0.09
Opp_AST	0.07	0.05	0.04	0.01	0.05	0.00	0.03	0.10	-0.00	0.04	0.51	0.12	0.29	-0.01	0.15	0.08	0.06	1.00	-0.02	0.16
Opp_PF	0.02	0.01	0.01	-0.00	0.05	-0.01	-0.02	-0.00	0.12	-0.02	-0.09	-0.02	-0.03	0.21	0.04	0.24	-0.01	-0.02	1.00	-0.02
Opp_Score_Diff	0.01	-0.02	-0.03	0.05	0.02	0.02	0.05	0.04	-0.02	0.13	0.05	-0.02	-0.01	0.07	0.03	0.11	0.09	0.16	-0.02	1.00

Correlation between features





Histogram of field goal rates for winning team versus losing team.



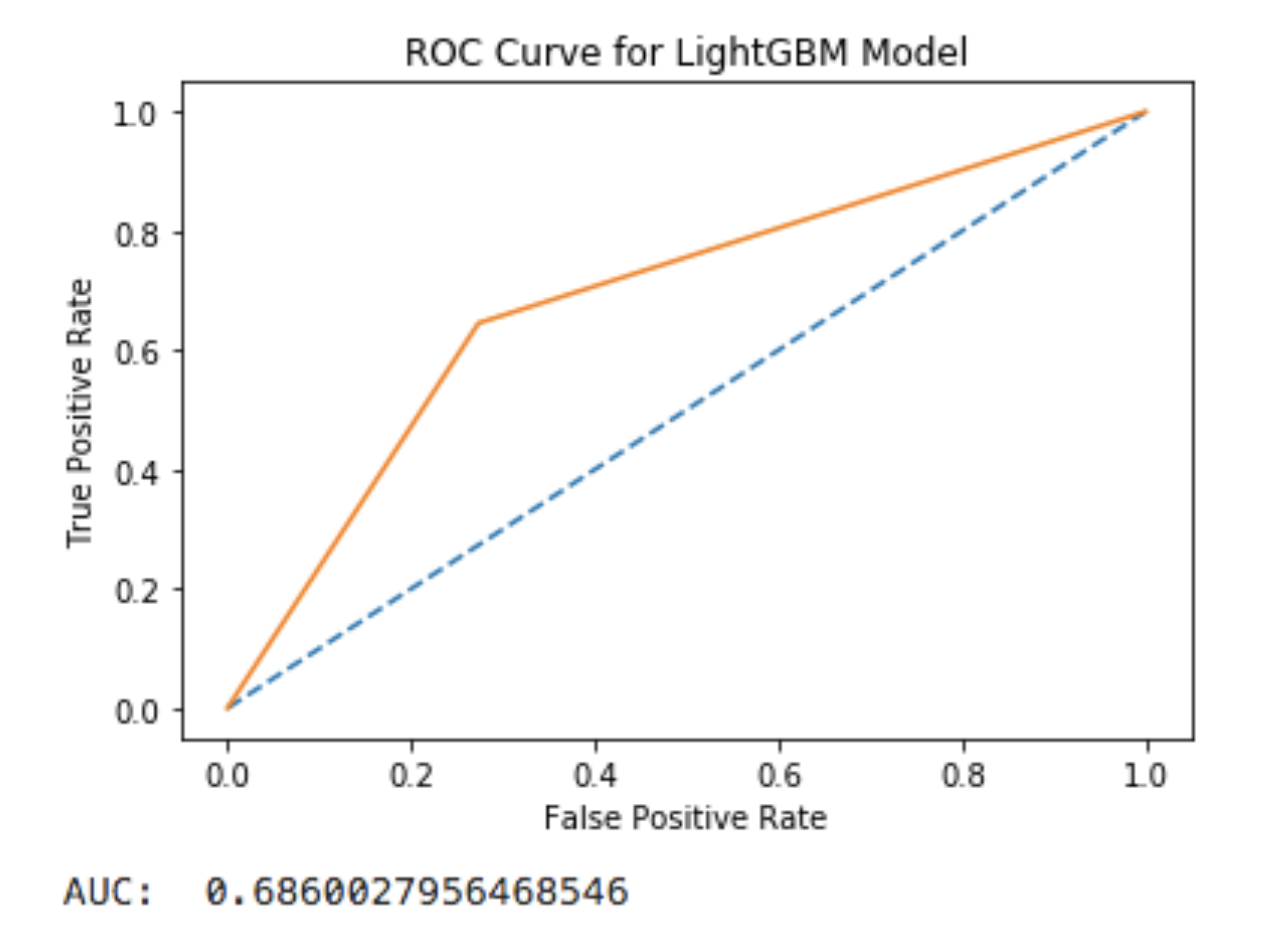
MACHINE LEARNING

- LightGBM (can handle missing data)
- Gradient boosting decision tree
- Binary classification
- Categorical features: team id, season, played in previous tournament (binary)
- Scikit-learn (to split into train and test sets)

FITTING THE MODEL

- Scikit-learn Randomized Search CV
- Randomized Search to find best values for the hyperparameters
- Used these parameters to train the model
- Converted probabilities to binary values (to compute accuracy)

Measuring the performance of the model



```
AUC:  0.6860027956468546
```

AUC: The model is about 68.6% likely to correctly predict the outcome of a game.

CONCLUSION

- We have discovered which features are useful in predicting the outcome of a game
- We can use data from the previous two years to predict whether a team will win or lose.
- We have a model that is about 68.6% accurate
- Knowing which features are useful for predicting wins is useful for those attempting to predict the outcome of the NCAA tournament

FUTURE WORK

- Create additional features
- Use individual player data (may be difficult to access)
- Model could be used for similar problems: predicting wins for other sports