NBA Betting Prediction

Raj Gosain, Angela Hu, Alan Zheng, Kuleen Sasse

Introduction

- Booming multibillion sports betting industry taking off
- Predicting the outcome of a game gives an edge to bettors
- Use Machine Learning for this task

Why US sports betting could become a \$45 billion business. (2024). Goldmansachs.com. https://www.goldmansachs.com/insights/articles/why-us-sports-betting-could-become-a-45-billion}

Betting Odds

Basketball / USA / NBA 2023/2024

17 Jun 2024 - Play Offs				1	2	B's
20:30	**Boston Celtics**	106 – 88	Dallas Mavericks	-270	+222	8
14 Jun 2024 - Play Offs				1	2	B's
20:30	**Dallas Mavericks**	122 – 84	Boston Celtics	-103	-115	8
12 Jun 2024 - Play Offs				1	2	B's
20:30	Dallas Mavericks	99 – 106	**Boston Celtics**	-147	+124	8
09 Jun 2024 - Play Offs				1	2	B's
20:00	**Boston Celtics**	105 – 98	Dallas Mavericks	-263	+214	8
06 Jun 2024 - Play Offs				1	2	B's

Basketball / USA / NBA 2022/2023

12 Jun 2023 - Play Offs				1	2	B's
20:30	**Denver Nuggets**	94 – 89	Miami Heat	-333	+270	7
09 Jun 2023 - Play Offs				1	2	B's
20:30	Miami Heat	95 – 108	**Denver Nuggets**	+124	-145	7
07 Jun 2023 - Play Offs				1	2	B's
20:30	Miami Heat	94 – 109	**Denver Nuggets**	+141	-161	7
04 Jun 2023 - Play Offs				1	2	B's
20:00	Denver Nuggets	108 – 111	**Miami Heat**	-333	+273	7
01 Jun 2023 - Play Offs				1	2	B's

Data Preprocessing

- Combined two datasets
  - Betting dataset (odds from each game)
  - Basketball stats per game
    - Individual Player (average and max over team)
    - Overall team stats
- Ended up with all games from 2016-2022
- Around 6539 games

Data Preprocessing (contd.)

- Problem
  - Temporal aspect of data: stats from end of game
- Solution
  - Running average up to that but not including that game

Training

- Supervised Binary Classification Task
- Used 2016-2021 as training/validation data
- Five models tested:
  - Logistic Regression
  - Decision Tree
  - Random Forest
  - SVM
  - MLP

Training (contd.)

- Two step pipeline
  - Feature selection
    - 136 features in the dataset
    - Too many features to efficiently run models on our hardware
    - Used cross validated forward feature selection
  - Hyperparameter optimization
    - Further fine tune the performance of our models
    - Use cross validated grid search across a couple select parameters
    - Take the parameters with best average accuracy across all folds

Evaluation

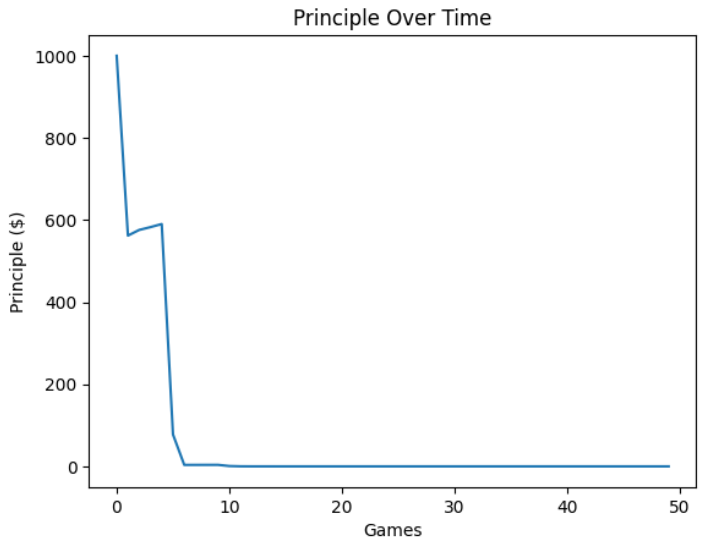
- Measured the out of sample performance in two ways
- Used games from 2022 as test set
- Accuracy
- Backtesting system
  - Give the model 1000$ to start
  - Model creates a bet based on predicted probabilities and external betting data

Results

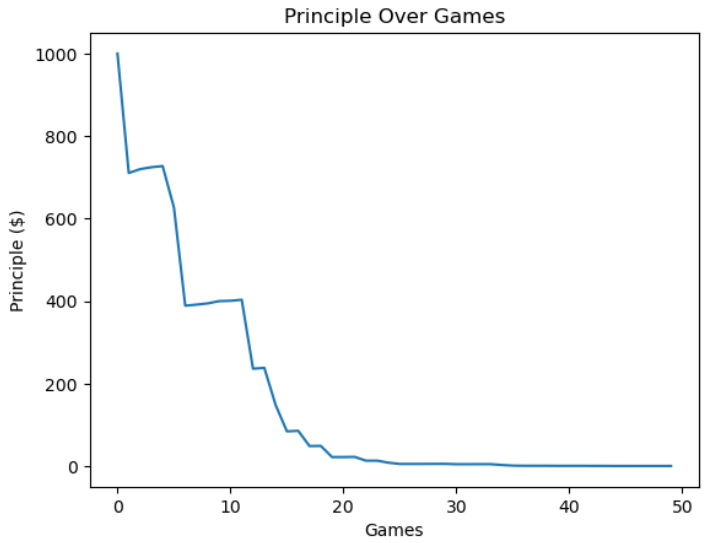
- No features were selected by all 5 models; max plus-minus, opponent max offensive rating, usage percentage selected by 4
- Random Forest and SVM took the longest to run out of money, ≈20 games
- Accuracy not correlated with betting success

Model Type	Accuracy
Baseline (Home Team Win%)	57%
Logistic Regression	58%
Random Forest	56%
Decision Tree	60%
Support Vector Machine	59%
Multi-Layer Perceptron	61%

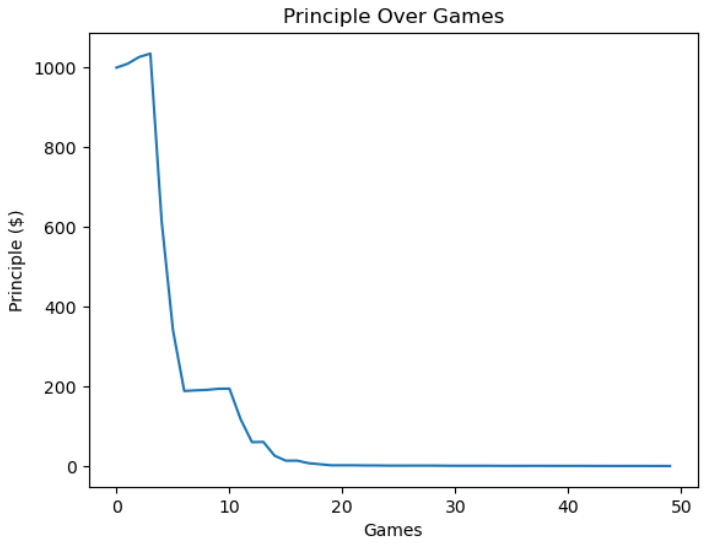
Table 1: Model Accuracies on Test Data



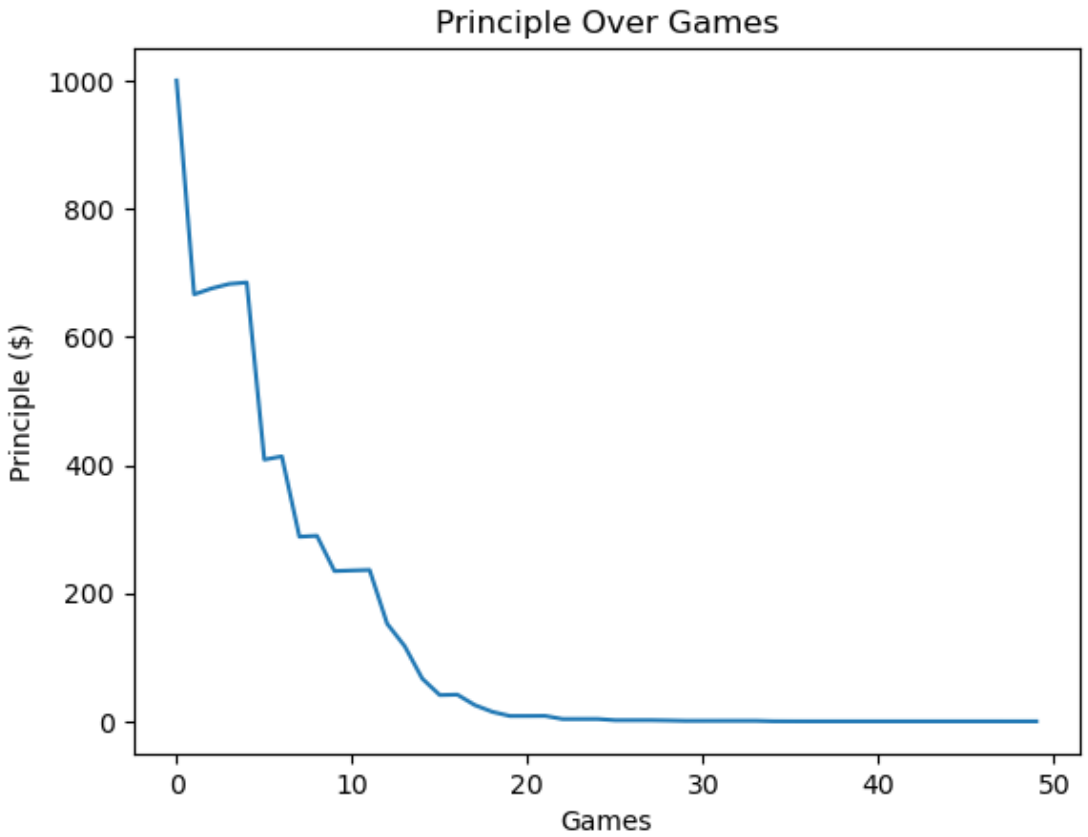
Logistic Regression



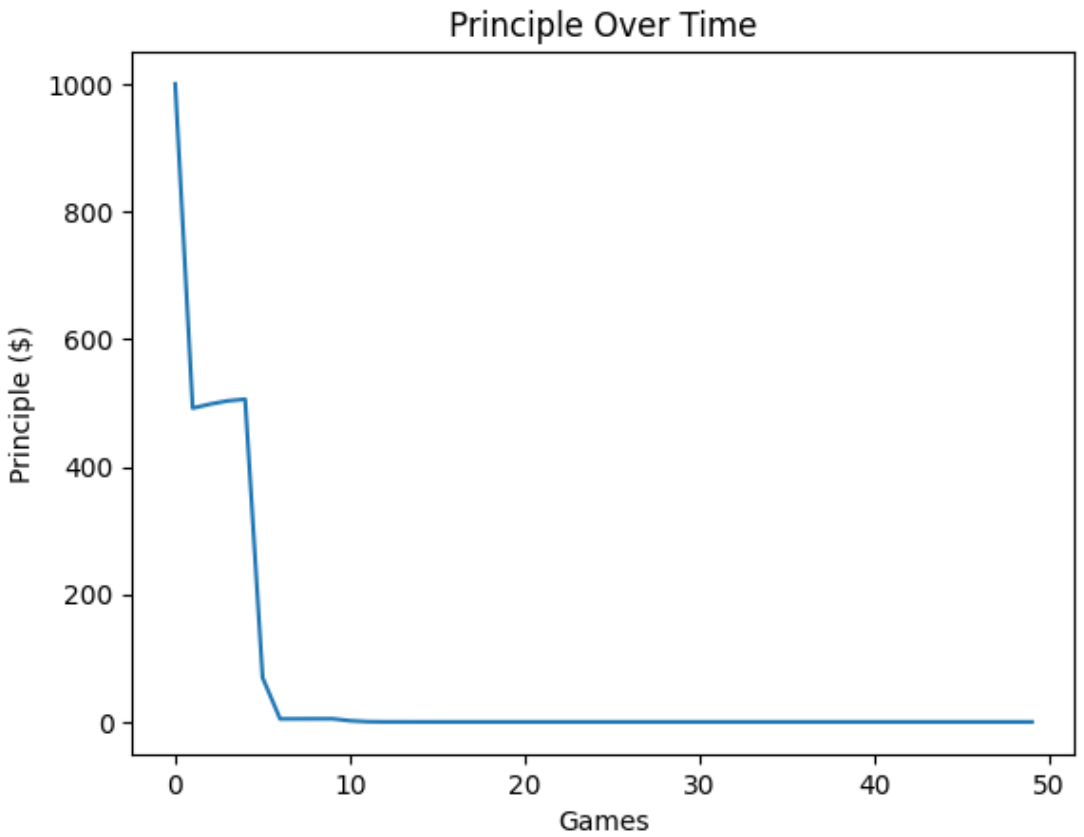
Random Forest



Decision Tree



Support Vector Machine



Neural Network

Future Work

- **Scrape additional data**
- **Incorporate ELO or ranking data** to capture the temporal dynamics of rankings
- **Model the impact of individual players**, considering factors like trades and injuries
- **Integrate detailed shooting statistics**, such as floor location data, to refine player-specific performance metrics
- **Leverage sports media sources** through NLP and sentiment analysis, including pre-game predictions and expert commentary
- **Explore more powerful models**, such as deeper neural networks