NBA Sports Betting Prediction Model

Gio Romero-Ruiz, Rahul Aneja, Anton Bilonog

Problem Definition

- We used box scores from the NBA 2012-2024 seasons to predict if a team will win or lose a game versus an opponent.
- We drew inspiration from a previous project which was done last year and we wanted to see whether their models to predict wins/losses were valid for the 2023-2024 season.
- Dive deep into the specific machine learning models and techniques used
- The data preprocessing and feature selection process

Related Works

- The Sports Betting tool that we took inspiration from used models such as Random Forest, Neural Network, Logistic Regression, Linear Regression,
- In "A Data-Driven Machine Learning Algorithm for Predicting the Outcomes of NBA Games"
  - It introduces a new team efficiency index, derived from player efficiency metrics and comparing performance against rivals.
- In "Sports Data Mining Technology Used in Basketball Outcome Prediction"
  - Out of Simple Logistics Classifier, Artificial Neural Networks, SVM, and Naïve Bayes, their best-performing model was the Simple Logistics Classifier, achieving an accuracy of 69.67%

How We Process The information

1. Scraped all the data from Basketball-reference.com
2. Pre-process the data to do time-series predictions
3. Created rolling averages over the last 10 games for both teams

Types of Training Models We Use

- Logistic Regression
  - Classic Classification model
  - *Sports data mining technology used in basketball outcome prediction by Chenjie Cao*
    - Best for binary outcomes (Wins and Losses)
- Neural Network
  - Great for recognizing complex patterns and scaling with large datasets
- Ridge Regression
  - Introduces a regularization term (L2 penalty) to the loss function to prevent overfitting
  - Useful when data suffers from multicollinearity
- XGBoost
  - Uses gradient boosted decision trees, designed for speed and performance
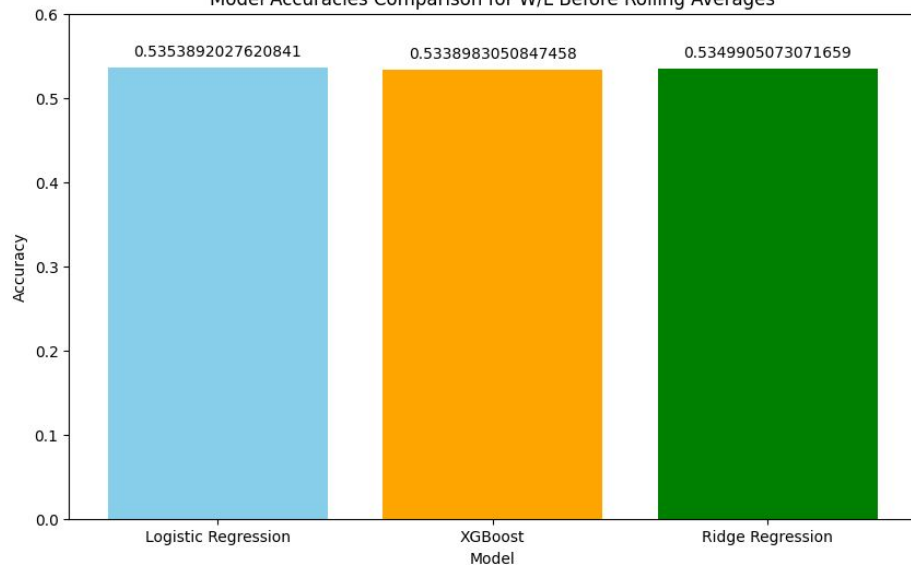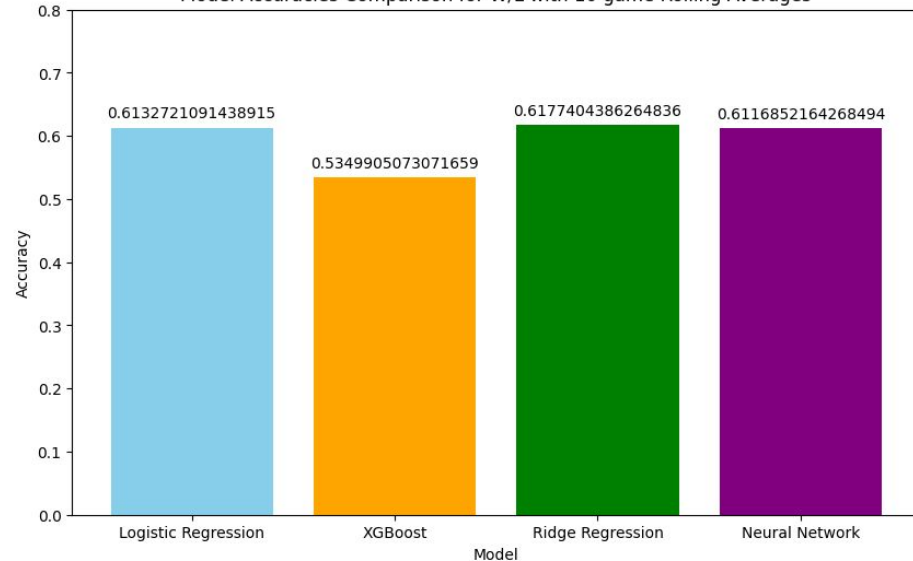
Results
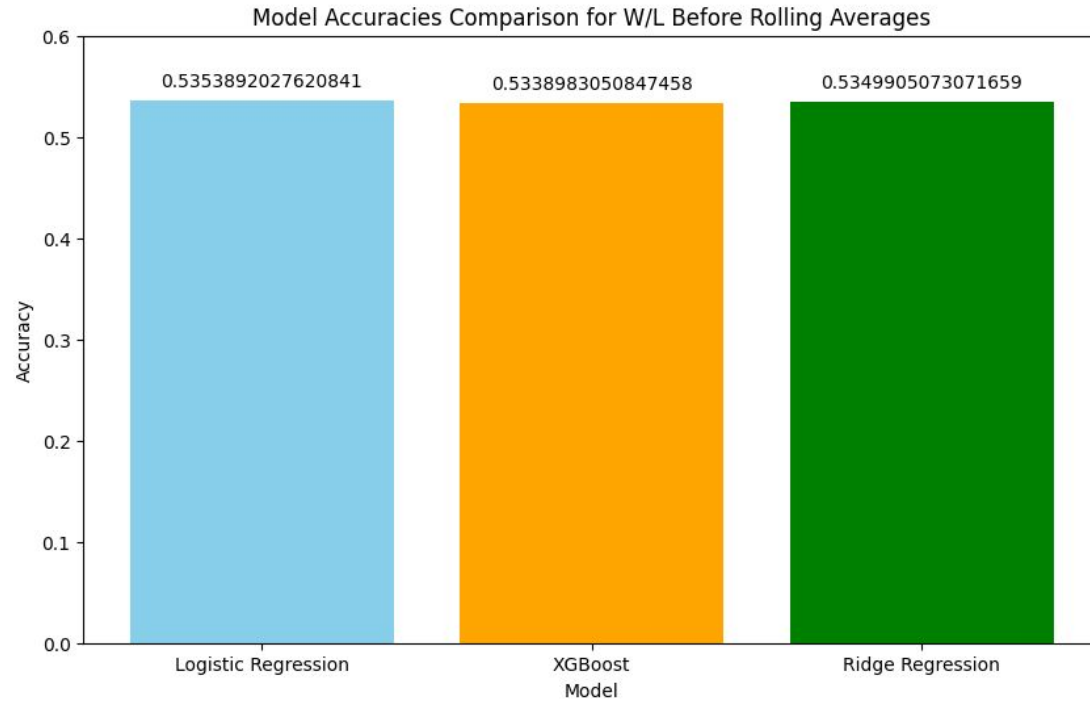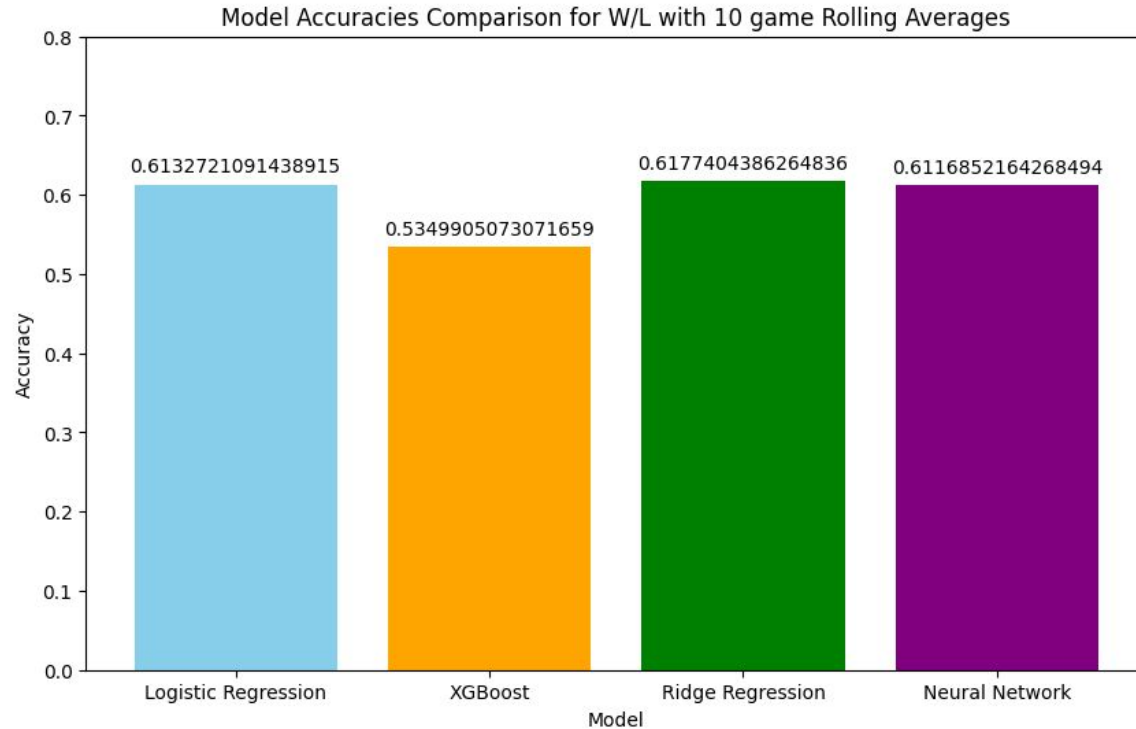
Model Accuracies Comparison for W/L Before Rolling Averages

Results Continued…

Model Accuracies Comparison for W/L with 10 game Rolling Averages

Side-by-side Comparison

Analysis

- Ridge Regression give us the most accurate results
- XGBoosting gave us the least accurate results
  - Using rolling averages did not have any effect on the accuracy of the model
- Implementing a 10 game rolling average for Logistic Regression, Ridge Regression made a big difference in our resulting accuracies

Analysis Continued

- Features used for Ridge Regression before using a 10 day rolling average prediction strategy
- Features used for Ridge Regression with 10 day rolling averages

```
['fta_rolling_x',
 'orb_opp_rolling_x',
 'trb_opp_rolling_x',
 'blk_opp_rolling_x',
 'won_rolling_x',
 'ft_rolling_y',
 'fga_opp_rolling_y',
 'orb_opp_rolling_y',
 'blk_opp_rolling_y',
 'won_rolling_y']
```

```
['fga',
 'ft',
 'fta',
 'pf',
 'fg%_opp',
 '3p%_opp',
 'orb_opp',
 'stl_opp',
 'blk_opp',
 'pf_opp']
```

Challenges

- Our models do not take into account of the individual players on each team.
  - Things not taken into account: injuries, trading, player efficiency, team chemistry, coaching staff, home court advantage
- Scraping information in the beginning was tedious
- Cleaning data to suit our needs was difficult
  - Finding the correct features to help predict our games
- We had to update the data collected for 2023-2024 box scores because it was an ongoing season

Conclusion and Future Work

- Incorporating more data
  - Player box scores and injuries
- Using sentiment analysis
  - Using tweets to gauge perception on the games' outcome
- Use sportsbook odds to find good betting opportunities
- Build a script that updates the models to use updated data during the season
- Adjust the models to factor in playoff games, which requires using a distinct set of features, as playoff games have unique characteristics compared to regular season games.

References

- Chenjie Cao. Sports data mining technology used in basketball outcome prediction prediction. September 2012.

- Robert Logozar ˇCaslav Livada Tomislav Horvat, Josip Job. A data-driven machine learning algorithm for predicting the outcomes of nba games. March 2023

- Project we drew inspiration from
    - https://tmwulff.github.io/#/Project