



Machine Learning in Sports Betting Final Presentation

Austin Fett, Davin Stephens, Tyler Wulff, Ron Yue



Problem Definition

- Due to its wide availability of data, we chose the NBA
 - 82 regular seasons games a year (not including covid year)
 - Limited number of players limits a little bit of the variation
 - Many datasets are available with in depth advanced metrics
- Initially want to predict wins and losses before looking at spreads
 - Want to see an accuracy that is significantly better than flipping of a coin or home/away win percentage
 - Over the last 10 or so years, win % of home teams is around 58%, but only about 55% for the 2023 season
- We want to see if we can create a model to meet or exceed an accuracy of 53% for spreads
 - We want to see success over past and current seasons
 - Will track our success using a unit size of 10 dollars
- Want to combine sentiment analysis with traditional box score analysis to predict outcomes



Datasets

OffPoss	Opp_OffPoss	Points	Opp_Points	FG2M	Opp_FG2M
106.0	106.0	117.0	107.0	38.0	33.0
104.0	103.0	112.5	102.5	32.5	29.0
103.0	102.333333	111.333333	110.333333	32.0	31.333333

- Focusing on PBP Stats, Basketball Reference, and ESPN to pull our data
 - Box scores/statistical information from PBP/Basketball Reference and lines/home team data from ESPN
- A kaggle dataset was combined with ESPN data to grab lines
 - ESPN only had lines from 2021 on, while the Kaggle dataset has data from offshore sites up until 2018
 - Removed the Covid season to account for missing lines and variance from the season as a whole
- Calculated statistics pertaining to last five game averages and full season averages
 - Used to help see how a team is currently playing and how they have played over the year
- Information such as Home/Away, Win/Loss and Cover was converted to categorical data so it could be used
- Data needed to be normalized
- NLP parses the data, tags parts of speech, utilizes Name-entity recognition and adapts TF_IDF
 - Goal of finding correlation between sentiment and game results
 - Use of discord to gather sentiment for teams

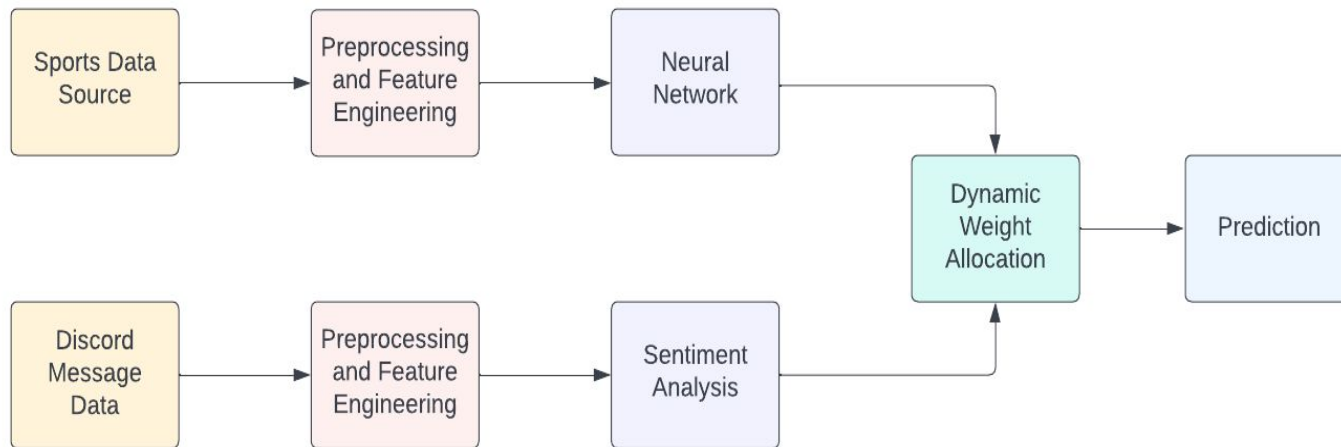


Feature Engineering

- Converted spread results, home/away, and win/loss into categorical data
- Added extra features such as Assist to turnover ratio to try to add more advanced metrics
- Win/loss percentage over the last five games was calculated to see how a team is playing at the moment
- 5 day, 10 day, and season averages were calculated to help account for changes for a team throughout a season and over seasons
 - The 5 day averages showed to be the most effective when predicting wins and losses and spread results

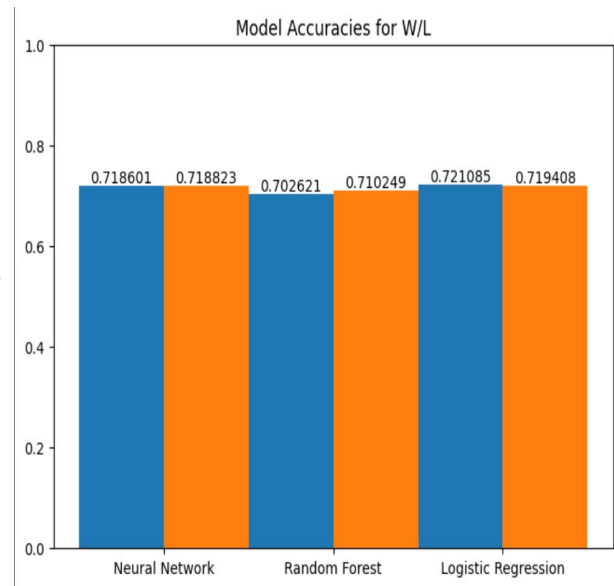


Pipeline



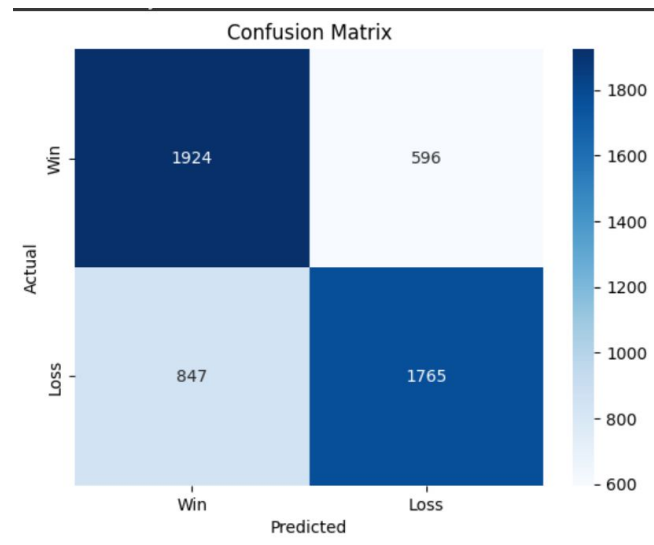
Final Models (Traditional Box Score)

- Looked at both classification and regression when approaching our problem
 - Classification was far more effective, with regression staying at an RMSE of around 11 and a spread result prediction accuracy hovering around 50%
- Trained Neural Network, Logistic Regression, Random Forest, Linear Regression, and XGBoost
- Used GridSearchCV to tune hyperparameters
 - Improved accuracy for logistic regression spread prediction and wins and losses around 3%
- Neural Network consists of 3 dense layers
 - Adam optimizer and binary cross entropy loss



Final Models (Traditional Box Score)

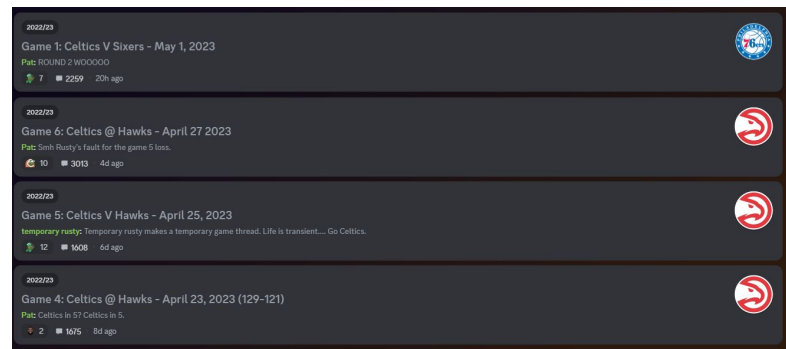
- Saw the most success with our Neural Network for wins and losses, spread results, and O/U results
 - ~72% for W/L, ~62% for spread results, ~60% for O/U results
- Very similar success for the 2023 season compared to previous seasons
- Some of the most influential features consisted of assists, defensive rebounds, and field goal percentage defense



Sentiment Analysis

- Data was scraped from discord servers
 - Most servers organized game threads
- There is a lot of data but a lot of it is useless
 - Getting rid of emojis, 1-2 word messages, links, @users
- Between 40-60% of messages were filtered out
- Smaller market teams had few messages
 - Celtics yielded 2000-3000 messages
 - Kings had 20-100 messages, mostly off topic

7070348095617761	WestCoastFUNko831	04/27/2023 12:12 AM	:CryingLaugh:	
6580567191126016	tea sea #5094	04/27/2023 12:12 AM	OH	
6580567191126016	tea sea #5094	04/27/2023 12:12 AM	OH MY	
7246952059572061	DefinitelyNotYosh.#3	04/27/2023 12:12 AM	OH HHH	
7070348095617761	WestCoastFUNko831	04/27/2023 12:12 AM	Omg	





Final Model (Sentiment Analysis) $\frac{|team1polarity| + |team2polarity|}{2}$

- We output a decimal point number to represent polarity
 - This number goes from -1.0 to 1.0, where 0 represents fairly neutral sentiment
 - Scores are inverted based on team viewpoints
 - We then average them out
- This number is then compared to the EV of a given line to see if there are any profitable lines
 - If positive odds, the formula is $EV = 1/(1 + x/100)$
 - If negative odds, the formula is $EV = 1/(1 + 100/|x|)$
 - So if EV is 55% team A and we have score of 0.85, betting on team A would be +EV
 - Model on its own loses \$0.08 per \$1 wagered over 433 games

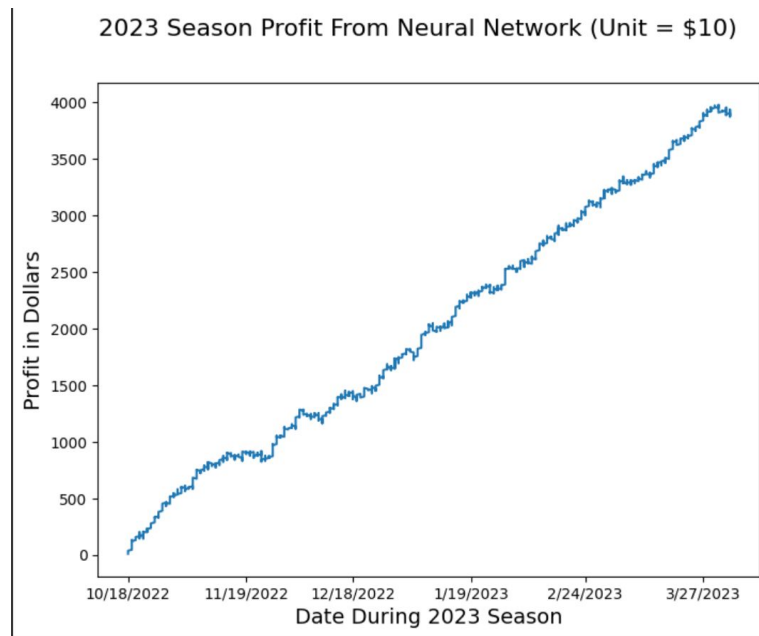


Combining models

- Since our traditional box score model was already very good at predicting spreads, we decided to simply add the sentiment analysis output as a feature for the model
 - This should help the model in predicting games where certain star players are not playing, showing how most fans believe the other players will step up into their new roles
- Sentiment Analysis is very different from our neural network working with more traditional player data.
 - Experimenting with ensemble learning but currently the model outputs are different

Final Model Results

- The results of the neural network model covering spreads for the 2023 season profited almost \$4000 using only \$10 bets on every game
- Some of the downturns in profit, such as near the end of the season, can be attributed to star players sitting out games ahead of the playoffs
- Possible quick fix for this could be to stay away from games where top player(s) are injured or resting





Website



Machine Learning in Sports Betting (NBA)

Live Games

Nuggets 33 @ Timberwolves 36

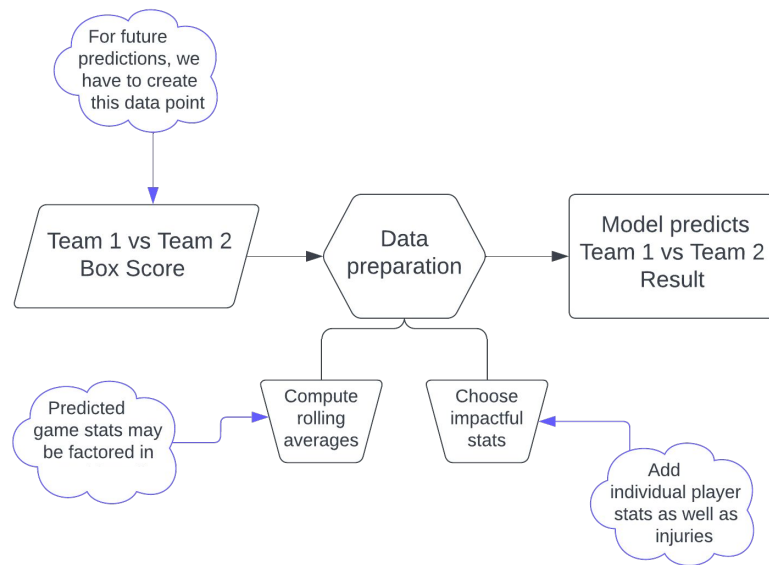
Cavaliers 93 @ Knicks 102

Kings 125 @ Warriors 126

Celtics 129 @ Hawks 121

Future Work

- Continue to fine tune the model to adjust for changes as the season moves along
 - Factor in home team vs away team history
 - Include individual player data to account for injuries and rest days
- Add capability for the model to automatically predict games yet to be played
 - As of now, can only be done manually
- Adjust model for playoffs
 - 7 game series where teams play each other back to back will result in different outcomes from regular season games





Considerations

- Changing this model to make future predictions would decrease accuracy (~5%)
 - Our win percent feature includes current game in average
- Met our final goal, which was to make a profit betting on the 2022-23 NBA season.
- We intend this model to be one in many factors for deciding on a bet.
- As models become more advanced, sports betting industry has potential to change.



Contributions

- Ron - Sentiment analysis
- Austin - Data scraping and preprocessing
- Tyler - Model/Neural network tuning and website building
- Davin - Neural network building and feature extraction



References

- [1] Tan, R. J. (2022, March 2). *Breaking down mean average precision (MAP)*. Medium. Retrieved March 13, 2023, from <https://towardsdatascience.com/breaking-down-mean-average-precision-map-ae462f623a52>
- [2] Mishra, A. (2020, May 28). *Metrics to evaluate your machine learning algorithm*. Medium. Retrieved March 13, 2023, from <https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234>
- [3] Bruce, A. (n.d.). *What percentage of sports bettors win?* Sports Betting News. Retrieved February 1, 2023, from <https://sitpicks.com/what-percentage-of-sports-bettors-win/#:~:text=Different%20studies%20spit%20out%20varying,system%20that%20works%20for%20them>
- [4] Culver, J. V. (2021, March 7). *Why 52.4% is the most important percentage in sports gambling*. Medium. Retrieved February 1, 2023, from <https://medium.com/the-intelligent-sports-wagerer/why-52-4-is-the-most-important-percentage-in-sports-gambling-16ade8003c04>
- [5] Dotan, G (2020). *Beating the Book: A Machine Learning Approach to Identifying an Edge in NBA Betting Markets*. Retrieved February 3, 2023, from <https://escholarship.org/uc/item/115957mb>