The Bank Is Open: AI In NBA Betting

Vishnu Sarukkai (sarukkai@stanford.edu) Alexandre Bucquet (bucqueta@stanford.edu)

Motivation and Problem Statement

- On May 14, the Supreme Court legalized sports betting, paving the way to a new market worth an estimated $150 billion.
- In this project, we attempt to apply Machine Learning algorithms to predict the outcomes of certain betting indicators in the NBA, such as the Money-Line, the Point Spread, or the Over Under.
- More specifically, we focus on estimating the number of points scored by both teams in every NBA.
- While we were able to closely approximate the number of points scored in a game, our estimates were not precise enough to allow us to "beat the house" on the long run.

Datasets

- The datasets acquired provide two forms of data: betting odds data, which informs us of the bets offered by sportsbooks, and game data, which gives us data summarizing NBA games.
  - Betting odds data: Sports Book Review Online offers betting odds for every NBA game since October 2007.
  - Game data: Basketball Reference provides game-by-game team- and player-level data, which we retrieved using Frank Goitia's NBA crawler for every season since 2007-2008.

Features

- For every game in our dataset, we extracted the following information:
  - Statistics for both teams' past three games. This includes simple statistics such as Points Scored or Total Rebounds, but also more complex features like Offensive Rating or Plus/Minus.
  - Season averages for both teams' respective opponents in the past three game in the same categories.
  - Number of days since the last game for both teams.
  - Distance traveled by both teams.

Models

Loss: MSE

Collaborative Filtering:

- Build a sparse matrix A containing past games.
- We factor A into $U\Sigma V^T$ using

$$min_{U,V,\Sigma} \sum_{ij} (A_{ij} - U\Sigma V^T{}_{ij})^2$$

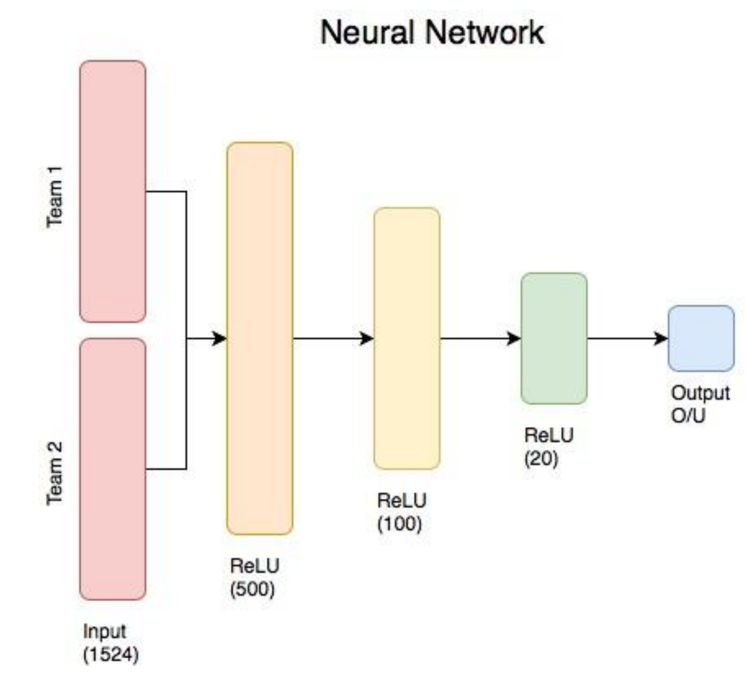
- Predict $u_i \Sigma v_j^T$.

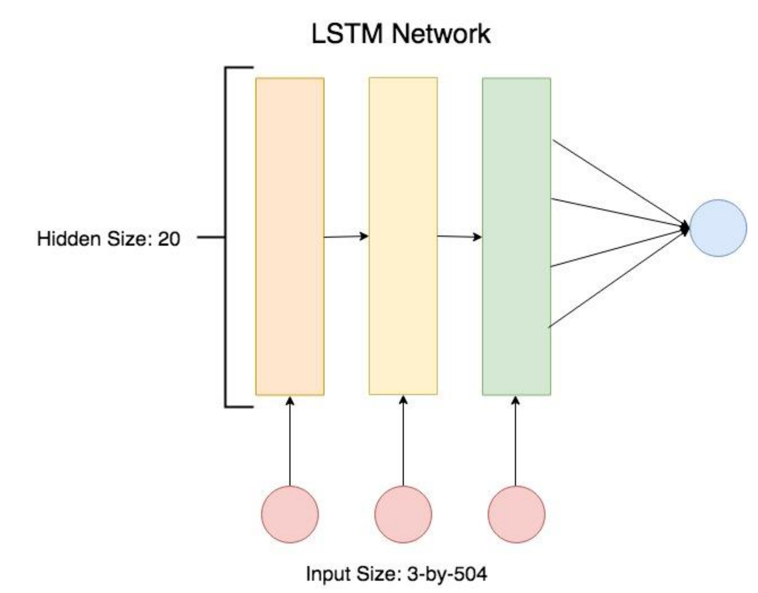
Neural Network:

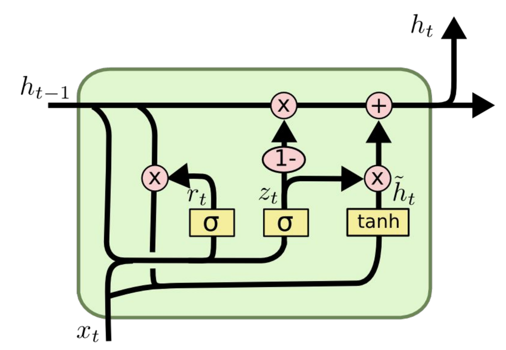
- Inputs: 762 features for each team.
- Architecture: three hidden layers (size 500, 100, 20) with ReLU activations.
- Output: predicted O/U.

LSTM Network

- Process the past three games sequentially.
- We add a fully connected layer to output the points scored.

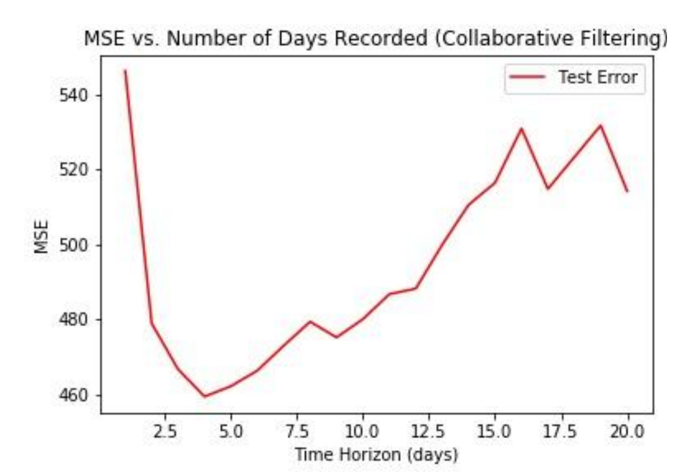


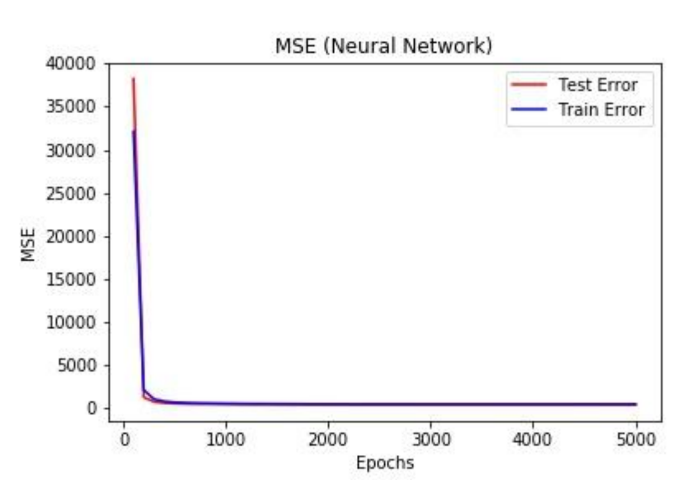




$$z_t = \sigma\left(W_z \cdot [h_{t-1}, x_t]\right)$$
$$r_t = \sigma\left(W_r \cdot [h_{t-1}, x_t]\right)$$
$$\tilde{h}_t = \tanh\left(W \cdot [r_t * h_{t-1}, x_t]\right)$$
$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t$$

Results





Model	Train MSE	Test MSE
Random Forest (baseline)	955.10	910.33
Collaborative filtering	2.95	459.85
Neural Network	349.17	369.84
LSTM Network	398.95	426.56

Table 1. Train and Test Errors for Various Models.

Architectures:

- Trained the Collaborative Filtering model with a hidden size of 10 for 1000 epochs.
- Trained the Neural Network for 5000 epochs with a learning rate of 10^-6 and weight decay of 1.
- Trained the LSTM model with a hidden size of 20 with dropout of 0.2 and SGD with learning rate 0.05.

Discussion

Discussion:

- Due to the rapidly changing nature of the NBA it is difficult to acquire sufficient training data that reflects the way the game is currently played.
- Note that the Collaborative Filtering model, which didn't use any team features, outperformed the Random Forest.
  - This shows the high variance in our data as well as the strong seasonal trends that a model needs to encompass in order to be accurate on this task.
- We achieved a test Mean Squared Error of 369.84 for our best model.
  - Encourages future work to be done on feature selection and engineering;
  - Current best models can beat the house around 51.5% of the time, but successful long-term betting patterns need to be correct at least 52-53% of the time.

Future Work

- Augment the dataset:
  - Incorporate the odds lines offered by various sportsbooks as features in our models (perhaps we can learn trends such as that the books tend to overestimate the performance of certain teams)
  - Incorporate player-level data, not just team-level data (should help account for when we know a player is injured before a game starts)
- Explore model architectures further:
  - Design novel neural network/LSTM architectures to take the additional features mentioned as input and train over longer sequences of games
  - Build architectures more similar to state-of-the-art rating prediction models (map user/item relationship to team/team relationship)

References

- Fran Goitia, Basketball Reference Scraper, (2017), GitHub repository. https://github.com/FranGoitia/basketball_reference
- *Historical NBA Scores and Odds Archives*. [Online]. Available: https://www.sportsbookreviewsonline.com/scoresoddsarchives/nba/nbaoddsarchives.htm
- Hochreiter, Sepp & Schmidhuber, Jurgen (1997). Long Short-Term Memory. *Neural Comput., 9*, 1735-1780.