A Dive into the World of Sports Betting

Author: Harrison Gu

Industry Overview

- Supreme Court struck down federal ban in 2018
- ~$150 billion handled in legal sports betting in 2020

Bottlenecks

1. 27 out of 50 states legalized
2. Many hoops to jump through
3. Negative connotation

Catalysts

1. More states are legalizing
    a. 89.9% CAGR since ban lift
2. As laws loosen, marketing will make betting through official books trendy

- Projected $250 billion handled in 2024

source: https://www.legalsportsbetting.com/how-much-money-do-americans-bet-on-sports/

Business Model

- Use machine learning and deep learning to predict the spread (difference in points) of NBA games
- Provide winning picks to our subscribers
- Vegas house edge for spread bets is typically 10%
  - No matter which side you bet, you are risking 1 unit to with 0.9 units
  - In order to offset the house edge, bettors need to win 52.4% of their bets

Goal: Create a model that can beat the Vegas spread for NBA games at least 52.5% of the time

Expected Value = P(win)*0.9 - P(lose)

The Data

Inputs

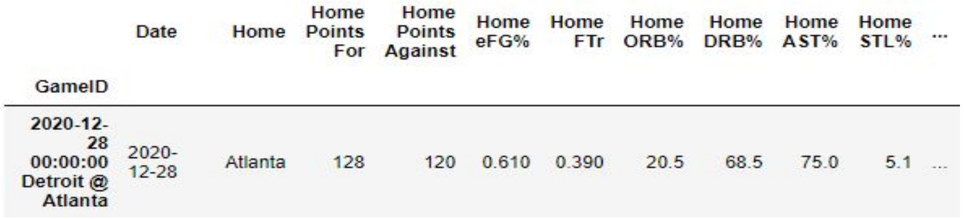
- Average basic stats for each team going into the game
    - Points for (PF)
    - Points allowed (PA)
- Average advanced stats
    - Effective field goal % (eFG%)
    - Offensive rating (Ortg)
    - Defensive rating (Drtg)
- Home and away records for each team
- Vegas spread
    - Serves as an input for exogenous factors

Output

- How many points will the home team win by?

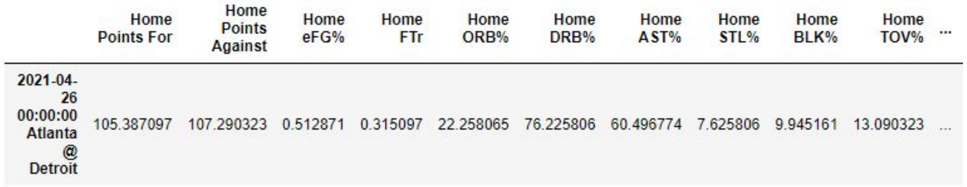
The Data

- The raw data for this model was scraped from
  - https://www.basketball-reference.com/ (stats by game)
  - https://www.sportsbookreviewsonline.com/ (odds by game)
  - 11,656 NBA regular season games from 2011-2021



	Date	Home	Home Points For	Home Points Against	Home eFG%	Home FTr	Home ORB%	Home DRB%	Home AST%	Home STL%	...
GameID											
2020-12-28 00:00:00 Detroit @ Atlanta	2020-12-28	Atlanta	128	120	0.610	0.390	20.5	68.5	75.0	5.1	...

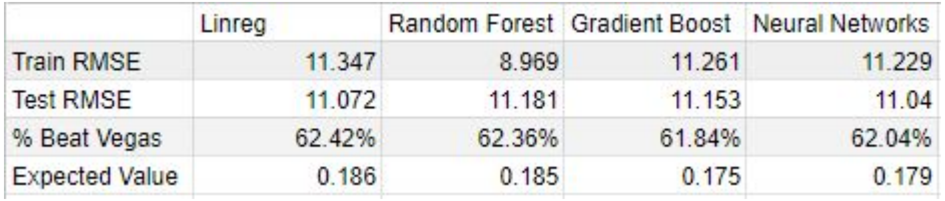
- Data used for model



	Home Points For	Home Points Against	Home eFG%	Home FTr	Home ORB%	Home DRB%	Home AST%	Home STL%	Home BLK%	Home TOV%	...
2021-04-26 00:00:00 Atlanta @ Detroit	105.387097	107.290323	0.512871	0.315097	22.258065	76.225806	60.496774	7.625806	9.945161	13.090323	...

Individual Models

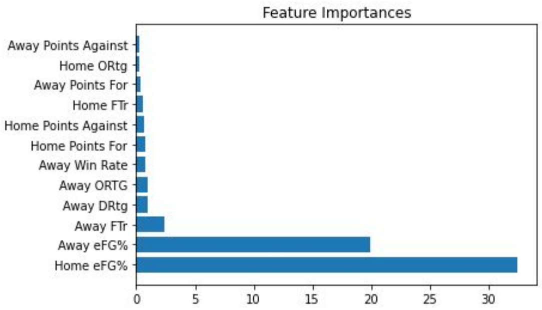
- Ran various regression models and chose the 3 best based on RMSE to tune
  - Linear regression
  - Random Forest regression
  - Gradient Boost regression
- Ran baseline neural networks model, and tuned using talos



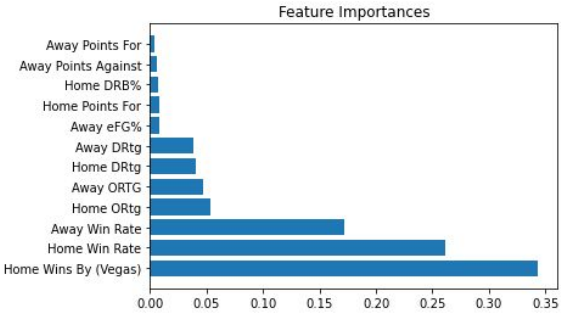
	Linreg	Random Forest	Gradient Boost	Neural Networks
Train RMSE	11.347	8.969	11.261	11.229
Test RMSE	11.072	11.181	11.153	11.04
% Beat Vegas	62.42%	62.36%	61.84%	62.04%
Expected Value	0.186	0.185	0.175	0.179

Feature Importances

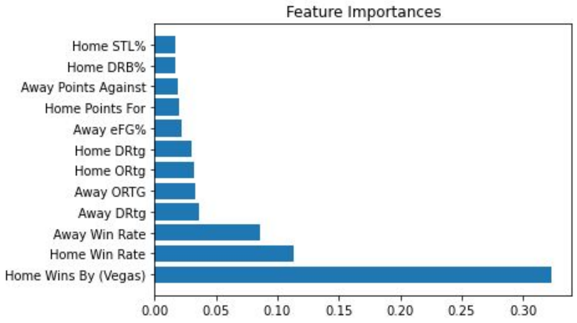
Linear Regression

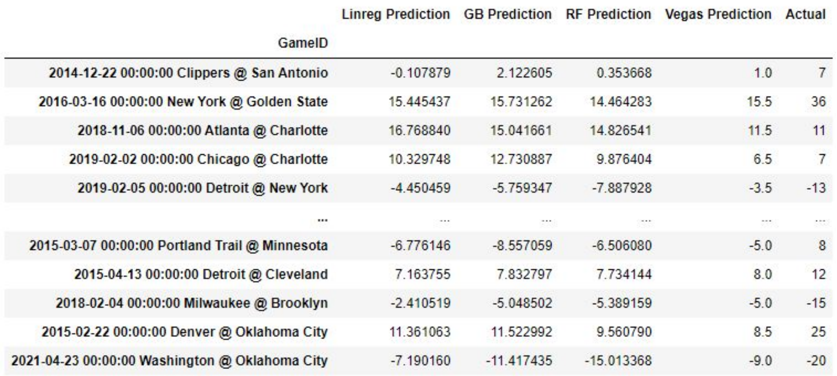


Gradient Boost



Random Forest

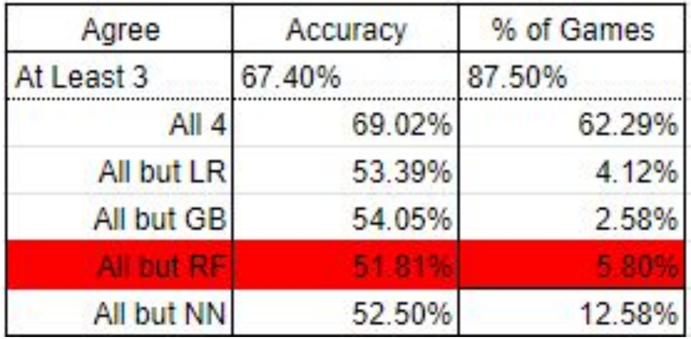




	Linreg Prediction	GB Prediction	RF Prediction	Vegas Prediction	Actual
GameID					
2014-12-22 00:00:00 Clippers @ San Antonio	-0.107879	2.122605	0.353668	1.0	7
2016-03-16 00:00:00 New York @ Golden State	15.445437	15.731262	14.464283	15.5	36
2018-11-06 00:00:00 Atlanta @ Charlotte	16.768840	15.041661	14.826541	11.5	11
2019-02-02 00:00:00 Chicago @ Charlotte	10.329748	12.730887	9.876404	6.5	7
2019-02-05 00:00:00 Detroit @ New York	-4.450459	-5.759347	-7.887928	-3.5	-13
...	...	...	...	...	...
2015-03-07 00:00:00 Portland Trail @ Minnesota	-6.776146	-8.557059	-6.506080	-5.0	8
2015-04-13 00:00:00 Detroit @ Cleveland	7.163755	7.832797	7.734144	8.0	12
2018-02-04 00:00:00 Milwaukee @ Brooklyn	-2.410519	-5.048502	-5.389159	-5.0	-15
2015-02-22 00:00:00 Denver @ Oklahoma City	11.361063	11.522992	9.560790	8.5	25
2021-04-23 00:00:00 Washington @ Oklahoma City	-7.190160	-11.417435	-15.013368	-9.0	-20

Model Stacking

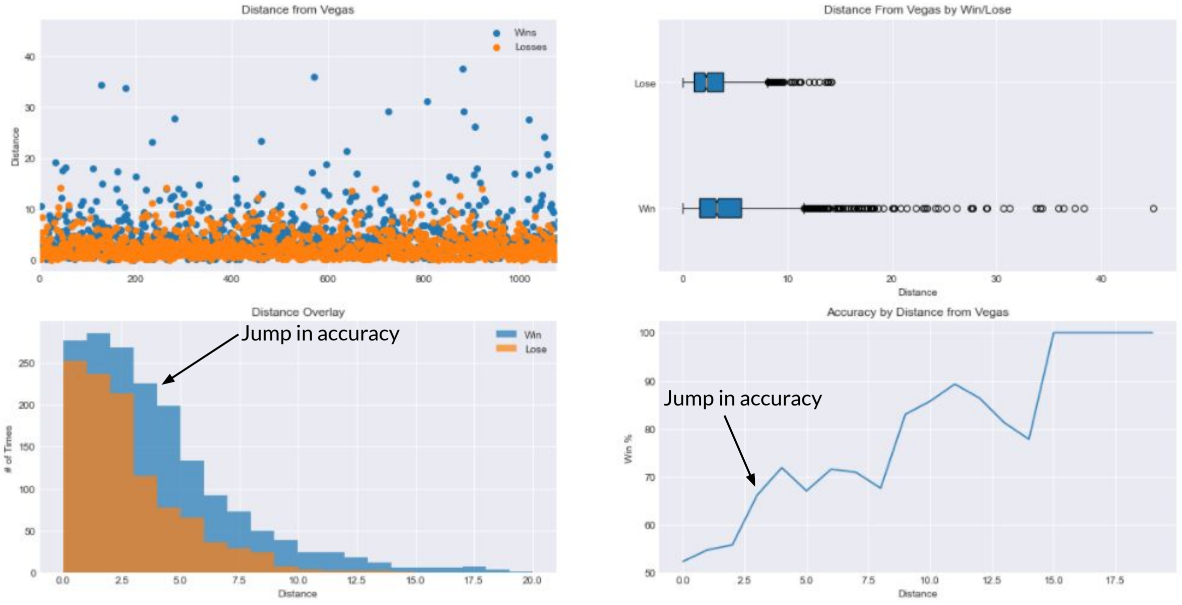
- We tested different combinations of models to see if predictions were more accurate when the majority of models agreed on one side
  - Accuracy falls off significantly when 1 out of 4 models disagree
  - Predictions are not reliable when RF disagrees



Agree	Accuracy	% of Games
At Least 3	67.40%	87.50%
All 4	69.02%	62.29%
All but LR	53.39%	4.12%
All but GB	54.05%	2.58%
All but RF	51.81%	5.80%
All but NN	52.50%	12.58%

Prediction vs Vegas

- We then explored if our predictions were more accurate based on their distances from the Vegas spread



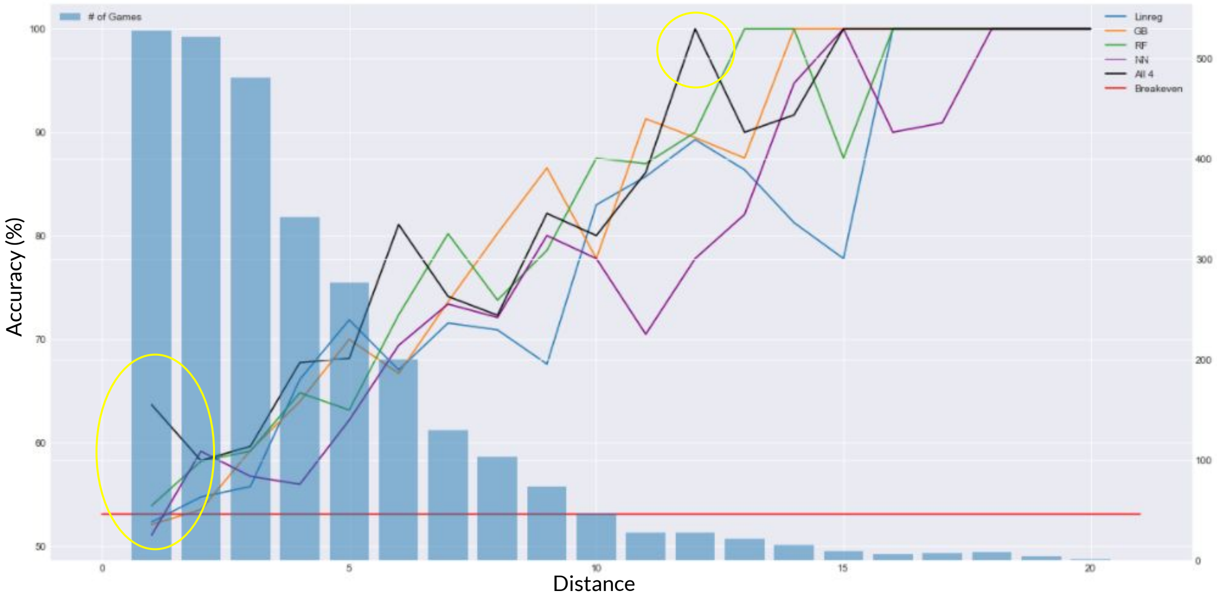
Jump in accuracy

Jump in accuracy

Shown is data for linear regression model only*

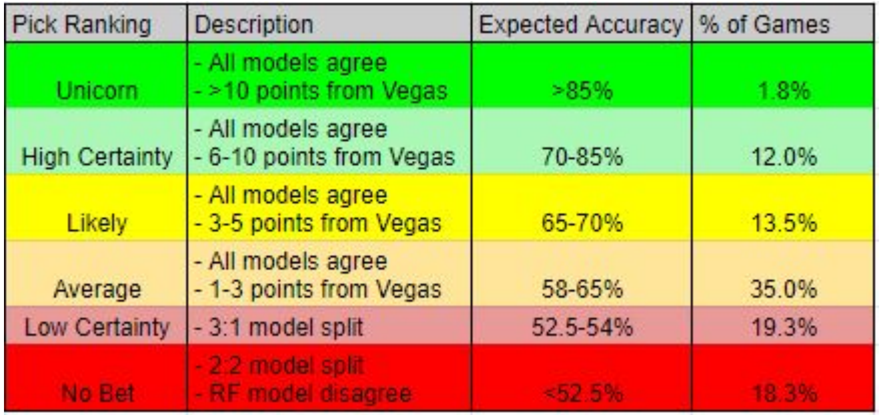
Individual vs Stacked

- Biggest divergence when distance is 0-1 (~23% of all games)
- Stacked model reaches max accuracy earlier



Product

- We will offer daily NBA spread picks to our subscribers
  - Expected pick accuracy is useful to bettors as it allows them to size their bets accordingly
  - Picks will be labeled as follows



Pick Ranking	Description	Expected Accuracy	% of Games
Unicorn	- All models agree - >10 points from Vegas	>85%	1.8%
High Certainty	- All models agree - 6-10 points from Vegas	70-85%	12.0%
Likely	- All models agree - 3-5 points from Vegas	65-70%	13.5%
Average	- All models agree - 1-3 points from Vegas	58-65%	35.0%
Low Certainty	- 3:1 model split	52.5-54%	19.3%
No Bet	- 2:2 model split - RF model disagree	<52.5%	18.3%

Subscription

- Average sports bettors bet $216 per month (prnewswire.com)
- Our model has an overall accuracy of 65.2%, giving us an expected value of 0.24
    - For every $1 bet, we are expected to make $0.24
- Our product gives the AVERAGE bettor $51.84 of value
    - Target audience is “serious” bettors
- We will price our monthly subscriptions at $50/month

Thank you for your time and attention!

Project Repository: Github link

Authors today can be reached using the following information:

Harrison Gu:
harrison.s.gu@gmail.com