Moneyball
(a true fintech story)

By Sean Brennan, John-Francis Kraemer, Andy Liang, Somya Panda, Faith Warari, and Priscilla Wong

Why Moneyball?!?

Do you want this guy on your team? Well… Neither do we!

Will Ferrell

John William Ferrell (Manimal) (@willferrell)
also known as Rojo Johnson

Position: P-C-1B-2B-SS-3B-LF-CF-RF-DH
"Bats:" Right, **Throws:** Right
Height: 6' 3", **Weight:** 220ish.

Motivation and Summary

- Use a deep learning methods to predict the 2021 OPS (On-base Plus Slugging) for the entire roster of MLB players and select a full roster of players with the highest OPS for their salary.

Backstory

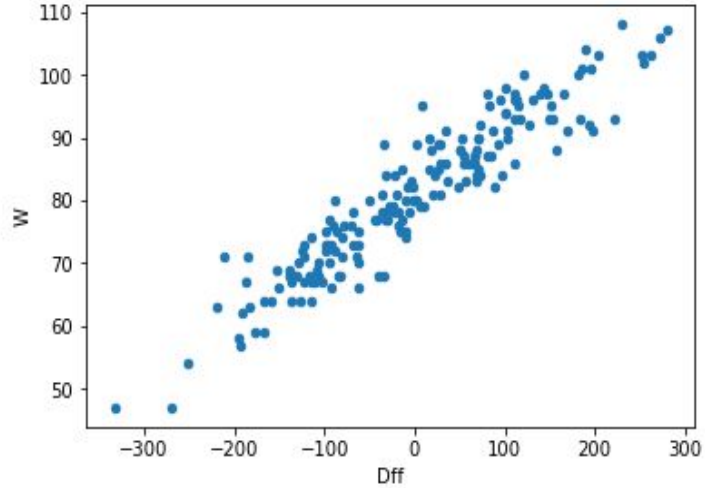
- Based on the 2003 Michael Lewis book, “Moneyball: The Art of Winning an Unfair Game”, about the Oakland Athletics baseball team and its general manager Billy Beane
- The concept genesis is the team's analytical, evidence-based, sabermetric (the empirical analysis of baseball) approach to assembling a competitive baseball team with the Oakland Athletics’ minimal budget
- Traditional methods of assessing a player's ability and signing long-term expensive contracts were deemed obsolete
- Utilizing sabermetrics to compile a competitive roster with players on free agency or under contract at an inexpensive salary

Assumptions

- The baseline assumption is teams that have more players on base, statistically score more runs, and teams that have a higher proportion of Runs Scored (RS) compared to Runs Allowed (RA) have a high correlation of winning (0.94)
- Taking MLB team stats from the past 6 seasons 2014 – 2019 (2020 was excluded due to the truncated season)

	Team	W	L	RS	RA	Dff
0	NY Yankees	103	59	943	739	204
1	Minnesota	101	61	939	754	185
2	Houston	107	55	920	640	280
3	Boston	84	78	901	828	73
4	Houston	101	61	896	700	196

	W	L	RS	RA	Dff
W	1.000000	-0.999710	0.585763	-0.728926	0.940690
L	-0.999710	1.000000	-0.584345	0.729098	-0.939881
RS	0.585763	-0.584345	1.000000	0.013667	0.652718
RA	-0.728926	0.729098	0.013667	1.000000	-0.748609
Dff	0.940690	-0.939881	0.652718	-0.748609	1.000000



Overview

- Traditional methods of assessing players were based on visual analysis from scouts, Batting Average, and Stolen Bases
- **The core metric to quantify using sabermetrics is the ability of a player to get on base and to facilitate existing players on base to score**
- Using the equations to quantify a players' On Base Percentage (OBP), Slugging Percentage (SLG), and On Base plus Slugging (OPS)

$$OBP = \frac{H + BB + HBP}{AB + BB + SF + HBP}$$

$$SLG = \frac{TB}{AB}$$

$$OPS = OBP + SLG$$

- H = hits; BB = bases on balls; HBP = times hit by pitch; AB = at bats; SF = sacrifice flies; TB = total bases

Note: Players with low at bats (< 50 career at bats) were excluded from this analysis that did not reflect an accurate representation of a player's batting ability

Data Clean Up and Exploration

- Sources:
  - Kaggle
  - mlb.com
  - spotrac.com
  - Rotowire
- Data Parameters:
  - MLB historical game/player data from 2010 through 2021
  - 2021 player salaries

Data Clean Up and Exploration Cont.

Pandas: Parsing, column name changes, dropping nulls, normalizing data types

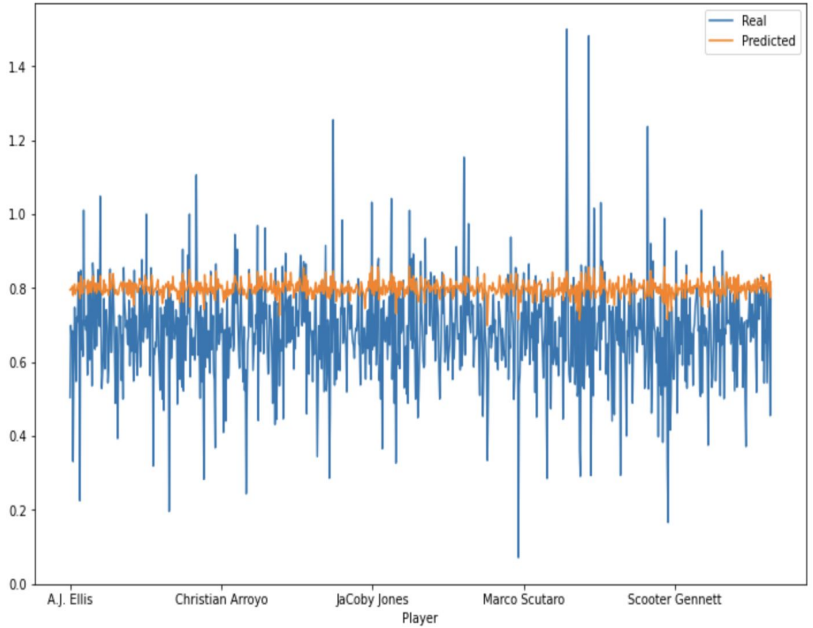
Year	Player	Team	Pos	Age	G	AB	R	H	2B	3B	...	CS	BB	SO	SH	SF	HBP	AVG	OBP	SLG	OPS
2019	Whit Merrifield	KC	2B	32	162	681	105	206	41	10	...	10	45	126	0	4	5	0.302	0.348	0.463	0.811
2019	Marcus Semien	OAK	SS	30	162	657	123	187	43	7	...	8	87	102	0	1	2	0.285	0.369	0.522	0.891
2019	Rafael Devers	BOS	3B	24	156	647	129	201	54	4	...	8	48	119	1	2	4	0.311	0.361	0.555	0.916
2019	Jonathan Villar	BAL	2B	30	162	642	111	176	33	5	...	9	61	176	2	4	4	0.274	0.339	0.453	0.792
2019	Ozzie Albies	ATL	2B	24	160	640	102	189	43	8	...	4	54	112	0	4	4	0.295	0.352	0.500	0.852
2019	Eduardo Escobar	ARI	3B	32	158	636	94	171	29	10	...	1	50	130	0	10	3	0.269	0.320	0.511	0.831

Model Parameters

- LSTM (Long short-term memory) RNN (recurrent neural network)
- Model Loss Function:
  - Binary cross-entropy (classification)
  - Root mean squared error (regression)
  - Did not greatly affect results either way
- Optimizer:
  - Adam: most popular choice for LSTM RNN models and regularly outperforms similar optimizers (RMSProp and AdaDelta)
- Epochs:
  - Tested with 50 and 100
  - Did more epochs do better?
    - Trended towards more pessimistic (and closer to “real”) predictions for OPS
    - Did result in different player roster

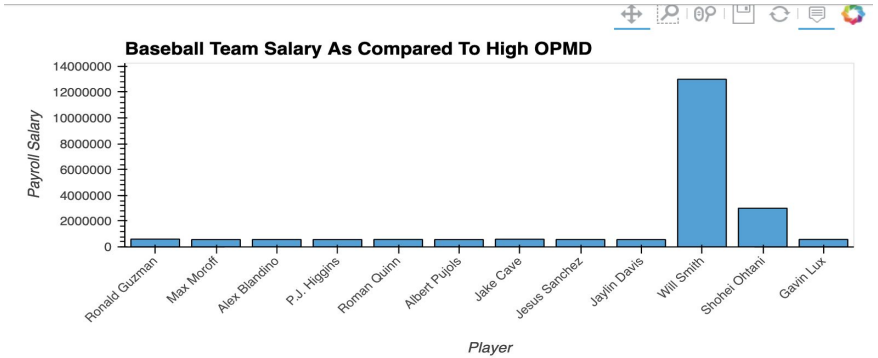
LSTM(Long short-term memory)

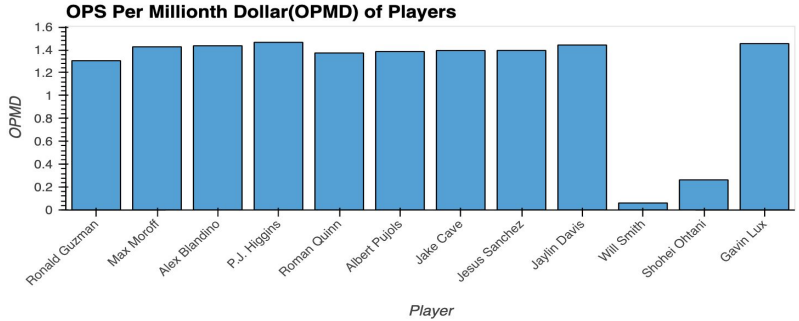
- An artificial recurrent neural network (RNN) architecture
- Use 6 years of baseball data
- Train and evaluate LSTM RNN model
- Performance
  - Loss = 0.5084630828841946
  - Accuracy = 0.005578093
  - Precision = 1.0
  - Recall = 0.9198783



Results and Conclusions

Position	Player	Payroll Salary	OPMD
1B	Ronald Guzman	606000	1.304727
2B	Max Moroff	570500	1.426082
3B	Alex Blandino	573000	1.435232
C	P.J. Higgins	570500	1.465250
CF	Roman Quinn	578000	1.372276
DH	Albert Pujols	570500	1.384780
LF	Jake Cave	597500	1.393563
OF	Jesus Sanchez	575000	1.394335
RF	Jaylin Davis	570500	1.441807
RP/CL	Will Smith	13000000	0.061147
SP	Shohei Ohtani	3000000	0.263346
SS	Gavin Lux	580500	1.454224





Opportunities & Challenges

- Assumptions to facilitate results may have missed value/hidden gems
  - Players with low at bats (< 50 career at bats) were excluded from this analysis that did not reflect an accurate representation of a player's batting ability - but was this correct?
- The theory was based on sabermetrics and nomenclature from 'Moneyball', which were largely unknown and under analyzed at the time, however, sabermetrics have gained notoriety and league adoption since the books release in 2003
- The project focused strictly on offense, it did not analyze starting/relief pitching to determine accretional/dilutional value to a team's Run Differential (RS-RA), which statistically has a high correlation (0.94) to a team winning and improving the probability of clinching a playoff berth

Additional Questions

- How is this plan implemented
  - For simplicity, this project did not include contract terms or trade negotiations to form a real world scenario to get an ‘ideal’ team compiled
    - Constraints on trade value (players on roster that have an intrinsic ‘trade value’ to another team) and budget for overall payroll
- Pitching/Defense - Incorporating a full pitching bullpen and a player’s fielding ability was not quantified in the project, additional data and analysis would be needed to complete a full 25 player roster

Reflections

- What incremental value could be added to this analysis?
  - With the evolution of data science, it is very important to discover what additional player attributes can be quantified into a determinable statistic in the model in assessing performance
- Forecasting performance/health in building a team?
  - Recent seasons have been plagued by covid related protocols (players unable to play due to sickness or close contact) or other real world challenges/opportunities
  - Players with a history of suspensions from HGH, MLBPA violations, fighting, etc.
  - Injury prone players statistically have frequent injuries or often sustaining injuries that would need to be quantified into a player's ability to remain healthy for an entire season (opportunity cost of replacement players)

Thank you!


WHEN? WHY? WHEN? WHY? WHO? WHAT? HOW? WHY? WHERE? WHO? WHAT? WHEN? HOW? HOW? WHERE? WHEN? WHERE? WHOSE? WHICH? HOW? WHEN? WHERE? HOW? WHAT? HOW? WHY? WHY? HOW? WHAT? HOW? WHY? WHO? WHERE? WHAT? HOW? WHAT? WHY? HOW? WHERE? WHO? WHOSE? WHERE? WHAT? HOW? WHY? HOW? WHERE? WHO? WHERE? WHAT? HOW? WHERE? WHICH? WHOSE? WHEN? WHY? WHERE? HOW? HOW? WHO? WHO? WHERE? WHAT? HOW? WHAT? HOW? WHY? WHERE? HOW? HOW? WHERE?

Questions?