Using Machine Learning To Predict Results of At-Bats in Baseball

MICHAEL POLONIO

Introduction

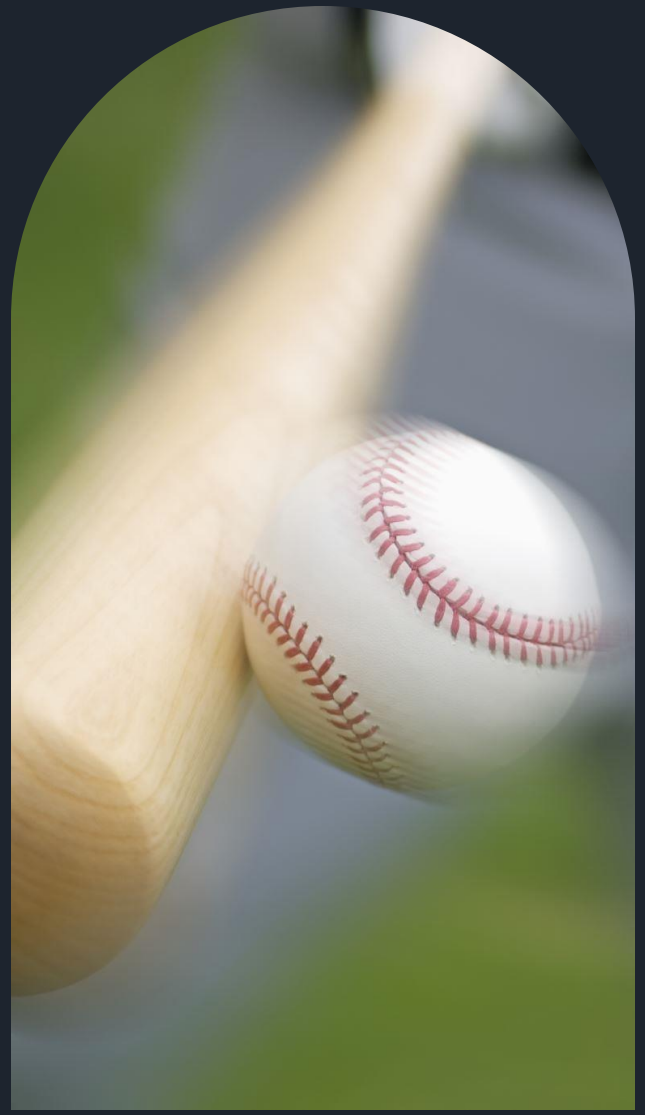
Baseball, known as the American Pastime, has been played for about 150 years and is one of the Major professional US sports. Major League Baseball (MLB) is the premier professional baseball league in the world.

Due to its discrete nature and large amount of data, baseball lends itself as a great subject of statistical analysis

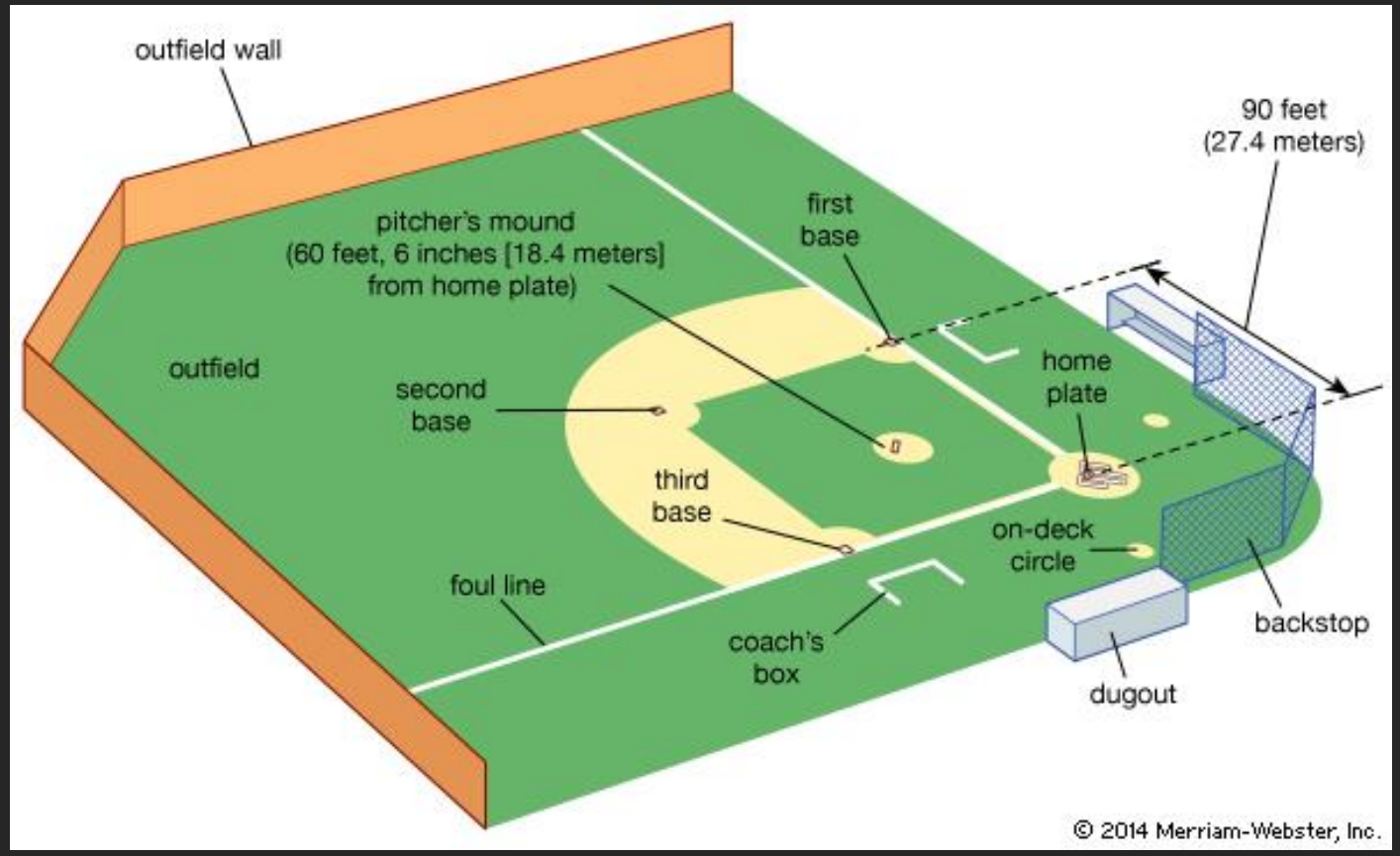
Analysts can now apply machine learning algorithms to large baseball datasets to derive meaningfull insights into player and team performance (Koseler & Stephan, 2017).

Why Predictive Models are Important in Baseball

- Big business - The cost of MLB with all its assets, revenue sources, and expenses is roughly $29.886 billion. Add in the cost of all 30 teams and baseball will cost $80.701 billion (Epstein, 2019)
- Sports betting is a $200 billion industry and gaining popularity. Now regulated in 18 states and counting (Perdum, 2020)
- Each MLB organization now has an analytics team, using data to gain a competitive advantage. Teams are unwilling to share exactly how they are using the data, engaging in an "arms race" of data analysis (Chen, 2016)
- Accurate prediction in Baseball is highly sought after



Baseball Rules & Terms



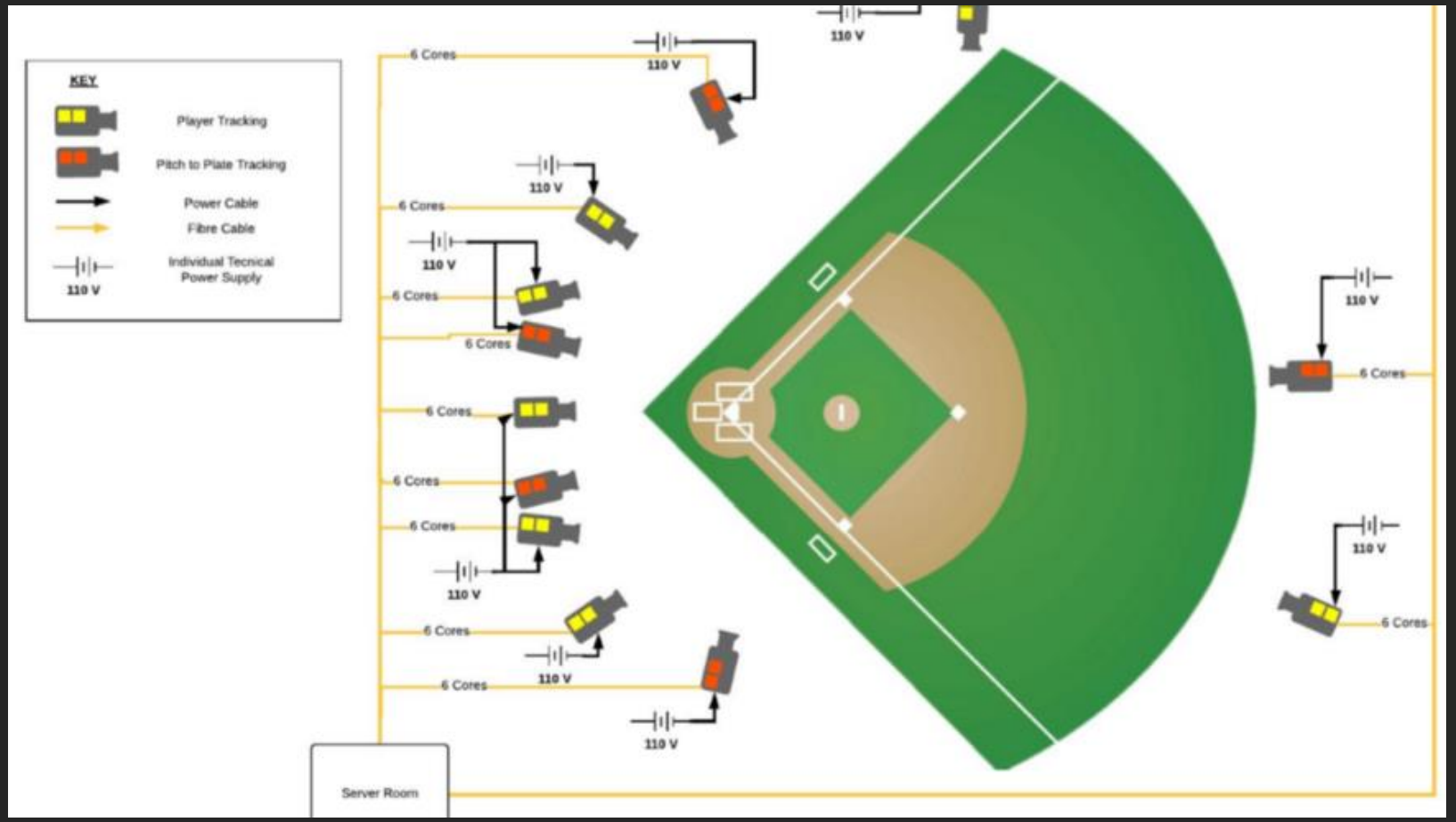
- Pitching team and batting team
- Pitching teams' goal is keep runners off the bases by getting them out
- Batting teams' goal is to get on base, and advance through the bases by getting hits, reaching back to home plate and scoring a run
- The pitcher tries to keep the batter guessing by throwing different pitches that vary in speed and trajectory
- At-bat: When the batter takes his turn against the pitcher

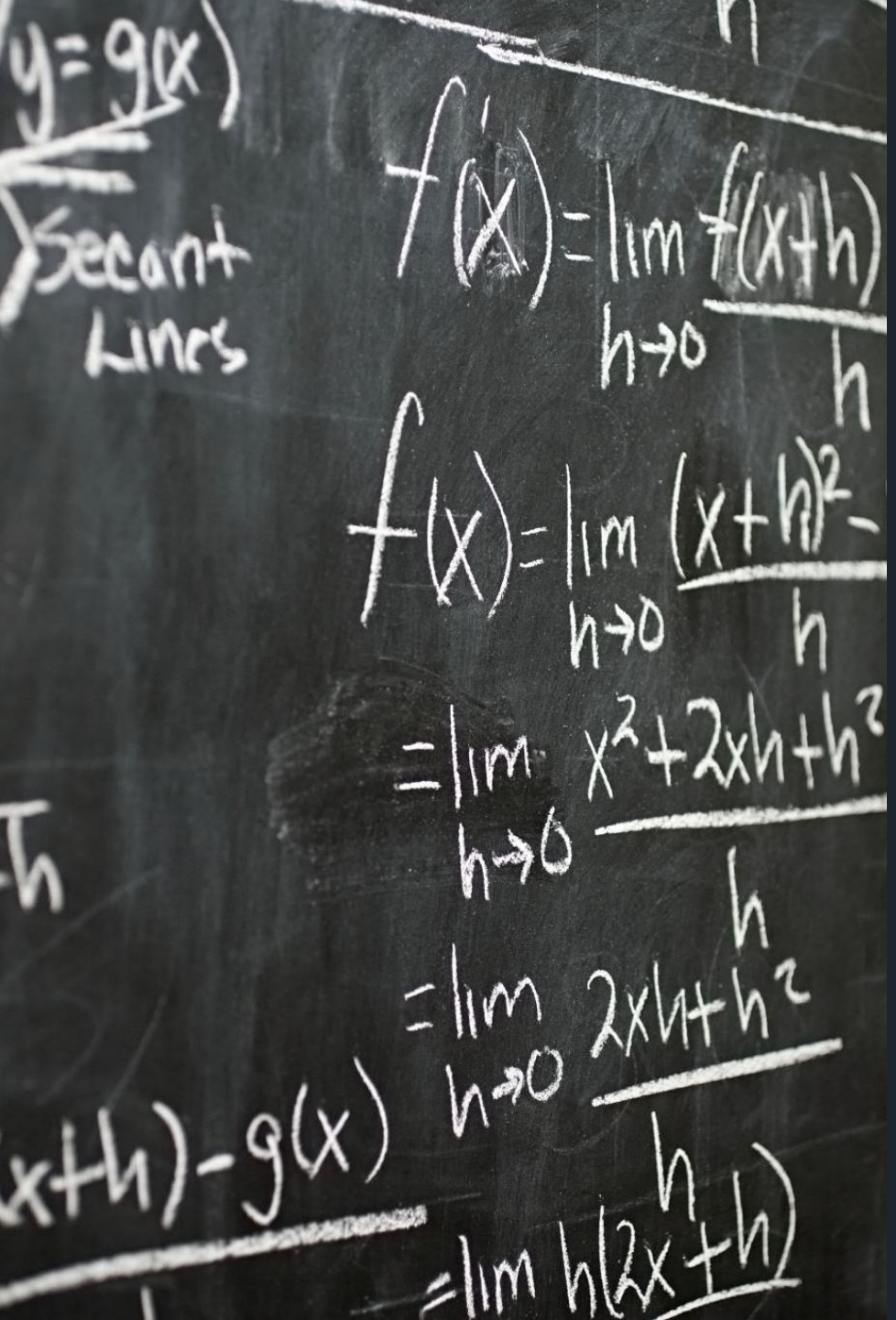
Statcast

-INTRODUCED IN EVERY MLB STADIUM IN 2015, STATCAST IS A SYSTEM OF CAMERAS USED FOR DATA COLLECTION

-THIS TECHNOLOGY INTEGRATES DOPPLER RADAR AND HIGH DEFINITION VIDEO TO MEASURE THE SPEED, ACCELERATION, AND OTHER ASPECTS FOR EVERY PLAYER ON THE FIELD.

- LICENSED TO ESPN AND USES GOOGLE CLOUD AS ITS CLOUD DATA AND ANALYTICS PARTNER





Machine Learning (Classification)

- Classification predictive modeling is the task of approximating a mapping function (f) from input variables (X) to discrete output variables (y) (Asiri, 2018).
- Uses input variables to predict class labels i.e. possible outcomes
- Binary or Multiclass
- Different classification methods performance can vary depending on the problem

Similar Studies in Baseball Analytics

- Koseler & Stephan (2017) performed a systematic literature review of the machine learning applications in baseball analytics. They studied the approaches used in Regression, Binary Classification, and Multiclass Classification problems and found that Support Vector Machines (SVM) were most widely used for Classification problems, being used in at least 25% of the studies. They also postulated the rise in popularity of neural networks in baseball analytics.
- Hamilton et al (2014) researched pitch prediction using a SVM and k-NN, and was able to produce models that accurately predict the pitch being thrown to the batter with a 79.76% and 80.88% accuracy, respectively.
- Tolbert and Trafalis (2016) aimed to predict the American League, National League, and World Series Champions. They used a SVM in their analysis and were able to achieve 69% accuracy in their model, which successfully predicted the winner and loser for the 2015 World Series.

Abstract

- As the game of baseball adapts to the information age, it becomes increasingly reliant on analytical and statistical methods in order to maintain a competitive edge. Predictive models in the MLB are useful and highly sought after by teams, coaches, and the sports betting industry. In this paper I will use classification-based machine learning techniques with the aim of predicting the result of an at-bat, whether it be an out or a hit, by using information about the pitch. Classification will be done using a Support Vector Machine, Naïve Bayes Classifier, Decision Tree, Random Forest, and a Deep Neural Network. The models will then be evaluated and compared to each other. Previous researchers have had success with these machine learning methods, so my goal is to find out if they perform comparably well when applied to the problem of predicting at-bat results.

Research Question/Hypotheses

- Is it possible to predict the result of an MLB at-bat using information about the pitch?
- Which model will perform to the highest degree of accuracy when predicting At-Bat Outcomes?
- The following classification models will be trained:
  - Support Vector Machine
  - Naïve Bayes
  - Decision Tree
  - Random forest
  - Deep Neural Network (Multilayer Perceptron)
- The models will then be evaluated for their accuracy, f1 score, precision, recall, and compared

Data

	pitch_type	speed	release_pos_x	release_pos_z	release_pos_y	events	balls	strikes	Plate_x	Plate_y	Outs_when_up	vx0	vy0	vz0	ax	ay	az	Spin_rate
0	SL	81.4	-2.95	6.07	54.27	single	3	2	-0.19	1.80	1	5.702548	-118.549653	-1.945689	1.307271	22.220532	-37.245128	2130.0
1	FF	93.6	-2.71	6.18	54.28	single	3	1	-0.04	1.79	1	8.627963	-135.846696	-8.883218	-11.843061	27.617710	-11.843779	2241.0
2	CH	85.3	-1.75	4.95	53.87	single	0	1	-0.03	1.92	0	5.524685	-124.264141	0.169049	-8.420712	23.074137	-37.898460	1363.0
3	FF	88.7	3.74	5.23	54.13	field_out	0	0	0.25	2.91	2	-10.754807	-128.856453	-1.907631	13.961439	25.443398	-19.721337	2255.0
4	FF	88.8	3.60	5.34	54.15	field_out	0	0	0.18	3.05	1	-10.579352	-128.945777	-2.049793	13.950409	25.071655	-18.623636	2205.0

Methods

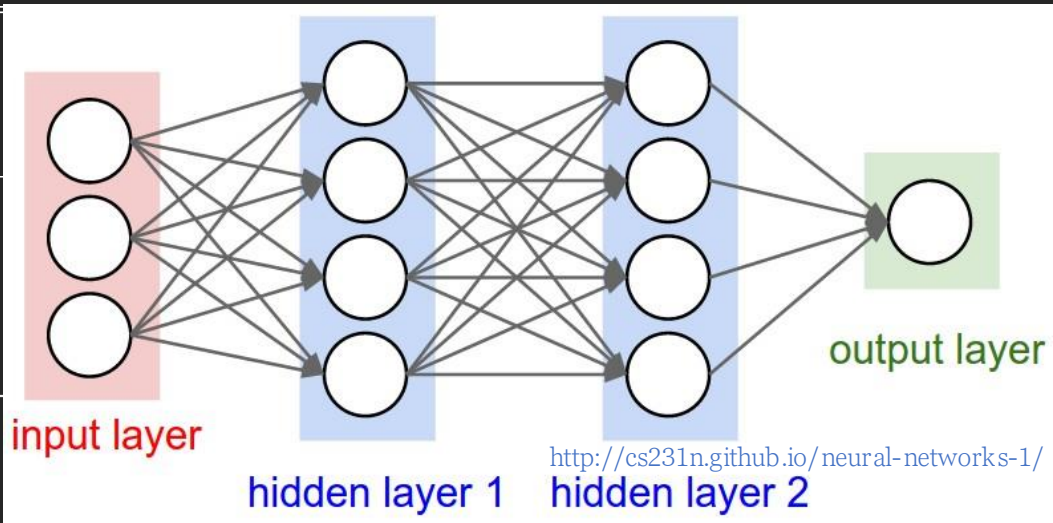
- Python (scikit-learn, Keras, Matplotlib)
- Use data collected by Statcast from Baseball Savant Database.
- Pitch variables are predictors: pitch type, release speed (mph), number of balls and strikes, pitch spin rate (rpm), x/y/z coordinate of the release point, x/y coordinate of the where the pitch crosses the plane at the plate, x/y/z velocity and acceleration of the pitch
- Target variable for classification: At-Bat Result (HIT or OUT)
- Standardize values and Perform feature selection (Correlation & Feature Importance)
- Build classification models: Support Vector Machine, Decision Tree, Random Forest, Naïve Bayes, and a Deep Neural Network (MLP)
- Evaluate models and compare results

Neural Network Architecture

- Multilayer Perceptron
- Two hidden layers with 64 nodes, each with a ReLU activation function
- Each has a Dropout of 40% to avoid overfitting
- Model is compiled with an Adam optimizer and a binary cross-entropy loss function
- Output layer has one node, as for binary classification

Layer (type)	Output Shape	Param #
dense_3 (Dense)	(None, 64)	960
dropout_2 (Dropout)	(None, 64)	0
dense_4 (Dense)	(None, 64)	4160
dropout_3 (Dropout)	(None, 64)	0
dense_5 (Dense)	(None, 1)	65

Total params: 5,185
Trainable params: 5,185
Non-trainable params: 0



http://cs231n.github.io/neural-networks-1/

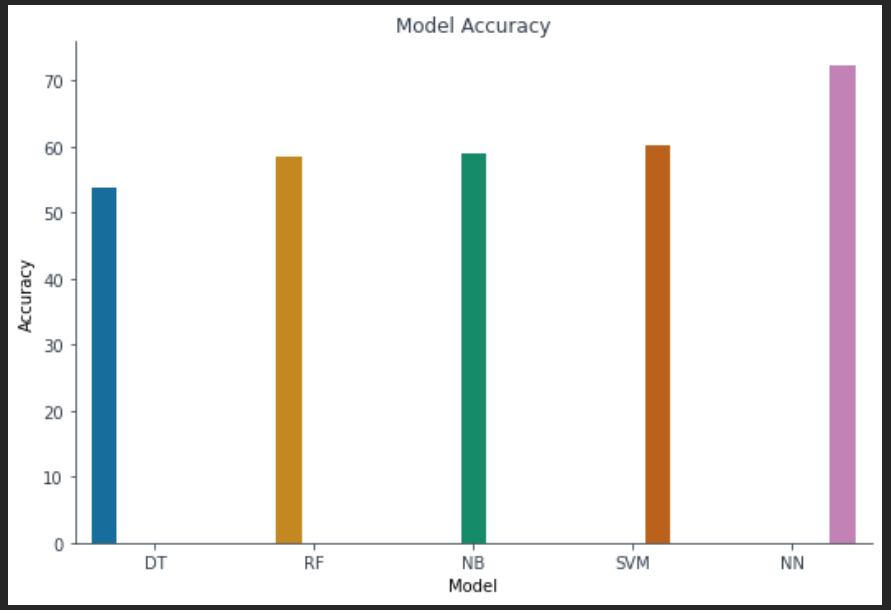
Results for naturally occurring hit/out

•Deep Neural Network outperformed all other classification methods with an accuracy of 72.1%.

• Although some models performed to a moderate accuracy, the F1 score is very low for the RF and DT (low precision and recall). This means that these models were only good at predicting one outcome, OUTS in this case.

• The low recall indicates most of one label of the binary classification is not correctly predicted

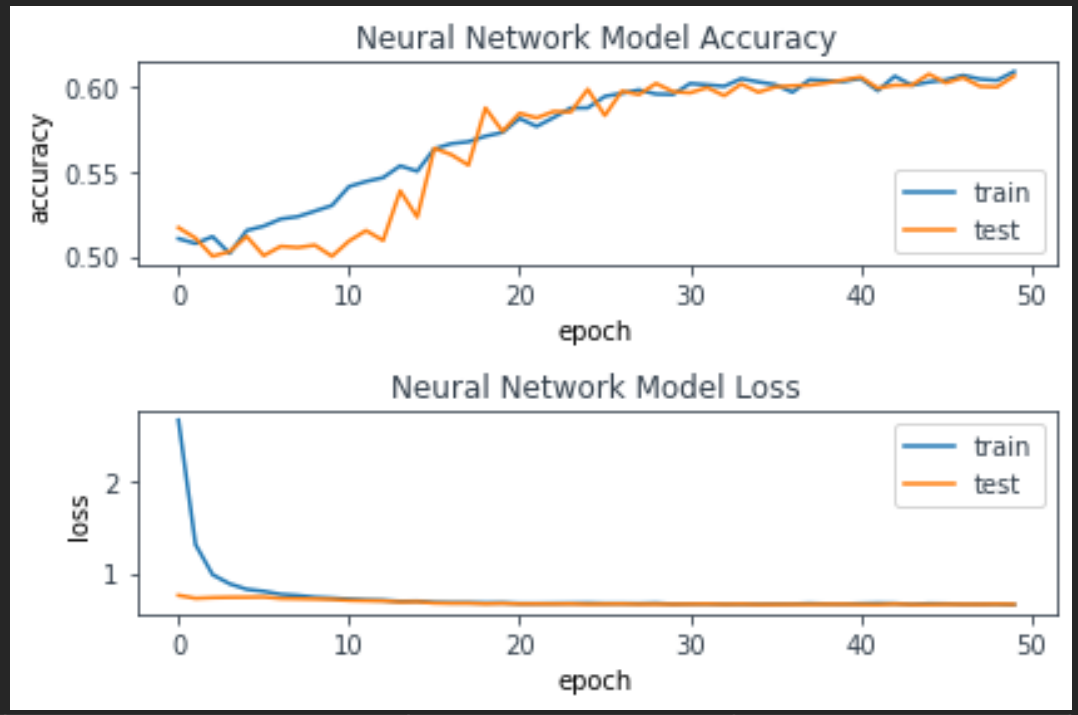
• These low metrics indicate that the models had a difficult time differentiating between hits and out



Model Name	Accuracy	CV Score	F1-Score	Precision	Recall
Decision Tree	62%	0.52	0.34	0.33	0.35
Random Forest	71%	0.52	0.09	0.41	0.05
Naive Bayes	69%	0.52	0.669	0.50	0.17
SVM (RBF kernel)	71.7%	0.42	0.34	0.33	0.35
Deep Neural Network	72%	0.52	0.4	0.41	0.39

Results for equal amount of hit/out (downsampled)

- The DNN outperforms all other models when looking at the balanced dataset as well
- The *precision* and *recall* are improved in these models, although the accuracies are lower. This indicates that the models are better at predicting both binary levels.
- The DNN improved up to 50 epochs before overfitting, with an accuracy of 62%.
- Although these models with an accuracy around 60% they are more reliable at predicting hits and outs than the previous models with a higher accuracy



Model Name	Accuracy	CV Score	F1-Score	Precision	Recall
Decision Tree	53%	0.54	0.52	0.53	0.52
Random Forest	58.5%	0.54	0.60	0.57	0.63
Naive Bayes	59%	0.54	0.61	0.57	0.65
SVM (RBF kernel)	60.3%	0.44	0.64	0.58	0.71
Deep Neural Network	62%	0. 54	0.65	0.60	0.72

Fig3

Conclusion

- Predicting whether an at-bat will result in a hit or an out based on information about the pitch proves to be a difficult task even for these machine learning techniques, but it was still possible to do so 62% of the time when analying the balanced dataset.
- The Support Vector Machine may have performed to the highest level of accuracy in several other studies on the topic, but not for the problem of predicting whether the result of the at-bat will be a hit or an out. The Deep Neural Network (MLP) performed the highest when looking at the naturally occurring hits/outs and the equal number of hits/outs with an accuracy of 72% and 62% respectively.
- I predict these models would perform better when trained on individual pitchers. Using pitch data from all MLB pitchers introduces a sort of randomness and high variability regarding pitch information. Training on individual pitchers would provide more easily identifiable differences between pitches.
- I think with some improvement, these models can be useful when analyzing a pitcher to gain a competitive advantage

References

- Epstein, Daniel (Jan 22, 2019). "The Purchase Price of MLB". Beyond The Boxscore. The purchase price of Major League Baseball - Beyond the Box Score
- Chen, Albert (August 26, 2016). "The Metrics System: How MLB's Statcast is creating baseball's new arms race". Sports Illustrated. Retrieved August 27, 2016
- Jason Dachman, C. E. (2019, March 28). *MLB 2019 preview: ESPN continues to up its virtual game with more K-zone 3D, Statcast Graphics*. Sports Video Group
- David Perdum, (May, 2020), "Sports betting's growth in US is 'extraordinary'". ESPN. Sports betting's growth in U.S. 'extraordinary' (espn.com)
- Kaan Koseler & Matthew Stephan (2017) Machine Learning Applications in Baseball: A Systematic Literature Review, Applied Artificia Koseler & Matthew Stephan l Intelligence
- Tolbert, B., & Trafalis, T. (2016). Predicting major league baseball championship winners through data mining. *Athens Journal of Sports*, *3*(4), 239-252.


THE END