Predicting the type of baseball pitch using supervised machine learning techniques

Ayden Salazar, Aaron Chow, Judy Moon, Michael Delvizis, Quoc Huynh

1

Motivation

Why is data-driven analysis important in predicting baseball pitches?

“Guessing what the pitcher is going to throw is 80 percent of being a successful hitter. The other 20 percent is just execution.”
-Hank Aaron

Problem Identification

- How difficult is it to hit a pitch?
- Can a hitter improve success in guessing which pitch comes next?
- What current methods are used by major league teams to train hitters in predicting pitches?

2

Dataset

What steps did we take to determine which features should or should not be used in our model? How did we clean/explore our raw dataset?

Dataset

MLB Statcast Data:

Starting in 2015, the MLB could track every aspect of a baseball play from spin rate of pitch, to exit velocity of hit and how fast a runner reacts

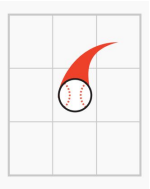
Implication:

- **Teams began an analytics arms race to determine which measurements in the sport determined success**

CH = Changeup
CU = Curveball
FC = Cutter
FF /FT = Four/Two seam Fastball
SI = Sinker
SL = Slider
IN = Intent ball
KC = Knuckleball Curve

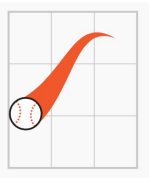
Dataset

Pitch Type



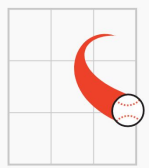
Four Seam Fastball

85-100 mph
Fastest, straightest pitch



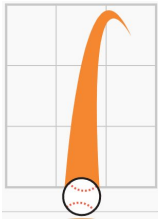
Two Seam Fastball

80-90 mph
Moves downward, aka Sinker



Cutter

85-95 mph
Mix of slider and fastball

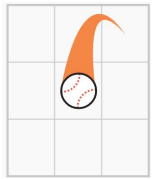


Curveball

70-80 mph
Commonly called a 12-6 curveball

Slider

80-90 mph
Between fastball and a curve



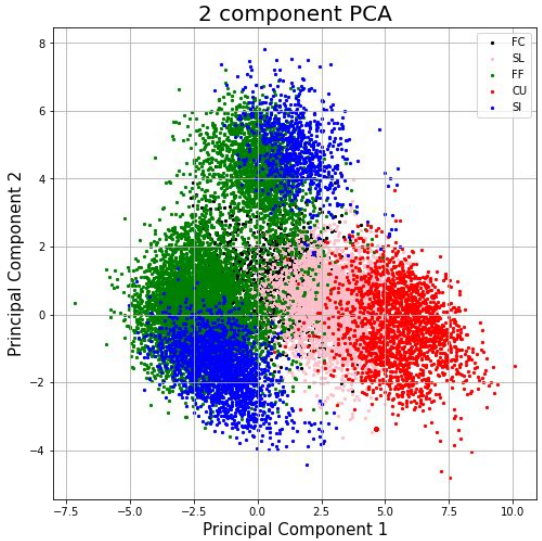
Changeup

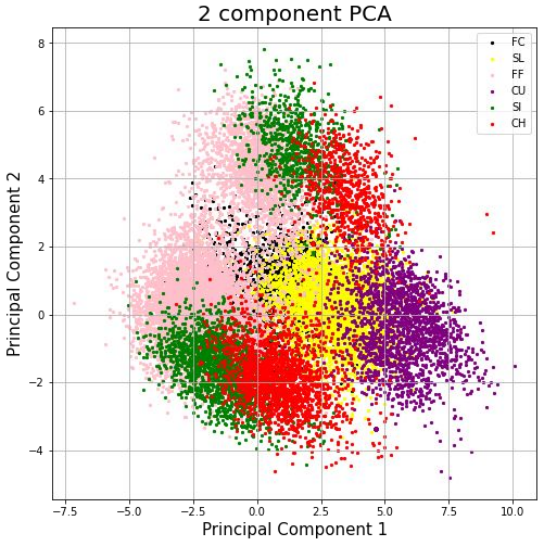
70-85 mph
Slower than a fastball, but thrown with the same arm motion

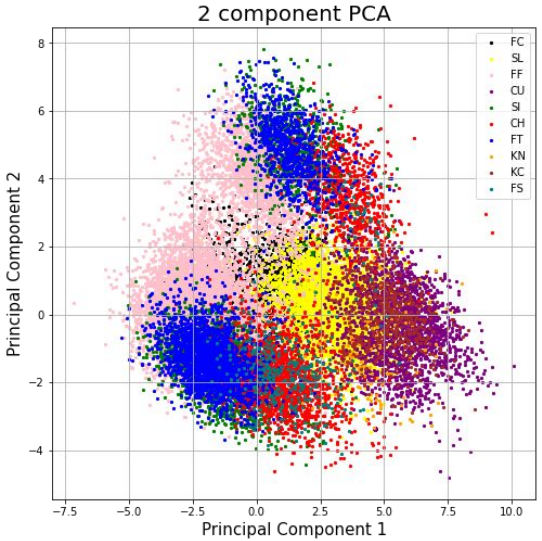
Data Visualizations

What information can we get about our dataset using visualizations? What features should be used for building our predictive model?

Principal Component Analysis (PCA)



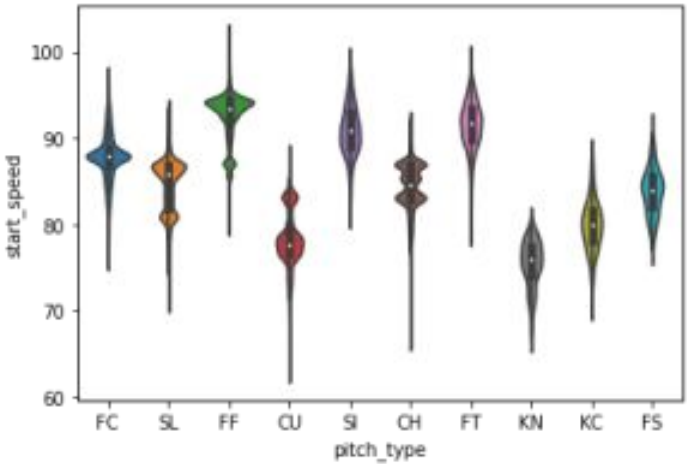




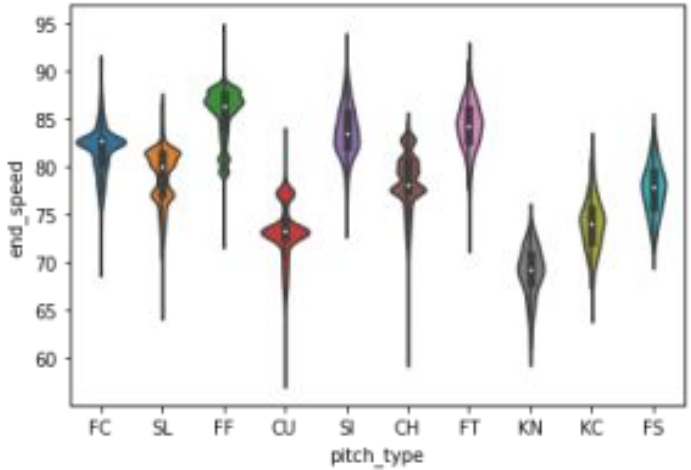
Feature Selection

Method: Using Seaborn Violin Plot to explore each feature per pitch type

Start speed

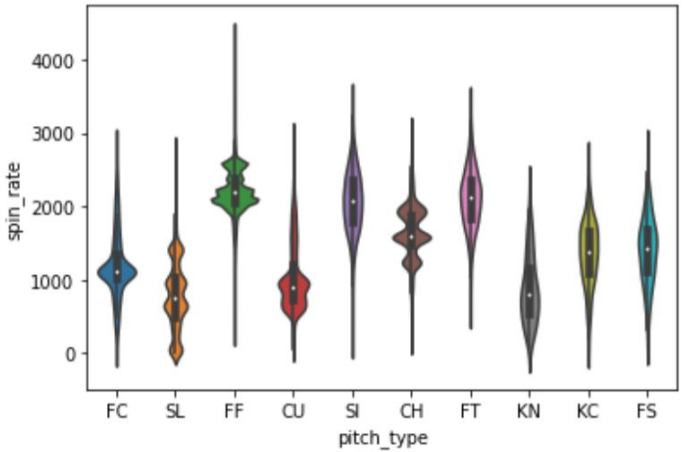


End speed



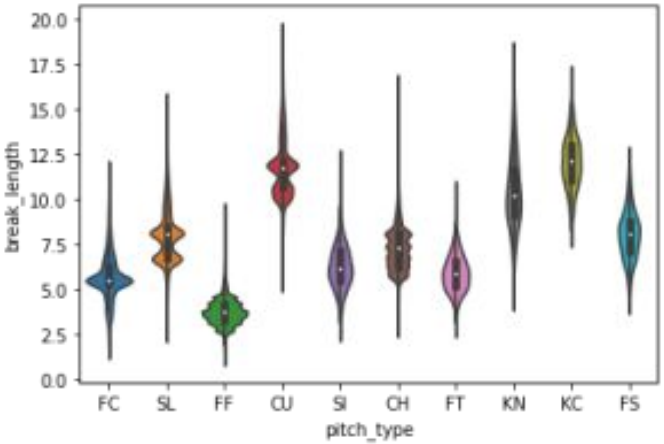
Feature Selection

Spin Rate



Increasing the spin rate of pitches make it harder to hit, thus important metric in predicting pitch type

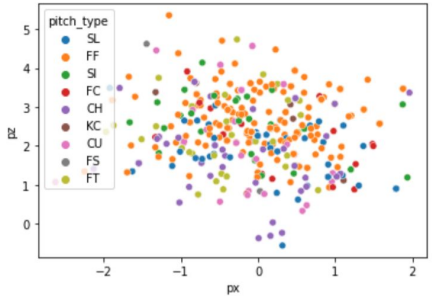
Break Length

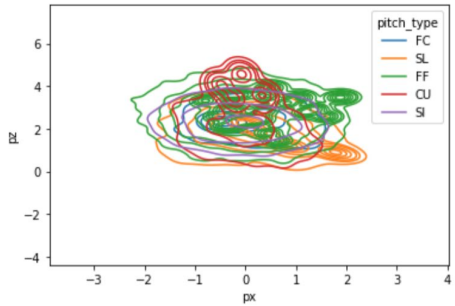


Greatest distance between the trajectory of the pitch at any point between the release point and the front of home plate

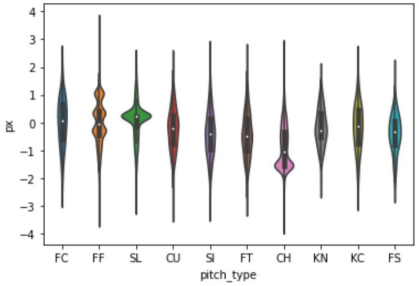
Feature Selection

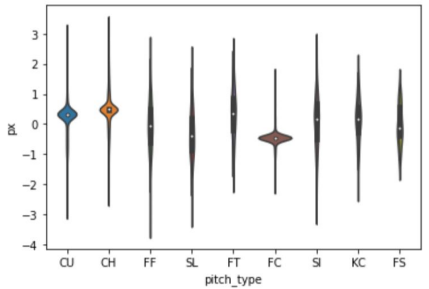
Location in strikezone



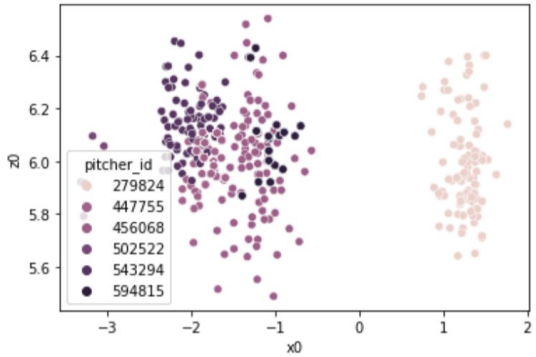


Left Hand v.s. Right Hand Pitches



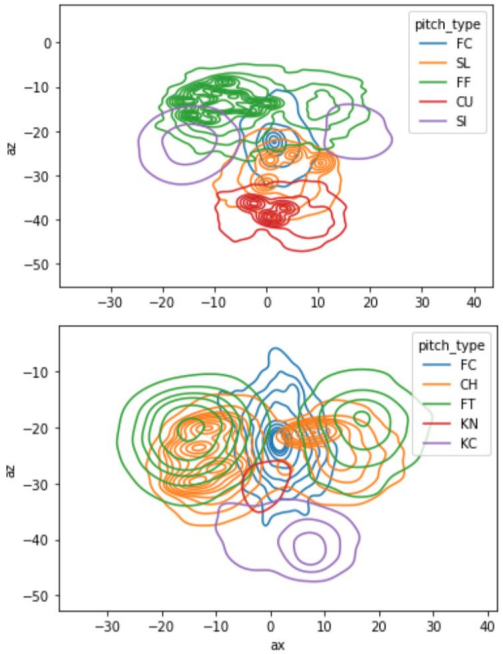


Pitch Release Position x, z

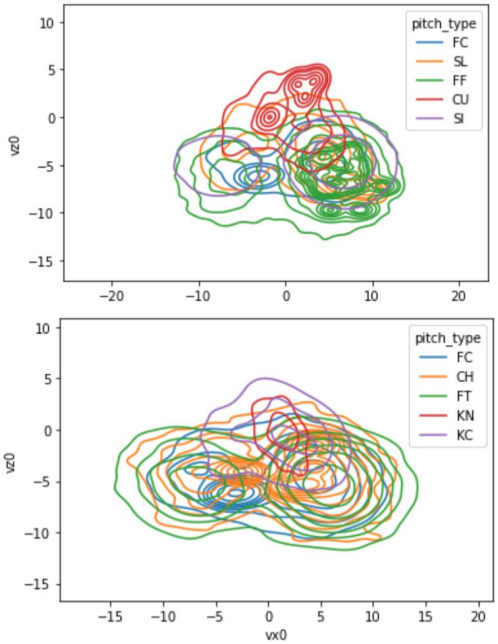


Feature Selection

Acceleration Attributes



Initial Velocity



3

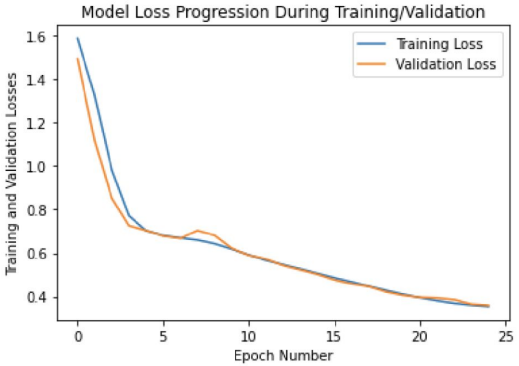
Methods / Modeling

Which supervised machine learning algorithms did we use?
Which algorithm produced the highest classification accuracy?
Which metrics did we use to determine “best predictive” strength?

Model Experimentation

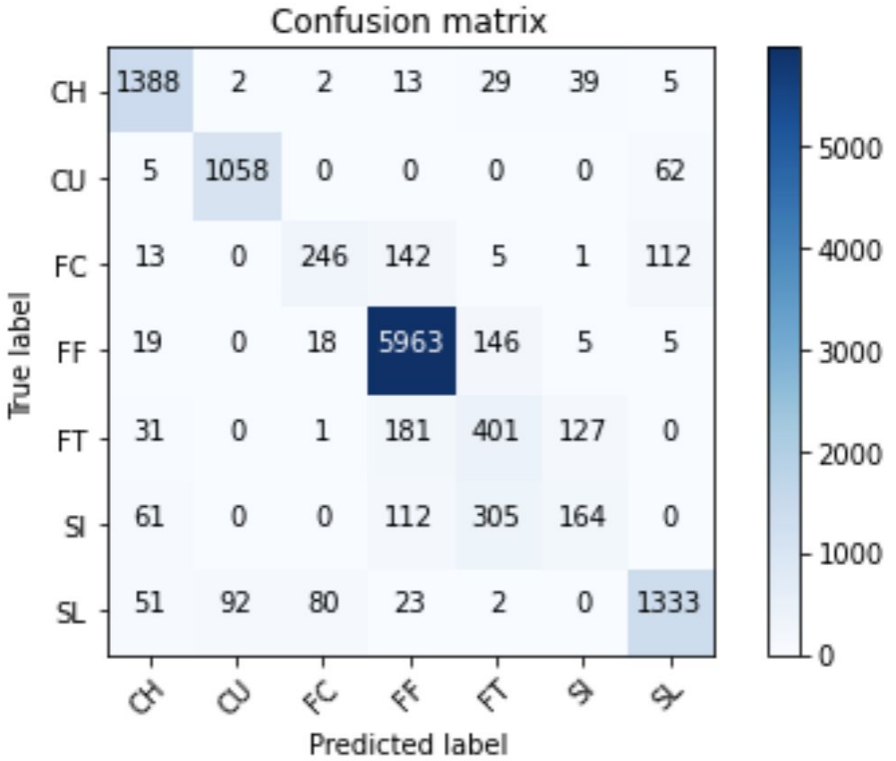
Neural Network Model

Test Set Accuracy: **86.2%**



Parameters/Hyperparameters:

- Two hidden layers with ReLu activation
- Output layer: Softmax function
- Stochastic Gradient Descent, α = .005
- Loss: categorical cross entropy

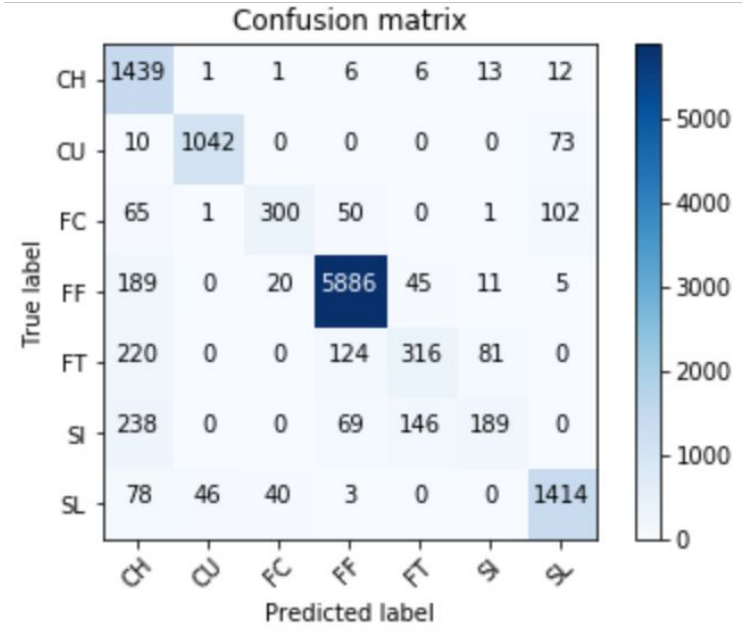


Random Forest Model

Test accuracy: **85.9%**

Parameters/Hyperparameters:

- max_depth=100
- n_estimators=20
- min_samples_split=4

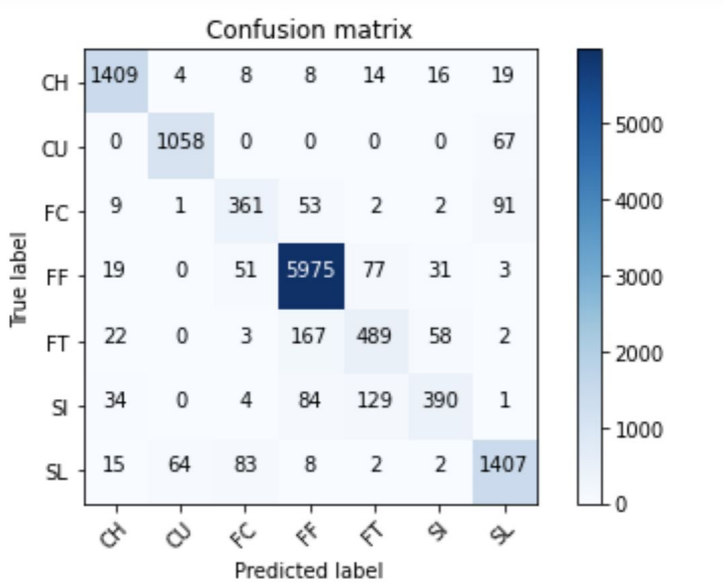


Decision Tree Bagging Classifier

Test accuracy: **90.6%**

Parameters/Hyperparameters:

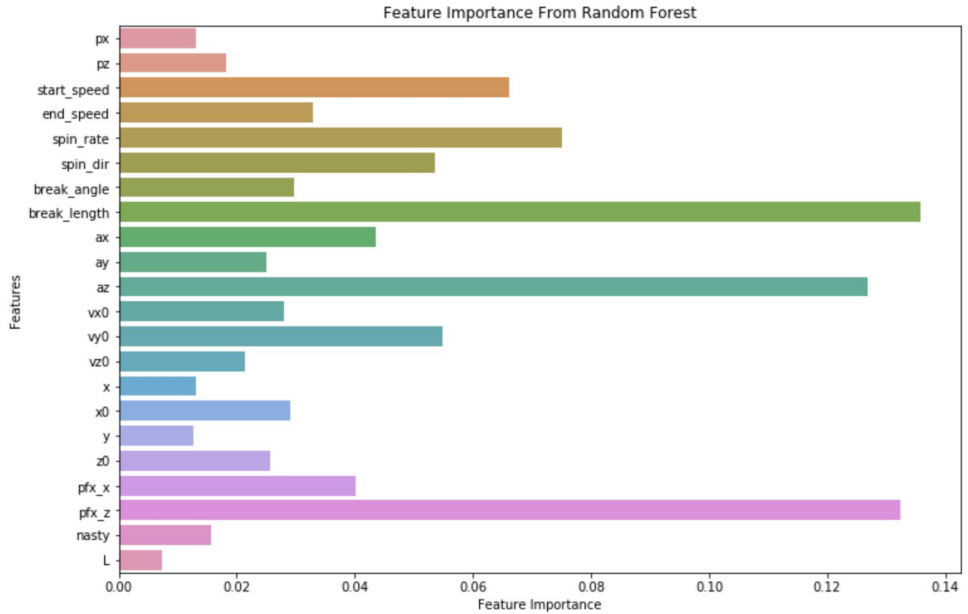
- base_estimator: Decision Tree Classifier
- n_estimators: 10



Model Metrics Comparison

	Neural Network	Random Forest	DTC Bagging
Train Accuracy	0.8704	0.987	0.9956
Test Accuracy	0.862	0.859	0.9058

Feature Importance



1. Break Length

2. Deviation of pitch trajectory of horizontal location

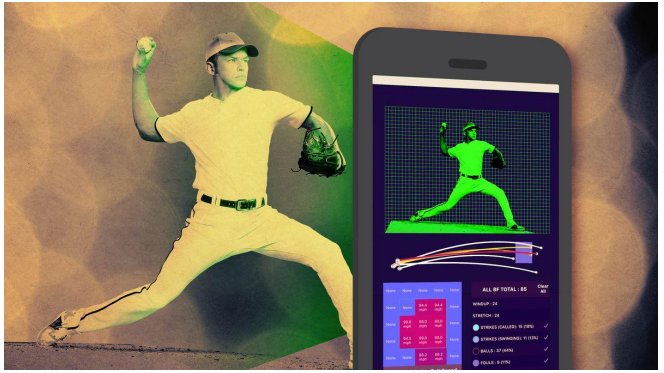
3. Acceleration of pitch, measured at the initial position

4

Results and Conclusions

What are the real world applications to our model?
What are the limitations to our project?
Some suggestions for future studies?

Real World Application



Predicting MLB Pitch Probability Based on the Game Situation

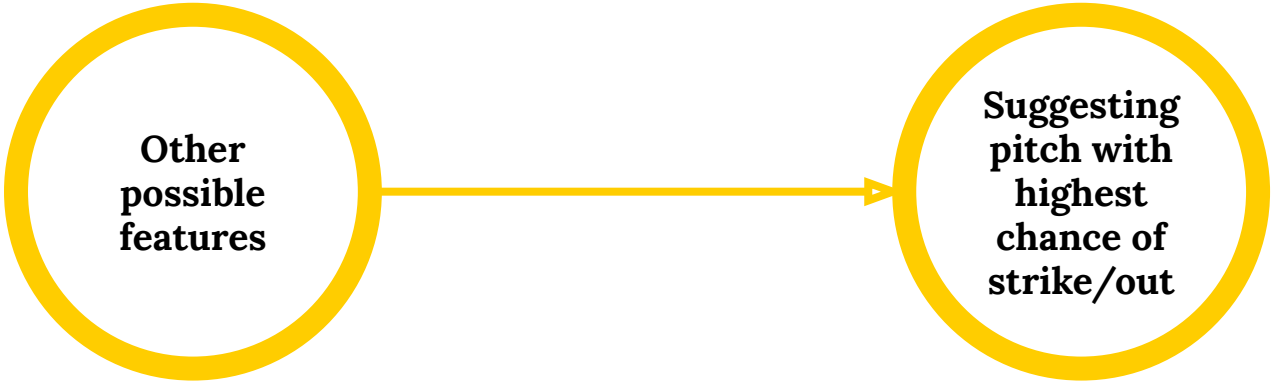
The Data Detective Dec 17, 2019 · 6 min read ★

Opinion | How AI and coaching can change player performance evaluation in football

The recent advancements in AI are shaping the world of sport. Here, Sankalp Chaudhary explains how CoachFirst and Quantiphi's AI-assisted coaching platform is developing the football ecosystem.

1. Scouting reports
2. Pitch simulations before and on-site
3. Player performance evaluation
4. Automated pitch feedback

Future Implementations



- Environmental factors (wind, temperature, humidity, etc.)
- Other entities with an effect on pitch selection (catcher, umpire, batter on deck)
- In-game information

Thank You!

Any questions ?