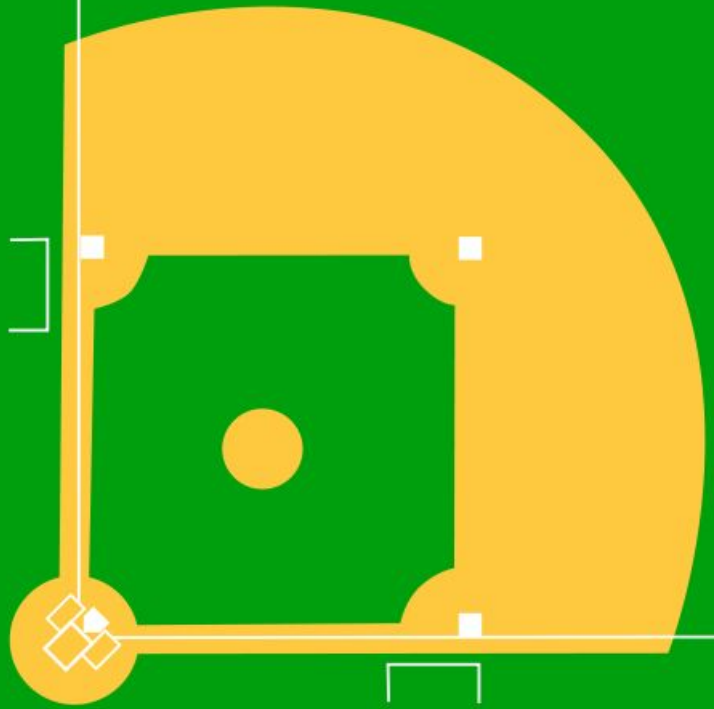


Single, Double, Triple, or Home Run?

Predicting Hits in Baseball

Background and Data Selection

- Team batting coaches are trying to determine what factors impact hitting
- Data comes from MLB
- Data includes:
  - ~90 columns for every pitch of every game
  - Both pitching data and hitting data
- We used 40,000 rows from 2019 regular season

Target and Feature Selection

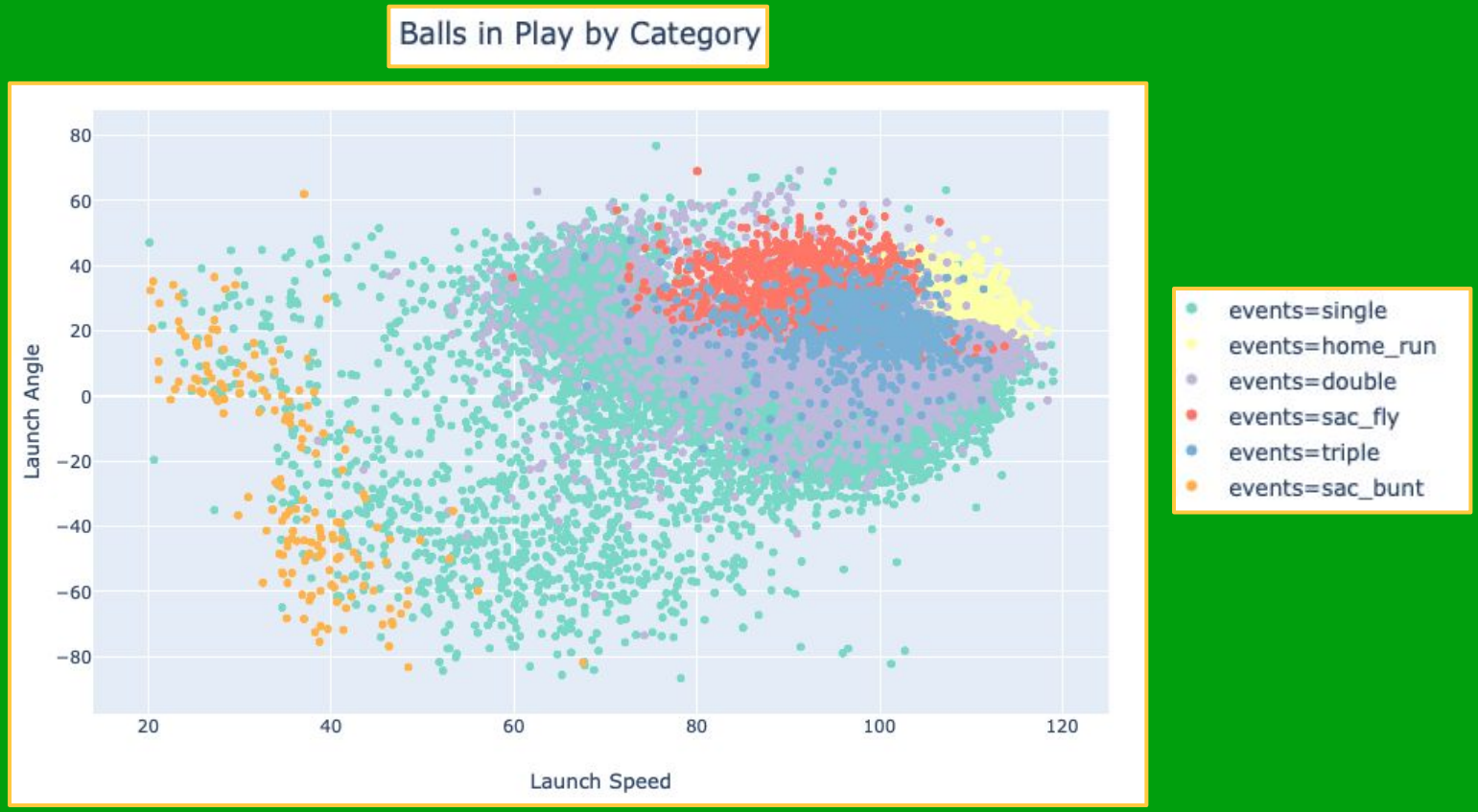
Target

- ‘Events’
  - Single
  - Double
  - Triple
  - Home run
  - Sacrifice bunt
  - Sacrifice fly

Features

- 29 continuous
  - Launch speed
  - Launch angle
  - Spray angle
- 6 categorical
  - Pitch type
  - Pitcher/batter handedness
  - Fielder alignment

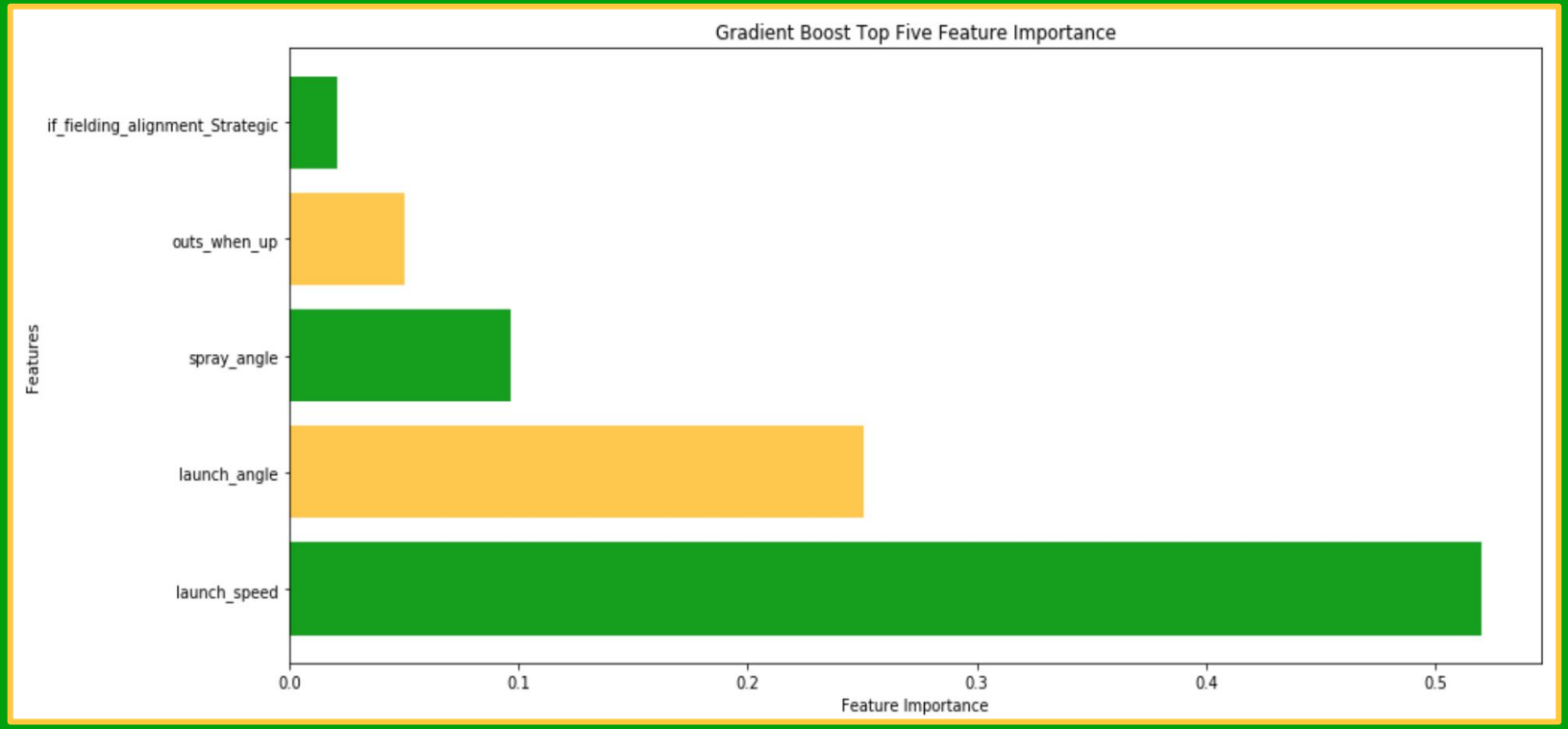
Initial Analysis



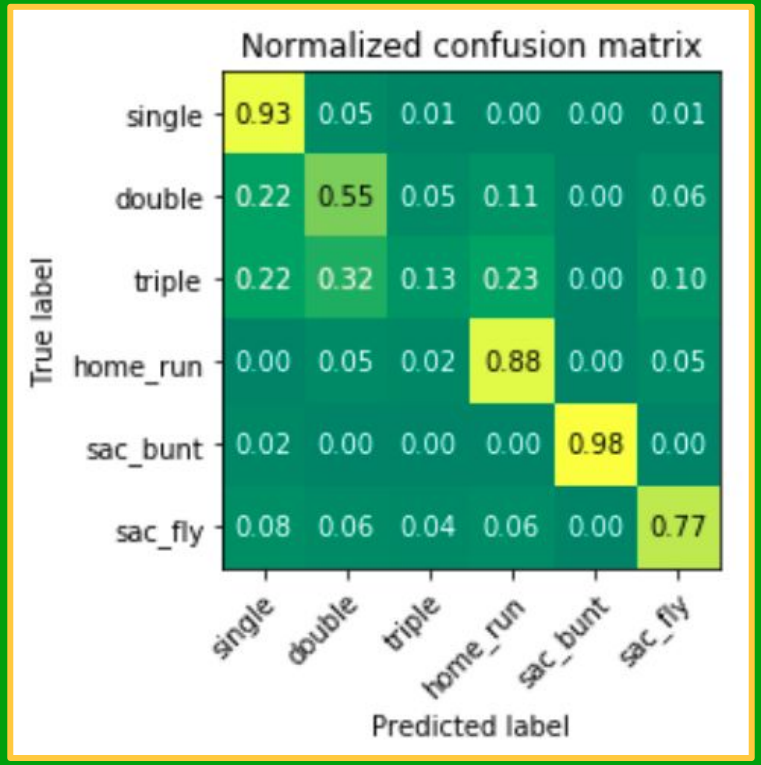
Model Performance

Model Name	Training Accuracy	Testing Accuracy
Decision Trees	80.0%	77.2%
Random Forest	83.6%	78.4%
KNN	N/A	47.6%
Gradient Boost	85.4%	**83.8%**
XG Boost	83.4%	82.9%

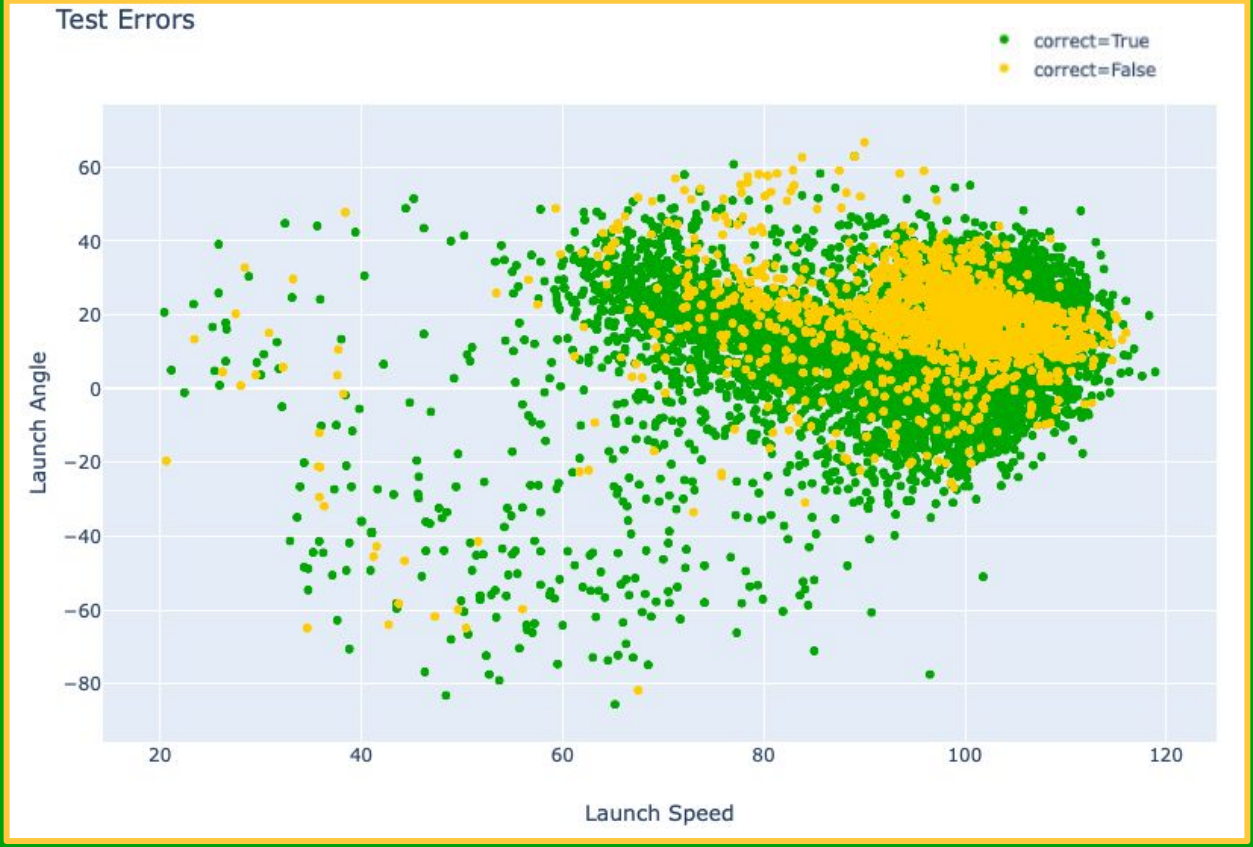
Which Features Matter the Most?



Where Does Our Model Struggle?

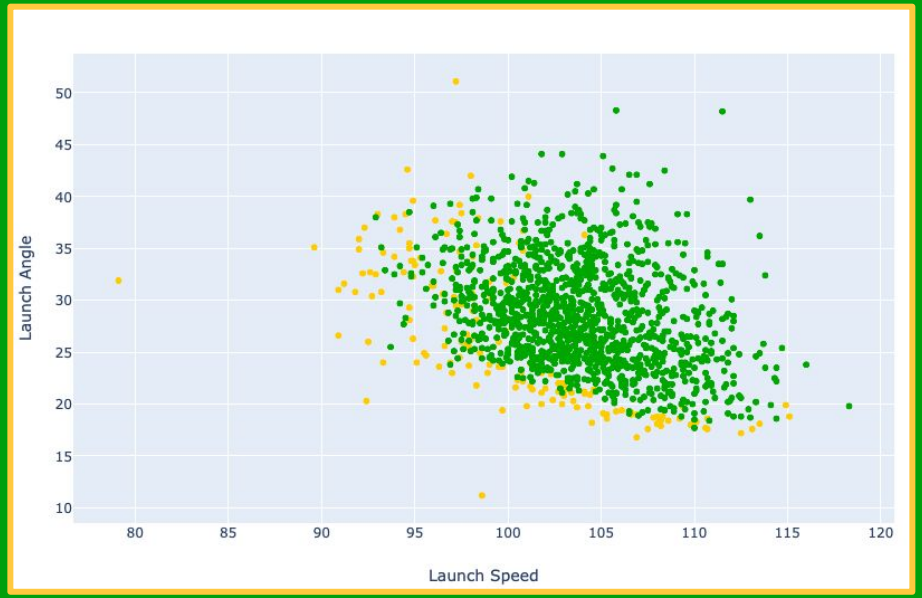


Where Does Our Model Struggle?

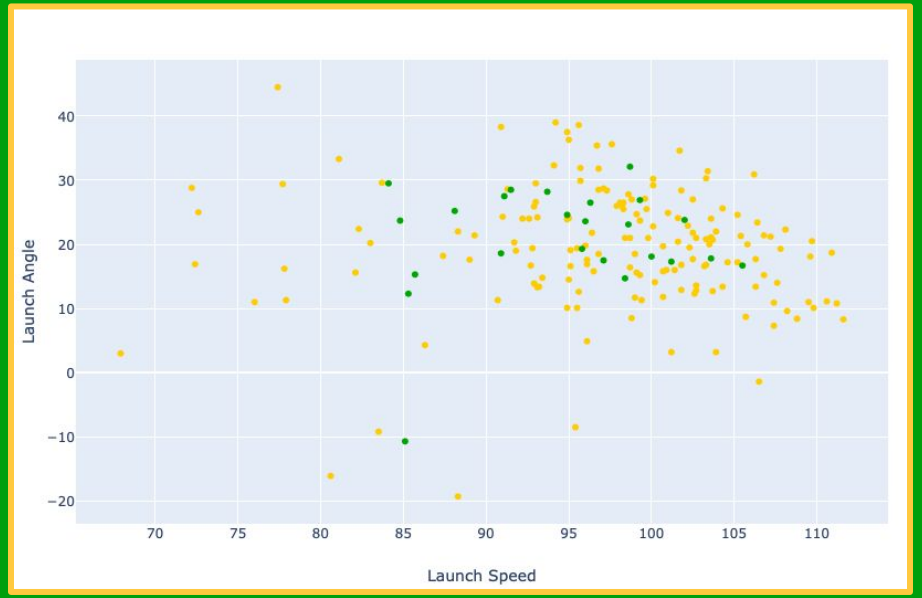


Where Does Our Model Struggle?

True Home Runs



True Triples



correct=True
correct=False

What Would Improve Our Model?

- Weather data
- Ballpark specific data
- True horizontal batting angle data
- Fielding data

Q&A