# Seamingly Accurate: a Machine Learning Approach to predict MLB player hit probability

Aimon Allouache, Akash Duseja, Ashwin Kommajesula, Justin Grotton, Manoj Desaraju, Roberto Morales

## Introduction

We developed an application that utilizes machine learning to discover players that are most likely to get a hit, ensuring it has **a higher accuracy** than current methods like Random Guessing (~60%) or Selections from Top 4 Picks (~67%).

The purpose is to get the longest possible hitting streak, as the MLB will reward the first to reach a streak of 57 games with $5.6 Million USD.

## Approaches

We built a Random Forest and Logistic Regression model using *scikit-learn*.

The Random Forest has 400 estimators, a max depth of 5 nodes, and requires at least 20 samples to be a leaf node.

The Logistic Regression model is trained with the Limited-memory BFGS solver, fits a binary problem and has 200 max iterations.

Once our models assign a hit probability for each matchup, we compare the results to see if they can agree on players for the day.

1. Both report the same player in their output and his average hit probability > .285: He becomes our pick.
2. Both report two or more players in their output with average hit probabilities > .285: “Double down” day, where we select the two with highest average hit probabilities.
3. Can’t agree on any players or their hit probabilities: “No pick day”.

A blog, which includes picks of the day and analysis is set up at [seaminglyaccurate.io/](seaminglyaccurate.io/)

## Data

Raw data for this experiment was obtained from Baseball Reference's Play Index. Models were trained using every at-bat from 2016 to 2018 (over 480,000).

The following features were computed for each batter/pitcher pair:

- At-Bat: times batter has faced pitcher
- Hits against pitcher
- Extra Base Hits against pitcher
- Home Runs against pitcher
- Strike Outs against pitcher
- Run Batted in: Runs scored as a result of batter's plate appearance
- Slugging Percentage: Total bases/At-bats
- Ballpark: Where the game will be played
- Batting Average: Hits/At-bats against the opposing pitcher

The method of identifying Beat the Streak picks involves downloading csv files from RotoWire that contain “Hot Batter Matchups”. These list batters who have performed well against the starting pitcher and matchup stats between the two.

## Results

Our Random Forest went 15/22 (68.2%) with its highest streak getting to 6. Not very good, but since we added 2018 data and re-tuned it, it has gone 9/11 (81.8%).

Our Logistic Regression model went 18/22 (81.8%) with its highest streak getting to 8.

Our program went 19/21 (90.4%) with its highest streak getting to 10 and its current streak at 8.