FAIR BALL !

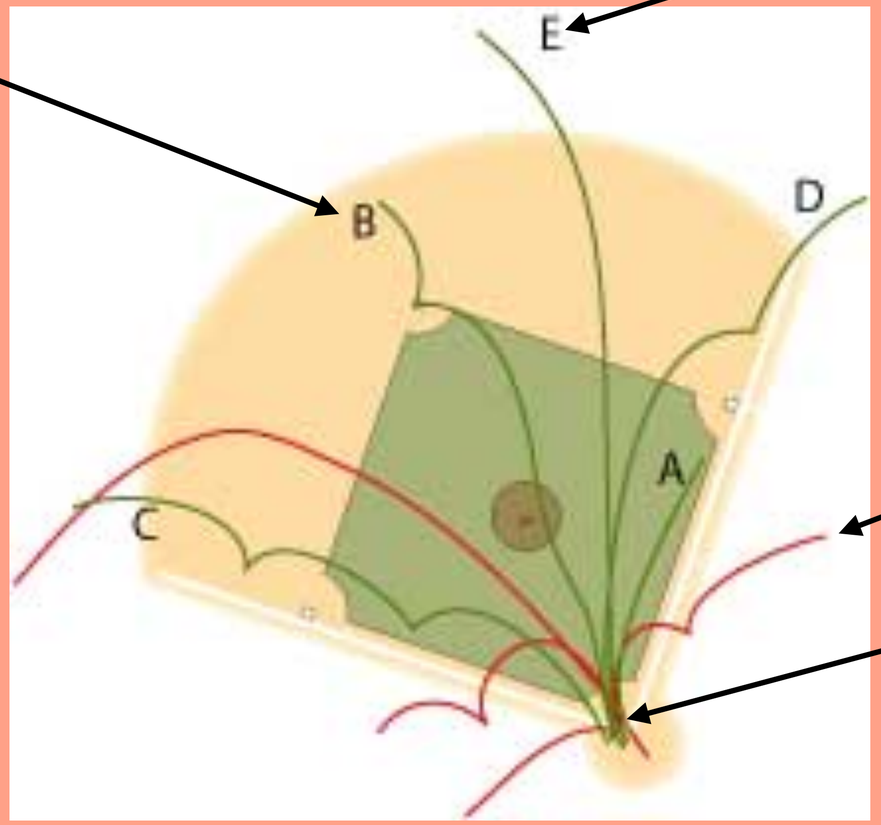
Predicting A Players BABIP

Jimmy Gardner
April 16, 2021

WHAT IS BABIP ?

BABIB IS A PLAYERS BATTING AVG ON
BALLS IN PLAY. FORMALLY , IT IS DEFINED
AS THE PERCENTAGE OF BALLS IN PLAY THAT RESULT
IN A HIT EXCLUDING THOSE THAT GO FOR HOME RUNS

Home Runs Do Not Qualify



Foul Balls Do Not Qualify

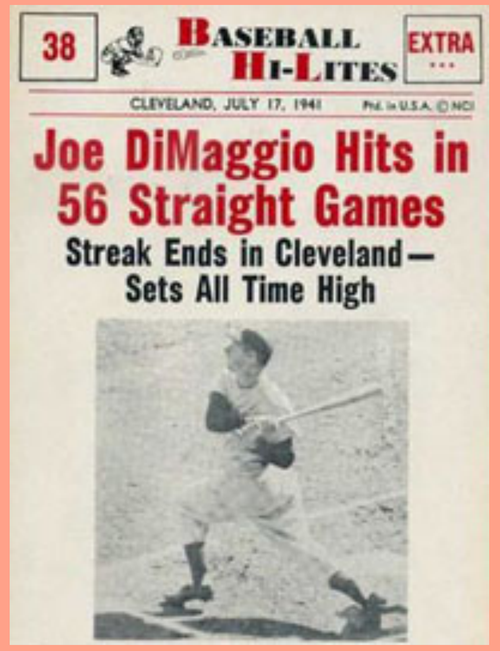
Strike Outs Do Not Qualify

MOTIVATION

IN THE 1941 SEASON JOE DIMAGGIO BROKE THE ALL TIME HITS RECORD, RECORDING A HIT IN 56 STRAIGHT GAMES

HIS BABIP THAT YEAR WAS .327 . THIS WAS 24 POINTS HIGHER THAN HIS CAREER BABIP OF .303

INTERESTINGLY ENOUGH HIS HOME RUN PERCENTAGE THAT YEAR WAS ONLY 4.8%. INDICATING THAT BEING A PLAYER THAT HITS A LOT OF HOME RUNS DOESN'T NECESSARILY MAKE YOU AN IDEAL CANDIDATE FOR GETTING A HIT ON ANY PARTICULAR NIGHT


38 BASEBALL HI-LITES EXTRA

CLEVELAND, JULY 17, 1941

Joe DiMaggio Hits in 56 Straight Games

Streak Ends in Cleveland— Sets All Time High

THE MLB OFFERS A PRIZE OF 5.6 MILLION DOLLARS THROUGH THEIR ‘BEAT THE STREAK’ APP TO ANYONE THAT CAN BEAT JOE DIMAGGIOS HITTING STREAK

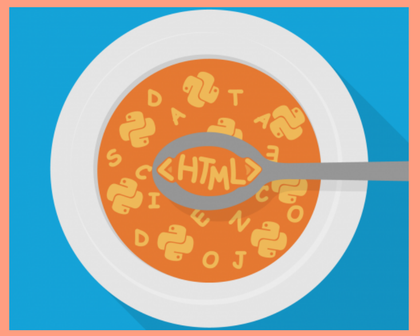
TERRY SIMS HOLDS THE ALL TIME BTS RECORD AT 49 GAMES AND WON HIM $10K FOR POSTING BEST STREAK

DATA ACQUISITION AND METHODOLOGY

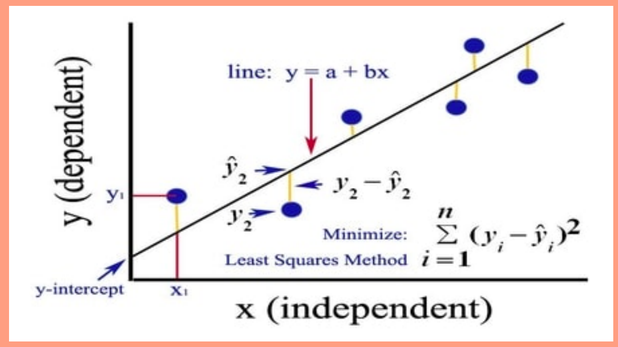
DATA SOURCES

BASEBALL REFERENCE

DATA TOOLS



REGRESSION USING LEAST SQUARES



WHAT DOES THE DATA REPRESENT ?

FEATURES

- GB%
- FB%
- FIH
- OPPPO%
- HARD%
- LD%

TARGET

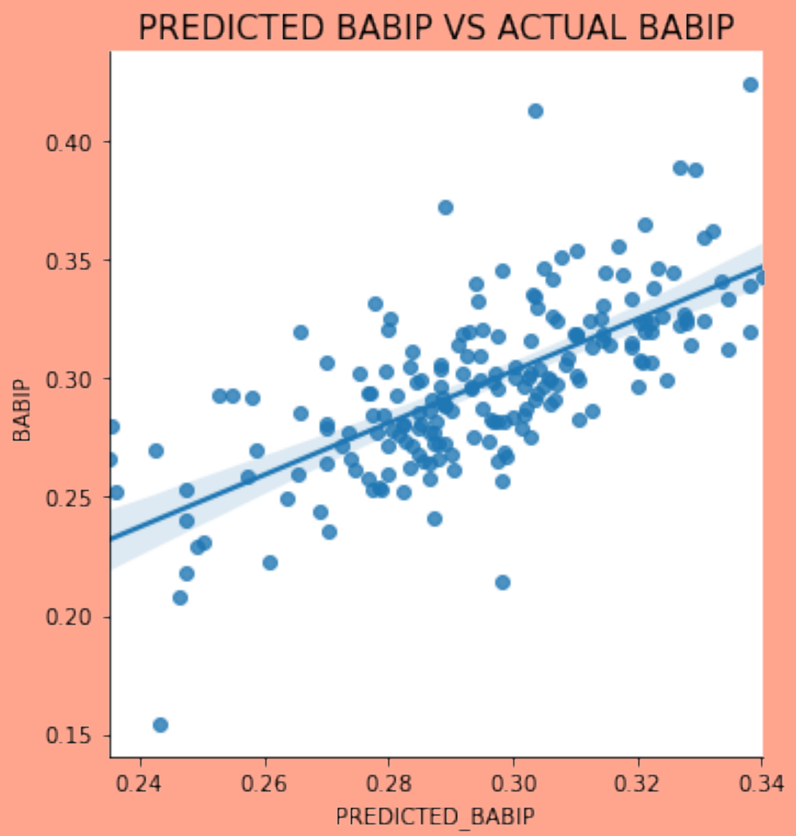
- BABIP

FORMULATION

- DATA COLLECTED FROM THE 2010-2018 SEASONS
- FEATURES ARE BASED ON THE MEAN OVER THIS TIME PERIOD
- TARGET IS BASED ON THE MEAN OVER THIS TIME PERIOD

AN INDIVIDUAL DATA POINT REPRESENTS THE COMBINATION OF FEATURES AND TARGET FOR 1 PLAYER, PLAYERS ONLY QUALIFY IF THEY HAVE A MINIMUM OF 100 AT BATS

RESULTS

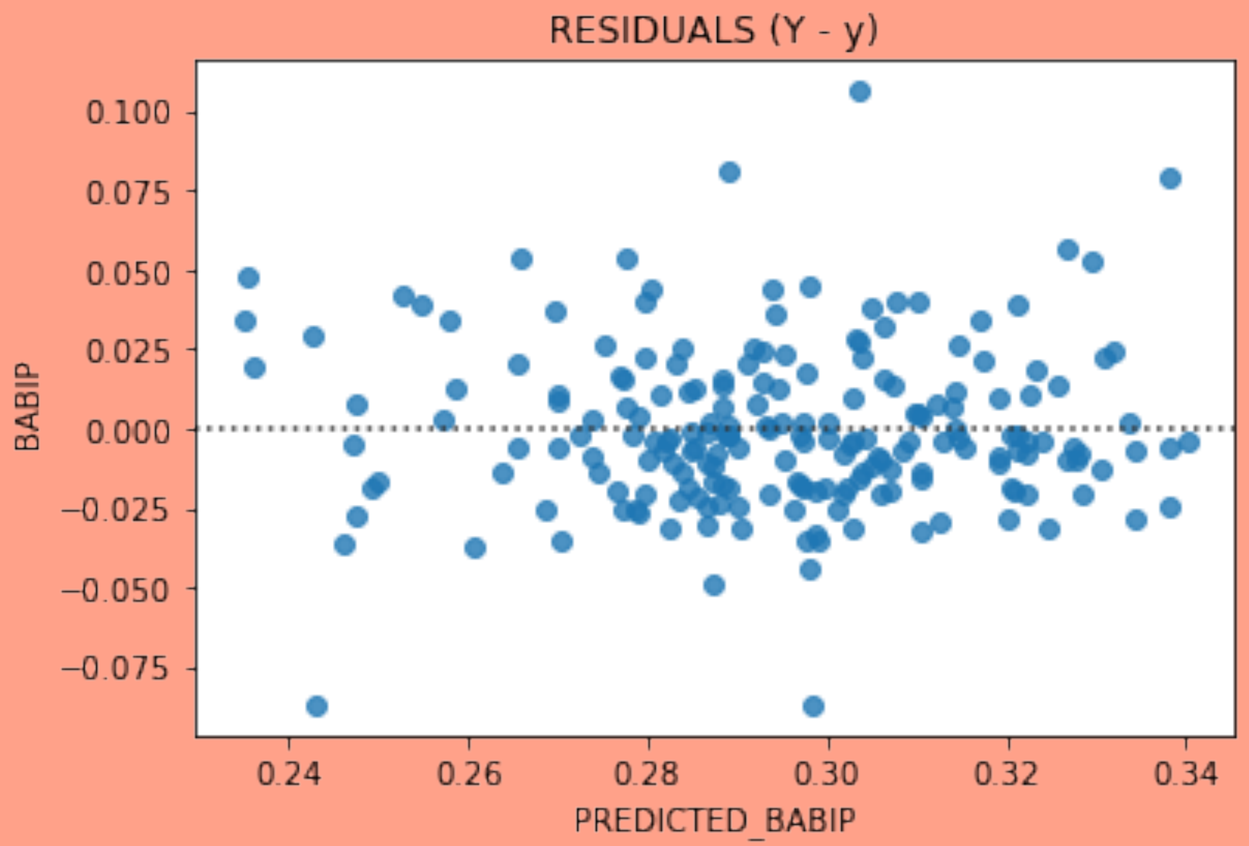


R^2 TEST (Simple Linear Regression): .467

MAE(Simple Linear Regression): 0.019

CANT EXPLAIN EVERYTHING….

LOOKING AT RESIDUALS



WHY MAE ?

MAE IS A MEASURE TO GAGE, ON AVERAGE THE MAGNITUDE OF THE DIFFERENCE
BETWEEN YOUR ACTUAL AND PREDICTED VALUES IF CHOOSING A SAMPLE AT RANDOM

HOW TO INTERPRET THIS IN THE CONTEXT OF BABIP ?

BABIP = .345
ONE BATTING POINT = .001
SO AN INCREASE OF ONE BATTING POINT WOULD MOVE BABIP FROM .345 TO .346

SO IF MAE = .019 , THAT WOULD MEAN ON AVERAGE A PLAYERS PREDICTED BABIB WAS +- 19 POINTS FROM HIS ACTUAL BABIP

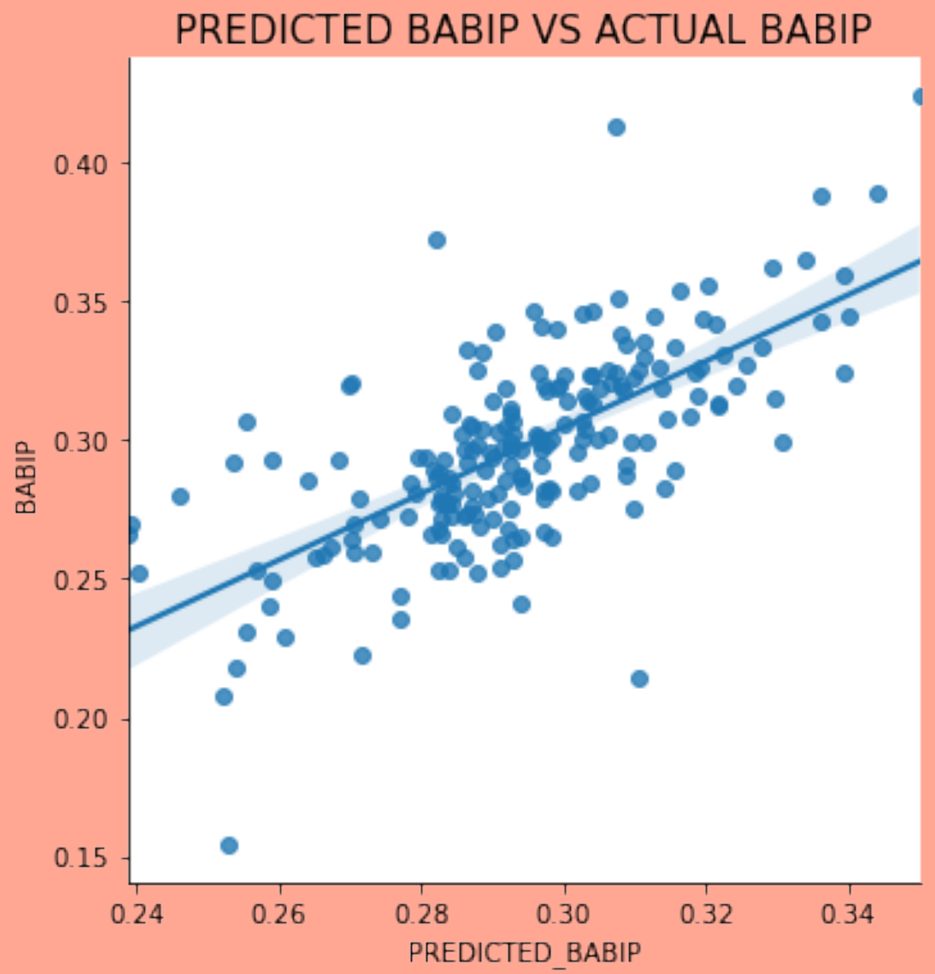
HOW DOES MAE RELATE TO A A PLAYERS PREDICTED BABIP ?

HI ! I'm Mookie Betts. I have a career BABIP of .314 and also hold a career bowling average of 204.6.

TAKE MOOKIE BETTS FOR EXAMPLE
MOOKIE HAD AN ACTUAL BABIP OF .319
AND THE MODEL PREDICTED HIS BABIP
WOULD BE .314

THIS IS A 5 PT DIFFERENCE WHICH FALLS
WELL WITHIN THE AVERAGE ERROR OF 18PTS

SPEED MATTERS !



R^2 on test: 0.439

W/OUT IFH

R^2 on test: .467

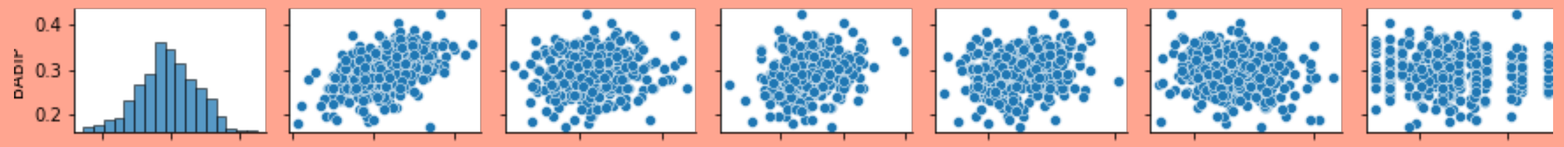
WITH IFH

MODEL PERFORMS WORSE W/OUT INCLUDING A FEATURE FOR SPEED OF PLAYER (IFH)

WRONG PATH…

YOU'D THINK QUALITY OF DEFENSE WOULD CONTRIBUTE TO THE SUCCESS OF A BATTER GETTING A HIT . BUT BASED ON A PRELIMINARY DATA EXPLORATION, DEFENSE IS LESS IMPORTANT THAN YOU WOULD THINK

UZR IS A MEASURE OF A TEAMS PARTICULAR DEFENSIVE PROWESS. AND EACH PLAYER HAS A UNIQUE SCHEDULE(EXCEPT FOR THOSE ON THE SAME TEAM) SO THIS WOULD MEAN THAT EACH PLAYER HAD A DIFFERENT AVERAGE UZR FACED OVER THE COURSE OF THE SEASON POTENTIALLY GIVING THEM AN ADVANTAGE IN GETTING A HIT



RELATIONSHIP BETWEEN BABIP AND AVGUZR
CORR = -0.032860

FUTURE WORK

- INCORPORATE EXIT VELOCITY IN THE MODEL
- COME UP WITH A BETTER DEFENSIVE FEATURE
- EXPLORE CONTRACT YEAR AND SALARY IMPLICATIONS
- EXPLORE MARKET OF PLAYER SMALL MARKET TEAMS TEND TO VALUE ANALYTICS MORE