What pitch is more likely to be batted?

Report for SMT Data Challange

Zhao Haojiao

Introduction

As we all know, pitch is the beginning of one play in a baseball game, and it decides whether the play will continue and the runner has chance to run to the base. So what factor contributes to a successful bat is a good question worth considering.

To start with, we'd like to see the basic situation of the game. From 1900 to 1903 season, the average plays per game and duration is relatively stable. But we can see that the bat rate (the proportion of plays in which the ball is hit by the bat) increases from 31.1% to 35.8%, implying that the teams are paying attention to batting success rate.

Table1: Bat Rate in each season

Season	Num of games	Num of play	Duration(ms)	Plays per game	Bat rate
1900	9	2223	1699.54	247.00	0.3108
1901	18	5024	1632.87	279.11	0.3268
1902	34	10359	1679.99	304.68	0.3513
1903	36	9627	1723.60	267.42	0.3584

Via the histogram, we can see that the frequency distribution of play duration is a thick-tail distribution, there are quite a few plays which continue for over 3000 milliseconds while most of plays last less than 1000 milliseconds (Figure1). And if we separately observe the play duration with and without bats via a box plot (Figure2), we can find that the play with bat has a significant long duration than those without bat.

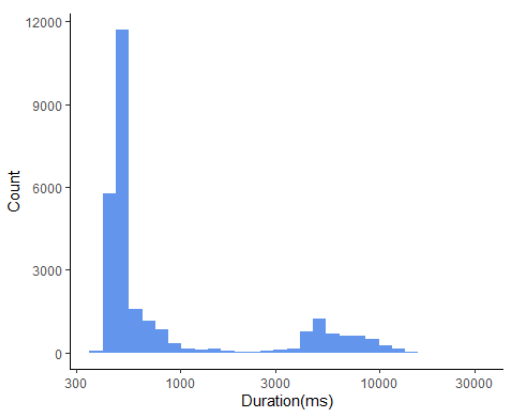

Figure1: Histogram of Play Duration

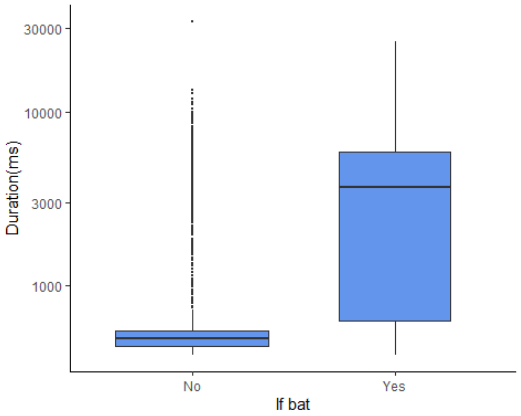

Figure2: Boxplot of Play Duration

To study which pitch is more likely to be hit, we need to extract some characteristics from the ball track data. For a pitch, two most important characteristics are speed and eagle. According to the first two record of the track.

Because the track of the ball is in a three-dimensional space, so we divide the track into horizontal and vertical way to get its angle, and for convenience we use qualitative categories to roughly character the directional information. Because the speed of the ball decides how long the ball takes to go to the home plate and how long the batter takes to prepare for bat.

Above all, from the track data of the ball, we can get three key variables:

- Initial speed describes the velocity of the ball just after the pitcher pitches (the speed between the moment of pitch and the next moment it is be recorded)
- Horizontal angle describes the Angle between the line between the pitcher and batter and the initial trajectory of the ball, it has been divided into left and right at the view of pitcher, and is represented as variable “Horizontal Direction”.
- Vertical Direction describes whether the ball goes up and down in comparison to its position at the moment it is pitched, it is divided into up and down two directions.

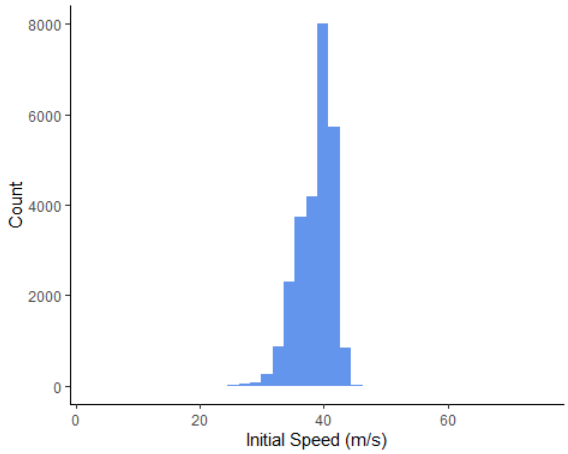

Figure 3: Histogram of ball's initial speed

Table 2: Table of ball's direction

Horizontal Direction	**Left**	**Right**
Num of pitches	8990	16979
Vertical Direction	**Down**	**Up**
Num of pitches	22413	3556

Then we'd like to look how these characteristics are distributed. Most pitches are around 30-40 m/s, a few of them are below 30m/s and over 40 m/s. In most situations, the ball goes down and goes right of the line between pitcher and batter.

Modeling and analysis

After previous work, Now we want to see whether these characteristics influence whether the ball will be hit.

So we set a logistic regression model (a special case of generalized linear model method, GLM) to connect our independent variables and dependent variable——whether the ball is hit. We put initial speed and vertical direction three independent variable into the regression function, the model is significant and the coefficients of independent variables are also significant.

Table 3: Regression Results with 3 independent variables

	Estimate	z value	z value	p-value
Intercept	-3.0851	0.2000	-15.426	$<2\times10^{-16}$
Initial speed	0.0619	0.0050	12.260	$<2\times10^{-16}$
Horizontal direction: Right	0.1131	0.0278	4.065	4.80×10^{-05}
Vertical direction: Up	-0.2726	0.0440	-6.202	5.57×10^{-10}

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

But that model is not enough, some variables haven't been considered into the model yet, we need introduce some control variables to avoid the influence of difference of individual ability and season difference. At start, we tried to put each batter into the model as dummy variable, but the model is too large and because the number of games are not large enough, the coefficients of dummy variables are not so meaningful. So we use another way to control individual ability, we calculated the batter's success rate of bat to show the batter's ability.

The results of Logistic Regression is as Table 4. The coefficient of initial speed is significantly positive, which means keep other variables constant, the ball is more likely to be hit if it is pitched at a high initial speed. The coefficients of horizontal and vertical direction are significant, indicating that keep other variables constant, the ball is more likely to be hit if it is pitched to the right than the left and the ball is more likely to be hit if it is pitched downward than upward.

Table 4: Regression Results with 3 independent variables and 2 control variables

	Estimate	z value	z value	p-value
Intercept	-4.7809	0.2163	-22.107	$<2\times10^{-16}$
Initial speed	0.0652	0.0051	12.690	$<2\times10^{-16}$
Horizontal direction: Right	0.0938	0.0283	3.314	0.0009
Vertical direction: Up	-0.2896	0.0446	-6.498	8.15×10^{-11}
Season: 1901	-0.0469	0.0592	-0.792	0.4285
Season: 1902	-0.0230	0.0548	-0.421	0.6739
Season: 1903	-0.0306	0.0552	-0.555	0.5792
Batter rate	4.6164	0.1759	26.243	$<2\times10^{-16}$

Significant codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Conclusion

After cleaning the data, deleting some abnormal records, observing distribution of variables and calculating the characteristics of ball's track, we get the fundamental information of the data and get a dataset for further modeling.

with trials and adjustments, we get a brief but significant model including 3 independent variables and 2 control variables, find that both initial speed and direction are important factor influencing the probability whether the ball can be hit by the batter. Keep other variables constant, a faster ball, pitch to the right and downward is more likely to be hit. The conclusion maybe can be used by the coaches and players to see how to perform better in attack and defense in the play of baseball game.