Practicum Report

Baltimore Orioles

**Project Scope**

**Objective:**

The Baltimore Orioles Baseball Analytics department is interested in creating a simple formula that can estimate when an American League manager removes his starting pitcher from a game. This model will mainly benefit the members of the front office and the coaching staff.

The manager has limited communication, so ideally the model would be able to be run manually by him or any member of the staff; to further emphasize that, the team has been using the codename of “clipboard model” .

**Hypothesis**

The likelihood that a generic American League manager would have removed his pitcher serves as a useful metric of how concerned the Baltimore Orioles Manager should be.

**Problem description:**

For each baseball game played, each team has an “starting” pitcher. As well as being the first one in a pitcher rotation for any game, he is expected to play most of it and giving the best possible odds to his team before giving his place to another pitcher.

When does a manager removes a starting pitcher? The decision is based on a variety of factors, such as:

* The pitcher’s performance
* His fatigue level
* The upcoming hitters he is scheduled to face
* Among others

The purpose of the project is to model the manager’s decision process for this situation. During each American League game, the team wants to estimate when the manager will remove his starting pitcher, given what has occurred to a certain point in the game as well as information about the upcoming hitters.

Ideally, this will be a simple counting system in which pre-defined points accumulate when certain events happen during a pitcher’s start, like him throwing a pitch, giving up a hit, a walk or an inning concluding.

Typically, managers pull a pitcher when they have started to perform poorly. However, the team thinks that managers try to get ahead of this performance drop and as a result the pitcher being removed typically represents the manager's best estimate of this change point.

There is no preference for reducing either the false positives or false negatives. The tradeoff would be very dependent on the game state and the availability of other members of the bullpen. For instance, if the team has a one run lead and the whole bullpen available, they would be much more aggressive in pulling the starter than if having had a large lead and the bullpen threw a bunch of innings the previous day.

**Constraints**

There will be some cases in which a starting pitcher struggles immediately in the game and is removed in the first or second inning because of poor performance. Since the decision to remove the pitcher was not due to fatigue or the hitters he is scheduled to face next, these situations will need to be accounted for.

There are also "openers" (pitchers that only throw for the very first plate appearances). MLB teams used them last year at a higher rate than they did in the past. Accounting for openers is important in properly modelling the removal of a starting pitcher from a game since the manager's logic behind an opener's removal is much different than the logic used to determine when to remove a traditional starter from the game.

**The dataset**

Baltimore Orioles Baseball Analytics department provided a dataset which contains play-by-play events, as well as information about the pitcher and the current batter. Traditional “starting” pitchers from every American League game in 2019 is included, although some further data cleaning may be needed.

**Descriptive Analysis**

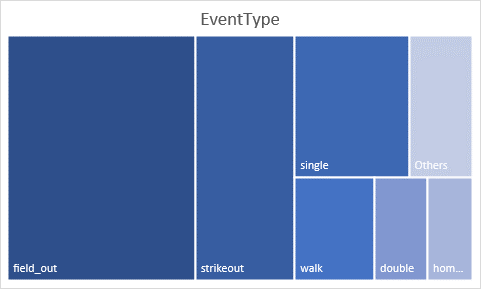
The dataset includes records for 49521 plate appearances, with 30 variables each. These include ids for the game, the starter, game, and plate appearance. In addition, there is an indicator variable, *last\_batter*, which tells whether it is the last pitch thrown by the starter or not. That variable is going to serve as our target in the classification problem.

There are 256 pitchers, which totalled 2177 different starts. In other words, 4.4% of the plate appearances (2177/49521) ended up being the last batter for the pitcher.

The dataset contained 11 missing values for the variables *next\_batter\_hand* and *opposite\_hand*. All of them correspond to the last batter for the pitcher. As they may have some importance into the decision of the manager for changing the pitcher, the correct values were researched and inputed. None of the other variables had missing values.

The distribution of the results from the plays is as follows:

**Graph 1.** The resulting *EventType* from each play on the dataset.

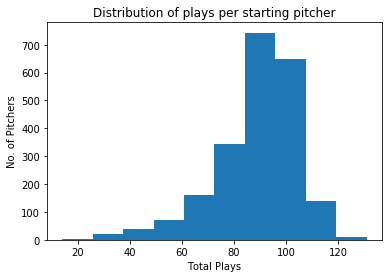


Two variables, *top\_bot* and *home\_away* have perfect collinearity. That makes total sense since the home team pitches on the top of the inning.

|  |  |  |
| --- | --- | --- |
| **home\_away** | **Away** | **Home** |
| **top\_bot** |  |  |
| **N** | 24890 | 0 |
| **Y** | 0 | 24631 |

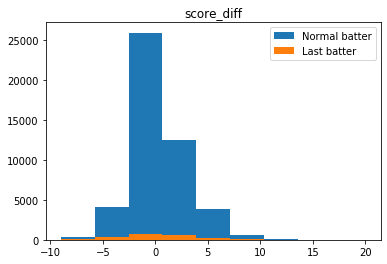
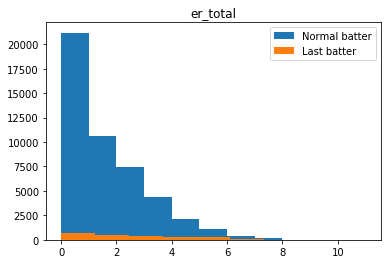
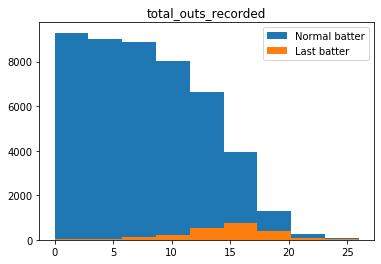
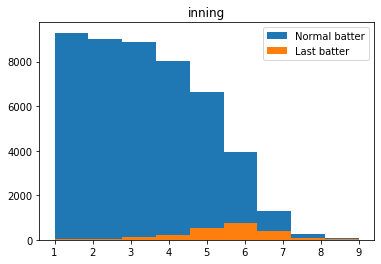
**Table 1.** *top\_bot* and *home\_away* have perfect collinearity

The number of pitches thrown by starter per game lies between 14 and 131, with a mean of 89. The distribution is as follows:



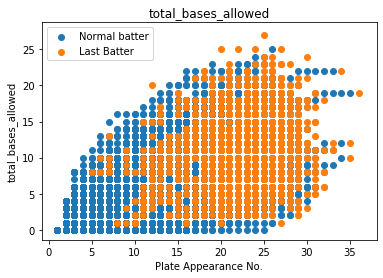
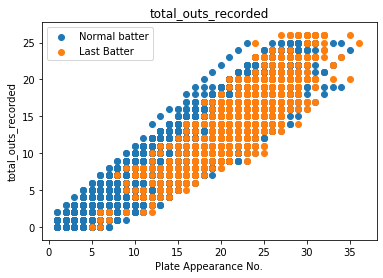
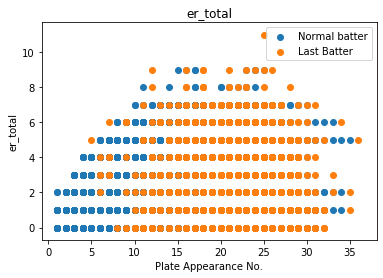
**Graph 2.** How many pitches did every starter threw?

The distribution of all numerical and categorical variables is shown. Here, some of the most notable are shown. In orange, the characteristics of the play that ended up being the last one for the pitcher. In blue, all other pitches.



**Graph 3.** Distribution of notable variables.

However, it is important to remark that the observations are not independent from each other. That is because they correspond to a row per play. Therefore, most of the numerical variables (for instance, *score\_diff*) will have the same value for several consecutive plays until there is a change. This also causes a direct relationship: with the more plate appearances, an increase on cumulative statistics. It will become more evident when plotting numerical variables vs the number of plate appearance.



**Graph 4.** Notable variables vs Number of Plate Appearance

The *opposite\_hand* (whether the next player batting side is the same as the pitcher throwing hand) may influence the decision of changing the player. 64.4% of the next players had opposing hand when changing the pitcher, but only 58.8% during “normal” plays.

Only 57% of the end of pitchers were changed at the inning that they were playing. Therefore, it may be interesting to develop new variables which are the cumulative not of the whole game, but also on the specific inning. This may be influential since pitchers have a small rest between innings; therefore, the performance may slightly reset.

Is there more pressure when playing as the Home team? At the time of change, the Home teams were performing worse than the Away teams when changing their starter.

Home team score difference: -0.0894

Away team score difference: 0.2747

**Table 2.** Mean score difference at the time of change

Finally, there is an influence on the actual batter. This may be seen as some batting orders had a higher probability of being the last one for the starter:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **last\_batter** | **0** | **1** | **All** | **percent** |
| **batter\_order** |  |  |  |  |
| **0** | 4235 | 237 | 4472 | 0.052996 |
| **1** | 6232 | 180 | 6412 | 0.028072 |
| **2** | 6016 | 216 | 6232 | 0.03466 |
| **3** | 5794 | 222 | 6016 | 0.036902 |
| **4** | 5570 | 224 | 5794 | 0.038661 |
| **5** | 5285 | 285 | 5570 | 0.051167 |
| **6** | 5011 | 274 | 5285 | 0.051845 |
| **7** | 4729 | 282 | 5011 | 0.056276 |
| **8** | 4472 | 257 | 4729 | 0.054346 |
| **All** | 47344 | 2177 | 49521 | 0.043961 |

**Table 3.** Batter Orders vs target Frequency table

**Feature Engineering**

The original dataset contained 30 variables, organized as follows:

* 4 id variables
* 15 categorical variables
* 10 numerical variables
* 1 Target variable

From these, derivated features will be created. However, it is important to note that features will be simple enough to be calculated on paper, as required for the clipboard model. This include:

An indicator variable for the most common values of *EventType*. That is, the most common results from the plays: field\_out, strikeout, single, walk, double or home\_run.

Relationship variables between the batting and pitching side of the players involved on the play.

Cumulative parameters, of both the indicator and numerical variables. For each one, these two variables were created:

* *Cum\_<var>*, which is the total <var> obtained during all the game, until the moment of the play.
* *Inning\_cum\_<var>*, which is the total <var> obtained during the current inning.

For instance, *cum\_strikeout* is the total number of strikeouts obtained until that point of the game, whereas *inning\_cum\_strikeouts* is the number of singles gained on that inning.

Plays between variables, of indicator variables. The hypothesis here is that certain events may occur more or less frequently when the player performance gets worse. For each one, these two variables were created:

* *Pas\_since\_<var>*, which is the number of plate appearances since <var> happened last.
* *has\_have\_<var>*, which tells if a <var> has been obtained at that point of the game.

For instance, *pas\_since\_walk* is how many plate appearances have passed since the last walk allowed, whereas *has\_have\_walk* becomes 1 after the first walk is allowed.

*Hot\_cold\_* variables, which are an indicator of the last result obtained by the next batter.

There are 99 variables at the end of the feature engineering.

**Missing values threatment**

As the *pas\_since\_ and hot\_cold* parameters depend onprevious results, they will have missing values at the beginning for each starter. Therefore, a value of 0 will be inputted. After all, *hot\_cold* are indicator features anyway, and *pas\_since* will be used in conjunction with the corresponding *has\_have* feature to avoid data contamination with erroneous values.

On the other side, the missing values for *next\_batter\_hand* were asked to the team Analytics Department, and they were manually inputted.