1 Introduction

In baseball, when a player does not swing at a pitch the home plate umpire must determine whether the pitch results in a ball or a called strike. Although there is a fairly well-defined strike zone that is intended to differentiate between balls and strikes, umpires are only human and must make the determination unaided. Whether the umpires are determining the pitch results base on the strike zone without being affected by other factors is an open question.

This work analyzes the pitch data with decision tree models, random forest models and logistic regression to study what factors would affect the umpires’ decision about pitching results. The trained models achieve roughly 88%-91% accuracy and precision. Unsurprisingly, the home plate entrance zone or equivalently the \((x, z)\)-coordinate of the ball entrance point have the largest impact while other features such as pitch type, release position and release velocity have certain but minor impact.

From the perspective of a baseball team, this project gives insight into the factors that affect an umpires decision of whether to call a ball or strike. Furthermore, this work hopefully give a characterization of how borderline pitches are usually decided.
<table>
<thead>
<tr>
<th>Feature</th>
<th>Type</th>
<th>tiny</th>
<th>small</th>
<th>medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pitch Type</td>
<td>Categorical</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>IDs</td>
<td>Categorical</td>
<td>Umpire</td>
<td>Umpire</td>
<td>Umpire, Pitcher and Batter</td>
</tr>
<tr>
<td>Ball-Plate Intersection Coordinate</td>
<td>Numerical</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>Ball-Plate Intersection Zone</td>
<td>Categorical</td>
<td>V</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>Strike Zone Range</td>
<td>Numerical</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>Effective Speed</td>
<td>Numerical</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>Release Position, Velocity and Spin Rate</td>
<td>Numerical</td>
<td>V</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>Pitcher and Batter Handedness</td>
<td>Categorical</td>
<td>V</td>
<td>V</td>
<td></td>
</tr>
<tr>
<td>Initial Acceleration</td>
<td>Numerical</td>
<td>V</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Selected Feature Sets

2 Data Preparation

(Data Source) We have acquired pitch and umpire data for the years 2014-2017 from Retrosheet\(^1\) and Baseball Savant\(^2\). The Retrosheet data contain the home plate umpire data while the Baseball Savant data include detailed pitch data each game within 2014-2017.

(Preprocessing) We joined the data from Retrosheet and Baseball Savant together by game ID to extend each pitch data with the home plate umpire ID. We also removed consecutive game records from the data since consecutive games have different home umpires while sharing the same game ID. This resulted in a dataset with roughly 89 attributes and 842644 instances.

(Splitting Data) We left out pitch data from year 2017 as Test Dataset I and further sampled 10% of the pitch data from year 2014-2016 as Test Dataset II. The remaining 90% of the data are used for training.

3 Training and Feature Selection

To study the impact of different factors in the pitch data, we selected 3 sets of features and trained one decision tree model, one random forest model and one logistic regression model for each feature set using the Orange framework\(^3\) as in Figure 2. The hyperparameters were turned using 10 fold cross-validation on the training data.

Due to time and computation resource limit, we were not able to explore the importance of each factor separately. Hence we only studied three sets of features summarized in Figure 1. The list of all features is given on the website.

4 Evaluation

For each of the tiny, small and medium feature set and each kind of model, Figure 3 depicts the learning curves and Figure 4 presents the accuracy and the precision result. The precision is calculated by taking weighted average over both outcomes.

\(^1\)Retrosheet: [http://www.retrosheet.org/](http://www.retrosheet.org/)

\(^2\)Baseball Savant, which provides an interface to MLB Statcast ([http://m.mlb.com/glossary/statcast](http://m.mlb.com/glossary/statcast)): [https://baseballsavant.mlb.com/](https://baseballsavant.mlb.com/)

\(^3\)https://orange.biolab.si/
From the evaluation result, we see that the learning curves for both accuracy and precision are mostly flat except for the logistic model on the tiny feature set. This shows that most models can predict the outcome fairly well and do not overfit training data.

Overall, the models trained with the small feature set perform the best. Including batter ID and pitcher ID in the medium feature set decreases performance. Excluding the zone attribute to form the tiny feature set stopped the logistic regression model from working, indicating that the logistic model is unable to determine the outcome using plate_x and plate_z alone.

5 Analysis, Discussion and Future Work

In this section, we investigate two models trained with the small feature set and study the importance of each feature. In the following models, the categorical attribute zone represents the number of the strike zone grid and its surrounding area:

Figure 6 shows the Nomogram of the logistic regression model. From the Nomogram, we can see that zone has the largest impact on the resulting probability. The second most impactful features include plate_z and plate_x, which can actually be seen as where zone is derived from. After that, pitch_type and release_speed have certain impact on the resulting probability.

Figure 5 shows the first 3 layers of the decision tree model. In (a), the decision tree first splits on zone. Following in (b), (c) and (d), the model further splits on plate_z or plate_x to determine the borderline cases.

---

4Baseball Savant Statcast Search: [https://baseballsavant.mlb.com/statcast_search](https://baseballsavant.mlb.com/statcast_search)
Figure 3: (a) The learning curve for model accuracy and (b) the learning curve for model precision. The models were trained using the tiny, small and medium feature sets from left to right and were tested on Test Data II.

<table>
<thead>
<tr>
<th></th>
<th>Tiny</th>
<th>Small</th>
<th>Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test Data I</td>
<td>0.669 0.669 0.91 0.906 0.669 0.885 0.911 0.908 0.669 0.893 0.895 0.906</td>
<td>0.661 0.661 0.908 0.911 0.661 0.874 0.908 0.911 0.661 0.879 0.893 0.903</td>
<td></td>
</tr>
<tr>
<td>Test Data II</td>
<td>0.448 0.38 0.911 0.906 0.448 0.884 0.911 0.907 0.448 0.892 0.894 0.905</td>
<td>0.436 0.436 0.908 0.911 0.437 0.873 0.908 0.911 0.437 0.878 0.892 0.902</td>
<td></td>
</tr>
</tbody>
</table>

Figure 4: Prediction accuracy and precision on Test Data I and Test Data II.
In (b) and (c), since the regions 1, 2, 8 and 9 are at the top and the bottom of the strike zone respectively, the decision tree tests plate_z to determine whether the z coordinate is out of border. In (d), the decision tree is splitting on both plate_z and plate_x since regions 11 and 12 both contain a corner.

To sum up, our models show that the features zone, plate_z and plate_x can largely characterize the pitching results. The logistic model further predicts that pitch_type and release_speed also affect pitching outcome.

Unfortunately, we did not have enough time to experiment with more feature sets to obtain a finer understanding on how each feature affects the pitch. Moreover, we were unable to prune the decision tree model enough to understand how the remaining features took place in the process. We would like to close this loophole in the future.

### A Work Allocation

Daniel Feltey  
data collection, feature selection and model tuning

Spencer Florence  
infrastructure setup, preprocessing and model tuning

Shu-Hung You  
model evaluation and analysis
Figure 6: Nomogram for the logistic regression model.