Amanda Kuznecov
anr431

Homework 7

DS-GA 3001.015
April 14, 2021

Question 1: Expanded Baseline Model

Please note that rows with no events were removed from the dataset. Also events 'intent_walk','ejection','game_advisory','wild_pitch' and 'pickoff_*' were removed as they were not actual pitches to the batter. Additionally, rows where 4 balls were recorded were removed.

Part a: bbrate model
(i) Intercept: 0.0332, Coeff: 0.5537
(ii) MSE: 4.610×10^{-4}

Part b: krate model
(i) Intercept: 0.0602, Coeff: 0.7447
(ii) MSE: 1.996×10^{-3}

Question 2: Called Strike Model

(a) Logit model

(b) List of Features:

*Note that missing numerical pitching data (such as velocities, positions, etc.) were imputed based on complete case mean estimates.

- **LeftMiss(plate_x, 0.75):** linear penalty for missing strike zone from left, with threshold of 0.75ft.
- **RightMiss(plate_x, 0.75) :** linear penalty for missing strike zone from right, with threshold of 0.75ft.
- **HighMiss(plate_z, sz_top, 0.25):** linear penalty for missing strike zone from top, with threshold of 0.25ft.
- **LowMiss(plate_z, sz_top, 0.25):** linear penalty for missing strike zone from bottom, with threshold of 0.25ft.
- **dist_mid:** distance from pitch location (when crossing the plate) to the centre of the strike zone.
- **plate_x:** directly from data.
- **outside_pitch_platex:** interaction term between plate_x and indicator variable for whether pitching hand is opposite of where the batter is standing.
- **b0_s2:** indicator variable for count of 0 balls, 2 strikes.
- **b1_s2:** indicator variable for count of 1 ball, 2 strikes.
- **b2_s2:** indicator variable for count of 2 balls, 2 strikes.
- **b3_s0:** indicator variable for count of 3 balls, 0 strikes.
- **b3_s1:** indicator variable for count of 3 balls, 1 strike.
- **num_mv:** indicator variable for missing numerical pitching data (ie. velocity, position, etc.)

(c) A logit model was initially developed with strike zone features, because it was assumed that these features would be very important in determining whether a pitch is called a strike. Features were engineered to linearly penalize pitches outside the strike zone. All features showed high significance and coefficients in the negative direction, which makes sense since the further we are outside the strike zone, the more likely the pitch is a ball and will not be called a strike.

Based on research by Brian Mills (University of Florida), we know that the actual strike zone is more concentric rather than a rectangular shape. This justifies the reasoning for adding 'dist_mid' to the model. Next, 'plate_x' and 'outside_pitch_platex' were included to provide the model with more information regarding horizontal positioning of the pitch. Additional research by Brian Mills suggested that the actual strike zone is asymmetrical and depends on the side where the batter is standing, and handedness of the pitcher. For example, if a right-handed pitcher is pitching to someone standing on the right side of home plate (from the catcher's perspective), the pitcher will be making an outside pitch. Outside pitches are more likely to be called strikes than inside pitches. Since batters are treated differently depending on their handedness, an interaction term, 'outside_pitch_platex' was engineered between an indicator variable for whether the pitcher's and batter's arms are on opposite sides, and the raw horizontal pitch coordinate. As expected, 'plate_x' and 'dist_mid' have negative coefficients since the further from the centre of the strike zone, the more likely the pitch is not a strike. The coordinate on 'outside_pitch_platex' was positive since outside pitches have a larger radius for being considered in the strike zone.

Finally, ball and strike counts were included as indicator variables in the last model iterations. These features significantly improved the model, however it was notable that low ball/strike counts did not provide much predictive power since there is more uncertainty about what the next pitch outcome will be when the counts are low. For this reason, low ball/strike counts were removed as features. We would expect that a 3-0 count would likely lead to a strike next, and an 0-2 count would likely lead to a ball next. However, an interesting finding which contrasted with DFL's findings is that these counts actually lead to the opposite. 3-0 counts had a negative coefficient, indicating that they are predictive of a ball next, and 0-2 counts had a positive coefficient, indicating that they are predictive of a strike next. This could possibly be due to the fact that a really bad pitcher (3-0 count) would continue to perform poorly, leading to another ball, and a really great pitcher (0-2 count) would continue to perform well, leading to another strike. Interestingly enough, after adding these ball/strike count features to the model, possible complete quasi-separation was achieved, where 0.32% of observations were perfectly predicted. This means that some of the included features were strong enough to be highly predictive on their own.

Additionally, as noted above, missing numerical pitching data (velocities, positions, etc) were imputed based on pitches where full data was available. The 'num_mv' feature was used to record which pitches this was done for. This feature was included in the final model and showed predictive power. The coefficient is negative, meaning that when these numerical values are missing in the raw data, there is likely to have been no called strike.

An attempt was made to include indicator variables for pitch type, but this did not add as much predictive power as the other features. It is interesting to note that fastballs and pitch outs never resulted in called strikes in our dataset. This is likely due to the fact that fastballs are always hit, because batters know they are good pitches to swing at. For pitch outs, these pitches result in balls since they are horrible throws.

Initial modeling yielded an adjusted R^2 of 0.5744, and the final model yielded an ajusted R^2 of 0.8760, showing significant improvement. The most significant features were counts where there were 2 strikes ('b0s2', 'b1s2', 'b2s2') and 'num_mv'.

(d) Out-of-sample Brier score: 0.02427

Question 3: Swinging Strike Model

(a) Logit model

(b) List of Features:

*Note that missing numerical pitching data (such as velocities, positions, etc.) were imputed based on complete case mean estimates.

- **LeftMiss(plate_x, 0.75):** linear penalty for missing strike zone from left, with threshold of 0.75ft.
- **RightMiss(plate_x, 0.75) :** linear penalty for missing strike zone from right, with threshold of 0.75ft.
- **HighMiss(plate_z, sz_top, 0.25):** linear penalty for missing strike zone from top, with threshold of 0.25ft.
- **LowMiss(plate_z, sz_top, 0.25):** linear penalty for missing strike zone from bottom, with threshold of 0.25ft.
- **dist_mid:** distance from pitch location (when crossing the plate) to the centre of the strike zone.
- **outside_pitch_platex:** interaction term between plate_x and indicator variable for whether pitching hand is opposite of where the batter is standing.
- **b0_s2:** indicator variable for count of 0 balls, 2 strikes.
- **b1_s2:** indicator variable for count of 1 ball, 2 strikes.
- **b2_s2:** indicator variable for count of 2 balls, 2 strikes.
- **b3_s2:** indicator variable for count of 3 balls, 2 strikes.
- **vx0:** directly from data.
- **vy0:** directly from data.
- **vz0:** directly from data.
- **release_speed:** directly from data.
- **release_spin_rate:** directly from data.
- **pitch_type_mv:** indicator variable for missing pitch type.
- **pitch_type_FF, SL, CH, CU, KC, FC, FS, KN, EP, FO:** indicator variables for these pitch types.
- **outs_when_up2:** indicator variable for whether team has 2 outs pre-pitch.
- **num_mv:** indicator variable for missing numerical pitching data (ie. velocity, position, etc.)
- **sz_top:** directly from data.

- **sz_bot:** directly from data.

(c) A logit model was initially developed with features similar to the ones used in the final model for called strikes. Pitch velocities were also included because it was assumed that batters would likely swing at fast-moving pitches. As expected, most of the included features provided significant predictive power. Features that linearly penalized pitches outside the strike zone all had negative coefficients just as before, except 'lower_sz_miss' which now has a positive coefficient. This could mean that batters tend to swing at pitches that may be a bit below the strike zone. Also interesting is that 'dist_mid' and 'plate_x now have positive coefficients as well. This could be interpreted as batters attempting to swing at pitches even when they are not close to the centre of the strike zone, for example, reaching further out to get to the pitch. This could indeed lead to a swinging strike because they could not get a piece of the ball, to get a good hit on it, although they extended to reach it with their bat. Most ball/strike counts are of importance, except for counts where number of strikes are low. For instance, in our dataset, nearly all 3-1 counts result in no swinging strike next, and all 3-0 counts never result in a swinging strike. All ball counts where number of strikes were less than 2 were removed as features. Pitch velocities 'vx0' and 'vy0' both had negative coefficients, meaning that faster horizontal velocities do not lead to swinging strikes, and faster velocities toward the batter do lead in swinging strikes (the negative component is due to the fact that it is measured from the catcher's perspective). Pitch velocity 'vz0' has a positive coefficient, indicating that pitches with increased vertical velocity result in swinging strikes.

Improving model performance for swinging strikes proved to be challenging, so somewhat of a mixed bag of additional features were added to test changes in the adjusted R^2 values. Various indicator variables for pitch types were added, however it was noted that sinkers and screwball pitches did not add much value, so these were removed from modeling efforts. All pitch types had positive coefficients, indicating that they aid in predicting swinging strikes. Feature 'pitch_mv' was also added indicating when pitch type information was missing. The feature had a positive coefficient, suggesting that when pitch type is missing, the pitch likely leads to a swinging strike. In addition to the velocity features, 'release_speed' was also added, which had a negative coefficient as expected since it is measured in the catcher's perspective. Featuer 'release_spin' had a positive coefficient, indicating that more spin on the pitch caused for swinging strikes - likely due to the directional change of the ball. Feature 'outs_when_up2', when the batting team had two outs pre-pitch, showed a positive coefficient, suggesting that when teams have two outs, they are likely to swing at a pitch because it is their last chance to get a good hit, however this often results in a strike. Raw 'sz_bot' and 'sz_top' features were also added - both providing some predictive power, with higher strike zones resulting in swinging strikes, and lower strike zones not. The addition of some of these features caused the predictive power attributed to 'outside_plate_x' to reduce quite a bit since some of its information is already baked into other features, and so, it was removed from the final model. Additionally, as noted above, missing numerical pitching data (velocities, positions, etc) were imputed from pitches where full data was available. The 'num_mv' feature was used to record which pitches this was done for. This feature was included in the final model and provided some predictive power.

Initial modeling yielded an adjusted R^2 of 0.3516, and the final model yielded an adjusted R^2 of 0.3624, showing significant improvement. The most significant features were all included ball counts (where 2 strikes were recorded). Features 'dist_mid', slider and changeup pitch types were also of great importance.

(d) Out-of-sample Brier score: 0.09126

Question 4a: Improving the Baseline: bb-rate

(i) Ordinary Least Squares

(ii) List of Features:

- **bbrate_prev:** walk rate from pitcher's previous season.
- **xtr1B_rate_prev:** rate of "walks missed out on" from pitcher's previous season.
- **ball_rate_prev:** number of balls per batter faced in pitcher's previous season.
- **b3_count_rate_prev:** frequency of 3 ball counts per batter faced in the pitcher's previous season.

(iii) An OLS model was developed to predict 'bbrate'. Using a suggestion from DFL, the concept of extra strikes was attempted by using the called strikes model to predict the probability of a called strike for each pitch, and then using this to compute net extra strikes for when the batter had 3 balls in their ball/strike count. This rate was computed as a pseudo 'walks missed out on'. It was expected that pitchers with positive net extra strikes would see an increase in walkout rate, and therefore, we would expect this feature to have a positive coefficient, which is observed in the model. 'ball_rate_prev', the rate of balls per batter faced, was computed for each pitcher for each season. It was expected that knowing the pitcher's previous seasonal rate of balls handed out per batter faced would give us a better sense of how many balls to expect the pitcher to hand out in the next year, and therefore give us a better idea of how many walkouts would be issued. When added to the model, this feature had a positive coefficient as expected, because more balls means the potential for more walkouts issued. The feature actually provided more predictive power than the rate of walks missed out on. Next, it was assumed that knowing how many three ball counts that the pitcher had in his previous season per batter faced, would allow us to better predict a walk, because we would have all the information leading up to the walk event itself. As expected, adding the feature 'b3_count_rate_prev' had a positive coefficient and improved performance of the model.

Initial modeling from our baseline in Question 1 yielded an adjusted R^2 of 0.293, and the final model yielded an adjusted R^2 of 0.325. The most significant features were 'bbrate_prev' and 'balls_rate_prev'.

(iv) MSE: 4.487×10^{-4}

Question 4b: Improving the Baseline: k-rate

(i) Ordinary Least Squares

(ii) List of Features:

- **krate_prev:** strikeout rate from pitcher's previous season.
- **s2_count_rate_prev:** frequency of 2 strike counts per batter faced in the pitcher's previous season.

(iii) An OLS model was developed to predict 'krate'. Improving model performance from the initial baseline, which only used 'krate_prev', proved to be challenging because 'krate' is most highly correlated with strikeout rate from the pitcher's previous season versus any other feature. Using a suggestion from DFL, the concept of extra strikes was attempted by using the called strikes model to predict the probability of a called strike for each pitch, and then using this to compute net extra strikes for when the batter had 2 strikes in their ball/strike count. This rate was computed as a pseudo 'extra strikeouts'. It was expected that pitchers with positive net extra strikes would see a decrease in strikeout rate, and therefore, we would expect this feature to have a negative coefficient. The expected coefficient sign was observed, however the feature was not statistically significant, and so it was removed from the model. The rate of strikes per batter faced was computed for each pitcher for each season as 'strikes_rate_prev'. Unfortunately this feature was not statistically significant either, although the coefficient was positive as expected. Next, it was assumed that knowing how many two strike counts that the pitcher had in his previous season per batter faced, would allow us to better predict a strikeout, because we would have all the information leading up to the strikeout event itself. As expected, adding the feature 's2_count_rate_prev' had a positive coefficient and improved performance of the model.

Initial modeling yielded an adjusted R^2 of 0.550, and the final model yielded an adjusted R^2 of 0.551. The most significant feature was still 'krate_prev' by a landslide.

(iv) MSE: 1.968×10^{-3}