1.A

i) Intercept: 0.036375
bbrate_prev: 0.526399

ii) MSE: 0.00045981374978 09986

1.B

i) Intercept: 0.060182
krate_prev: 0.744610

ii) MSE: 0.002001949418894451

2

a) I used a logit model for the called strikes predictor

b) The following features were used:
   a) Zone: This feature is the location of the ball when it crosses the plate area. The lower zones generally correspond to a strike
   b) HighMiss (0.25 Threshold): As noted in lecture, this variable looks at the height of the ball when it crosses the plate area (towards the high end) with some threshold of variability to show the uncertainty of the zone
   c) LowMiss (0.25 Threshold): As noted in lecture, this variable looks at the height of the ball when it crosses the plate area (towards the low end) with some threshold of variability to show the uncertainty of the zone
   d) LeftMiss (0.6 Threshold): As noted in lecture, this variable looks at the x-axis of the ball when it crosses the plate area (towards the left end) with some threshold of variability to show the uncertainty of the zone
   e) RightMiss (0.6 Threshold): As noted in lecture, this variable looks at the x-axis of the ball when it crosses the plate area (towards the right end) with some threshold of variability to show the uncertainty of the zone
   f) Strikes: The current number of strikes before the pitch
   g) Balls: The current number of balls before the pitch
   h) Runners: A variable that looks at the number of runners that are on the field for the batting team
   i) Stand: The stance of the batter
   j) Plate_x: Where the ball crosses the plate in the x-dimension
   k) Plate_z: Where the ball crosses the plate in the z-dimension
   l) sz_top: The top edge of the strike zone
   m) sz_bot: The bottom edge of the strike zone
   n) Release_pos_x: Where the pitcher lets go of the ball in the x direction
   o) Release_speed: The speed when the pitcher lets go of the ball
   p) Pitch_Type: What type of pitch the pitcher threw

c) I originally started off with mostly looking at metrics to show where the ball crossed the plate (plate x/z, sz top/bot, zone) and then, as done in Lecture 9, I added in the high/low/left/right miss variables. I tuned the thresholds to see which one would have the lowest brier score and while the High and Low Miss variables kept the 0.25, I found that a 0.6 threshold for the left and right misses were slightly better than the 0.75 from lecture. Instead of adding a variable for the count, I felt that the number of strikes and balls were a good enough indicator so I used those instead. Additionally, similar to the count metric, I created a variable to see the number of runners on base, hypothesizing that the number of runners might also influence close calls (e.g. if the bases are loaded, the umpire might be hesitant to give a walk on a 3-1 pitch).

   For the next few metrics, I decided to run a forward/backward selection via Logistic

Regression and Random Forest to see what other variables might be significant and noticed that release metrics seemed to help the model. From the release metric, I figured that the type of pitch might also influence called strikes. One of the variables I tried to create, which I was not able to, tried to measure the amount of break after the halfway point of the pitch - a way to mark the pitcher's deception. I wanted to test if the pitcher's deception had any influence on the way that the umpire would call a strike. As a proxy, I decided to use pitch type instead. Pitch type might be able to somewhat measure the effect of deception for pitches like a cutter which starts off similar to a 4-seam fastball and moves closer to the plate.

For this model, the directional miss variables had the highest z scores, as well as the count variables, and the ball position upon crossing the plate.

d) Brier Score: 0.057536

3

a) I used a logit model for the swinging strikes predictor

b) The following features were used:

   a) Zone: This feature is the location of the ball when it crosses the plate area. The lower zones generally correspond to a strike

   b) HighMiss (0.25 Threshold): As noted in lecture, this variable looks at the height of the ball when it crosses the plate area (towards the high end) with some threshold of variability to show the uncertainty of the zone

   c) LowMiss (0.25 Threshold): As noted in lecture, this variable looks at the height of the ball when it crosses the plate area (towards the low end) with some threshold of variability to show the uncertainty of the zone

   d) LeftMiss (0.6 Threshold): As noted in lecture, this variable looks at the x-axis of the ball when it crosses the plate area (towards the left end) with some threshold of variability to show the uncertainty of the zone

   e) RightMiss (0.6 Threshold): As noted in lecture, this variable looks at the x-axis of the ball when it crosses the plate area (towards the right end) with some threshold of variability to show the uncertainty of the zone

   f) Strikes: The current number of strikes before the pitch

   g) Runners: A variable that looks at the number of runners that are on the field for the batting team

   h) PA (Pitcher Advantage): An indicator variable for if the pitcher throws the same side as the batter stands - generally a favorable matchup for the pitcher

   i) Plate_x: Where the ball crosses the plate in the x-dimension

   j) Plate_z: Where the ball crosses the plate in the z-dimension

   k) sz_top: The top edge of the strike zone

   l) sz_bot: The bottom edge of the strike zone

   m) Release_pos_x: Where the pitcher lets go of the ball in the x direction

   n) Release_pos_z: Where the pitcher lets go of the ball in the z direction

   o) Release_speed: The speed when the pitcher lets go of the ball

   p) Pitch_Type: What type of pitch the pitcher threw

   q) pfx_x: The horizontal movement of the pitch

   r) az: acceleration in the z direction at 50 feet from the catcher

   s) ax: acceleration in the x direction at 50 feet from the catcher

c) I started this one similar to the called strikes model above and used all the same features. From there, I removed the ones that had high z-scores such as the “balls” feature. I again ran a forward/backward selection via Logistic Regression and Random Forest to see what other variables might be significant and noticed that acceleration metrics seemed to help the model. I again want to create a variable to measure the amount of break after the halfway point of the pitch - but was not able to successfully figure out that variable. Again, pitch type could be a proxy for that variable so I included that in the final model.

   The most significant features in this model were the top and bottom of the strike zone, the number of strikes, and certain pitches like the four seamer and the sinker.

d) Brier Score: 0.091115

4a

i) I used an OLS model for predicting bbrate

ii) The following features were used:

   i) bbrate_prev: Last year's bbrate

   ii) bfp_prev: Last year's batters faced per game

   iii) fps_prev: Last year's proportion of first pitch strikes

   iv) maxspeed_prev: The fastest pitch thrown by the pitcher in the previous season (based on release speed)

iii) For this part, and the krate part, I created a couple of features that I would think have some part to play in predicting strike rate and walk rate. I created variables from the previous parts to see the called strike percentage and the swinging strike percentage (swinging strikes vs called strikes). I also looked at speed based metrics, hypothesizing that certain pitchers that throw too hard might sacrifice some control - potentially having a higher walk rate than the average pitcher. I also create variables to see how often the pitcher's first pitch was a strike or a ball and a variable to see how many batters a pitcher faced per game. I assumed that relievers don't come on to walk batters, and the pitchers that face fewer batters might have a lower bbrate.

   Similar to the last parts, I ran forward and backward selection on a linear regression and a random forest, to see what variables could minimize mean squared error. In that, a bunch of variables showed up. When I put them into the model, I decided to only go for the ones that had low p-values, even if they lowered the MSE.

   The bbrate_prev was the highest indicator for bbrate - having both a high coefficient and a high t-value. The next big indication was the first pitch strike metric. There was an inverse relationship between first pitch strike and walk rate, which makes sense as the pitcher had a higher chance at starting at a 1-0 count.

iv) MSE: 0.00042937110963 39291

4b

i) I used an OLS model for predicting krate

ii) The following features were used:

   i) krate_prev: Last year's krate

   ii) Variety_prev: The number of pitches in a pitcher's arsenal

   iii) maxspeed_prev: The fastest pitch thrown by the pitcher in the previous season (based on release speed)

iii) Similar to the bbrate, I created a couple of features that I would think have some part to play in predicting strike rate and walk rate. I created variables to see the called strike percentage, the swinging strike percentage (swinging strikes vs called strikes),speed based metrics, and a variable to see how often the pitcher's first pitch was a strike or a ball. I also had a variable to see how many pitches a pitcher could use - a pitcher with a high variety could keep the batter guessing more than one that always threw fastballs.

Similar to the last parts, I ran forward and backward selection on a linear regression and a random forest, to see what variables could minimize mean squared error. In that, a bunch of variables showed up. When I put them into the model, I decided to only go for the ones that had low p-values, even if they lowered the MSE.

The krate_prev was the highest indicator for krate - having both a high coefficient and a high t-value. While the pitchers variety and maxspeed did not have a high coefficient, they had high t-values.

iv) MSE: 0.001910619315889745