Mariners Technical Assessment

Question 1 Methodology:
1. Data Cleaning:
    a. Filled missing values of relevant columns (hit distance, speed, etc.) with the mean value based on player, year, and slg
    b. Replaced missing pitch_type values with 'Undefined'
    c. Replace the rest of the pitcher side missing variables with the average based on pitcher id
    d. Removed any potential outliers

2. Feature Engineering:
    a. Slugging Percentage vs Same Hand / Opposite hand: if the hitter is batting from the same side the pitcher is throwing from: slugging percentage of a Right hand hitter vs a right handed pitcher
    b. Ground Ball Rate: the percentage of balls hit in play that are hit on the ground
    c. Barrel Rate: the percentage of balls hit in play that are considered barrels. A barrel is if the ball is hit greater than 98 mph and is within some threshold of launch angle
    d. Slugging Percentage by pitch type: what is a players percentage on each type of pitch after we group all pitches into fastballs, breaking balls, and changeups
    e. Interaction terms: combinations of slugging percentage vs pitches (ex: slugging on FB * slugging on BB). Combinations of same hand/ opposite hand slugging and all types of pitches

3. Modeling
    a. Group all variables by player id and season. Get the 90th percentile of hit distance and exit speed. Then get the mean of all other columns. Use these features to predict the second half season slugging percentage
    b. Used Cross Validation on 4 different models
        i. Linear Model RMSE: 0.101884
        ii. Random Forest RMSE: 0.113719
        iii. Lasso Regularization RMSE: 0.099682
        iv. XGBoost: 0.101135

4. Evaluation:
    a. The best model was found to be a Linear Model with Lasso Regularization. It had an R-squared value of 0.405 and a 5-fold cross validation RMSE of 0.0997
    b. The features selected by lasso regularization were: 'hit_exit_speed', 'hit_distance', 'hit_vert_exit_angle', 'slg', 'barrel', ''FB_CH_slg', 'SameHand_BB_slg', 'OppHand_FB_slg', 'OppHand_CH_slg'
    c. In the summary shown in the code, we see that exit speed, hit distance, and hit_vert_exit_angle are the only predictors that are significant at a 0.05 alpha threshold. This means that essentially most of the variance in second half

slugging can be explained by just these 3 variables. This makes sense because high slugging percentage is attributed to being able to hit the ball for extra bases and the features that most influence this are how hard and how far you hit the ball. Additionally, a high launch angle means you hit the ball in the air more and have a better chance of hitting the ball out of the park.

d. Looking at the coefficients, we can see that increasing your 90th percentile exit velocity by one mph can increase your slugging percentage by .011.

e. Barrel is a feature engineered variable included in this model that is not significant with a p-value of 0.3, but it reinforces the idea that hitting the ball at high speeds at the right angle will lead to higher slugging percentage.

f. First half slugging makes sense as it sets a precedent for what a player will do in the second half. Most times, a player with a lower slugging percentage in the first half won't have a high slugging percentage in the second half and vice versa.

g. The interaction terms included are interesting. Most of them have very high p-values and likely do not have a lot of predictive power, but one to look at is SameHand_BB_slg which has a p-value of 0.193. The positive coefficient for this interaction term indicates that the effect of one predictor on the response variable increases as the other predictor increases. If I would have to guess, this would be because hitters often perform worse against same handed pitchers. To add to this, sliders and curveballs coming from the same side are usually the hardest pitches to hit. If a player is able to slug well on both, they are likely very good hitters and will have a high slugging percentage in the second half.

5. Next Steps
   a. To improve this model there are a couple of other variables I would like to look at
      i. Contact Rate: to see how much a player makes contact with the ball based on how many at bats they get
      ii. Strikeout Rate: which is closely related to Contact Rate
      iii. Walk Rate
      iv. Park Factors such as playing at Coors Field, lefty at Yankees Stadium, etc.

   b. Now that we know what features influence a player's second half slugging, the next steps would be helping our own players increase their slugging percentage. Based on our model, they can do this by increasing their exit speed and launch angle. We could investigate biomechanical features and find which ones influence these specific variables and create plans to adjust players swings in order to increase these variables.

Question 2:

       If I were tasked with choosing which player to trade for assuming all other factors to be equal such as years of control, position, players given up, and clubhouse presence, I would choose player x084802d3 (player 1). At first glance player xe54ef656 (player 2) looks like he has player one beat in many categories. However, taking a deeper look, we see that player one has many more balls hit in play and more hits than player 2 in the first half of the 2023 season. This is similar to comparing a contact hitter like Luis Arraez to a three true outcomes player like Kyle Schwarber. If these two played the same position, there could be an argument for either one, but given what the Mariners need in July 2023, I would argue that player 1 would be a better fit.

       At the All-Star break, the Mariners had the second most strikeouts in all of the MLB. They desperately needed someone who could get on base and could make consistent contact in front run-producing guys like Julio Rodriguez, Ty France, and Eugenio Suarez or even to have towards the bottom of the lineup. With players in the lineup like Eugenio Suarez and Mike Ford who already encapsulate the "swing for the fences" mentality, the team needs someone whose main purpose is to get on base. Player 1s Batting Average on Balls in Play (BABIP calculated by hits - HR/ balls in play - HR) was .372 while player 2s BABIP was .212. This indicates that the success of player 1 when putting the ball in play was much better than the success of player 2. One aspect of this could be luck, but again we would like someone with a high chance of getting on base. Another statistic that stands out is the mean exit velocity. In 2023, player 1's exit velocity is nearly 5 mph less than player 2 at 84.9 mph compared to 89.2 mph. While this may seem like a large difference, the exit velocity of player 2 is right around league average, so we would not even be getting a player with elite exit velocity to make up for the lack of contact.

       Finally, another statistic we can look at is clutch. While there isn't a lot of data to create a variable for this, we can look at the number of two out hits as a proxy for "clutch". Getting a hit with 2-outs is essential because it can either extend an inning so that our big hitters can get an extra at bat. It can also drive in a run where just making contact won't do the job. Player 1 has 21 2-out hits which is more than double that of Player 1 who had 10, indicating that player 2 could not get the job done when hitting with 2 outs.

       If we were to trade for Player 1, a player development plan for him could be to increase his launch angle. Since the model from the previous question showed that this directly influences slugging percentage, we could work with the player to get more launch on his hits so that he can get some more extra base hits while also hitting for contact.

       In conclusion, player 1 (id: x084802d3) fits the needs of the Mariners at the All Star break and would fill a hole in the lineup that will set the team up for a postseason run.

- Code that backs up any statements above is located at the the bottom of the Mariners_Assessment_Code.ipynb file

Question 3:

I think the most recent mistake the Mariners made is getting rid of Tommy LaStella. I love that guy and I can't believe he made an all star team in 2019. He really exemplifies what it means to hustle.

All jokes aside, I think a recent mistake the team made was signing Robbie Ray in 2022. Ray had an incredible season in 2021, but in my opinion this was an outlier season. His strikeout numbers were elite but he was prone to home runs. This along with the fact that he pitched over 175 innings for the first time led to signs of decline in following seasons. In his first season as a Mariner, he pitched well but nothing close to his 2021 numbers. He also showed that he could not perform under the bright lights in the playoffs. He then got Tommy John the following year and I suspect that he will never live up to his Cy Young season again. Hindsight is 20/20, but I think the team should have made a harder push for hitters instead because of how strong the young starting pitchers for the team already are with Logan Gilbert, George Kirby, Bryce Miller, and Byan Woo looking to be the faces of an exceptional, young starting rotation. Marcus Semien or Corey Seager would have been great additions that would have solidified the middle infield for years to come, especially with second base being such an uncertainty now.