Texas Rangers

Analytics Survey

Part 1:

1. **What do you hope to gain from a position with the Texas Rangers? (200 word max)**

Short version: To be a part of a team that helps the Texas Rangers win more baseball games.

To be a part of a collaborative team that works in the trenches with data from Major League Baseball. My goal in working as a part of that team is to draw insights from data in order to deliver custom-tailored solutions to players, coaches, and staff that focus on improving player efficiency.

2. **Assuming you are given all the time and resources you need, what baseball question or project would you most like to answer? What would your approach be? (500 word max)**

Given unlimited time and resources there are many questions I could look to answer, however; the question that interests me the most is: "Are draft picks or trades more effective at building a winning franchise?" There are many factors to consider here: (1) does a "good draft pick" increase trade value above the expected performance of that draft pick? (2) A veteran player may impact winning rate immediately, but is "mortgaging the future" by trading away prospects more impactful to a franchise in the long run? (3) Should you draft players with the goal of trading them? The list of factors to consider would be extensive to start.

My approach to this question first and foremost would be to gather the data of all draft picks and trades between major league teams. From there creating a metric (or just using WAR) to track player development over time may point to valuable insights based on historical trades and draft picks. With that data in hand, it would serve as a good starting point for exploratory analysis. Some questions a model could answer are: "What teams draft the best?" "Who 'Won' a specific trade?" "What teams are the best at trading?"

After an initial model is created, I would look to make adjustments in order to improve either the model or the metric used to measure "success" of a draft pick/trade. This might include some assumptions, such as: (1) If a player was NEVER traded, would they have the same success in their original draft team's ballpark? i.e. If a power hitting lefty remained at Wrigley Field would they have more/less success than if they played at Yankee Stadium? (2) Adjusted WAR if all ballpark conditions remained the same. (3) If a player from "Team A" is traded to "Team B" and subsequently traded to "Team C" in a few years after promising development, would the player have developed at the same rate if they had remained on "Team A" or "Team B." Further exploration of the data would be needed to determine the direction of this analysis.

After refining a metric or model that measures draft pick and trade value, my goal would be to create a short summary or visualization that could capture the main points: (1) When is the best time to trade a player? (2) What current players in an organization would have the best "return" if they were to be traded. (3) What players in the draft would be the best trade pieces in 3-5 years.

With Major League Baseball having such a large amount of data, there are thousands of questions that could be answered with unlimited time and resources. I believe that determining trade/draft pick value (and which is more important) can give the most valuable insights to an organization.

3. **What interest and/or hobbies do you have outside of sports? (200 word max)**

Outside of sports I enjoy anything outdoors, whether it is visiting national parks, landscaping in my backyard, or traveling I find all of it interesting and fun.

Additionally, I love cooking. I usually find myself in charge of making the "signature dish" for friends/family get togethers. I love it because it doesn't become a "task to-do," and instead it is an opportunity to share why I love to cook with others.

Lastly, I'm a huge movie/film buff. Any movie that has some sort of psychological/thriller aspect (anything that makes me think throughout really) will keep me tuned in until the final credits roll.

4. **Rank the following areas of baseball operations by your level of interest:**

1. Player Evaluation
2. Player Development
3. Major League Operations
4. Sport Science

Part 2: Baseball Knowledge & Thought Process

A. **Discuss the relative importance of subjective and objective analysis as well as their relationship to one another in player evaluation.**

In baseball terms, both subjective and objective analysis play an important role in player development, player safety, and player evaluation. To understand the importance of each, first we have to define them. ***Subjective Analysis*** is where the feeling of the individual taking part in the analysis process determines the outcome. ***Objective Analysis*** on the other hand is the opposite. It is fact-based, measurable, and observable. If using one type of analysis for decision making, objective analysis is preferred as there tends to be more usable data to back up a decision.

While these are opposite forms of analysis, they can be used in conjunction and deliver actionable insights. For example, say a batter is in a slump. A hitting coach may notice the batter dropping their back shoulder slightly on off-speed pitches. The coach might not have any data on-hand to back this up, but he has seen thousands of swings and it is the coach's opinion that something is different with this batter. The hitting coach's subjective analysis could prompt a further investigation to confirm this theory. Objective analysis can confirm that hypothesis by identifying motion trends when the batter sees off-speed pitches. Alternatively, objective analysis could point out that the batter is in fact batting at a much lower percentage on off-speed pitches. The example mentioned can work in reverse as well, where data points out that a batter is struggling on off-speed pitches and is delivered to the player's hitting coach. The coach focuses in on swings the batter is making against off-speed pitches and might notice the batter's shoulder drop.

While objective and subjective analysis can work in harmony, one might find that the baseball world struggles with this concept and views the different approaches as mutually exclusive. Baseball historically, has been a "feel" type of game (subjective analysis). There are scouts and key decision makers that have built careers on "an eye for talent" so it is only natural that the rise of data and analytics (objective analysis) is met with some pushback. On the other hand, some organizations have seen success by embracing data and making objective decisions – Moneyball as an example. In terms of player evaluation it may be impossible to determine what is the correct "in-game" decision since some of the decisions can be so polarizing to players and fans. A recent example of this is Tampa Bay's decision to pull pitcher Blake Snell from Game 6 of the 2020 World Series. For reference, Snell was dominant over 5.1 innings, striking out nine batters on just 73 pitches. Subjectively, "he was on fire." However, over the course of the season Blake Snell made it to the sixth inning four times in 2020. His ERA was 13.50 in the sixth inning alone and teams facing Snell for the third time in a single game hit .304 with a .609 OBP and .913 OPS. Objectively, the data said pulling Snell from the game would give Tampa Bay the best chance at winning. Tampa Bay went on to lose 3-1 in that game. Hindsight is 20/20 and it is unknown if the correct decision was made.

In summary, an ideal scenario would use both objective and subjective analysis to evaluate players in order to make the best-informed decision. Objective analysis is preferred for delivering actionable insights, but there are times when subjective analysis must be considered as well. At the end of the day, both types of analysis can only partially explain what the best decision might be. To quote Ron Washington, "That's the way baseball go."

Part 2B: Data Visualization & Storytelling

https://public.tableau.com/profile/brett.thompson7992#!/vizhome/texasRangersSalaryViz/Dashboard1

(Adding description to walk through the options.)

Part 2C: Data Science & Coding Skills

- **Predict the top 10 HR hitters for next season and the number of HRs they will hit. Discuss your methodology, data sets, assumptions, and include any code your wrote.**

Overview

This dataset contains standard and statcast baseball statistics from the 2017 – 2020 Major League Baseball season for individual batters. Initially, there are a total of 847 unique batters included in the dataset. This data was collected directly from Lahman's Baseball Database, as well as Baseball Reference and Baseball Savant.

Questions to be asked from this dataset include: (1) Who will be the top 10 Home Run hitters for the 2021 MLB season? (2) How many home runs will they hit?

Data and Methodology

The data was first cleaned to ensure accurate projections for 2021 home runs. First, any player considered as a "pitcher" was removed from the dataset.

Second, the 2020 baseball season was shortened to a 60-game games due to COVID-19 so all data from 2020 was adjusted to reflect the standard 162-game season. A simple adjustment of multiplying each hitting statistic by 2.7 for 2020 season was used.

Next, batters that were not considered as "qualifying" were filtered and removed. To determine qualified batters, Major League Baseball denotes that a minimum of 502 plate appearances is required to be considered a "qualified batter." Plate appearances (PA) were not listed in the original dataset, but can be calculated as follows:

$$PA = (AtBats) + (Walks) + (SacBunts) + (SacFlys) + (HitByPitch) + (CatcherInterference)$$

This calculation was performed per player and added as an individual column to conclude the initial cleaning of data.

The Marcel Projection system was performed on the individual 2018, 2019, and 2020 (adjusted) baseball seasons. Marcel is a system of player projections developed by statistician Tom Tango. After weighting the 2018, 2019, and 2020 seasons based off Marcel, aging adjustments were added to the data to refine the projections.

Alternatively, a random forest regressor was used on Statcast data to model a regression of predicted home runs. The results of this model did not end up being significant and it should be considered less reliable than the Marcel Projections.

Results

Marcel Top 10 HR for 2021

```
In [86]: final_df.head(10)
```

Out[86]:

	nameFirst	nameLast	proj_HR_2021	playerID
1	Mike	Trout	38	troutmi01
2	Eugenio	Suarez	36	suareeu01
3	Christian	Yelich	33	yelicch01
4	Jose	Ramirez	32	ramirjo01
5	Nelson	Cruz	32	cruzne02
6	Matt	Olson	32	olsonma02
7	Cody	Bellinger	31	bellico01
8	Mookie	Betts	31	bettsmo01
9	Marcell	Ozuna	31	ozunama01
10	Jose	Abreu	30	abreujo02

Figure 1: Marcel 2021 HR Results

RF Regression for Season HR

```
In [52]: forest_df.head(10).reset_index(drop=True)
```

Out[52]:

	first_name	last_name	Predicted_HR
0	Giancarlo	Stanton	51
1	Aaron	Judge	48
2	Jorge	Soler	45
3	Khris	Davis	44
4	Pete	Alonso	44
5	Christian	Yelich	41
6	Cody	Bellinger	41
7	Khris	Davis	41
8	Nelson	Cruz	41
9	J.D.	Martinez	40

Conclusions

The marcel system based on qualified batters projects Mike Trout to lead all batters in HR next year.

Marcel is a basic forecasting system that regresses toward the mean and adds age as a factor. Given the little amount of intelligence the model uses the projections are not certified, reviewed or confirmed.

Recommendations

Marcel weights each season individually, with the most recent season being weighted the highest. With the 2020 season being shortened, further analysis might swap the 2019 weights with 2020 weights. This would put more emphasis on a full season of data and may improve projection performance.

The random forest model may improve with alternative features. The features utilized in the model generated the highest r^2 value, but not ALL statcast data was utilized.

Park factors may lead to further adjustments in both models for total predicted home runs. Creating a metric that measures performance would be recommended for improving the model.

Limitations/Bias

Overall, this was an amazing learning experience. I was introduced to Machine Learning two weeks ago and applied what I could through the random forest model. With continued practice and education in the machine learning field, the regressions will improve. The random forest model in particular is something that I learned can be more accurate than a linear regression or simple decision tree regressor.

Marcel is also something I was introduced to last week. I am confident in the projections, but understand there are some areas that I can improve on in the model.

Players that were injured or did not qualify for the batting title were removed from the dataset making the data limited to those with a significant number of plate appearance. This may make the model less accurate for predicting home runs if players were to qualify the following season or remained healthy.

Works Cited

TBD