New York Rangers Skater Analysis

MDA_620 Final Capstone

Brian Albertsen

12.10.2023

Background

The data set that I chose to work on for this project was from a website called Hockey Reference, this site collects and holds all the stats for the National Hockey league. They have just about every hockey stat for every player who has played a game in the NHL. I decided to dive into data regarding my favorite team, The New York Rangers, specifically all the Rangers Skaters not including goalies dating back to the start of the franchise in 1927. This data set turned out to be 1020 columns (players) and 22 rows (various stats)

Problem Scenario/ Business Issue

For this project I decided that it would be interesting for me to dive into a topic that means something to me as this would help me understand the concepts and solve problems efficiently. With this project I was struck with the problem that the New York Rangers analytic department was having issues with determining which statistics were most beneficial to use when determining if a player was more impactful than another. Using this data set I was to determine which player statistics influenced the statistic Plus minus. This is a Statistics that does not directly measure personal goal but instead it measures the players overall impact on the team, a player can achieve a plus when their team scores a goal either even strength or short handed when they are on the ice. A player will receive a minus when the other team scores a shorthanded or even strength goal when they are on the ice. These pluses and minuses add up over the players career so the higher number plus/minus a player has the more impactful the more negative or lower number the worse.

Objective / Goals of this project

1. The First goal I had for this project was to create performance predictions for players based off different statistics, this allows for a proper introduction of how different factors and statistics contribute to a players success
2. Secondly, after these predictions had been created we had to determine which statistics correlate the most when determining which are the most influential
3. Third, we wanted to create models that would help in determining these factors both visually and conceptually
4. Lastly we wanted to compute model accuracy to determine whether these factors and outcomes are correct
5. All of these objective allow for us to help determine what statistics are most vital when it comes to overall impact of a player, which will not only help us to manage a team in a more effective way and win more games

Data - Time Stamp November 16, 2023

Columns	Description	Data Type
Player	Player Name	Object
Start	Career start year	int64
End	Career end year	int64
Years	Total Years played	int64
Games Played	Total games played	int64
Goals	Total Goals	int64
Assists	Total Assists	int64
Points	Total Points	int64

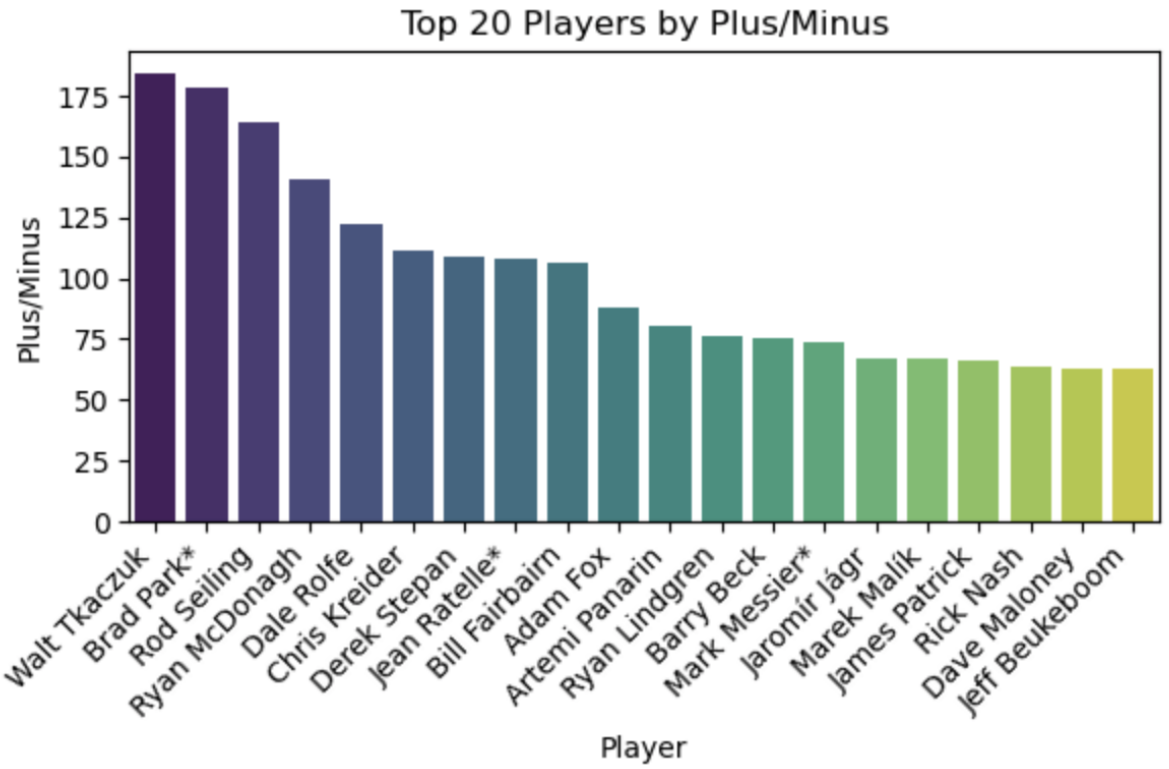
+/-	Plus Minus	float64
Penalty Minutes	Total Penalty Minutes	int64
Even Strength Goals	Total Even Strength Goals	float64
Power Play Goals	Total Power Play Goals	float64
Short Handed Goals	Total Short Handed Goals	float64
Game Winning Goals	Total Game Winning Goals	int64
Even Strength Assists	Total Even Strength Assists	float64
Power Play Assists	Total Power Play Assists	float64
Short Handed Assists	Total Short Handed Assists	float64
Shots on Goal	Total Shots on Goal	float64
Shooting Percentage	Total Shooting Percentage	float64

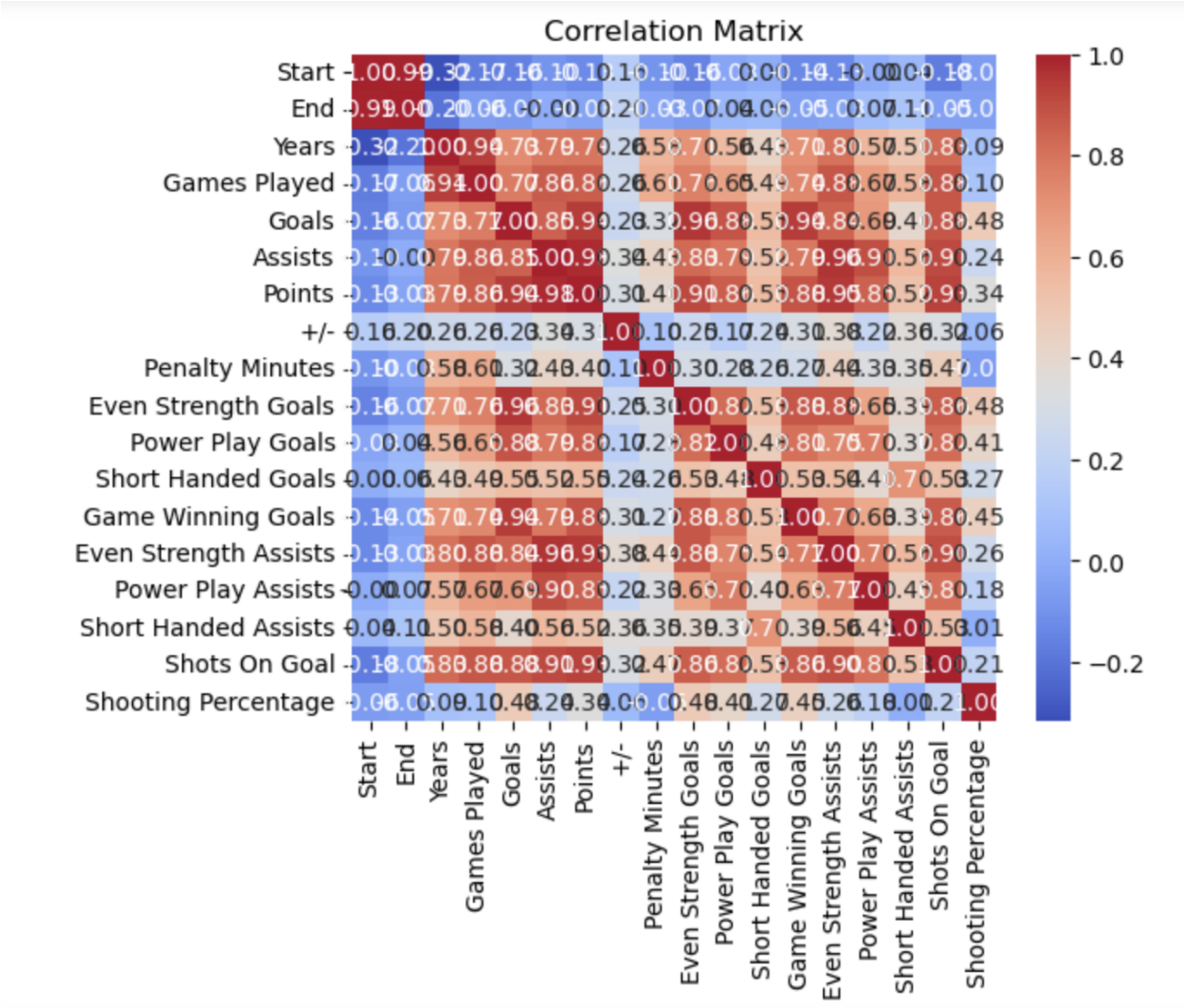
Data Exploration / Data Visualization

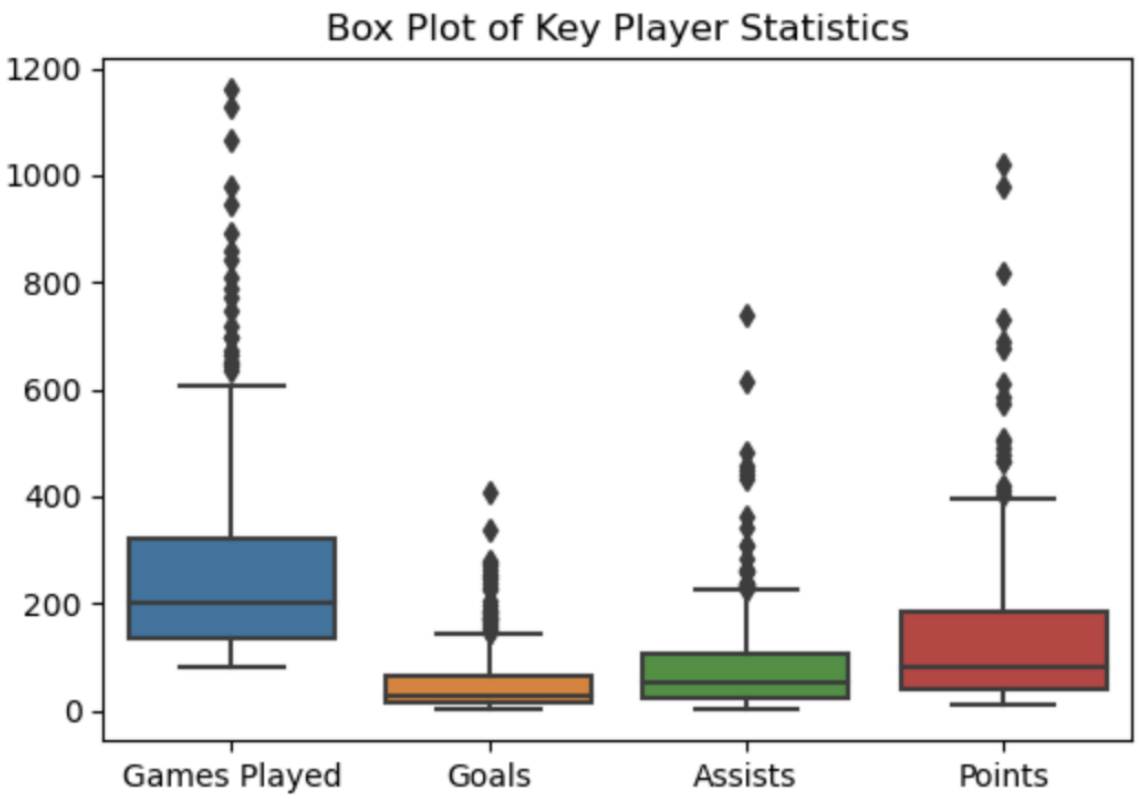
When it came to finding this data, I had done quite a bit of research to find a data set that was as complete as this one. Having the data ranging all the way back to 1927, although yes there were missing columns and data but with this large of a sample size it allows for a justified and high quality data set after manipulation. This data set also contained more stats then most of the others which also allows for a more precise and in depth analysis. This data set was also changing by the day which allows for me the most up to date analysis, when we had retrieved this set it was November 16th, 2023 so that is the date included in the data.

From here I wanted to explore key player statistics such as goals, assists, points, penalty minutes etc and determine and understand their effects on overall player impactfulness, and then check the impactfulness and correlation by using correlation matrix and visualizations to show their positive or negative impact.

Through a series of visualizations we were able to determine which stats we were going to focus on and which ones we deemed unfit for our analysis.




Correlation Matrix
Start
End
Years
Games Played
Goals
Assists
Points
+/-
Penalty Minutes
Even Strength Goals
Power Play Goals
Short Handed Goals
Game Winning Goals
Even Strength Assists
Power Play Assists
Short Handed Assists
Shots On Goal
Shooting Percentage
1.0
0.8
0.6
0.4
0.2
0.0
−0.2



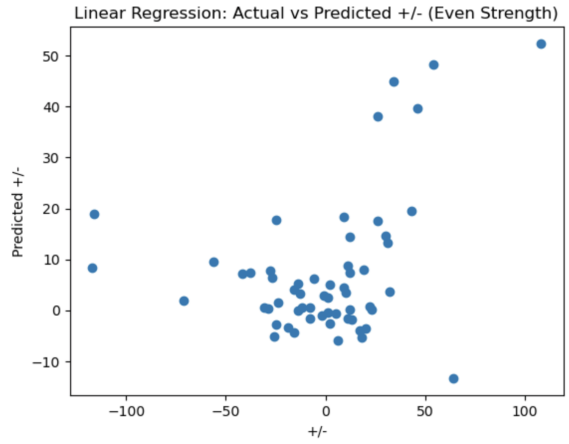
Data Manipulation

After finding this data set I knew that it was going to need some work to clean it up not only to make it easier to work with but also to make it more readable in the end. First I started by dropping columns that were unnecessary and did not have sufficient data. I also realized this was a good point to actually look at the types of data we had, and how many missing values and what their impacts were. After determining their impactfulness to my analysis I was able to decide what I wanted to do with these players that had missing data. With these players that had missing values especially in the plus/minus column I had determined it would be best to drop these players. This was done for a few different reasons, first most of these players that had missing values for plus minus where players that didn't play a whole season with the rangers and secondly with such a larger data set it allowed for a better analysis rather then inputting filler data such as the mean which would skew the data and lead to inaccuracy. One other step I took to make the analysis as accurate as I could sort the player by games played. I only included players who played 82 games or more which equates to one season. I then was able to rename the columns to ease the viewers understanding and increase the clarity of the data.

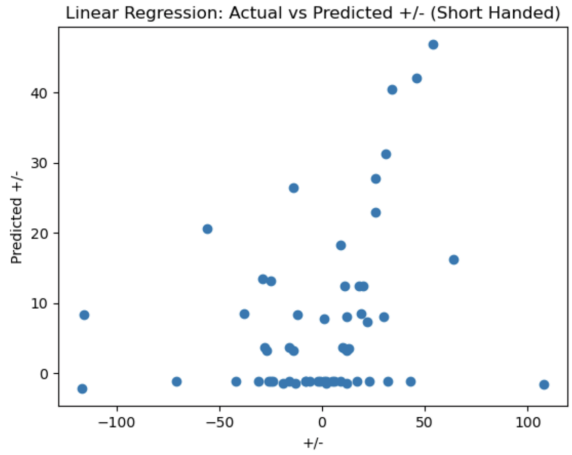
Methodology / Model Building

After I had gone through data manipulation to clean up the data, I had decided to use linear regression to better understand the relationship between key statistics and overall player impactfulness which was determined by the Plus Minus statistic. After this I had split the data into training and testing sets which allowed for generalization in the models. After I analyzed the coefficients of the features I chose to understand their impact on the Plus Minus Statistic. This provided us with insights into which player statistics are more influential. Along with this I evaluated the model's performance using metrics like Mean Absolute Error, Mean Squared Error, and Root Mean Squared Error. These metrics support the accuracy of the models. Followed by visualization, like a scatter plot to visually compare actual plus/minus values with predicted values.

```
Linear Regression Model Evaluation:
Mean Absolute Error: 23.39292939900237
Mean Squared Error: 1251.9555763240232
Root Mean Squared Error: 35.38298427668338
```



```
Linear Regression Model Evaluation:
Mean Absolute Error: 23.057578493314985
Mean Squared Error: 1286.0818283782673
Root Mean Squared Error: 35.86198305139117
```



In these models I had decided to compare the effects of a “Offensive minded” player and a “Defensive minded” player and what that effect had on plus/minus.

Model Selection

The model I chose for my analysis was a linear regression and there were many reasons for this, first off was the effectiveness of linear regression for finding and predicting a target variable based on other variables. I also found the ability to interpret the coefficients that are created from these models very useful in my analysis allowing me compare different key statistics and their relationship to players overall impact, not only is this produce good findings for teams but also players in determining what skills they need to work on to increase there impactfulness. Lastly, I tested and evaluated the linear regression model, decision tree and even random forest models, from this I was able to see that the linear regression models had the lower MSE, MAE and RMSE which helped me settle on linear regressions.

Conclusion

As I began to work on this project I had discovered many different combinations and thoughts that could possibly help me come to more of an accurate conclusion. After a few combinations I had realized that these statistics provided may not provide the best correlations to the target Plus Minus, things like matchups, line pairs and other factors that cannot be categorized play a role in determining a statistic like this. But this doesn't mean that these general statistics do not have an impact at all, they do. But the question is to what degree. After determining this I was able to lock in on two different variables that contribute to Plus Minus, first I looked at Even strength goals and Even strength assists which both are common statistics that help determine the impact of an average player, the next set was with the variables of Short handed goals and short handed assists. Yes these may seem like similar statistics but what's deeper about these stars is the overall impact these have on the game. Scoring a short handed goal or having a short handed assist is a stat that has great magnitude as these are more hard to come by and definitely swing momentum in games which in result increase your impact as a player.

Getting into the results,

	Feature	Coefficient
0	Games Played	-0.169428
1	Goals	-0.972363
2	Assists	1.130750
3	Points	0.158387
4	Penalty Minutes	0.014469
5	Even Strength Goals	0.155244
6	Power Play Goals	0.281899
7	Short Handed Goals	-1.409507
8	Game Winning Goals	3.285440
9	Even Strength Assists	-0.516192
10	Power Play Assists	-1.632414
11	Short Handed Assists	3.279356
12	Shots On Goal	0.024015
13	Shooting Percentage	0.085780

Seeing these Coefficients provides us with information that helps determine the impact of these statistics. First we can look at Even strength goals, which has a positive impact on overall Plus Minus. This makes sense in hockey terms, the more you score the more positive impact it has on your team. But when it comes to Even strength assists, it has a stronger negative impact on Plus Minus. This is strange but in a way makes sense, a

player more prone to making extra passes or more risky passes has a higher chance of turning the puck over in turn giving your opponent a chance to score and decrease your Plus Minus.

Moving over to the other model involving Short handed Goals and short handed assists, these both have the opposite effect to its even strength counterpart. Short handed Goals has a strong negative effect on Plus Minus whereas Short handed assists has a strong positive effect. After looking at these it can also be made sense as a player attempting to score a goal short handed is more likely to be forcing a opportunity and then a result of an even better chance at scoring for the other team. And for short handed goals it as well can make sense as a player who sets the play up short handed and isn't pushing a bad opportunity takes less risk of the other team having a better opportunity at scoring.

After these insights it is important to look at a few things, such as MSE, MAE and RMSE. These all help to determine the effectiveness and accuracy of the model. For the linear regression for Even strength goals and assists with the target on Plus Minus. The numbers are as follows: MAE = 23.392, MSE = 1251.96 and RMSE =35.38, what these numbers tell us is that these models are not quite accurate and need to be taken with a grain of salt. Same goes for Short handed goals and assists as the numbers are similar yet different MAE = 23.05, MSE = 1286.08 and RMSE =35.86. Although they are similar and both high it shows that Short handed goals and assists has a tiny bit more impact on overall Plus Minus, but again these are larger numbers that don't provide great accuracy. And this makes sense as discussed earlier there are way more outside factors that go into determining the impactfulness of a player and many of these cannot be determined by simple statistics.

References

https://www.hockey-reference.com/teams/NYR/skaters.html

Python